White Paper
FUJITSU Server PRIMERGY
Memory performance of Xeon Scalable Processor (Cooper Lake) based Systems

In the 3rd Generation Xeon Scalable Processor (Cooper Lake) based FUJITSU Server PRIMERGY, with the introduction of the processor frequency, the Ultra Path Interconnect (UPI) and memory architecture improvement, performance has improved. 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
2021-04-02

Table of contents

Document History

Version 1.0 (2021-04-02)

Initial version

Introduction

The 3rd Generation Xeon Scalable Processor (Cooper Lake) inherits the features which previous Xeon Scalable Processor generations (Skylake-SP and Cascade Lake-SP) have.

The processor is equipped with two *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. The processor has six memory channels and even if the number of DIMMs per channel increases, the memory frequency no longer decreases.

When this processor requests the contents of the memory (remote memory) of the adjacent processor, it uses an Ultra Path Interconnect (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.

Moreover, the higher processor frequency and the following improvement to the 3rd Gen Xeon Scalable Processor (Cooper Lake) realizes the better performance than the previous generation processors. While the maximum memory frequency of the previous generation Cascade Lake-SP systems was 2933 MHz, the new Cooper Lake system supports 3200 MHz and its peak memory bandwidth when the maximum memory is installed is 154 GB/s per processor. The support of high capacity 256 GB 3DS RDIMM enables to equip 3 TB of memory per processor[1]. The UPI link up to 10.4 GT/s is still used for the connection between CPUs but for the configuration with four sockets or more, the remote memory access bandwidth is improved due to doubled UPI links.

The cache coherence protocol option known as *Sub-NUMA Clusters (SNC)* is available for the 3rd Generation Xeon Scalable Processors (Cooper Lake) as well. 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 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 3200, 2933, 2667, 2400 or 1867 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.

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 ADDDC sparing, make up the end of this document.

[1] The maximum memory capacity depends on the processor type.

Memory architecture

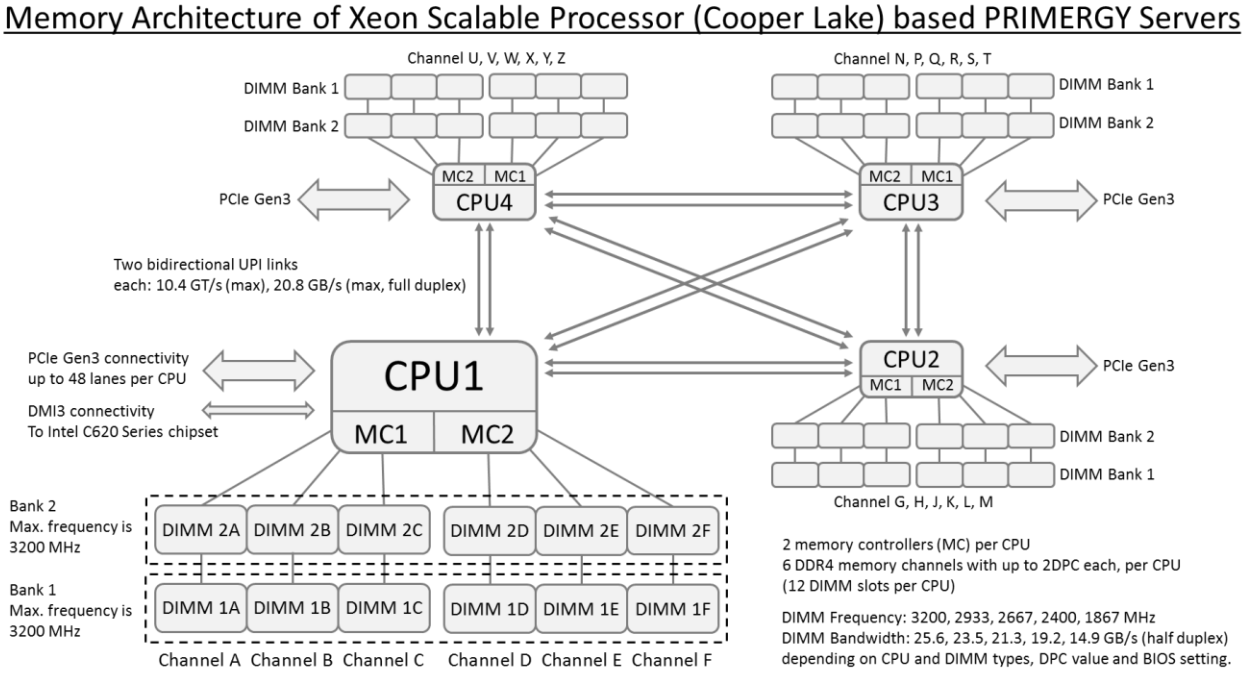
This section explains the outline of the memory system with five parts. First, we will explain the arrangement of available DIMM slots in the block diagram. The second section shows the available DIMM types. The following third section describes the effect on the effective memory frequency. The fourth 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 of the 3rd Gen Xeon Scalable Processor (Cooper Lake) based systems.

The Cooper Lake based PRIMERGY servers have 12 DIMM slots per processor. The data path width is 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.

Although the previous generations (Cascade Lake-SP and Skylake-SP) have one UPI link between processors for four or more processors configuration, the number of the new Cooper Lake generation is doubled. It is expected to improve the performance of the applications which have frequent memory accesses between processors, such as the database processing.



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 Xeon Scalable Processors based PRIMEGY servers, DPC does not reduce the memory frequency.

We also use the term “memory bank” in the following. In the figure, a group of six 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 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 8380HL	28	56	38.50	10.4	2.9	4.3	3200	250
Platinum 8380H	28	56	38.50	10.4	2.9	4.3	3200	250
Platinum 8376HL	28	56	38.50	10.4	2.6	4.3	3200	205
Platinum 8376H	28	56	38.50	10.4	2.6	4.3	3200	205
Platinum 8360HL	24	48	33.00	10.4	3.0	4.2	3200	225
Platinum 8360H	24	48	33.00	10.4	3.0	4.2	3200	225
Platinum 8356H	8	16	35.75	10.4	3.9	4.4	2933	190
Platinum 8354H	18	36	24.75	10.4	3.1	4.3	3200	205
Gold 6348H	24	48	33.00	10.4	2.3	4.2	2933	165
Gold 6330H	24	48	33.00	10.4	2.0	3.7	2933	150
Gold 6328HL	16	32	22.00	10.4	2.8	4.3	2933	165
Gold 6328H	16	32	22.00	10.4	2.8	4.3	2933	165
Gold 5320H	20	40	27.50	10.4	2.4	4.2	2667	150
Gold 5318H	18	36	24.75	10.4	2.5	3.8	2667	150

The quantitative memory performance tests were performed based on the supported memory frequency as listed in the second-to-last column of the table according to the topic.

DDR4 topics and available DIMM types

The Cooper Lake based PRIMERGY servers use the DDR4 SDRAM memory module as well as the previous Xeon Scalable Processor based PRIMERGY servers use. The Cooper Lake based systems have the following improvement.

- DDR4 supports a memory frequency up to 3200 MHz. For Cooper Lake based systems, this has reached 3200 MHz. The previous generation systems with Cascade Lake-SP was supported up to 2933 MHz..
- The new Cooper Lake based system can be equipped with up to 3 TB of DRAM per socket with 256 GB 3DS RDIMMs[2]. The old Cascade Lake-SP based systems can be equipped with up to 1.5 TB of DRAM.

The following table shows the DIMMs supported by the Cooper Lake based PRIMERGY servers. In DIMM, there are Registered DIMM (RDIMM), Load Reduced DIMM (LRDIMM), 3DS Registered DIMM (3DS RDIMM) types. The mixed configurations are only possible within the four sections of the table. RDIMM, LRDIMM and 3DS RDIMM cannot be mixed.,

DIMM type	Control	Maximum frequency (MHz)	Volt (V)	# of Ranks	Capacity	Rel. price per GB
8GB (1x8GB) 1Rx8 DDR4-3200 R ECC	Registered	3200	1.2	1	8 GB	0.92
16GB (1x16GB) 2Rx8 DDR4-3200 R ECC	Registered	3200	1.2	1	16 GB	0.98
16GB (1x16GB) 1Rx4 DDR4-3200 R ECC	Registered	3200	1.2	2	16 GB	0.98
32GB (1x32GB) 2Rx4 DDR4-3200 R ECC	Registered	3200	1.2	2	32 GB	**1.00**
64GB (1x64GB) 2Rx4 DDR4-3200 R ECC	Registered	3200	1.2	2	64 GB	1.00
64GB (1x64GB) 4Rx4 DDR4-3200 LR ECC	Load Reduced	3200	1.2	4	64 GB	1.34
128GB (1x128GB) 4Rx4 DDR4-3200 LR ECC	Load Reduced	3200	1.2	4	128 GB	1.34
128GB (1x128GB) 4Rx4 DDR4-3200 3DS R ECC	3DS Registered	3200	1.2	4	128 GB	1.10
256GB (1x256GB) 8Rx4 DDR4-3200 3DS R ECC	3DS Registered	3200	1.2	8	256 GB	1.10

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 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).

[2] The maximum memory capacity depends on the processor type.

- 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 RDIMM: This is a RDIMM 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.

Which type of RDIMM, LRDIMM, or 3DS RDIMM is desirable is usually determined by the memory capacity required. But LRDIMM and 3DS RDIMM 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 RX4770 M6's price list as of Dec 2020. 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 sales area, the price may differ. 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: 3200, 2933, 2667, 2400 and 1867 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 8380HL	28	56	38.50	10.4	2.9	4.3	3200	250
Platinum 8380H	28	56	38.50	10.4	2.9	4.3	3200	250
Platinum 8376HL	28	56	38.50	10.4	2.6	4.3	3200	205
Platinum 8376H	28	56	38.50	10.4	2.6	4.3	3200	205
Platinum 8360HL	24	48	33.00	10.4	3.0	4.2	3200	225
Platinum 8360H	24	48	33.00	10.4	3.0	4.2	3200	225
Platinum 8356H	8	16	35.75	10.4	3.9	4.4	2933	190
Platinum 8354H	18	36	24.75	10.4	3.1	4.3	3200	205
Gold 6348H	24	48	33.00	10.4	2.3	4.2	2933	165
Gold 6330H	24	48	33.00	10.4	2.0	3.7	2933	150
Gold 6328HL	16	32	22.00	10.4	2.8	4.3	2933	165
Gold 6328H	16	32	22.00	10.4	2.8	4.3	2933	165
Gold 5320H	20	40	27.50	10.4	2.4	4.2	2667	150
Gold 5318H	18	36	24.75	10.4	2.5	3.8	2667	150

In Cooper Lake, 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 RDIMM	
	1DPC	2DPC	1DPC	2DPC	1DPC	2DPC
DDR4-3200	3200	3200	3200	3200	3200	3200
DDR4-2933	2933	2933	2933	2933	2933	2933
DDR4-2667	2667	2667	2667	2667	2667	2667

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.

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		LRDIMM		3DS RDIMM	
	1DPC	2DPC	1DPC	2DPC	1DPC	2DPC
DDR4-3200	1867	1867	1867	1867	1867	1867
DDR4-2933	1867	1867	1867	1867	1867	1867
DDR4-3200	1867	1867	1867	1867	1867	1867

By selecting the BIOS parameter Power balanced, you can balance between the performance and the power consumption.

DDR Performance = Power balanced						
CPU type	RDIMM		LRDIMM		3DS RDIMM	
	1DPC	2DPC	1DPC	2DPC	1DPC	2DPC
DDR4-3200	2400	2400	2400	2400	2400	2400
DDR4-2933	2400	2400	2400	2400	2400	2400
DDR4-2667	2400	2400	2400	2400	2400	2400

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 7 parameters. The default is underlined each time.

- Memory Mode : Independent / Mirroring / Address Range Mirroring
- ADDDC Sparing : Disabled / Enabled
- NUMA : Disabled / Enabled
- DDR Performance : Performance optimized / Energy optimized / Power balanced
- PPR Type : Hard PPR / Soft PPR / PPR Disabled
- Patrol Scrub : Disabled / Enabled
- SNC (Sub NUMA) : Disabled / Enabled / Auto

The first parameter Memory Mode and the second parameter ADDDC (Adaptive Double Device Data Correction) Sparing handle the redundancy function. They are part of the RAS (Reliability, Availability, Serviceability) functionality.

Memory Mode specifies whether to duplicate the data in the memory (mirroring). With Memory Mode set to *Mirroring*, mirroring is enabled and it halves the memory capacity. The option *Address Range Mirroring* mirrors a part of system memory. It needs the operating system support.

ADDDC Sparing activates the spare areas at the level of DIMM ranks or banks to increase fail-safety, if memory errors become frequent. ADDC Sparing will be disabled automatically by enabling mirroring with Memory Mode. Please refer to the respective configurator for the configuration available for ADDDC Sparing.

If these functions are requested during the configuration in SystemArchitect, appropriate default settings are made in the factory. Otherwise, the parameters are set to *Independent* and *Disabled* (no redundancy). Quantitative statements about the effect of the redundancy functions on system performance are to be found below.

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

The fifth parameter PPR type treats Post Package Repair (PPR), which is the feature of DDR4. PPR replaces fault memory cells with spare cells in DRAM chips at system boot. With *Soft PPR* set, the replacement will be lost when the system is powered off or reset. With *Hard PPR* set, the replacement holds permanently. If *PPR Disabled* is set, the system doesn't replace them.

The sixth parameter is the Patrol Scrub parameter. The default setting is *Disabled*. In the main memory, a correctable error is searched periodically, 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 seventh, SNC (Sub NUMA) 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 particularly recommended for NUMA optimized applications because it can minimize local memory latency and maximize local memory bandwidth.

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 256 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 (Cooper Lake) Family equipped PRIMERGY server Performance Mode configuration						
1 CPU system	**2 CPU system**	**DIMM type**	**DIMM size (GB) bank 1**	**DIMM size (GB) bank 2**	**CPU per maximum MHz**	**Comment**
96 GB	192 GB	DDR4-3200 R	8		3200	6-way rank interleave
192 GB	384 GB	DDR4-3200 R	16		3200	6-way rank interleave (++)
288 GB	576 GB	DDR4-3200 R	16	8	3200	Mixed configuration (-)
384 GB	768 GB	DDR4-3200 R	16	16	3200	6-way rank interleave
384 GB	768 GB	DDR4-3200 R	32		3200	6-way rank interleave
288 GB	576 GB	DDR4-3200 R	32	16	3200	Mixed configuration (-)
768 GB	1536 GB	DDR4-3200 R	32	32	3200	6-way rank interleave (++)
1152 GB	2304 GB	DDR4-3200 R	64	32	3200	Mixed configuration (-)
1536 GB	3072 GB	DDR4-3200 R DDR4-3200 LR	64	64	3200	6-way rank interleave (++)
2304 GB	4608 GB	DDR4-3200 LR	128	64	3200	Mixed configuration (-)
3072 GB	6144 GB	DDR4-3200 LR DDR4-3200 3DS R	128	128	3200	6-way rank interleave
4608 GB	9216 GB	DDR4-3200 3DS R	256	128	3200	Mixed configuration (-)
6144 GB	12288 GB	DDR4-3200 3DS R	256	256	3200	Maximum configuration

The table is organized according to the total memory capacity of the left end. The total capacity is defined in two or four processor configurations. It is assumed that the memory configuration is the same for all the processors. The next column is the DIMM type used. RDIMM, or 3DS RDIMM 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 96 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 RDIMMs 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 followings:

- RDIMMs, LRDIMMs and 3DS RDIMMs 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 number of the DIMM on a processor is limited to one, four, six or twelve.

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. 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 four, six or twelve DIMMs per processor can lead to balanced results as regards performance and energy consumption. Operation with one DIMM configuration is not recommended.

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, one out of every two memory accesses 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 RX4770 M6 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 RX4770 M6
Processor	Xeon Platinum 8360HL (24 cores, 3.0GHz, DDR4-3200) x 2 Xeon Gold 6348H (24 cores, 2.3GHz, DDR4-2933) x 2 Xeon Gold 5318H (18 cores, 2.5GHz, DDR4-2667) x 2
Memory types	8GB (1x8GB) 1Rx8 DDR4-3200 R ECC 16GB (1x16GB) 1Rx4 DDR4-3200 R ECC 16GB (1x16GB) 2Rx8 DDR4-3200 R ECC 32GB (1x32GB) 2Rx4 DDR4-3200 R ECC 64GB (1x64GB) 2Rx4 DDR4-3200 R ECC 64GB (1x64GB) 4Rx4 DDR4-3200 LR ECC 128GB (1x128GB) 4Rx4 DDR4-3200 LR ECC 256GB (1x256GB) 8Rx4 DDR4-3200 3DS R ECC
Disk subsystem	1 × SATA 6G SSD 480GB (via SAS RAID controller)
Software	
BIOS	R1.5.0
Operating system	Red Hat Enterprise Linux 8.2

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 Scalable Processor based PRIMERGY server.

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.

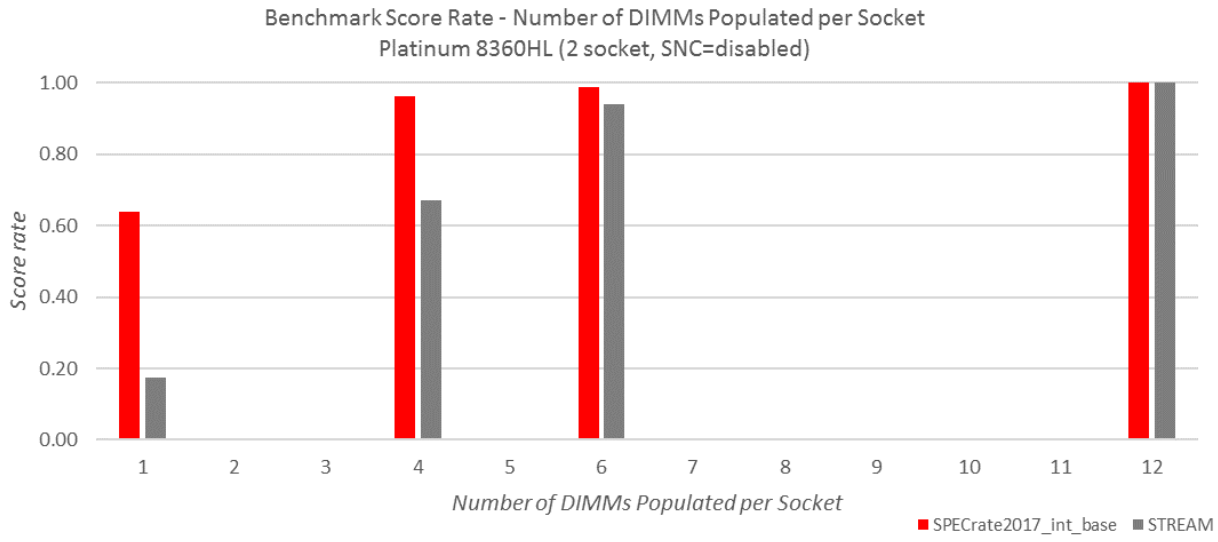
A measurement that is compliant with the regulations requires three runs, and the median result is evaluated for each individual benchmark. This was not complied with in the technical investigation described here. To simplify matters only one run was performed at all times.

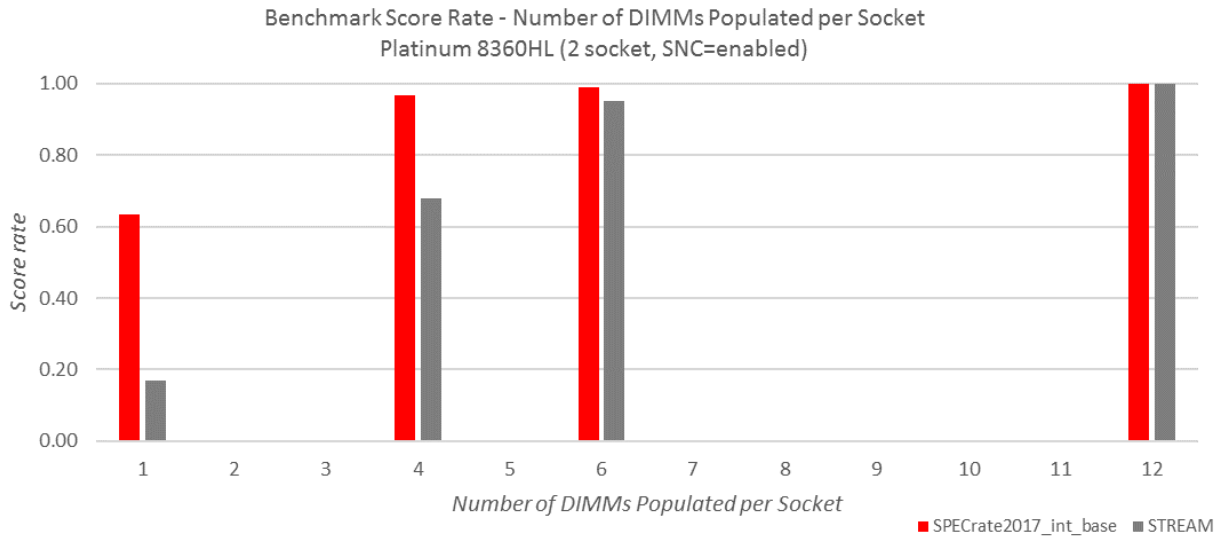
Interleaving across the memory channels

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. The number of DIMMs populated per one processor is limited to one, four, six or twelve for the Cooper Lake based PRIMERGY servers.

In particular, marked declines are seen in the STREAM index that measures memory throughput. When the number of DIMMs is one, four and six, the performance is improved according to the increase in the number of DIMMs.





The processor model used for this test (and the later test of the same category) is a Xeon Platinum 8360HL. 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 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.

Memory frequency

The memory frequency of the Xeon Scalable Processor based PRIMERGY servers does not change depending on DPC. 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 1867 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	1867 MHz	2400 MHz	2667 MHz	2933 MHz	3200 MHz
STREAM	Platinum 8360HL	3200 MHz	1.00	1.24			1.59
	Gold 6348H	2933 MHz	1.00	1.24		1.48	
	Gold 5318H	2667 MHz	1.00	1.22	1.32		
SPECrate2017 _int_base	Platinum 8360HL	3200 MHz	1.00	1.03			1.06
	Gold 6348H	2933 MHz	1.00	1.02		1.03	
	Gold 5318H	2667 MHz	1.00	1.02	1.03		

Benchmark	Processor type	DIMM Max frequency	1867 MHz	2400 MHz	2667 MHz	2933 MHz	3200 MHz
STREAM	Platinum 8360HL	3200 MHz	0.63	0.78			1.00
	Gold 6348H	2933 MHz	0.68	0.84		1.00	
	Gold 5318H	2667 MHz	0.76	0.93	1.00		
SPECrate2017 _int_base	Platinum 8360HL	3200 MHz	0.95	0.98			1.00
	Gold 6348H	2933 MHz	0.97	0.99		1.00	
	Gold 5318H	2667 MHz	0.97	0.99	1.00		

The processor models used in this test are the Xeon Platinum 8360HL (DDR4-3200), Xeon Gold 6348H (DDR 4-2933) and Xeon Gold 5318H (DDR4-2667). 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 1867 MHz. In addition, with “DDR Performance = *Power balanced*”, it will be set at 2400 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 with DDR4 modules, of which the voltage is always 1.2 V, is small. That is why we don't recommend setting *Energy optimized*.

Influence of the DIMM types

Nine types of DIMMs are planned when the Cooper Lake based PRIMERGY servers 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 carried out using Xeon Platinum 8360HL.
- 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 3200 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	Configuration	STREAM	SPECrate2017_int_base
8GB (1x8GB) 1Rx8 DDR4-3200 R ECC	1DPC	0.82	0.96
	2DPC	0.95	0.99
16GB (1x16GB) 2Rx8 DDR4-3200 R ECC	1DPC	0.95	0.99
	2DPC	1.00	1.00
16GB (1x16GB) 1Rx4 DDR4-3200 R ECC	1DPC	0.82	0.97
	2DPC	0.95	0.99
32GB (1x32GB) 2Rx4 DDR4-3200 R ECC	1DPC	0.95	0.99
	2DPC	**1.00**	**1.00**
64GB (1x64GB) 2Rx4 DDR4-3200 R ECC	1DPC	0.95	0.99
	2DPC	1.00	1.00
64GB (1x64GB) 4Rx4 DDR4-3200 LR ECC	1DPC	0.98	0.99
	2DPC	0.95	0.96
128GB (1x128GB) 4Rx4 DDR4-3200 LR ECC	1DPC	0.98	0.99
	2DPC	0.95	0.96
256GB (1x256GB) 8Rx4 DDR4-3200 3DS R ECC	1DPC	0.72	0.91
	2DPC	0.84	0.90

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 a 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 described above. The influence of the refresh becomes more remarkable for higher capacity DIMMs.

Optimization of the cache coherence protocol

SNC (Sub NUMA) setting in Xeon Scalable Processor can select the protocol of cache coherence. 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.

Benchmark	Processor type	SNC=Enabled	SNC=Disabled
STREAM	Platinum 8360HL	1.00	0.98
SPECrate2017_int_base	Platinum 8360HL	1.00	0.99

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 one out of every two memory accesses 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 8360HL	10.4GT/s	1.00	0.46
SPECrate2017_int_base	Platinum 8360HL	10.4GT/s	1.00	0.89

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.

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 ADDDC sparing, there is no decrease in capacity because it replace faulty DRAM cells with the spare areas in DIMM devices.

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 and ADDC Sparing.

The loss that occurred under mirroring is smaller than a half of the performance at default settings, because both halves of the mirror can be used for read access. In the case of ADDDC sparing, a little loss of performance is observed.

Benchmark	Processor type	Normal (Memory Mode = Independent, ADDDC Sparing = Disabled)	Mirroring (Memory Mode = Mirroring, ADDDC Sparing = Disabled)	Sparing (Memory Mode = Independent, ADDDC Sparing = Enabled)
STREAM	Platinum 8360HL	1.00	0.66	0.95
SPECrate2017_int_base	Platinum 8360HL	1.00	0.98	0.99

Literature

PRIMERGY Servers

[L1] https://www.fujitsu.com/global/products/computing/servers/primergy/index.html

Memory Performance

[L2] This white paper:
https://docs.ts.fujitsu.com/dl.aspx?id=5c3e63aa-2736-4ed5-86ae-1c6cc1eed8fe
https://docs.ts.fujitsu.com/dl.aspx?id=f9678119-7b94-4ec8-81cc-254630e46ce6

[L3] Memory Performance of Xeon scalable processor (Skylake-SP) based Systems
https://docs.ts.fujitsu.com/dl.aspx?id=914e6c8a-8bc8-4441-bcbe-e33bbb4c7a3c
Memory Performance of Xeon scalable processor (Cascade Lake-SP) based Systems
https://docs.ts.fujitsu.com/dl.aspx?id=543b9166-f047-4442-b506-b0acb7ba0c46

Benchmark

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

[L5] SPECcpu2017
https://docs.ts.fujitsu.com/dl.aspx?id=20f1f4e2-5b3c-454a-947f-c169fca51eb1

BIOS Settings

[L6] BIOS optimizations for Xeon Scalable Processor based systems
https://docs.ts.fujitsu.com/dl.aspx?id=7a93f0a9-5faf-47c6-9f4d-698debde7f95

PRIMERGY Performance

[L7] https://www.fujitsu.com/global/products/computing/servers/primergy/benchmarks/

Contact

FUJITSU

Website: https://www.fujitsu.com/global/

PRIMERGY Product Marketing

mailto:Primergy-PM@ts.fujitsu.com

PRIMERGY Performance and Benchmarks

mailto:primergy.benchmark@ts.fujitsu.com

© Copyright 2021 Fujitsu Limited. Fujitsu and the Fujitsu logo are trademarks or registered trademarks of Fujitsu Limited in Japan and other countries. Other company, product and service names may be trademarks or registered trademarks of their respective owners. All rights including intellectual property rights belong to our company. Product data may be subject to change. The time until delivery depends on stock status. We are not responsible for the completeness, facts or accuracy of data and figures. The names of the hardware and software described in this manual may be trademarks of the respective manufacturers. If third parties use these for their own purposes, they may infringe the owner's rights.
For details, please refer to http://www.fujitsu.com/fts/resources/navigation/terms-of-use.html.
2021-04-02 WW EN