

White Paper

FUJITSU Server PRIMERGY & PRIMEQUEST Memory Performance of Xeon scalable processor (Cascade Lake-SP) based system memory performance

In Xeon Scalable Processor based FUJITSU Server PRIMERGY and PRIMEQUEST, with the introduction of the new Ultra Path Interconnect (UPI) and memory architecture improvement, performance has improved remarkably. 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.

Version

1.0

2019-07-05



Table of contents

Document History	2
Introduction	3
Memory architecture	4
DIMM slots and memory controllers	4
DDR4 topics and available DIMM types	8
Definition of the memory frequency	10
BIOS parameters	13
Memory parameters under Memory Configuration	13
Performant memory configurations	14
Performance Mode configurations	14
Independent Mode configurations	15
Symmetric memory configurations	16
Quantitative effects on memory performance	17
The measuring tools	18
STREAM Benchmark	18
SPECrate2017_int_base Benchmark	18
Interleaving across the memory channels	19
Memory frequency	21
Influence of the DIMM types	22
Optimization of the cache coherence protocol	24
Access to remote memory	24
Memory performance under redundancy	25
Literature	26
Contact	26

Document History

Version 1.0 (2019-07-05)

Initial version

Introduction

The Intel Xeon Scalable Processor (Skylake-SP and Cascade Lake-SP) installed in the current PRIMERGY and PRIMEQUEST servers is produced using the same 14 nm manufacturing process as the old generation Broadwell-EP, but the processor microarchitecture and chipset have been updated to improve performance.

A major factor in the Xeon Scalable Processor generations (Skylake-SP and Cascade Lake-SP) performance improvement is that the maximum number of cores per processor has increased from 22 (Broadwell-EP) to 28. In the memory system as well, there is a new function that contributes to the improvement of the new generation performance.

The DDR4 memory technology introduced in the Haswell-EP generation three generations ago is still used in the current Cascade Lake-SP system, but while the maximum memory frequency for the Broadwell-EP generation two generations ago was 2400 MHz, and 2666 MHz for the previous generation Skylake-SP, the Cascade Lake-SP system supports 2933 MHz, which is a new feature. The connection between CPUs was improved to Ultra Path Interconnect (UPI) from Quickpath Interconnect (QPI) used in the previous generation Broadwell-EP, the link maximum frequency that was up to 9.6 GT/s in Broadwell-EP is supported up to 10.4 GT/s.

The cache coherence protocol option known as *Cluster-on-die* in the old generation is available as Sub-NUMA Clusters (SNC) for the Xeon Scalable Processor generations (Skylake-SP and Cascade Lake-SP). This option handles latency and bandwidth trade-offs for local and remote memory access differently, but in most applications, except for particularities in the tests for small performance differences as well, it is not necessary to have settings that deviate from the default settings.

In other respects, by replacing the old generation QPI with UPI, idle power consumption and data efficiency have been improved. The processor is equipped with *on-chip* memory controllers, and each processor controls a group of memory modules allocated to each processor. The performance of this local memory access is very high. When this processor requests the contents of the memory (remote memory) of the adjacent processor, it uses the UPI link. The performance of remote memory access is not quite high. This architecture, which distinguishes between local memory and remote memory access, is a Non-Uniform Memory Access (NUMA) type of architecture.

Compared to the memory system function of the old generation Broadwell-EP, the Xeon Scalable processor generations (Skylake-SP and Cascade Lake-SP) has the number of memory channels of each processor increased from four to six. Since the number of DIMM slots in each channel has been reduced from three to two, the maximum number of DIMMs per processor remains at 12. However, unlike the old generation, even if the number of DIMMs per channel increases, the memory frequency no longer decreases¹, therefore there is no need to consider the trade-off between memory capacity and maximum memory bandwidth, and the peak memory bandwidth when the maximum memory is installed is 141 GB/s.

In this document, we will look at the new memory system function of the latest server generation. On the other hand, as in the earlier issues, this document also provides basic knowledge about the UPI-based memory architecture which is essential when configuring powerful systems. We are dealing with the following points here:

- Due to the NUMA architecture each processor 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 and further speed it up, the adjacent area of the physical address space is distributed to several components of the memory system. In technical terms, this is called *interleaving*. Interleaving is done in two dimensions. First, there are six memory channels per processor in a horizontal direction. Optimal interleaving in this direction is achieved by setting the number of DIMMs installed in each processor to a multiple of six. In addition, interleaving among individual memory channels is realized. The definitive memory resource for this is the so-called number of ranks. The number of ranks is a DIMM sub-structure, and a group of DRAM (Dynamic Random Access Memory) chips are integrated here. Individual memory access always refers to such groups.
- Memory frequency affects performance. Depending on the processor type, DIMM type, memory capacity, and BIOS settings, they can be either 2933, 2666, 2400, 2133 or 1866 MHz.

In this white paper, factors that affect memory performance are taken up and quantified. For quantification, we use the STREAM and SPECrate2017_int_base benchmarks. STREAM measures the memory bandwidth. SPECrate2017_int_base is used as a model for the performance of commercial applications.

¹ In the case of 2DPC configuration with large-capacity 3DS LRDIMMs, the memory frequency drops to 2666 MHz.

Results show that the influences depend on the performance of the processors by ratio. 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

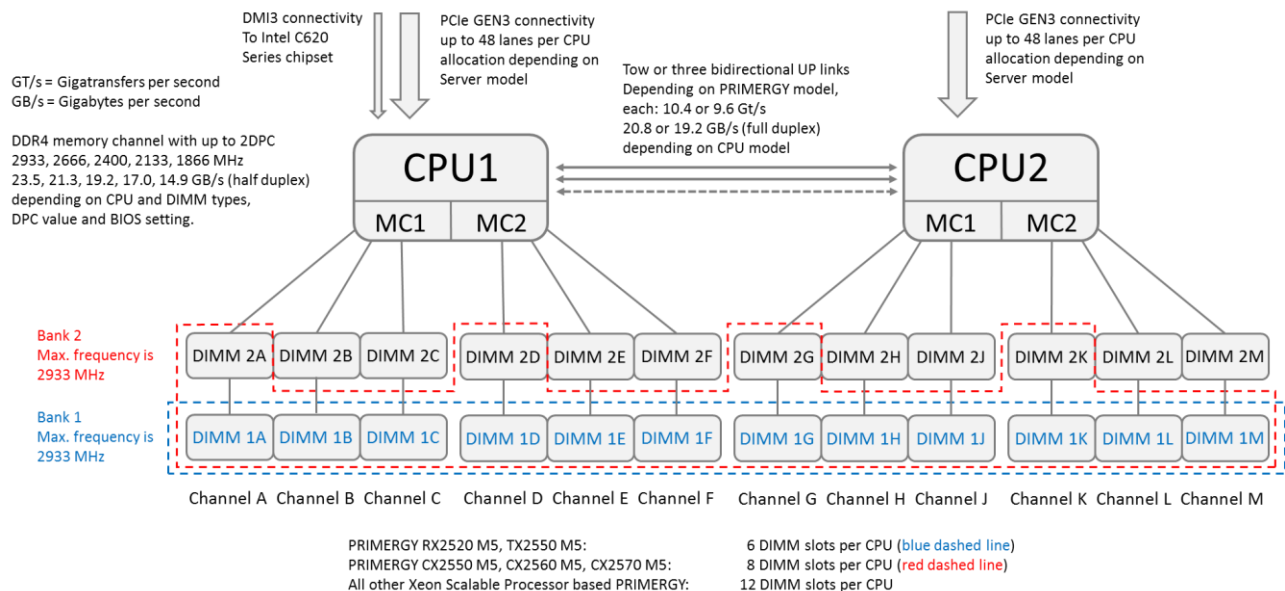
This section explains the outline of the memory system with five parts. First, we will explain the arrangement of available DIMM slots in the block diagram. The second section shows the available DIMM types. The following third section describes the effect on the effective memory frequency. The fourth section describes the BIOS parameters that affect the memory system. The last section lists examples of memory performance optimized DIMM configuration.

DIMM slots and memory controllers

The following figure shows the memory system architecture.

PRIMERGY and PRIMEQUEST servers with Xeon Scalable Processors usually have 12 DIMM slots per processor. The exception is the PRIMERGY CX2550 M4 model, the CX2560 M4 model, and the CX2570 M4 model, and because of the high density of the form factor, the number of slots is eight. Another exception is the PRIMERGY RX2520 M4 model and the TX2550 M4 model. Since these systems are cost optimized, there is only 1 DIMM slot in each of the six memory channels per processor. As a result, this becomes six DIMM slots per processor. Looking at the memory channel of the resource and the UPI link in the figure below, there is a correlation between frequency and bandwidth, which depends on the actual data path width of each of these. They are 64 bits on the DDR4 memory channel and 20 bits on the UPI link. For bidirectional UPI links, the bandwidth is called full duplex because it is valid in each direction. On memory channels, read/write accesses must share data path, therefore it is called a half duplex method.

Memory Architecture of Xeon Scalable Processor based PRIMERGY Servers



For frequency and bandwidth details, consider the type of processor shown in the figure below.

There are six memory channels in one processor. In the Broadwell-EP generation, when the value of DPC (the term is used hereinafter), which is the number of DIMMs per channel, changes, the memory frequency changes, and furthermore the memory performance is affected. However, with Skylake-SP and Cascade Lake-SP, DPC does not reduce the memory frequency².

² In the case of 2DPC configuration with large-capacity 3DS LRDIMMs, the memory frequency drops to 2666 MHz.

We also use the term “memory bank” in the following. In the figure, a group of six or two DIMMs (PRIMERGY CX2550 M5, CX2560 M5, and CX2570 M5) distributed to multiple channels forms one bank. When distributing DIMMs via available slots per processor, allocating them sequentially from bank 1 provides optimal interleaving across the entire channel. Interleaving is the main factor affecting memory performance.

The corresponding processor must be available in order to use the DIMM slots. If CPU installation does not have the maximum configuration, slots assigned to empty CPU sockets cannot be used.

The number of memory channels per processor is common to all Xeon Scalable Processor families. In the Xeon E5 processor family, from Nehalem to Sandy Bridge, there was only one memory controller in the dual socket server EP processor. In Ivy Bridge, two memory controllers were installed in the two most powerful processor models for the first time (though there were very few groups). In the Haswell and Broadwell generations, two memory controllers are installed in all the processors of the model whose number of processor cores is medium to high, and in the latest Xeon Scalable Processor Family, two memory controllers are installed in all processors, including subordinate models.

Refer to the following table for the exact classification of processors.

Processors (since system release)								
Processor	Cores	Threads	Cache [MB]	UPI Speed [GT/s]	Nominal Frequency [GHz]	Max. Turbo Frequency [GHz]	Max. Memory Frequency [MHz]	TDP [Watt]
Platinum 8280L	28	56	38.5	10.4	2.7	4.0	2933	205
Platinum 8280M	28	56	38.5	10.4	2.7	4.0	2933	205
Platinum 8280	28	56	38.5	10.4	2.7	4.0	2933	205
Platinum 8276L	28	56	38.5	10.4	2.3	4.0	2933	165
Platinum 8276M	28	56	38.5	10.4	2.3	4.0	2933	165
Platinum 8276	28	56	38.5	10.4	2.3	4.0	2933	165
Platinum 8270	26	52	35.8	10.4	2.7	4.0	2933	205
Platinum 8268	24	48	35.8	10.4	2.9	3.9	2933	205
Platinum 8260L	24	48	35.8	10.4	2.4	3.9	2933	165
Platinum 8260M	24	48	35.8	10.4	2.4	3.9	2933	165
Platinum 8260Y	24	48	33.0	10.4	2.4	3.9	2933	165
	20	40						
	16	32						
Platinum 8260	24	48	35.8	10.4	2.4	3.9	2933	165
Platinum 8256	4	8	16.5	10.4	3.8	3.9	2933	105
Platinum 8253	16	32	22.0	10.4	2.2	3.0	2933	125
Gold 6262V	24	48	33.0	10.4	1.9	3.6	2933	130
Gold 6254	18	36	24.8	10.4	3.1	4.0	2933	200
Gold 6252	24	48	33.0	10.4	2.1	3.7	2933	150
Gold 6248	20	40	27.5	10.4	2.5	3.9	2933	150
Gold 6246	12	24	24.8	10.4	3.3	4.2	2933	165
Gold 6244	8	16	11.0	10.4	3.6	4.4	2933	150
Gold 6242	16	32	22.0	10.4	2.8	3.9	2933	150
Gold 6240L	18	36	24.8	10.4	2.6	3.9	2933	150
Gold 6240M	18	36	24.8	10.4	2.6	3.9	2933	150
Gold 6240Y	18	36	24.8	10.4	2.6	3.9	2933	150
	14	28						
	8	16						
Gold 6240	18	36	24.8	10.4	2.6	3.9	2933	140
Gold 6238M	22	44	30.3	10.4	2.1	3.7	2933	140
Gold 6238L	22	44	30.3	10.4	2.1	3.7	2933	140
Gold 6238	22	44	30.3	10.4	2.1	3.7	2933	140
Gold 6234	8	16	24.8	10.4	3.4	4.0	2933	130
Gold 6230	20	40	27.5	10.4	2.1	3.9	2933	125
Gold 6226	12	24	19.3	10.4	2.8	3.7	2933	125
Gold 6222V	20	40	27.5	10.4	1.8	3.6	2933	115
Gold 6212U	24	48	33.0	10.4	2.4	3.9	2933	165
Gold 6210U	20	40	27.5	10.4	2.5	3.9	2933	150
Gold 6209U	20	40	27.5	10.4	2.1	3.9	2933	125

Gold 5222	4	8	16.5	10.4	3.8	3.9	2666	105
Gold 5220S	18	36	24.8	10.4	2.2	3.9	2666	125
Gold 5220	18	36	24.8	10.4	2.2	3.9	2666	125
Gold 5218B	16	32	22.0	10.4	2.3	3.9	2666	125
Gold 5218	16	32	22.0	10.4	2.3	3.9	2666	125
Gold 5217	8	16	11.0	10.4	3.0	3.7	2666	115
Gold 5215L	10	20	13.8	10.4	2.5	3.4	2666	85
Gold 5215M	10	20	13.8	10.4	2.5	3.4	2666	85
Gold 5215	10	20	13.8	10.4	2.5	3.4	2666	85
Silver 4216	16	32	22.0	9.6	2.1	3.2	2400	100
Silver 4215	8	16	11.0	9.6	2.5	3.5	2400	85
Silver 4214Y	12	24	16.5	9.6	2.2	3.2	2400	85
	10	20						
	8	16						
Silver 4214	12	24	16.5	9.6	2.2	3.2	2400	85
Silver 4210	10	20	13.8	9.6	2.2	3.2	2400	85
Silver 4208	8	16	11.0	9.6	2.1	3.2	2400	85
Bronze 3204	6	6	8.3	9.6	1.9		2133	85

The quantitative memory performance tests were performed based on processor class—platinum, gold, silver, or bronze, according to the topic—or based on the supported memory frequency as listed in the second-to-last column of the table.

DDR4 topics and available DIMM types

The DDR4 SDRAM memory module is used for PRIMERGY and PRIMEQUEST servers with Cascade Lake-SP installed. The shift from DDR3 to DDR4 was done with Haswell-EP three generations ago. 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.

DDR4 technology, which is mainstream now, has an important difference compared with 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 a memory frequency up to 3200 MHz. This frequency range will be used in several generations of servers in the future. As with the DDR3-based server generation, the frequency continually rises by 266 MHz. For Cascade Lake-SP, this has reached 2933 MHz. The system with Broadwell-EP was supported up to 2400 MHz and the system with Skylake-SP was supported up to 2666 MHz..
- An important merit of DDR4 is that DIMMs operate at only 1.2V. This was 1.5V or 1.35V (low voltage version) in DDR3. This corresponds to a saving of about 30 % power consumption when the data transfer rate is the same.
- Just as in the first phase of DDR3 technology, there is currently no low voltage version in 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. Nevertheless, these trade-offs play an important role for the processor.

The following table shows the DIMMs supported by PRIMERGY and PRIMEQUEST servers with Xeon Scalable Processors. In DIMM, there are Registered DIMM (RDIMM), Load Reduced DIMM (LRDIMM), 3DS Load Reduced DIMM (3DS LRDIMM) types. The mixed configurations are only possible within the three sections of the table. RDIMM, LRDIMM and 3DS LRDIMM cannot be mixed.,

DIMM type	Control	Maximum frequency (MHz)	Volt (V)	# of Ranks	Capacity	Rel. price per GB
8GB (1x8GB) 1Rx8 DDR4-2933 R ECC	Registered	2933	1.2	1	8 GB	0.92
16GB (1x16GB) 2Rx8 DDR4-2933 R ECC	Registered	2933	1.2	1	16 GB	0.98
16GB (1x16GB) 1Rx4 DDR4-2933 R ECC	Registered	2933	1.2	2	16 GB	0.98
32GB (1x32GB) 2Rx4 DDR4-2933 R ECC	Registered	2933	1.2	2	32 GB	1.00
64GB (1x64GB) 2Rx4 DDR4-2933 R ECC	Registered	2933	1.2	2	64 GB	1.00
64GB (1x64GB) 4Rx4 DDR4-2933 LR ECC	Load Reduced	2933	1.2	4	64 GB	1.34
128GB (1x128GB) 8Rx4 DDR4-2933 3DS LR ECC	3DS Load Reduced	2933	1.2	8	128 GB	- ³

For any DIMM type, the data is transferred in 64-bit units. This is a feature of the DDR-SDRAM memory technology. A 64-bit bandwidth memory area is set on the DIMM from a group of DRAM chips. This individual chip is responsible for 4 bits or 8 bits (see code x4 or x8 for type name). Such a chip group is called a *rank*. As shown in the table, there are DIMM types of one, two, four, or eight ranks. While the

³ This DIMM type is not supported by PRIMERGY RX2540 M5.

advantage of the eight rank DIMM is its maximum capacity, at the same time the DDR4 specification only support up to eight ranks per memory channel. The number of available ranks per memory channel has a certain effect on performance. This will be described later.

That being said, the essential features of the three 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 2DPC (DIMMs per channel).
- LRDIMM: Apart from control commands, the data itself is also buffered in the components on the DIMM. In addition, with this DIMM type "*rank multiplication*" function, you can map some physical ranks to virtual ranks. Therefore, the memory controller only monitors the virtual rank. This function is valid when the number of physical ranks in the memory channel exceeds eight.
- 3DS DIMM: This is a DIMM with multiple silicon dies laminated by Through Silicon Via technology based on the Three Dimensional Stack (3DS) standard. Only one die called a master exchanges signals with the outside, and the other dies adopt an architecture that exchanges signals only with the master as a slave, enabling higher capacity and higher speed.

3DS LRDIMM described in the table is a DIMM with combination of the 3DS DIMM technology and the LRDIMM technology.

Which type of RDIMM, LRDIMM, or 3DS LRDIMM is desirable is usually determined by the memory capacity required. But LRDIMM and 3DS LRDIMM have a little overhead in performance.

The last column of the table shows the price of each DIMM in relative ratio. This price is based on PRIMERGY RX 2540 M4's price list as of June 2019. Here we show the price ratio per GB based on the 32 GB 2Rx4 RDIMM (highlighted as 1.0). Compared with the previous issues of this document series, you can see that the relative memory price has always changed.

Depending on the PRIMERGY and PRIMEQUEST models, some DIMM types may not be available. Always refer to the latest configurator. In addition, depending on the sales area, there are DIMM types that cannot be used.

Definition of the memory frequency

There are five types of memory frequencies: 2933, 2666, 2400, 2133 and 1866 MHz. 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 Scalable Processor models according to the last but one column of the following table already shown above. The column shows the maximum supported memory frequency.

Processors (since system release)								
Processor	Cores	Threads	Cache [MB]	UPI Speed [GT/s]	Nominal Frequency [GHz]	Max. Turbo Frequency [GHz]	Max. Memory Frequency [MHz]	TDP [Watt]
Platinum 8280L	28	56	38.5	10.4	2.7	4.0	2933	205
Platinum 8280M	28	56	38.5	10.4	2.7	4.0	2933	205
Platinum 8280	28	56	38.5	10.4	2.7	4.0	2933	205
Platinum 8276L	28	56	38.5	10.4	2.3	4.0	2933	165
Platinum 8276M	28	56	38.5	10.4	2.3	4.0	2933	165
Platinum 8276	28	56	38.5	10.4	2.3	4.0	2933	165
Platinum 8270	26	52	35.8	10.4	2.7	4.0	2933	205
Platinum 8268	24	48	35.8	10.4	2.9	3.9	2933	205
Platinum 8260L	24	48	35.8	10.4	2.4	3.9	2933	165
Platinum 8260M	24	48	35.8	10.4	2.4	3.9	2933	165
Platinum 8260Y	24	48	33.0	10.4	2.4	3.9	2933	165
	20	40						
	16	32						
Platinum 8260	24	48	35.8	10.4	2.4	3.9	2933	165
Platinum 8256	4	8	16.5	10.4	3.8	3.9	2933	105
Platinum 8253	16	32	22.0	10.4	2.2	3.0	2933	125
Gold 6262V	24	48	33.0	10.4	1.9	3.6	2933	130
Gold 6254	18	36	24.8	10.4	3.1	4.0	2933	200
Gold 6252	24	48	33.0	10.4	2.1	3.7	2933	150
Gold 6248	20	40	27.5	10.4	2.5	3.9	2933	150
Gold 6246	12	24	24.8	10.4	3.3	4.2	2933	165
Gold 6244	8	16	11.0	10.4	3.6	4.4	2933	150
Gold 6242	16	32	22.0	10.4	2.8	3.9	2933	150
Gold 6240L	18	36	24.8	10.4	2.6	3.9	2933	150
Gold 6240M	18	36	24.8	10.4	2.6	3.9	2933	150
Gold 6240Y	18	36	24.8	10.4	2.6	3.9	2933	150
	14	28						
	8	16						
Gold 6240	18	36	24.8	10.4	2.6	3.9	2933	140
Gold 6238M	22	44	30.3	10.4	2.1	3.7	2933	140
Gold 6238L	22	44	30.3	10.4	2.1	3.7	2933	140
Gold 6238	22	44	30.3	10.4	2.1	3.7	2933	140
Gold 6234	8	16	24.8	10.4	3.4	4.0	2933	130
Gold 6230	20	40	27.5	10.4	2.1	3.9	2933	125

Gold 6226	12	24	19.3	10.4	2.8	3.7	2933	125
Gold 6222V	20	40	27.5	10.4	1.8	3.6	2933	115
Gold 6212U	24	48	33.0	10.4	2.4	3.9	2933	165
Gold 6210U	20	40	27.5	10.4	2.5	3.9	2933	150
Gold 6209U	20	40	27.5	10.4	2.1	3.9	2933	125
Gold 5222	4	8	16.5	10.4	3.8	3.9	2666	105
Gold 5220S	18	36	24.8	10.4	2.2	3.9	2666	125
Gold 5220	18	36	24.8	10.4	2.2	3.9	2666	125
Gold 5218B	16	32	22.0	10.4	2.3	3.9	2666	125
Gold 5218	16	32	22.0	10.4	2.3	3.9	2666	125
Gold 5217	8	16	11.0	10.4	3.0	3.7	2666	115
Gold 5215L	10	20	13.8	10.4	2.5	3.4	2666	85
Gold 5215M	10	20	13.8	10.4	2.5	3.4	2666	85
Gold 5215	10	20	13.8	10.4	2.5	3.4	2666	85
Silver 4216	16	32	22.0	9.6	2.1	3.2	2400	100
Silver 4215	8	16	11.0	9.6	2.5	3.5	2400	85
Silver 4214Y	12	24	16.5	9.6	2.2	3.2	2400	85
	10	20						
	8	16						
Silver 4214	12	24	16.5	9.6	2.2	3.2	2400	85
Silver 4210	10	20	13.8	9.6	2.2	3.2	2400	85
Silver 4208	8	16	11.0	9.6	2.1	3.2	2400	85
Bronze 3204	6	6	8.3	9.6	1.9		2133	85

In Cascade Lake-SP, except the case using 128 GB 3DS LRDIMMs, the DPC value of the memory configuration does not affect the memory frequency, while the processor type has a big influence on the memory frequency. This cannot be disabled in the BIOS. However, by using the BIOS parameter DDR Performance, you can choose whether to give priority to either performance or power consumption, although limited, as described in detail later. When you select performance, the valid memory frequency is as shown in the following table. This is the default BIOS setting.

DDR Performance = Performance optimized (Default)						
CPU type	RDIMM		LRDIMM		3DS LRDIMM	
	1DPC	2DPC	1DPC	2DPC	1DPC	2DPC
DDR4-2933	2933	2933	2933	2933	2933	2666
DDR4-2666	2666	2666	2666	2666	2666	2666
DDR4-2400	2400	2400	2400	2400	2400	2400
DDR4-2133	2133	2133	2133	2133	2133	2133

As mentioned earlier, DDR4 memory modules do not currently have a low voltage version. The DDR4 module always operates at a voltage of 1.2 V. This is lower than the low voltage version 1.35 V DDR3. The setting of “DDR Performance = Low-voltage optimized” introduced in the previous issue of this document series is not found in PRIMERGY and PRIMEQUEST servers with Xeon Scalable Processors.

Slight power consumption can be saved by lowering the memory frequency, but be aware that the power consumption of the memory module is affected mainly by 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. Pay attention to the impact to the test before production.

DDR Performance = Energy optimized						
CPU type	RDIMM		3DS RDIMM		LRDIMM	
	1DPC	2DPC	1DPC	2DPC	1DPC	2DPC
DDR4-2933	1866	1866	1866	1866	1866	1866
DDR4-2666	1866	1866	1866	1866	1866	1866
DDR4-2400	1866	1866	1866	1866	1866	1866
DDR4-2133	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. This parameter is in the Memory Configuration submenu under Advanced.

Memory parameters under Memory Configuration

There are 6 parameters. The default is underlined each time.

- Memory Mode : Independent / Mirroring / Sparing
- NUMA: Disabled / Enabled
- DDR Performance : Performance optimized / Energy optimized / Power balanced
- Patrol Scrub : Disabled / Enabled
- IMC Interleaving : Auto / 1-Way / 2-Way
- Sub NUMA Clustering : Disabled / Enabled / Auto
- WR CRC feature control : Disabled / Enabled / Auto

The first parameter Memory Mode handles the redundancy function. 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 System Architect, an appropriate default setting is made in the factory. Otherwise, the parameter is set to Independent (no redundancy). Quantitative statements about the effect of the redundancy functions on system performance are to be found below.

The second parameter NUMA defines whether to build the physical address space from a segment of local memory or to notify the operating system of the structure. The default setting is *Enabled*. This setting should not be changed as long as there is no clear reason. Quantitative aspects of this topic will be discussed later.

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

The fourth parameter is the Patrol Scrub parameter. The default setting is *Enabled*. In the main memory, a correctable error is searched in the 24-hour cycle, and correction is started as necessary. In this way, it prevents the accumulation of memory errors that will make automatic correction impossible (counted in the corresponding register). If you have sensitive performance indicators, you can temporarily disable this feature. However, it may be difficult to demonstrate the effect on performance.

The fifth, IMC Interleaving, is a parameter that controls the interleaving of the on-chip memory controllers and the default setting is *Auto*. The Xeon Scalable Processor has two on-chip memory controllers, and if this parameter is set to [2-way], the memory address is interleaved between the two memory controllers. When it is set to [1-way], interleaving between memory controllers is not performed. When Sub NUMA Clustering as described below is enabled, the [1-way] setting is recommended as the processor is divided into two Sub-NUMA domains. We recommend the [2-way] setting when SNC is *disabled*. In the case of [Auto], the recommended setting according to the Sub NUMA Clustering setting is automatically selected.

The sixth, Sub NUMA Clustering (SNC) setting, is a parameter for dividing the L3 cache into two clusters according to the address range, which default setting is *Enabled*. Each of the divided clusters are attached to either of the two memory controllers of the processor. In addition, it is treated as one NUMA domain from the operating system, and access to the L3 cache and memory in NUMA mode improves its latency.

SNC is a substitute for Cluster on Die (COD), which was the CPU of the old generation. Just like COD, SNC is particularly recommended for NUMA optimized applications because it can minimize local memory latency and maximize local memory bandwidth.

The seventh, WR CRC feature control setting, is a new feature added for the Cascade Lake-SP generation. It controls the Write CRC feature of DDR4. If it is enabled, the memory controller in the processor send the generated CRC code as well as written data to the DRAM when writing data to the DRAM. By checking the CRC code, the DIMM can detect 1-bit error, 2-bit error, odd bit and horizontal column multi-bit error. Although it improves the reliability of the memory bus, the latency will get worse due to generating CRC and the memory bandwidth will drop since it uses extra data bus.

Performant memory configurations

The memory frequency and the number of memory channels used greatly affect memory performance. Since the memory frequency depends on the type of processor installed, each user should keep track of the memory frequency of their environment. In addition, Xeon Scalable Processor has six memory channels in total for each processor. In order to realize high memory performance, it is necessary to place DIMMs in as many memory channels as possible.

Furthermore, there are several configuration features that affect memory performance. The number of ranks, activation of redundancy functions, and invalidation of the NUMA function, etc. In the Part 2 of this document we will report the test results of these topics.

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 128 GB DIMMs) which are ideal regarding memory performance. The following table lists the particularly interesting configurations of this type (it is not necessarily complete).

With these configurations, all six memory channels per processor are the same. In each bank configuration, the same type of six DIMMs set is used. This ensures that memory accesses are evenly distributed among these memory system resources. Technically speaking, the optimum 6-way interleaving is realized via the memory channel. In this document, this is called Performance Mode configuration.

Xeon Scalable Processor Family equipped PRIMERGY and PRIMEQUEST server Performance Mode configuration (Configuration in which the memory is allocated to the second bank is only possible with a PRIMERGY model that can mount 12 DIMMs per processor)						
1 CPU system	2 CPU system	DIMM type	DIMM size (GB) bank 1	DIMM size (GB) bank 2	CPU per maximum MHz	Comment
48 GB	96 GB	DDR4-2933 R	8		2933	6-way rank interleave
96 GB	192 GB	DDR4-2933 R	16		2933	6-way rank interleave (++)
144 GB	288 GB	DDR4-2933 R	16	8	2933	Mixed configuration (-)
192 GB	384 GB	DDR4-2933 R	16	16	2933	6-way rank interleave
192 GB	384 GB	DDR4-2933 R	32		2933	6-way rank interleave (++)
288 GB	576 GB	DDR4-2933 R	32	16	2933	Mixed configuration (-)
384 GB	768 GB	DDR4-2933 R	32	32	2933	6-way rank interleave (++)
768 GB	1536 GB	DDR4-2933 R DDR4-2933 LR	64	64	2933	6-way rank interleave
768 GB	1536 GB	DDR4-2933 3DS LR	128		2933	6-way rank interleave
1536 GB	3072 GB	DDR4-2933 3DS LR	128	128	2666	Maximum configuration

The table is organized according to the total memory capacity of the left end. The total capacity is defined in one or two processor configurations. In the case of the two processor configuration, it is assumed that the memory configuration is the same for both processors. The next column is the DIMM type used. RDIMM, or 3DS LRDIMM technology is the determinant. The next two columns show the DIMM size by bank. This is because it is using the Performance Mode configuration and therefore groups the DIMMs into sets of 6 per bank.

The smallest configuration in the table has 48 GB for two processors because the six 8 GB DIMMs (i.e. 48 GB) must be counted for each processor.

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

- RDIMMs, LRDIMMs and 3DS LRDIMMs must not be mixed.
- RDIMMs of type x4 and x8 must not be mixed.
- The configuration is incrementing from bank 1 to 2 with decreasing DIMM sizes. The larger modules are installed first.

The last column has caution notes. For example, the information that mixed configurations (-) can have the decisive performance factor of the -6-way channel interleave, but can drop slightly in comparison to the configurations with a single DIMM type. This is due to complex addressing within individual memory channels.

Of course the table also contains the memory configurations from the standard benchmarks executed for the Xeon Scalable Processor Family based PRIMERGY servers. They are highlighted in the comments column with ++.

The second column from the right of the table shows the maximum memory frequency that can be achieved with each configuration. However, whether or not that value is reached depends on the processor model to be used.

Independent Mode configurations

This covers all the configurations that are not in Performance Mode. There are no restrictions other than the “banning of mixed use” rule between RDIMM, LRDIMM and 3DS LRDIMM and between RDIMMs of type x4 and x8..

You also need to pay attention to configurations where the number of DIMMs per processor does not become a multiple of six, that is, less than the minimum number required for the Performance Mode configuration, or seven to 11 configurations. This configuration may be done for reasons such as power saving and a low memory capacity. Cost savings may be realized by minimizing the number of DIMMs. From the quantitative evaluation showing the influence of the interleave configuration to the memory channel on the system performance introduced below, the following items are recommended.

- Operation with two, four, six or twelve DIMMs per processor can lead to balanced results as regards performance and energy consumption. Operation with other configurations is not recommended.

The non-recommended configurations are one DIMM, three DIMMs and five DIMMs per processor, which cannot equally allocate DIMMs to two memory controllers if the number of DIMMs is six or less. If the number of DIMMs exceeds six, even if 6-way interleaving is possible in the first bank, interleaving less than 6-way is configured in the second bank, therefore in applications requiring DIMM throughput, performance may be worse than the six DIMMs configuration.

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 default setting of the BIOS is not to be changed without a convincing reason. Only in this way is the UPI-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 the performance merit of this, although the memory of the 2-way server is configured symmetrically, the measurement results when the BIOS option is set to *NUMA = disabled* are shown below. Statistically, each secondary memory access is done to the remote memory. In an asymmetric memory configuration where the application is executed by 100 % remote memory, or in a one-sided memory configuration, it is necessary to estimate double the performance loss when local memory and remote memory are executed at a ratio of 50 %/50 %.

In addition, the configuration of 12 DIMMs in the first processor and six DIMMs in the second processor satisfies the Performance Mode criteria. This is because the memory channels *per processor* are handled in the same way. This is a way of thinking in PRIMERGY's ordering 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 respect to the Xeon Scalable Processors, the maximum supported memory frequency varies according to the processor type. For that reason, quantitative testing was performed with processors selected based on the maximum memory frequency supported by them, with some exceptions.

The measurements were made on a PRIMERGY RX2530 M5 with two processors under the Linux operating system. The following table shows the details of the configuration used for quantitative testing, particularly the representatives used for the processor classes.

System Under Test (SUT)	
Hardware	
Model	PRIMERGY RX2530 M5
Processor	Xeon Platinum 8260L (24 cores, 2.4GHz, XCC, DDR4-2933) x 2 Xeon Gold 6242 (16 cores, 2.8GHz, XCC, DDR4-2933) x 2 Xeon Gold 5220 (18 cores, 2.2GHz, HCC, DDR4-2666) x 2 Xeon Silver 4210 (10 cores, 2.2GHz, LCC, DDR4-2400) x 2 Xeon Bronze 3204 (6 cores, 1.9GHz, LCC, DDR4-2133) x 2
Memory types	8GB (1x8GB) 1Rx8 DDR4-2933 R ECC 16GB (1x16GB) 1Rx4 DDR4-2933 R ECC 16GB (1x16GB) 2Rx8 DDR4-2933 R ECC 32GB (1x32GB) 2Rx4 DDR4-2933 R ECC 64GB (1x64GB) 2Rx4 DDR4-2933 R ECC 64GB (1x64GB) 4Rx4 DDR4-2933 LR ECC 128GB (1x128GB) 8Rx4 DDR4-2933 3DS LR ECC
Disk subsystem	1 x SAS 12G HDD 450GB 15Krpm (via onboard SATA controller)
Software	
BIOS	R1.2.0
Operating system	SUSE Linux Enterprise Server 15

The 32 GB 2Rx4 RDIMM was usually used for the test set described below. In the testing of interleaving across memory channels, with only one or two DIMMs per processor in order to achieve the minimum capacity of main memory required for the tests. All of the DIMMs listed in the table were used only in the test set for the impact of the DIMM type.

The following table shows relative performance. The absolute measurement values for the STREAM and SPECrate2017_int_base 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.

First let's clarify one important result obtained from this test. 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. There are essential considerations for the most powerful and expensive processors of the Platinum class as well, and for the Bronze class, these can be ignored in many cases.

The measuring tools

Measurements were made using the benchmarks STREAM and SPECrate2017_int_base.

STREAM Benchmark

The STREAM benchmark (Developer: Mr. John McCalpin) [Related documents 4] is a tool to measure memory throughput. This benchmark implements copying and arithmetic operations on a large array of double type data, and provides four types of access results: Copy, Scale, Add and Triad. For access types other than Copy, arithmetic operations are included. Results are always indicated with throughput in GB/s. In general, the value of Triad is best quoted. Afterwards, the measured value of STREAM's benchmark is the Triad access value, and the unit is GB/s.

STREAM is the industry standard for measuring the memory bandwidth of a server, and can apply a large load to the memory system using a simple method. In particular, this benchmark is suitable for investigating the effect on memory performance in complex configurations. STREAM shows the effect of the configuration on memory and the resulting performance (degradation or improvement) caused by it. The value related to the STREAM benchmark described below shows the degree of influence on performance.

The memory impact on application performance is distinguished by the latency of each access and the bandwidth required by the application. Since the latency increases as the memory bandwidth increases, both are related. The degree to which the latency is canceled by parallel memory access also depends on the application and the quality of the machine code created by the compiler. For this reason it is very difficult to make a general forecast for all application scenarios.

SPECrate2017_int_base Benchmark

The SPECrate2017_int_base benchmark has been added as a model for commercial application performance. This is part of the Standard Performance Evaluation Corporation (SPEC) SPECcpu2017 [Related documents 5]. SPECcpu2017 is the industry standard for evaluating system processors, memory and compilers. It is the most important benchmark in the server field because a large number of measurement results are released and used for sales projects and technical investigation.

SPECcpu2017 consists of two independent test sets that use a lot of *integer* operations and *floating point* operations. The integer operation portion is equivalent to a commercial application and consists of 10 types of benchmarks. The floating point operation portion is equivalent to a scientific application and consists of 10 or 13 types of benchmarks. In either case, the benchmark execution result is 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.

In addition, depending on the type of measurement, the optimization allowed for the compiler differs. 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.

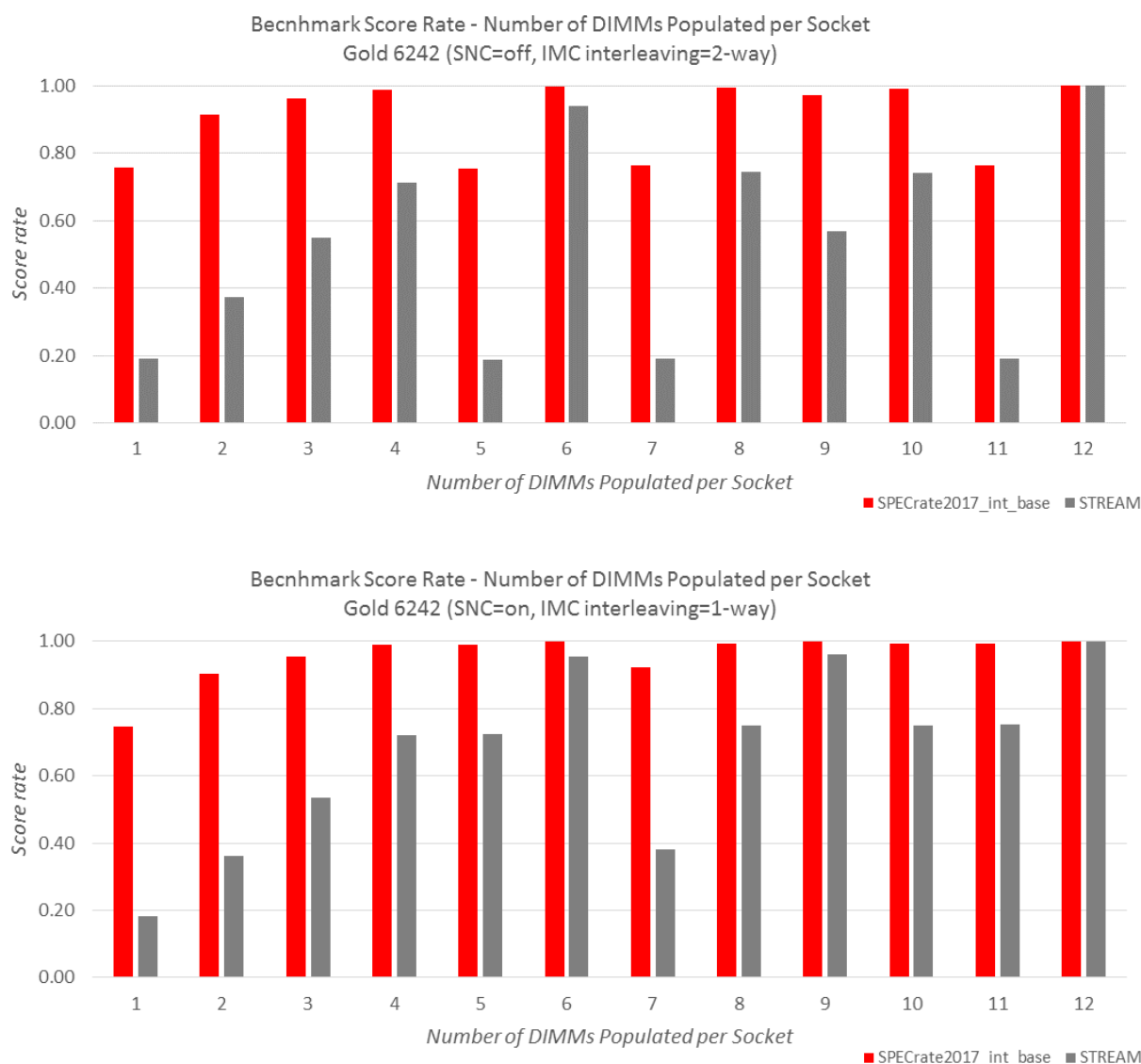
This is the summary of SPECrate2017_int_base. The integer suite was selected, because commercial applications predominate in the use of PRIMERGY servers.

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 is a method of setting a physical address area so that six memory channels are alternately used for each processor, such that the first block is on the first channel, the second block is on the second channel, and so on. Memory access is mainly done in the adjacent memory area according to the locality principle, and as a result it is spread over all of the channels. This performance gain situation results from parallelism. The channel interleave block size is based on a cache line size of 64 bytes. The *cache line size* is a unit of memory access in terms of the processor.

The following figure shows the ratio of the performance, when DIMMs are not mounted in a set of six pieces per processor and the ideal 6-way interleave is not performed; the value is considered as 1 when the number of DIMMs is 12. In particular, marked declines are seen in the STREAM index that measures memory throughput. When the number of DIMMs is two, three, four, and six, the performance is improved according to the increase in the number of DIMMs, but when the number of DIMMs is five, seven, eight, nine, 10, and 11, there is not enough performance compared with the case where the number of DIMMs is six. This is because the DIMMs cannot be evenly allocated to the memory channels under the two memory controllers, resulting in actual interleaving divided and the throughput of the interleave group with a small number of WAYs being a bottleneck.



The processor model used for this test (and the later test of the same category) is a Xeon Gold 6242. The DIMM type used is 32 GB 2Rx4 RDIMM.

Evaluation on SPECrate2017_int_base concerns the performance of commercial applications. 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 SPECrate2017_int_base not only applies for the performance aspect dealt with in this section, but also for all following sections.

There may be good reasons for choosing a 2-way, or a 4-way interleave, where performance degradation is gentle. In other words, the required memory capacity is small or the number of DIMMs is kept to a minimum because of low power consumption. 1-way interleaving is not recommended. Strictly speaking this is not interleaving, it is only called as such in the classification. In this case, the performance of the processor and the memory system are not well balanced.

These are examples of cases where the physical address area needs to be divided into multiple segments with different interleaving. Another example that needs to be divided is a configuration with different capacity per memory channel (GB per channel). Such configurations may occur when using DIMMs of different sizes or when using five or more DIMMs of the same size. Common to all examples is that you cannot set standardized address area segments by switching between memory channels. Switching is always “evenly” done. In such cases, by grouping existing DIMMs, segments are generated with the highest possible interleave. The following table shows two examples. The left column of the table shows the number of DIMMs connected to each of the six channels per CPU.

You cannot control which address area segment is assigned to an application, but this will result in different memory performance. In sensitive use cases this phenomenon may be a reason for avoiding configurations with a need for segmenting.

Where division is necessary DIMM configuration example (per CPU)	Address area segments	Size / Interleave
1 - 1 - 1, 1 - 1 - 0	1 - 1 - 0, 1 - 1 - 0	80 % of the address area / 4-way
	0 - 0 - 1, 0 - 0 - 0	20 % of the address area / 1-way
2 - 1 - 1, 2 - 1 - 1	1 - 1 - 1, 1 - 1 - 1	75 % of the address area / 6-way
	1 - 0 - 0, 1 - 0 - 0	25 % of the address area / 2-way

Memory frequency

The influences on effective memory frequency is explained in detail in the previous sections. 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 table will help you compare and balance the impact. The values in the first table are based on the minimum memory frequency of 1866 MHz common to all series of measurements. The second table captures the same information from different perspectives. Values are based on an ideal case, in other words, the maximum frequency in the processor class.

Benchmark	Processor type	DIMM Max frequency	1866 MHz	2133 MHz	2400 MHz	2666 MHz	2933 MHz
STREAM	Platinum 8260L	2933 MHz	1.00				1.50
	Gold 5220	2666 MHz	1.00			1.33	
	Silver 4210	2400 MHz	1.00		1.10		
	Bronze 3204	2133 MHz	1.00	1.01			
SPECrate2017_int_base	Platinum 8153	2933 MHz	1.00				1.05
	Gold 5220	2666 MHz	1.00			1.03	
	Silver 4210	2400 MHz	1.00		1.01		
	Bronze 3204	2133 MHz	1.00	1.00			

Benchmark	Processor type	DIMM Max frequency	1866 MHz	2133 MHz	2400 MHz	2666 MHz	2933 MHz
STREAM	Platinum 8260L	2933 MHz	0.67				1.00
	Gold 5220	2666 MHz	0.75			1.00	
	Silver 4210	2400 MHz	0.91		1.00		
	Bronze 3204	2133 MHz	0.99	1.00			
SPECrate2017_int_base	Platinum 8153	2933 MHz	0.96				1.00
	Gold 5220	2666 MHz	0.97			1.00	
	Silver 4210	2400 MHz	0.99		1.00		
	Bronze 3204	2133 MHz	1.00	1.00			

The processor models used in this test are the Xeon Platinum 8260L (DDR4-2933), Xeon Gold 5220 (DDR4-2666), Xeon Silver 4210 (DDR4-2400) and Xeon Bronze 3106 (DDR4-2133). The DIMM type used is 32 GB 2Rx4 RDIMM. It is used with a 2DPC configuration.

If you set "DDR Performance = *Energy optimized*" in the BIOS, the frequency will always be 1866 MHz. However, the effect of the voltage of the DIMM is large and the influence of the memory frequency is small, therefore the power saving effect obtained is small. Since the voltage of the new DDR4 module is always 1.2 V, it has a power saving effect compared with the DDR3 generation, which was at least 1.35 V. That is why we don't recommend setting *Energy optimized*.

Influence of the DIMM types

Seven types of DIMMs are planned when PRIMERGY and PRIMEQUEST servers with Xeon Scalable Processors are opened to the public. 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:

- The measurement was basically carried out using Xeon Platinum 8260L and some results of DIMM types which aren't supported by the PRIMERGY servers were converted from the result measured with the PRIMEQUEST server with 2-socket configuration.
- It is evident that with these measurements all the memory channels were equally configured, i.e. Performance Mode configurations were compared. The number of installed DIMMs was 12 for 1DPC measurement and 24 for 2DPC measurement.
- All the measurements were carried out with the consistent memory frequency 2933 MHz except the 2DPC configuration with the 128 GB 8Rx4 3DS LRDIMM, which was measured at the speed of 2666 MHz..
- The table is standardized to the 2DPC configuration with the 32 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	Config uration	STREAM	SPECrate2017_int_base
8GB (1x8GB) 1Rx8 DDR4-2933 R ECC	1DPC	0.84	0.97
	2DPC	0.96	1.00
16GB (1x16GB) 2Rx8 DDR4-2933 R ECC	1DPC	0.96	1.00
	2DPC	1.00	1.00
16GB (1x16GB) 1Rx4 DDR4-2933 R ECC	1DPC	0.84	0.98
	2DPC	0.96	1.00
32GB (1x32GB) 2Rx4 DDR4-2933 R ECC	1DPC	0.95	1.00
	2DPC	1.00	1.00
64GB (1x64GB) 2Rx4 DDR4-2933 R ECC	1DPC	0.90	0.97
	2DPC	0.97	0.97
64GB (1x64GB) 4Rx4 DDR4-2933 LR ECC	1DPC	0.99	1.00
	2DPC	0.97	0.97
128GB (1x128GB) 8Rx4 DDR4-2933 3DS LR ECC	1DPC	0.83 ¹⁾	0.93 ¹⁾
	2DPC	0.77 ¹⁾	0.89 ¹⁾

1) Convert from the measured value of the PRIMEQUEST server

The difference in performance shown here is mainly due to the difference in the number of rank interleaves. The rank interleave number is equal to the number of ranks per memory channel and follows the DIMM type and DPC value. 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 interleave is greater than the interleaving on the channel. Channel interleaving is used for 64 byte cache line sizes. Rank interleaving is towards the 4 KB page size of the operating system and is related to the physical characteristics of the DRAM memory. Memory cells are roughly arranged in two dimensions. During access, a line (also called a page) is opened and the column item is read. While the page is open you can also read the values of other columns with much lower latency. Furthermore, rough rank interleaving is optimized for this function.

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.

For example, the most noticeable performance degradation in this table is in the 8 GB 1Rx8 RDIMM, but this is explained as it is missing rank interleaving. Except for a 1DPC configuration using single rank DIMMs, this case can also occur in a mixed configuration using, for example, a 32 GB 2Rx4 RDIMM in the first bank and

an 16 GB 1Rx4 RDIMM in the second bank. In the case of this missing or 1-way rank interleaving, it is necessary to pay attention to some degree of performance degradation. In situations where performance is emphasized and a powerful processor model is used in particular, this needs to be avoided.

The resulting table contains a number of other subtle effects as well as a major impact of rank interleaving. For example, because the memory channel has more than four ranks, the overhead per rank performed to refresh the DRAM becomes prominent in a bad way. This refresh corresponds to a constant basic load per address line of the memory channel shared by all ranks. This can explain the relationship with the case where the results of the 2DPC configuration are worse than the corresponding 1DPC configuration results in the 4Rx4 LRDIMM and 8Rx4 3DS LRDIMM described above.

Optimization of the cache coherence protocol

The function to select the Cluster on Die (COD) setting as the protocol of cache coherence that existed in Broadwell-EP was changed to the Sub NUMA Clustering setting in Xeon Scalable Processor. For details, refer to the section on memory system BIOS options.

The following table shows the effect on the two loads or benchmarks examined in this document.

The measurements are made in 2DPC configurations with 32 GB 2Rx4 RDIMMs.

The table shows that performance is affected in the 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 SPECrate2017_int_base for commercial application performance therefore only applies at this stage in a restricted manner.

Keep in mind that Bronze and some Silver processor types do not support SNC.

Benchmark	Processor type	SNC=Enabled (IMC Interleaving1-Way)	SNC=Disabled (IMC Interleaving2-Way)
STREAM	Platinum 8260L	1.00	0.96
	Gold 5220	1.00	0.98
	Silver 4210	(Not supported)	N/A
SPECrate2017_int_base	Platinum 8153	1.00	0.98
	Gold 5120	1.00	0.99
	Silver 4210	(Not supported)	N/A

Access to remote memory

For the tests using the STREAM and SPECrate2017_int_base benchmarks mentioned above, only the local memory was targeted (the processor accesses the DIMM module of its own memory channel). Modules of adjacent processors are not accessed at all, or only rarely accessed via the UPI 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 6-way rank-interleaved Performance Mode configuration with 32 GB 2Rx4 RDIMMs operating at 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 UPI link.

Benchmark	Processor type	UPI frequency	NUMA = enabled	NUMA = disabled
STREAM	Platinum 8260L	10.4GT/s	1.00	0.46
	Gold 5220	10.4GT/s	1.00	0.50
	Silver 4210	9.6GT/s	1.00	0.51
SPECrate2017_int_base	Platinum 8260L	10.4GT/s	1.00	0.87
	Gold 5220	10.4GT/s	1.00	0.90
	Silver 4210	9.6GT/s	1.00	0.91

In *NUMA = disabled*, the physical address space is set by detailed processor mesh switching. This switching assumes that both processors have the same memory capacity. 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.

Since NUMA is not supported or insufficient in the system software or system related software, measurements on *NUMA = disabled* were performed in a narrow range as an exceptional case where setting is recommended. All of the above measurements are useful for estimating the impact of most or all accesses to remote memory. This situation occurs when the configuration memory capacity of each processor is

significantly different. Performance degradation compared to local access can be up to twice the drop shown in the table.

Memory performance under redundancy and reliability

There are two redundancy options for the Xeon Scalable Processor based PRIMERGY servers and PRIMEQUEST. And for the Cascade Lake-SP generation a setting related to the reliability of the memory bus was added.

In mirroring, mirrors are configured between two memory channels within one processor's memory controller. The operating system can utilize 50% of the memory that is actually configured.

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.

WR CRC feature control setting controls the Write CRC feature of DDR4. If it is enabled, the memory controller in the processor send the generated CRC code as well as written data to the DRAM when writing data to the DRAM. By checking the CRC code, the DIMM can detect 1-bit error, 2-bit error, odd bit and horizontal column multi-bit error.

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 32 GB 2Rx4 RDIMMs 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 3-way channel interleaving, because both halves of the mirror can be used for read access.

Relationship under sparing is derived from the rank interleaving described in the previous section on the DIMM type impact. Normally, the reservation rank becomes an odd number of active ranks per memory channel, and it becomes a 1-way rank interleave. The sparing column is due to the difference between the 6-way rank interleave and 1-way rank interleave.

Benchmark	Processor type	Normal	Mirroring	Sparing
STREAM	Platinum 8260L	1.00	0.64	0.84
	Gold 5220	1.00	0.69	0.84
	Silver 4210	1.00	0.84	0.85
SPECrate2017_int_base	Platinum 8260L	1.00	0.99	0.98
	Gold 5220	1.00	0.99	0.99
	Silver 4210	1.00	1.00	1.00

The following table shows the effect of the Write CRC feature in the event of an otherwise ideal memory configuration, i.e. a Performance Mode 2DPC configuration with 32 GB 2Rx4 RDIMMs in each case.

As mentioned in the BIOS parameter section, enabling Write CRC improves the reliability of the memory bus but the latency will get worse due to generating CRC and the memory bandwidth will drop since it uses extra data bus.

As the tables shows, disabling Write CRC is one of the option for the performance improvement of bandwidth-intensive applications.

Benchmark	Processor type	Write CRC = disable	Write CRC = enabled
STREAM	Platinum 8260L	1.00	0.90
	Gold 5220	1.00	0.91
	Silver 4210	1.00	0.97
SPECrate2017_int_base	Platinum 8260L	1.00	1.00
	Gold 5220	1.00	1.00
	Silver 4210	1.00	1.00


Literature


PRIMERGY Servers

[L1] <http://jp.fujitsu.com/primergy>

Memory Performance

[L2] This white paper:

 <http://docs.ts.fujitsu.com/dl.aspx?id=543b9166-f047-4442-b506-b0acb7ba0c46>

 <http://docs.ts.fujitsu.com/dl.aspx?id=ade521ff-45c7-408d-9b36-a88b248497ca>

[L3] Xeon E5-2600 v3 (Haswell-EP) equipped system memory performance

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

Memory Performance of Xeon scalable processor (Skylake-SP) based Systems

<http://docs.ts.fujitsu.com/dl.aspx?id=914e6c8a-8bc8-4441-bcbe-e33bbb4c7a3c>

Benchmark

[L4] STREAM

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

[L5] SPECcpu2017

<http://docs.ts.fujitsu.com/dl.aspx?id=20f1f4e2-5b3c-454a-947f-c169fca51eb1>

BIOS Settings

[L6] BIOS optimizations for Xeon Scalable Processor based systems

<http://docs.ts.fujitsu.com/dl.aspx?id=7a93f0a9-5faf-47c6-9f4d-698debde7f95>

PRIMERGY Performance

[L7] <http://jp.fujitsu.com/platform/server/primergy/performance/>

Contact

FUJITSU

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

PRIMERGY Product Marketing

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

PRIMERGY Performance and Benchmarks

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