White Paper
FUJITSU Server PRIMERGY
Memory Performance of Xeon E5-2600 v4 (Broadwell-EP) based Systems

The Xeon E5-2600 v4 (Broadwell-EP) based FUJITSU Server PRIMERGY models also acquire their impressive increase in performance from an improvement in the QuickPath Interconnect (QPI) memory architecture, which has proved itself now for five generations of systems. This white paper explains the essential features of the architecture as well as the latest improvements and quantifies their effect on the performance of commercial applications.

ersion
1.0
2016-03-31

Contents

Document History

Version 1.0 (2016-03-31)

Initial version

Introduction

The Intel Xeon E5-2600 v4 (Broadwell-EP) processors of the current dual socket PRIMERGY servers are produced in a new 14 nm manufacturing process, from which the increase in performance over the predecessor generation, the Haswell-EP (produced in a 22 nm manufacturing process), basically results. Microarchitecture, chipset and Grantley-EP platform are retained from the predecessor generation.

The new generation provides a 20 to 30% increase in performance in comparison to the predecessor generation with regard to most of the load scenarios. A large element in this impressive improvement is the result of a maximum of 22 cores per processor instead of the previous 18. The memory system also has new features which contribute to the generational increase in performance.

The Haswell-EP generation was accompanied by the introduction of DDR4 memory technology. The current Broadwell-EP systems also use this technology, but support as a new feature memory frequencies up to 2400 MHz, compared with a maximum of 2133 MHz with Haswell-EP. The maximum frequency of the QPI (QuickPath Interconnect) links remains unchanged. Both Haswell-EP and Broadwell-EP support a maximum of 9.6 GT/s here.

A further new feature of the memory system concerns the ability to select the cache coherence protocol, which was introduced with the predecessor generation. We are familiar with the three versions from Haswell-EP *Early Snoop*, *Home Snoop* and *Cluster-on-die*. Added to these now is the option *Home Snoop with Directory and OSB*. At the same time, this option becomes the new BIOS default, compared with the default *Early Snoop* in the predecessor generation. The versions differ in terms of their trade-offs between the latencies and bandwidths for local and remote memory access. In most applications, there will be no need for any action that deviates from the default setting. It would be sensible to look at this topic more closely with regard to sensitive performance expectations and the corresponding tests.

Otherwise, the proven basic features of the QPI-based memory architecture of the predecessor generations are retained. The processors have *on-chip* memory controllers, i.e. every processor controls a group of memory modules that has been allocated to it. The performance of this local memory access is very high. At the same time, the processor is able to provide the neighboring processor with memory content via unidirectional, serial QPI links and itself request such content. The performance of the remote access is not quite so high. This architecture with its distinction between local and remote memory access is of the NUMA (Non-Uniform Memory Access) type.

However, when it comes to detail many memory system features of the immediate predecessor generation Haswell-EP [L3] have also been retained. There are four memory channels per processor each with three DIMM slots. Thus, the maximum number of 12 DIMMs per processor is unchanged. What is new, however, is the already mentioned increase in the maximum memory frequency from 2133 to 2400 MHz. The most elementary indicator of memory performance, the maximum memory bandwidth, increases as a result of this measure for the dual socket server by more than 10% to about 130 GB/s.

On the one hand, this document looks at the new memory system features in the current server generation. On the other hand, as in the earlier issues, this document also provides basic knowledge about the QPI-based memory architecture which is essential when configuring powerful systems. We are dealing with the following points here:

- Due to the NUMA architecture both 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 and thus accelerate 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, widthwise across the four memory channels per processor. The *Performance Mode* configuration of the PRIMERGY configurator in groups of four DIMMs of the same type on each processor ensures optimal interleaving in this direction. There is also interleaving in the depth of the individual memory channel. The decisive memory resources for this are the so-called 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 2400, 2133 or 1866 MHz depending on the processor type, DIMM type and quantity as well as the BIOS setting. Very large memory capacities limit the memory frequency. For this reason, we must balance features such as performance and capacity against each other.

Influencing factors 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.

Results show that the percentage influences depend on the performance of the processors. The more powerful the configured processor model, the more thoroughly the issues of memory configuration dealt with in this document should be considered.

Statements about memory performance under redundancy, i.e. with enabled mirroring or rank 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 influences on the effective memory frequency. The fourth section deals with the BIOS parameters that affect the memory system. The last section lists DIMM configuration examples which are ideal regarding memory performance.

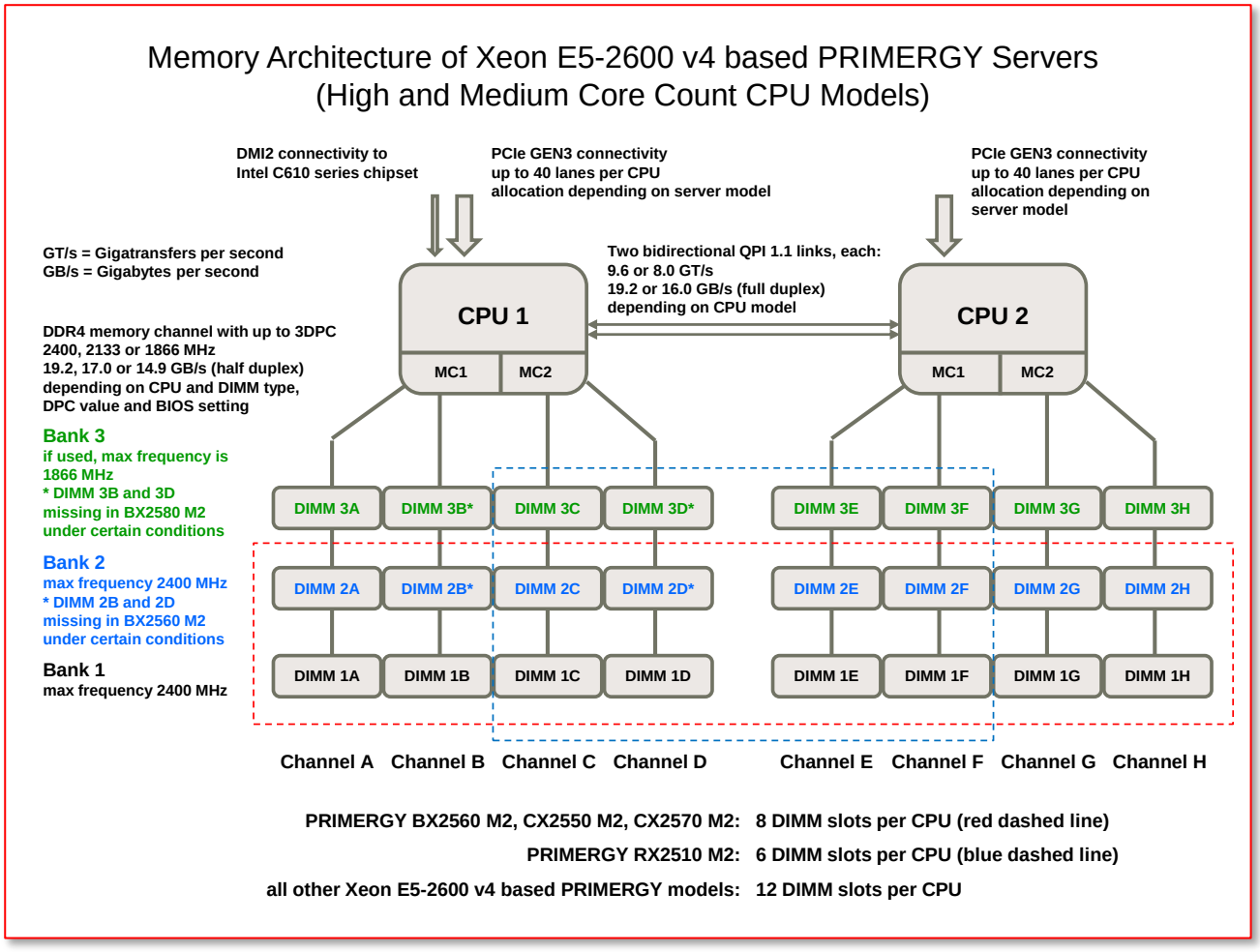
DIMM slots and memory controllers

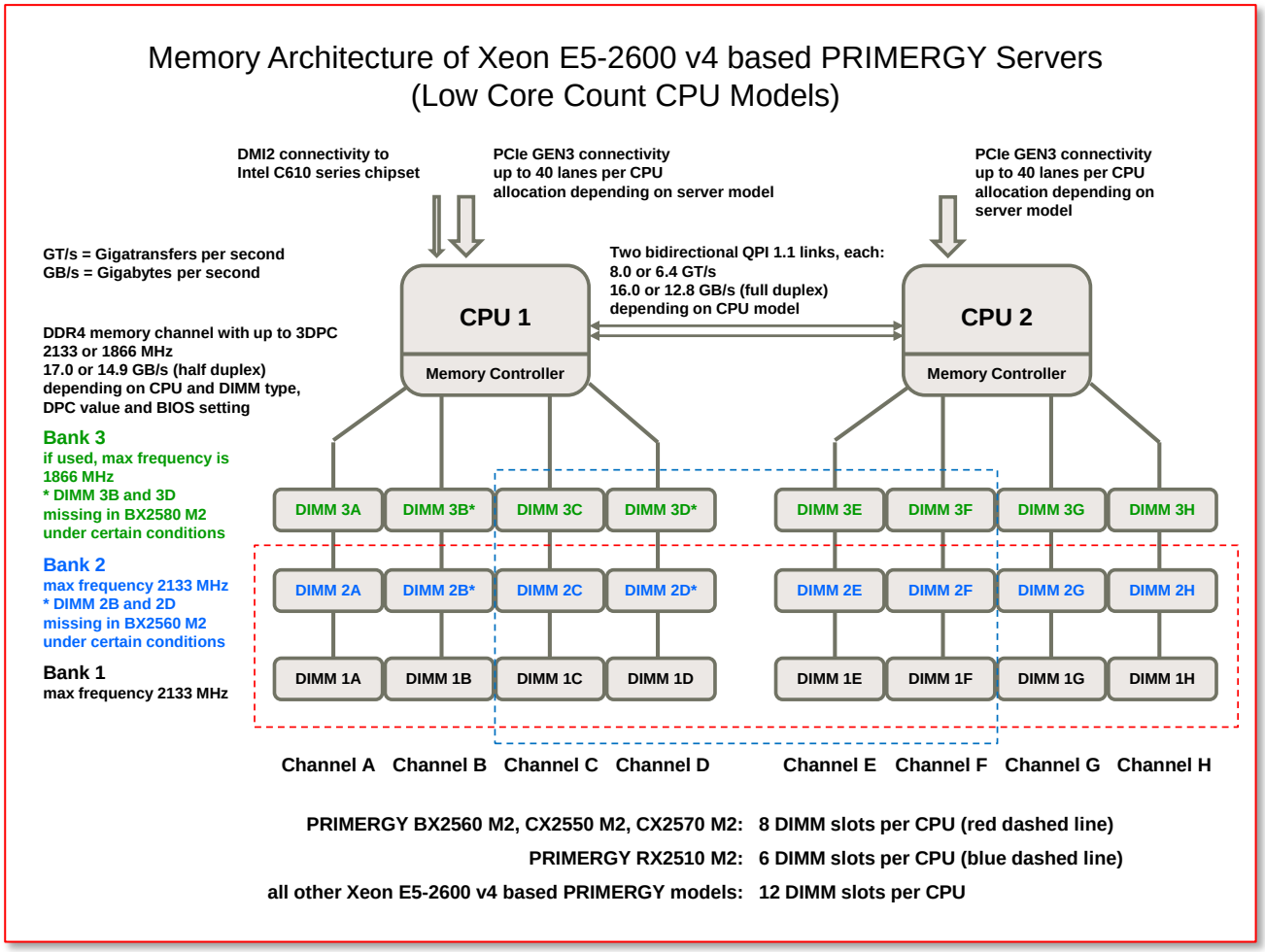
The following diagrams show the structure of the memory system. Before explaining the subtle differences between both diagrams, here first of all are the essential elements they have in common.

The Xeon E5-2600 v4 based PRIMERGY servers usually have 12 DIMM slots per processor. An exception to this are the models PRIMERGY BX2560 M2, CX2550 M2, und CX2570 M2 with 8 slots each due to a high density form factor. Another exception is the PRIMERGY RX2510 M2. This system is optimized according to cost aspects, which is why two out of four possible memory channels are missing per processor. The result of this is 6 DIMM slots per processor.

There can be a third exception in the blade servers PRIMERGY BX2560 M2 and BX2580 M2. A group of high-performance processors – see the configurator for details – requires a larger heat sink for CPU 1, and consequently 2 DIMM slots are then omitted.

For the resources memory channel and QPI link the diagrams show the connection between frequency and bandwidth, which results from the respective net data path widths. They are 64 bits for the DDR4 memory channel and 16 bits for the QPI link. In the case of the bidirectional QPI link the bandwidth is valid for each direction, hence the name full duplex. And with memory channels the read and write accesses have to share the data paths, thus the name here is half duplex.





The details about frequencies and bandwidths also take the processor class viewed in the respective diagram into consideration.

There are usually four memory channels per processor. The number of DIMM strips configured per channel influences the memory frequency and thus the memory performance. This value, often referred to below, is known as DPC (DIMMs per channel). If the channels are differently configured, the largest occurring DPC value is decisive for the effect of the memory configuration on the frequency.

Another term used below is "memory bank". As shown in the diagram, a group of four or two (PRIMERGY RX2510 M2) DIMM strips distributed across the channels forms a bank. The colors in the diagram (black, blue, green) correspond to the colored marking of the banks on the system boards of the servers, which is aimed at preventing configuration errors. When distributing the DIMM strips via the slots available per processor, it is desirable to start with bank 1 and to proceed bank-by-bank in order to attain the best possible interleaving across the channels. Interleaving is a main influence on memory performance.

The corresponding processor must be available in order to use the DIMM slots. If there is no maximum configuration, the slots allocated to the empty CPU socket cannot be used.

The number of memory channels per processor is always the same for the Xeon E5-2600 v4 processor family. However, there is a difference in the number of memory controllers per processor, hence the two diagrams in this section.

It continues an evolutionary development in the Xeon E5 processor families. EP processors for dual socket servers only had one memory controller from Nehalem to Sandy Bridge. With Ivy Bridge the two most powerful processor models – a very small group – had two controllers for the first time. With Haswell and Broadwell this feature is extended to all processor models with a high and medium number of processor cores – that is about half the respective processor family.

The exact classification is in the following table, column *Die design*.

- HCC (high-core count) and MCC (medium-core count) processors have two memory controllers (first diagram).
- LCC (low-core count) processors have one controller (second diagram).

Processors (since system release)									
Processor	Cores	Threads	Die Design	Cache [MB]	QPI Speed [GT/s]	Nominal Frequency [GHz]	Max. Turbo Frequency [GHz]	Max. Memory Frequency [MHz]	TDP [Watt]
Xeon E5-2623 v4	4	8	LCC	10	8.00	2.60	3.20	2133	85
Xeon E5-2637 v4	4	8	LCC	15	9.60	3.50	3.70	2400	135
Xeon E5-2603 v4	6	6	LCC	15	6.40	1.70	n.a.	1866	85
Xeon E5-2643 v4	6	12	LCC	20	9.60	3.40	3.70	2400	135
Xeon E5-2609 v4	8	8	LCC	20	6.40	1.70	n.a.	1866	85
Xeon E5-2620 v4	8	16	LCC	20	8.00	2.10	3.00	2133	85
Xeon E5-2667 v4	8	16	LCC	25	9.60	3.20	3.60	2400	135
Xeon E5-2630L v4	10	20	LCC	25	8.00	1.80	2.90	2133	55
Xeon E5-2630 v4	10	20	LCC	25	8.00	2.20	3.10	2133	85
Xeon E5-2640 v4	10	20	LCC	25	8.00	2.40	3.40	2133	90
Xeon E5-2650 v4	12	24	MCC	30	9.60	2.20	2.90	2400	105
Xeon E5-2650L v4	14	28	MCC	35	9.60	1.70	2.50	2400	65
Xeon E5-2660 v4	14	28	MCC	35	9.60	2.00	3.20	2400	105
Xeon E5-2680 v4	14	28	MCC	35	9.60	2.40	3.30	2400	120
Xeon E5-2683 v4	16	32	HCC	40	9.60	2.10	3.00	2400	120
Xeon E5-2690 v4	16	32	MCC	40	9.60	2.60	3.50	2400	135
Xeon E5-2697A v4	16	32	HCC	40	9.60	2.60	3.60	2400	145
Xeon E5-2695 v4	18	36	HCC	45	9.60	2.10	3.30	2400	120
Xeon E5-2697 v4	18	36	HCC	45	9.60	2.30	3.60	2400	145
Xeon E5-2698 v4	20	40	HCC	50	9.60	2.20	3.60	2400	135
Xeon E5-2699 v4	22	44	HCC	55	9.60	2.20	3.60	2400	145

The difference between HCC and MCC depends on the topologies in which the processor cores are organized inside the chip. The cores and L3 cache shares are arranged as if in matrix. HCC models have four columns, MCC have three and LCC have two. LCC has a ring-shaped interconnect for all cores to which the memory controller is connected as well. HCC and MCC have two interconnects, each with a controller.

The *die design* classification is basically nothing more than a classification of the performance level of the processor models. The quantitative memory performance tests have been carried out separately according to processor classes whereby either the HCC, MCC or LCC classification has been used depending on the topic or the one according to the supported memory frequency (the last but one column in the table).

DDR4 topics and available DIMM types

The Broadwell-EP based PRIMERGY servers use DDR4 SDRAM memory modules. The transition from DDR3 to DDR4 took place with the predecessor generation Haswell-EP. 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. It continues the increase in memory frequency in steps of 266 MHz as known with DDR3-based server generations. In the case of Broadwell-EP this use has now reached 2400 MHz. The Haswell-EP based systems have supported a maximum of 2133 MHz.
- 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, insofar as it concerns the memory system. These trade-offs still play a major role for the processors.

DIMM stripes are considered for the memory configuration of the Xeon E5-2600 v4 based PRIMERGY servers according to the following table. There are *registered* (RDIMM) and *load-reduced* (LRDIMM) DIMMs. Mixed configurations are in each case only possible within the three sections of the table, i.e. only between RDIMMs of type x4 on the one hand and RDIMMs of type x8 on the other hand. As a matter of principle LRDIMMs also cannot be mixed with RDIMMs.

DIMM type	Control	Max fre-quency (MHz)	Volt	Ranks	Capa-city	SDDC	Rel. price per GB
8GB (1x8GB) 1Rx4 DDR4-2400 R ECC	registered	2400	1.2	1	8 GB	Yes	1.2
16GB (1x16GB) 1Rx4 DDR4-2400 R ECC	registered	2400	1.2	1	16 GB	Yes	[1]
16GB (1x16GB) 2Rx4 DDR4-2400 R ECC	registered	2400	1.2	2	16 GB	Yes	**1.0**
32GB (1x32GB) 2Rx4 DDR4-2400 R ECC	registered	2400	1.2	2	32 GB	Yes	1.2
8GB (1x8GB) 2Rx8 DDR4-2400 R ECC	registered	2400	1.2	2	8 GB	No	[1]
16GB (1x16GB) 2Rx8 DDR4-2400 R ECC	registered	2400	1.2	2	16 GB	No	1.1
64GB (1x64GB) 4Rx4 DDR4-2400 LR ECC	load reduced	2400	1.2	4	64 GB	Yes	3.0

[1] Not yet available when this document was published

Data is transferred in units of 64 bits for all DIMM types. This is a feature of DDR-SDRAM memory technology. 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 or x8 in the type name). Such a chip group is referred to as a *rank*. According to the table there are DIMM types with 1, 2 or 4 ranks. Maximum capacities are 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. The number of available ranks per memory channel has a certain influence on performance, which is explained below.

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).
- 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 virtual one. The memory controller then only sees virtual ranks. This function is enabled if the number of physical ranks in the memory channel is greater than 8.

The x4 or x8 structure of the DIMMs influences the ECC detectability of memory errors that either can or cannot be corrected. This is the reason why RDIMMs of type x4 cannot be mixed with RDIMMs of type x8. SDDC (Single Device Data Correction, see the last but one table column) refers to the extended ECC functionality, which is restricted to x4 modules and which can compensate for the failure of an entire DRAM chip.

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 type-specific influences are to be given here. However, a systematic quantitative evaluation does not take place below due to insignificance:

- The increasing complexity of the DIMM types RDIMM and LRDIMM due to additional components on the DIMM is connected with a slight increase in access latency in the order of a few nanoseconds.
- Rank Multiplication in the case of configurations with LRDIMMs with more than 8 physical ranks per memory channel results in a small reduction in the maximum memory bandwidth and the application performance – in comparison to configurations with RDIMMs – of less than 5%.

The effective frequency of a given configuration depends on a series of influences. 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 table shows the relative price differences. The list prices from April 2016 for the PRIMERGY RX2540 M2 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). A comparison with earlier versions of this document series shows that the idea of relative memory prices has been undergoing a constant change.

Depending on the 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.

Definition of the memory frequency

There are three possible values 2400, 2133 and 1866 MHz for the frequency of the memory. The frequency is defined by the BIOS when the system is switched on and applies per system, not per processor. Initially, the configured processor model is of significance for the definition.

This section recommends the classification of Xeon E5-2600 v4 models according to the last but one column of the following table already shown above. The column shows the maximum supported memory frequency. Furthermore, this classification covers that frequency according to the QPI clock rate.

Processors (since system release)									
Processor	Cores	Threads	Die Design	Cache [MB]	QPI Speed [GT/s]	Nominal Frequency [GHz]	Max. Turbo Frequency [GHz]	Max. Memory Frequency [MHz]	TDP [Watt]
Xeon E5-2623 v4	4	8	LCC	10	8.00	2.60	3.20	2133	85
Xeon E5-2637 v4	4	8	LCC	15	9.60	3.50	3.70	2400	135
Xeon E5-2603 v4	6	6	LCC	15	6.40	1.70	n.a.	1866	85
Xeon E5-2643 v4	6	12	LCC	20	9.60	3.40	3.70	2400	135
Xeon E5-2609 v4	8	8	LCC	20	6.40	1.70	n.a.	1866	85
Xeon E5-2620 v4	8	16	LCC	20	8.00	2.10	3.00	2133	85
Xeon E5-2667 v4	8	16	LCC	25	9.60	3.20	3.60	2400	135
Xeon E5-2630L v4	10	20	LCC	25	8.00	1.80	2.90	2133	55
Xeon E5-2630 v4	10	20	LCC	25	8.00	2.20	3.10	2133	85
Xeon E5-2640 v4	10	20	LCC	25	8.00	2.40	3.40	2133	90
Xeon E5-2650 v4	12	24	MCC	30	9.60	2.20	2.90	2400	105
Xeon E5-2650L v4	14	28	MCC	35	9.60	1.70	2.50	2400	65
Xeon E5-2660 v4	14	28	MCC	35	9.60	2.00	3.20	2400	105
Xeon E5-2680 v4	14	28	MCC	35	9.60	2.40	3.30	2400	120
Xeon E5-2683 v4	16	32	HCC	40	9.60	2.10	3.00	2400	120
Xeon E5-2690 v4	16	32	MCC	40	9.60	2.60	3.50	2400	135
Xeon E5-2697A v4	16	32	HCC	40	9.60	2.60	3.60	2400	145
Xeon E5-2695 v4	18	36	HCC	45	9.60	2.10	3.30	2400	120
Xeon E5-2697 v4	18	36	HCC	45	9.60	2.30	3.60	2400	145
Xeon E5-2698 v4	20	40	HCC	50	9.60	2.20	3.60	2400	135
Xeon E5-2699 v4	22	44	HCC	55	9.60	2.20	3.60	2400	145

The DIMM type and the DPC value of the memory configuration also restrict the frequency. Processor type, DIMM type and DPC value are strong influences on the memory frequency, which cannot be overridden via BIOS. However, the BIOS parameter *DDR Performance* permits a consideration between performance and energy consumption - in a restricted sense that must be explained later. If you decide in favor of performance, the result is the effective memory frequency according to the following table. This is in particular the BIOS default.

DDR Performance = Performance optimized (Default)						
CPU type	**RDIMM**			**LRDIMM**		
	1DPC	2DPC	3DPC	1DPC	2DPC	3DPC
DDR4-2400	2400	2400	1866	2400	2400	1866
DDR4-2133	2133	2133	1866	2133	2133	1866
DDR4-1866	1866	1866	1866	1866	1866	1866

It is clear that the 3DPC columns for PRIMERGY BX2560 M2, CX2550 M2 and CX2570 M2 which do not have these slots are not relevant.

The support of 2400 MHz for 2DPC and of 1866 MHz for 3DPC in the case of RDIMMs (marked in red) is a special feature of the Xeon E5-2600 v4 based PRIMERGY servers. The Intel specification makes allowance for 2133 MHz (2DPC) and 1600 MHz (3DPC) in these cases. The current configurator and data sheet of the BIOS version are always decisive for each PRIMERGY model. In order to refer to this special feature, the differentiation between RDIMMs and LRDIMMs has been retained in the table although they are actually unnecessary for the shown target frequencies.

The explanations about DDR4 already pointed out that there are currently no low-voltage versions for DDR4 memory modules. DDR4 modules always run with the voltage 1.2 V. A value which is also lower than the 1.35 V DDR3 low-voltage! The setting *DDR Performance = Low-voltage optimized,* known from earlier versions in this document series, does not exist with Xeon E5-2600 v4 based PRIMERGY servers.

However, a slight amount of power can be saved by reducing the memory frequency. But, please note that the energy consumption of the memory modules is primarily based on the voltage. As the reduction in memory frequency also influences system performance (the scope is described in the second part of this document), a certain care is recommended when making the setting according to the following table. Care means that the effect should be tested before it becomes productive.

DDR Performance = Energy optimized						
CPU type	**RDIMM**			**LRDIMM**		
	1DPC	2DPC	3DPC	1DPC	2DPC	3DPC
DDR4-2400	1866	1866	1866	1866	1866	1866
DDR4-2133	1866	1866	1866	1866	1866	1866
DDR4-1866	1866	1866	1866	1866	1866	1866

BIOS parameters

Having looked at the BIOS parameter *DDR Performance* in the previous section, we now turn to the other BIOS options that affect the memory system. The parameters are in the submenus *Memory Configuration* and *CPU Configuration* underneath *Advanced*.

Memory parameters under Memory Configuration

The following four parameters apply: The default is underlined each time.

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

The first parameter Memory Mode handles the redundancy functions. They are part of the RAS (Reliability, Availability, Serviceability) functionality and increase fail-safety by mirroring the memory (mirroring) or activating the memory spare at the level of DIMM ranks, if memory errors become frequent (sparing). If these functions are requested during the configuration in SystemArchitect, an appropriate default setting is made in the factory. Otherwise, the parameter is set to *Normal* (no redundancy). Quantitative statements about the effect of the redundancy functions on system performance are to be found below.

The NUMA parameter defines whether the physical address space is built from segments of the local memory and whether the operating system is notified about its structure. The default setting is *Enabled* and should not be changed without a convincing reason. There are quantitative statements below for this topic as well.

The third parameter DDR Performance concerns memory frequency and was dealt with in the last section in detail.

The Patrol Scrub parameter is preset with *Enabled*. The main memory is searched in 24-hour cycles looking for correctable memory errors and the correction is initiated where necessary. This prevents memory errors from accumulating (they are counted in the corresponding registers) which can result in states which can no longer be automatically corrected. Highly sensitive performance measurements may be a reason for temporarily disabling this functionality. However, establishing proof of an effect on performance may be difficult.

Memory parameters under CPU Configuration

The submenu *CPU Configuration* is comprehensive and is dealt with in detail in the document *BIOS optimization for Xeon E5-2600 v4 based systems* [L6]. The following three parameters are of particular interest in conjunction with the memory system. The default setting is underlined.

- COD Enable: Disabled / Enabled / Auto
- Early Snoop: Disabled / Enabled / Auto
- Home Snoop Dir OSB: Disabled / Enabled / Auto

These parameters are used to make a selection from the four versions *Home Snoop with Directory and OSB, Cluster-on-die, Home Snoop* and *Early Snoop* of the cache coherence protocol. Such a protocol ensures that - in multiprocessor systems with a shared address area - no inconsistencies result due to the caching of the same memory address in several processor caches.

In the case of the Xeon E5-2600 v4 based systems the default setting results in *Home Snoop with Directory and OSB* and should not be changed without a convincing reason. The difficulty in optimizing the protocol is that the impact on the performance of an application can only be clarified via careful testing. A mere qualitative comparison (rough estimate) with the guidance detailed below is not sufficient.

If there is a deviation from the default for such a test, the protocol versions result by setting the three BIOS parameters according to the following plan:

Protocol	COD Enable	Early Snoop	Home Snoop Dir OSB
Home Snoop Dir OSB	Auto	Auto	Auto
Cluster-on-die	Enabled	Auto	Auto
Home Snoop	Disabled	Disabled	Disabled
Early Snoop	Disabled	Enabled	Disabled

Basically the L3 caches in the system are involved in the cache coherence protocol as well as the memory controller responsible for the respective address. To be more exact, these are subcomponents called Caching Agents (CA) or Home Agents (HA), which exchange messages inside the chip and via the QPI network. The message type, which includes a caching agent that is possibly affected, is referred to as a snoop. Bearing this in mind, the four protocols can be described as follows:

Protocol	Description
Home Snoop Dir OSB	Home Snoop by the responsible HA, who has an extended in-memory directory. Supplemented by precautionary snoops (OSB = Opportunistic Snoop Broadcast), depending on the available QPI bandwidth and other heuristics.
Cluster-on-die	Only for processor models with two memory controllers. For each processor there are two NUMA nodes, each with half the processor cores, half the L3 cache and a memory controller. Snoop by HA with in-memory directory and on-die directory cache. Suitable for loads with excellent NUMA features.
Home Snoop	The activating CA contacts the responsible HA, who specifically sends snoops. Classic behavior for systems with many processors. Relief for the QPI network, but increased latency.
Early Snoop	Precautionary snoops by the activating CA at the earliest possible point in time. Classic behavior for systems with a few processors. High QPI network load, short latencies.

The restriction to processor models with two memory controllers mentioned in the COD (cluster-on-die) description is based on the HCC, MCC, LCC classification (discussed above) according to the number of processor cores. COD is only for processors in the HCC/MCC classes. The processor table is not inserted here as it was shown earlier in the document.

COD was configured for the majority of the standard benchmarks carried out for the Broadwell-EP based PRIMERGY servers; but the other protocols also occurred occasionally. However, we do not recommend any untested copy of the deviations from the BIOS default.

Finally, here is a qualitative statement about how the four protocols affect the features latency and bandwidth of memory performance. A distinction must be made between local and remote memory access. Lower values are better for latencies; higher values are better for bandwidths. If the application requirements regarding memory latencies and bandwidths are known, the table can be helpful in finding the optimal protocol.

Relative Snoop Mode Performance				
Performance metrics	Home Snoop Dir OSB	COD (only HCC and MCC CPUs)	Home Snoop	Early Snoop
L3 Cache hit latency	Low	Lowest	Low	Low
Local memory latency	Low	Lowest	High	Medium
Remote memory latency	Low	Low – High	Low	Lowest
Local memory bandwidth	High	High	High	Low
Remote memory bandwidth	High	Medium	High	Low - Medium

Performant memory configurations

In summary, the memory frequency has so far proven to have the main effect on memory performance. A range of dependencies have been defined which affect the memory frequency of a configuration. Each user should be clear about the memory frequency of his installation.

Furthermore, there are some configuration features which also affect the memory performance: the special features of individual DIMM types, such as the number of ranks; the activation of the redundancy functions; the deactivation of the NUMA functionality; the options of the cache coherence protocol. Although the second part of this document also supplies test results for these topics, they are presumably of little significance for most customer installations and can be bypassed by many users.

Performance Mode configurations

The second factor which should always be observed is the influence of the DIMM placement. There are a range of memory configurations between the minimum configuration (an 8 GB DIMM per configured processor) and the maximum configuration (full configuration with 64 GB DIMMs) which are ideal regarding memory performance. The following table lists the particularly interesting configurations of this type (it is not necessarily complete).

These configurations have identical configurations for all four memory channels per processor. Configuring bank-by-bank is in groups of four DIMMs of the same type. Memory access is then equally distributed over these memory system resources. Technically speaking, the optimum 4-way interleaving is achieved via the memory channels. The PRIMERGY configurator calls these *Performance Mode* configurations.

Performance Mode configurations of Xeon E5-2600 v4 based PRIMERGY servers (configurations with assigned 3rd bank are only possible in PRIMERGY models with 12 DIMMs per processor)							
1 CPU system	**2 CPU system**	**DIMM type**	**DIMM size (GB) bank 1**	**DIMM size (GB) bank 2**	**DIMM size (GB) bank 3**	**Max. MHz**	**Comment**
32 GB	64 GB	DDR4-2400 R	8			2400	2Rx8 variant available (++)
64 GB	128 GB	DDR4-2400 R	16			2400	
96 GB	192 GB	DDR4-2400 R	16	8		2400	Mixed configuration (-)
128 GB	256 GB	DDR4-2400 R	16	16		2400	4-way rank interleave (++)
192 GB	384 GB	DDR4-2400 R	16	16	16	1866	Reduced frequency 3DPC (-)
		DDR4-2400 R	32	16		2400	Mixed configuration (-)
256 GB	512 GB	DDR4-2400 R	32	32		2400	4-way rank interleave (++)
320 GB	640 GB	DDR4-2400 R	32	32	16	1866	Mixed configuration (-)
384 GB	768 GB	DDR4-2400 R	32	32	32	1866	Reduced frequency 3DPC (-)
512 GB	1024 GB	DDR4-2400 LR	64	64		2400	(++)
768 GB	1536 GB	DDR4-2400 LR	64	64	64	1866	Maximum configuration

The table is structured according to the total memory capacity on the far left. The total capacity is defined for the configuration with one or two processors. In a 2-processor situation it is assumed that the memory configuration is the same for both processors. The next columns have the used DIMM type, whereby RDIMM or LRDIMM technology is decisive as well as the GB DIMM size. A specification of the DIMM size per bank is sufficient as these are Performance Mode configurations that are configured in groups of four DIMMs, i.e. bank-by-bank.

The smallest configuration in the table has 64 GB for two processors because the four 8 GB DIMMs (i.e. 32 GB) must be counted for each processor.

The Performance Mode configuration requires an identical DIMM group of four per bank, but it does not forbid different DIMM sizes in different banks if the following restrictions are observed:

- RDIMMs and LRDIMMs must not be mixed.
- RDIMMs of type x4 and x8 must not be mixed.
- The configuration is incrementing from bank 1 to 3 with decreasing DIMM sizes. The larger modules are installed first.

The last column contains comments. For example, the information that mixed configurations can have the decisive performance factor of the 4-way channel interleave, but can drop slightly in comparison to the configurations with a single DIMM type. This is due to the somewhat more complicated addressing within the individual memory channel.

Of course the table also contains the memory configurations from the standard benchmarks executed for the Xeon E5-2600 v4 based PRIMERGY servers. They are highlighted in the comments column with ++.

The last but one table column states the maximum memory frequency that can be reached with the respective configuration. However, reaching the value also depends on the processor model used. This column shows the trade-off between the memory capacity and the memory performance (represented here by the frequency).

Independent Mode configurations

This covers all the configurations that are neither in Performance Mode nor are redundant. There are no restrictions apart from the "Don't mix" ruling for RDIMMs and LRDIMMs as well as for x4 and x8 RDIMMs.

Special attention is also given to configurations with less than four DIMMs per processor, i.e. less than the minimum number that is required for Performance Mode configurations. The reason for such configurations can be energy-saving considerations as well as a low amount of required memory capacity. Savings also result from a minimization of the number of DIMMs. The quantitative assessment that follows below of how a configuration of less than four memory channels impacts on system performance suggests the following recommendations:

- With regard to the LCC processor class (low-core count), operation with only one DIMM per processor (minimum configuration) is not recommended. Operation with two or three DIMMs per processor can on the other hand lead to balanced results as regards performance and energy consumption.
- In the HCC (high-core count) and MCC (medium-core count) processor classes, operation with one or three DIMMs per processor is not recommended. Operation with two DIMMs per processor can on the other hand lead to balanced results as regards performance and energy consumption.

The non-recommended configurations mean entire (1 DIMM per processor) or partial (3 DIMMs per processor for the HCC and MCC processors) 1-way interleaving via the memory channels with the clear performance disadvantage of up to 30%, as shown below, for the commercial application performance. The special feature regarding three DIMMs with HCC and LCC processors results from the configuration with two memory controllers over which three DIMMs cannot be equally distributed.

Symmetric memory configurations

Finally, a separate section is to once again highlight that all configured processors are to be equally configured with memory if possible and the *NUMA = enabled* default setting of the BIOS is not to be changed without a convincing reason. Only in this way is the QPI-based architecture of the systems taken into consideration.

It goes without saying that preinstallation at the factory takes this circumstance into account. The ordered memory modules are distributed as equally as possible across the processors.

These measures and the related operating system support create the prerequisite to run applications as far as possible with a local, high-performance memory. The memory accesses of the processor cores are usually made to DIMM modules, which are directly allocated to the respective processor.

In order to estimate what performance advantage this means, measurement results are listed below in the event that the memory of a 2-way server is indeed symmetrically configured, but where the BIOS option *NUMA = disabled* is set. Statistically, every second memory access is then made to a remote memory. The possible case for asymmetric or single-sided memory configuration that an application is run 100% with a remote memory should be estimated at the double loss in performance of the 50/50% case.

Incidentally, configurations with for example eight DIMMs on the first processor and four on the second processor fulfil the Performance Mode criteria because the memory channels *per processor* are handled identically; that is the thinking of the PRIMERGY order and configuration process. However, such configurations are not recommended.

Quantitative effects on memory performance

After the functional description of the memory system with qualitative information, we now have specific statements about with which gain or loss in performance differences are connected in the memory configuration. As a means of preparation the first section deals with the two benchmarks that were used to characterize memory performance.

This is followed - in order of their impact - by the already mentioned features interleaving of the memory channels, memory frequency, influence of the DIMM types and cache coherence protocol. At the end we then have measurements for the case of *NUMA = disabled* and memory performance under redundancy.

With one exception the quantitative testing is performed separately for the performance classes of the Xeon E5-2600 v4 processor family. The HCC, MCC and LCC classification according to the number of processor cores is usually used; however, when testing the memory frequency the DDR4 frequency classification is more adequate. The exception: The measurements for the various DIMM types were carried out with only one processor of type HCC.

The measurements were made on a PRIMERGY RX2530 M2 with two processors under the Linux operating system. The following table shows the details of the measurement configuration, particularly the representatives used for the processor classes.

System Under Test (SUT)	
Hardware	
Model	PRIMERGY RX2530 M2
Processors	2 × Xeon E5-2698 v4 (HCC, DDR4-2400) 2 × Xeon E5-2660 v4 (MCC, DDR4-2400) 2 × Xeon E5-2630 v4 (LCC, DDR4-2133) 2 × Xeon E5-2609 v4 (MCC, DDR4-1866)
Memory types	8GB (1x8GB) 1Rx4 DDR4-2400 R ECC 8GB (1x8GB) 2Rx8 DDR4-2400 R ECC 16GB (1x16GB) 2Rx4 DDR4-2400 R ECC 16GB (1x16GB) 2Rx8 DDR4-2400 R ECC 16GB (1x16GB) 1Rx4 DDR4-2400 R ECC 32GB (1x32GB) 2Rx4 DDR4-2400 R ECC 64GB (1x64GB) 4Rx4 DDR4-2400 LR ECC
Disk subsystem	1 × HD SATA 6G 1TB 5.4Krpm (via onboard controller for SATA / SAS)
Software	
BIOS	1.4.0
Operating system	Red Hat Enterprise Linux Server release 6.7

The 16 GB 2Rx4 RDIMM was normally used for the test sets described below. In the testing of interleaving across memory channels the 32 GB 2Rx4 RDIMM was used for the test cases with only one or two DIMMs per processor in order to achieve the minimum capacity of main memory required for the tests. All the DIMMS in the table were only used in the test sets on the influence of the DIMM types.

The following tables 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 1.0 measurement of the tables, are included in the Performance Reports of each Xeon E5-2600 v4 based PRIMERGY server.

One essential result of the testing should be made clear from the very beginning. The more powerful the processor model that is used, the greater the performance influence and the more carefully you should weigh up the configuration details. Considerations that are imperative for the most powerful and most expensive processors of the HCC class are frequently negligible for the LCC class.

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 specified in the following to quantify memory performance are based on this practice and are GB/s for the access type Triad.

STREAM is the industry standard for measuring the memory bandwidth of servers, known for its ability to put memory systems under immense stress using simple means. It is clear that this benchmark is particularly suitable for the purpose of studying effects on memory performance in a complex configuration space. In each situation STREAM shows the maximum effect on performance caused by a configuration action which affects the memory, be it deterioration or improvement. The percentages specified below regarding the STREAM benchmark are thus to be understood as bounds for performance effects.

The memory effect on application performance is differentiated between the latency of each access and the bandwidth required by the application. The quantities are interlinked, as real latency increases with increasing bandwidth. The scope in which the latency can be "hidden" by parallel memory access also depends on the application and the quality of the machine codes created by the compiler. As a result, making general forecasts for all application scenarios is very difficult.

SPECint_rate_base2006 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 the memory channels

Interleaving in this conjunction is the set-up of the physical address area by alternating between the four memory channels per processor: the first block is in the first channel, the second in the second, etc. Memory access, which according to the locality principle is mainly to adjacent memory areas, is thus distributed across all channels. This performance gain situation results from parallelism. The blocking size of channel interleaving is based on the *cache line size* of 64 bytes, the unit of memory accesses from the point of view of the processor.

The following table shows the performance disadvantage in the event that the ideal 4-way interleaving, which is achieved with memory configurations in Performance Mode, is not given. The table shows the already highlighted fact that the performance influence is more significant the more powerful the processor.

Benchmark	Processor type	4-way	3-way	2-way	1-way
STREAM	High-core count	1.00		0.52	0.26
	Medium-core count	1.00		0.52	0.26
	Low-core count	1.00	0.84	0.59	0.30
SPECint_rate_base2006	High-core count	1.00		0.87	0.66
	Medium-core count	1.00		0.93	0.77
	Low-core count	1.00	0.99	0.96	0.82

The processor models used for this test (and for the test below with the same classification) were Xeon E5-2698 v4 for HCC, Xeon E5-2660 v4 for MCC and Xeon E5-2630 v4 for LCC. The DIMM type used was the 16 GB 2Rx4 RDIMM in 1DPC configuration.

The statements about SPECint_rate_base2006 are representative for the commercial application performance. 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.

There may be good reasons for 2-way and 3-way interleaving with a moderate loss in performance: a low memory capacity that is needed or minimization in the number of DIMMs in order to save energy. We advise against 1-way interleaving, which is not strictly speaking interleaving and is only referred to as such for the sake of the systematics involved. In this case, the performance potential of processors and memory system are not in a well-balanced relationship to each other.

Due to their configuration with two memory controllers the processors of the classes HCC and MCC do not support 3-way interleaving. Hence the question as to what happens if three DIMM strips per processor are configured with these processor models.

These are then examples for the need to segment the physical address area into segments with different interleaving. Further examples of this necessity are configurations with different partial capacities per memory channel (GB per channel). These can occur if the configuration has DIMMs of a different size, or in the case of configurations with five or more DIMMs of the same size. A common trait of all these examples is that a standardized address area segment cannot be set up by alternating between the memory channels. The alternating must always "work out even". By grouping the existing DIMMs an attempt is made in these cases to generate segments with as high interleaving as possible. The following table provides two examples of this. The notation in the left column of the table shows the number of connected DIMMs for each of the four channels per CPU.

In the case of segmenting the memory performance of an application can then vary, depending on the segment from which the application is provided with memory. In sensitive use cases this phenomenon may be a reason for avoiding configurations with a need for segmenting.

DIMM configuration examples (per CPU) with a need for segmenting	Address area segments	Size / Interleave
1 – 1 – 1 – 0 (HCC and MCC)	1 – 0 – 1 – 0	66% of the address area / 2-way
	0 – 1 – 0 – 0	33% of the address area / 1-way
2 – 1 – 1 – 1	1 – 1 – 1 – 1	80% of the address area / 4-way
	1 – 0 – 0 – 0	20% of the address area / 1-way

Memory frequency

The influences on the effective memory frequency have been dealt with in detail above. Large memory configurations (3DPC configurations) and power-saving (managed via the BIOS parameter *DDR Performance*) can be the reasons why the effective frequency is lower than the maximum one supported by the processor type.

The following tables should be helpful when weighing up these influences against each other. The quantitative statements in the first table are related to the lowest memory frequency of 1866 MHz that is common to all series of measurements. The second table shows the same information from a different perspective. Here the statements refer to the respective ideal case, the highest possible frequency per processor class.

Benchmark	Processor type	1866 MHz	2133 MHz	2400 MHz
STREAM	DDR4-2400	1.00	1.11	1.21
	DDR4-2133	1.00	1.07	
	DDR4-1866	1.00		
SPECint_rate_base2006	DDR4-2400	1.00	1.01	1.02
	DDR4-2133	1.00	1.01	
	DDR4-1866	1.00		

Benchmark	Processor type	1866 MHz	2133 MHz	2400 MHz
STREAM	DDR4-2400	0.83	0.92	1.00
	DDR4-2133	0.93	1.00	
	DDR4-1866	1.00		
SPECint_rate_base2006	DDR4-2400	0.98	0.99	1.00
	DDR4-2133	0.99	1.00	
	DDR4-1866	1.00		

The processor models used for this testing were Xeon E5-2660 v4 for DDR4-2400, Xeon E5-2630 v4 for DDR4-2133 and Xeon E5-2609 v4 for DDR4-1866. The DIMM type used was the 16 GB 2Rx4 RDIMM in 1DPC configuration.

The BIOS setting *DDR Performance* = *Energy optimized* always results in a frequency with 1866 MHz. However, the potential for the power savings that can thus be obtained is very low, as the energy consumption results primarily from the DIMM voltage and not so much from the memory frequency. The voltage with the new DDR4 modules is always 1.2 V, a lower value and thus more energy-saving than the minimum 1.35 V of the DDR3 generation. That is why the *Energy optimized* setting is not to be recommended.

If a reduced memory frequency is connected to the memory capacity, one issue should for the sake of completeness also be mentioned. The memory capacity can have an implicit influence on application performance, for example in the form of I/O rates. Such an influence is of course not taken into account in

the testing on which this section is based. In the comparisons in the table the different memory frequency is the only influence on performance.

Influence of the DIMM types

At the time of the general release seven DIMM types are planned for the Xeon E5-2600 v4 based PRIMERGY servers. However, reference is made to the respective configurator for exceptions and special features of specific servers.

The following table shows the differences in performance between these DIMM types under otherwise identical conditions:

- All measurements were carried out with the HCC processor model Xeon E5-2698 v4. This powerful processor model highlights the differences in performance most clearly. Since it is a matter of comparatively fine differences, the series of measurements was by way of exception carried out with this model only.
- It is evident that with these measurements all the memory channels were equally configured, i.e. *Performance Mode* configurations were compared. 8 DIMMs were installed for 1DPC measurements and 16 DIMMs for 2DPC measurements.
- All the measurements were carried out with the consistent memory frequency 2400 MHz. On account of this general condition it was only possible to consider 1DPC and 2DPC configurations.
- The table is standardized to the 2DPC configuration with the 16 GB 2Rx4 RDIMM (highlighted in bold print), which currently provides the best memory performance. This DIMM is preferred in benchmarking as long as the memory capacity that can be achieved with it is sufficient.

DIMM type	Configuration	STREAM	SPECint_rate_base2006
8GB (1x8GB) 1Rx4 DDR4-2400 R ECC	1DPC	0.87	0.97
	2DPC	0.98	0.99
8GB (1x8GB) 2Rx8 DDR4-2400 R ECC	1DPC	0.98	0.99
	2DPC	1.00	1.00
16GB (1x16GB) 1Rx4 DDR4-2400 R ECC	1DPC	0.86	0.96
	2DPC	0.97	0.99
16GB (1x16GB) 2Rx4 DDR4-2400 R ECC	1DPC	0.98	0.99
	2DPC	**1.00**	**1.00**
16GB (1x16GB) 2Rx8 DDR4-2400 R ECC	1DPC	0.96	0.99
	2DPC	0.99	0.99
32GB (1x32GB) 2Rx4 DDR4-2400 R ECC	1DPC	0.97	0.99
	2DPC	0.99	0.99
64GB (1x64GB) 4Rx4 DDR4-2400 LR ECC	1DPC	0.95	0.98
	2DPC	0.91	0.98

The main cause of the differences in performance shown here is another form of interleaving. The method of alternating across memory resources when setting up the physical address space can be continued from interleaving across the memory channels to interleaving across the ranks in a channel.

Rank interleaving is controlled directly via address bits. The bit arithmetic performed in channel interleaving to establish the 3-way case is not carried out. 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 always 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 number of ranks per memory channel follows from the DIMM type and the DPC value of the configuration. The 1DPC configurations with dual-rank DIMMs in the table, for example, allow a 2-way rank interleave, whereas 2DPC configurations allow a 4-way interleave.

The granularity of the rank interleaving is larger than with interleaving across the 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.

2-way and 4-way rank interleaving provides very good memory performance. The minute additional advantage of 4-way interleaving only plays a role if we are dealing with the very last ounce of performance. It can usually be ignored.

The most striking performance disadvantages in the table, for example with the 8 GB 1Rx4 RDIMM in 1DPC, can be explained by the lack of a rank interleave. Except for 1DPC configurations with single-rank DIMMs this case can also occur with mixed configurations, for example with the 16 GB 2Rx4 RDIMM in the first bank and the 8 GB 1Rx4 RDIMM in the second bank. In these cases of a missing or 1-way rank interleave you should be aware of a certain performance disadvantage. This case should be avoided in sensitive use cases, particularly with powerful processor models.

In addition to the main influence of rank interleave, a few other subtle influences are incorporated in the results table. For example, with more than four ranks in the memory channel the overhead that is to be performed per rank for the DRAM refresh 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. This explains the above mentioned relationships with the 4Rx4 LRDIMM, the only case in which a 2DPC configuration is poorer than the corresponding 1DPC.

Optimization of the cache coherence protocol

The ability to select between the three versions *Early Snoop*, *Home Snoop* and *Cluster-on-die* of the protocol for the cache coherence belonged to the new features of the predecessor generation Haswell-EP. The default in Haswell-EP was *Early Snoop*. There is an additional fourth option in Broadwell-EP, namely *Home Snoop with Directory and OSB*, which is simultaneously the new default. The explanations are in the section about the BIOS options for the memory system.

The following table shows the effect on the two loads or benchmarks examined in this document. The table is standardized to the performance of the new BIOS default *Home Snoop with Directory and OSB*.

The measurements are made in 1DPC configurations with 16 GB 2Rx4 RDIMMs.

The table shows that it is fine tuning within a range of a few percentage points. When evaluating this table it should be considered that both benchmarks are extremely NUMA friendly due to careful process binding during test setup. The model character of SPECint_rate_base2006 for commercial application performance therefore only applies at this stage in a restricted manner.

It should be recalled that processors of type LCC with only one memory controller do not support the protocol *Cluster-on-die*.

Benchmark	Processor type	Home Snoop Dir OSB	Cluster-on-die	Home Snoop	Early Snoop
STREAM	High-core count	1.00	1.02	0.99	0.87
	Medium-core count	1.00	1.02	0.99	0.94
	Low-core count	1.00		0.82	0.80
SPECint_rate_base2006	High-core count	1.00	1.02	0.98	0.98
	Medium-core count	1.00	1.01	0.98	0.99
	Low-core count	1.00		0.98	1.00

Access to remote memory

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

The following table shows the effect of the BIOS setting *NUMA = disabled* in the case of an otherwise ideal memory configuration, i.e. a 4-way rank-interleaved Performance Mode configuration with 16 GB 2Rx4 RDIMMs under the highest possible memory frequency per processor type. The deterioration in performance occurs because statistically every second memory access is to a remote DIMM, i.e. a DIMM allocated to the neighboring processor, and the data must make a detour via the QPI link.

Benchmark	Processor type	NUMA = enabled	NUMA = disabled
STREAM	High-core count	1.00	0.78
	Medium-core count	1.00	0.77
	Low-core count	1.00	0.80
SPECint_rate_base2006	High-core count	1.00	0.93
	Medium-core count	1.00	0.94
	Low-core count	1.00	0.94

The physical address space is set up for *NUMA = disabled* by means of a fine-mesh alternating between the processors. This alternating presumes the same memory capacity in both processors. If this general condition does not exist, the address space is then split into a main part, which permits the inter-socket interleaving, and a processor-local remaining part.

The experiment with the setting *NUMA = disabled* was performed to a lesser extent because of the exceptional cases, in which this setting is recommended, because the NUMA support in system software or system-related software is missing or unsatisfactory. The experiment is above all useful in estimating the effect when most or all accesses are to remote memory. This situation can occur if the memory capacities configured per processor greatly vary. The loss in performance compared with local access can then be up to twice the amount of the loss specified in the table.

This table shows that the percentage loss is not quite as regular as in earlier tables. The rule of thumb - the more powerful the processor, the greater the loss - does not necessarily apply here. This is due to the influence of the different QPI frequencies between 6.4 and 9.6 GT/s for the processor types. The QPI links are the bottleneck of bandwidth measurement with disabled NUMA support.

Memory performance under redundancy

There are two redundancy options for the Xeon E5-2600 v4 based PRIMERGY servers.

All four memory channels of a processor are identically configured with mirroring (the configuration rule matches the one for the Performance Mode configurations), but two channels mirror the other two. 50% of the actually configured memory is available to the operating system.

For sparing, or more precisely rank sparing, one rank per memory channel is the unused reserve in case an active rank is taken out of operation as a precaution because of accumulating memory errors. The net memory capacity available for the operating system depends in this case on the DIMM type and DPC value. The exact calculation as well as the general conditions of the sparing mode DIMM configurations are in the configurators of the respective PRIMERGY models.

The table shows the effect if the redundancy options are activated in the event of an otherwise ideal memory configuration, i.e. a Performance Mode 2DPC configuration with 16 GB 2Rx4 RDIMMs under maximum memory frequency in each case. The columns in the table correspond to the options of the BIOS parameter *Memory Mode*.

The loss that occurred under mirroring is smaller than with 2-way channel interleaving, because both halves of the mirror can be used for read access.

The relationships under sparing are drawn from the rank interleave that was dealt with above in the section on the influence of DIMM types. The reserve rank usually results in an odd number of active ranks per memory channel and thus in the 1-way rank interleave. The column Sparing is due to the difference between a 4-way and a 1-way rank interleave.

Benchmark	Processor type	Normal	Mirroring	Sparing
STREAM	High-core count	1.00	0.68	0.87
	Medium-core count	1.00	0.69	0.90
	Low-core count	1.00	0.77	0.93
SPECint_rate_base2006	High-core count	1.00	0.96	0.97
	Medium-core count	1.00	0.99	0.98
	Low-core count	1.00	0.99	0.99

Literature

PRIMERGY Servers

[L1] http://primergy.com/

Memory Performance

[L2] This White Paper:
http://docs.ts.fujitsu.com/dl.aspx?id=8f372445-ee63-4369-8683-da9557673357
http://docs.ts.fujitsu.com/dl.aspx?id=5c5111c5-7d23-4c2c-aa3f-88c7053fe41a
http://docs.ts.fujitsu.com/dl.aspx?id=3ce313c6-6713-4350-880c-16959489a510

[L3] Memory Performance of Xeon E5-2600 v3 (Haswell-EP) based Systems
http://docs.ts.fujitsu.com/dl.aspx?id=74eb62e6-4487-4d93-be34-5c05c3b528a6

Benchmarks

[L4] STREAM
http://www.cs.virginia.edu/stream/

[L5] SPECcpu2006
http://docs.ts.fujitsu.com/dl.aspx?id=1a427c16-12bf-41b0-9ca3-4cc360ef14ce

BIOS Settings

[L6] BIOS optimizations for Xeon E5-2600 v4 based systems
http://docs.ts.fujitsu.com/dl.aspx?id=eb90c352-8d98-4f5a-9eed-b5aade5ccae1

PRIMERGY Performance

[L7] http://www.fujitsu.com/fts/x86-server-benchmarks

Contact

FUJITSU

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

PRIMERGY Product Marketing

mailto:Primergy-PM@ts.fujitsu.com

PRIMERGY Performance and Benchmarks

mailto:primergy.benchmark@ts.fujitsu.com

2016-03-31 WW EN