White Paper
Fujitsu Server PRIMERGY
Memory performance of Xeon Scalable Processor (Sapphire Rapids) based Systems

This white paper explains the essential features of the memory architecture and the latest improvements in the 4th Generation Xeon Scalable Processor (Sapphire Rapids) based FUJITSU Server PRIMERGY and quantifies their effect on the performance of commercial applications.

Version
1.0
2023-07-04

Contents

Introduction

The 4th Generation Xeon Scalable Processor (Sapphire Rapids) inherits the features which previous Xeon Scalable Processor generations (Ice Lake) have and has significantly improved its performance over the previous processor by using Intel's latest manufacturing process. The top model of the processor achieves over 50% performance improvement compared to the previous processor genration in many scenarios. The main reasons for this achievement are the increase in the number of cores per processor from a maximum of 40 to a maximum of 60, and the introduction of a more advanced micro architecture.

In terms of the memory architecture, the processor has evolved significantly. In addition to the increased cache memory, it now supports the latest DDR5 memory. While the maximum memory transfer rate of the previous generation was 3200 MT/s, the new Sapphire Rapids generation supports 4800 MT/s. With these improvements, the theoretical memory bandwidth reaches up to 300 GB/s per processor. In addition, the support of high capacity 256 GB 3DS RDIMM enables to equip 4 TB of memory per processor.

When this processor requests the contents of the memory of the adjacent processor (remote memory), it uses an Ultra Path Interconnect (UPI) link. The performance of remote memory access is not quite high as that of local memory access. This architecture, which distinguishes between local memory and remote memory access, is a Non-Uniform Memory Access (NUMA) type of architecture. The speed of this connection between processors has been raised from 11.2 GT/s in the previous generation to 16.0 GT/s. In addition, the number of links has been increased from a maximum of three links to a maximum of four links.

In the Ice Lake generation, the new option called UMA-Based Clustering was added in addition to the option for the clustering in the processor called SNC (Sub-NUMA Clustering). These options were improved in the Sapphire Rapids generation. Note that 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 of this white paper, 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 eight 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 eight. 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 tranfer rate affects performance. Depending on the processor type, DIMM type, memory capacity, and BIOS settings, they can be either 4800, 4400, 4000 or 3200 MT/s.

In this white paper, factors that affect memory performance are taken up and quantified. For quantification, we use the STREAM and SPECrate2017 Integer benchmarks. STREAM measures the memory bandwidth. SPECrate2017 Integer 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.

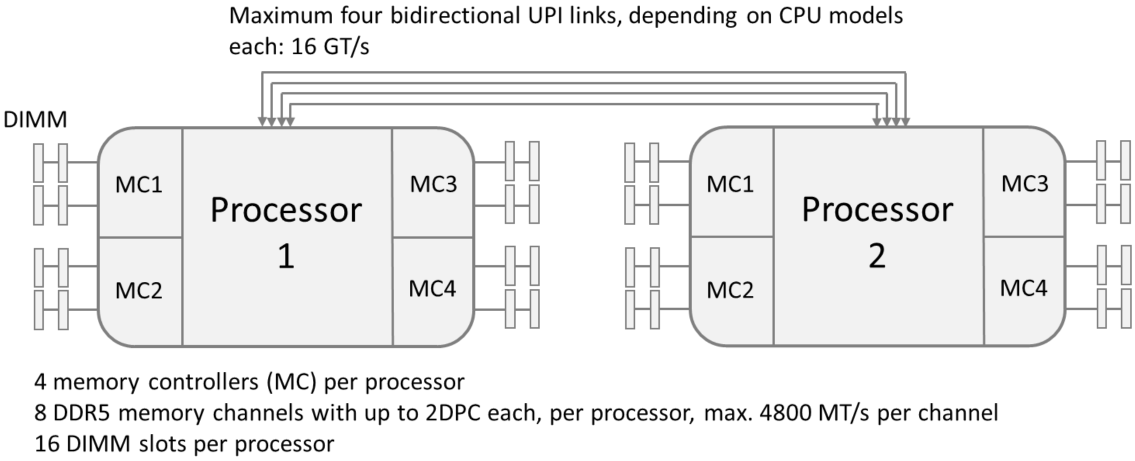
Statements about memory performance under redundancy, i.e. with enabled mirroring or ADDDC sparing, make up the end of this document.

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 transfer rate. 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 4th Generation Xeon Scalable Processor (Sapphire Rapids) based systems.


Maximum four bidirectional UPI links, depending on CPU models
each: 16 GT/s
DIMM
MC1
MC2
Processor 1
MC3
MC4
MC1
MC2
Processor 2
MC3
MC4
4 memory controllers (MC) per processor
8 DDR5 memory channels with up to 2DPC each, per processor, max. 4800 MT/s per channel
16 DIMM slots per processor

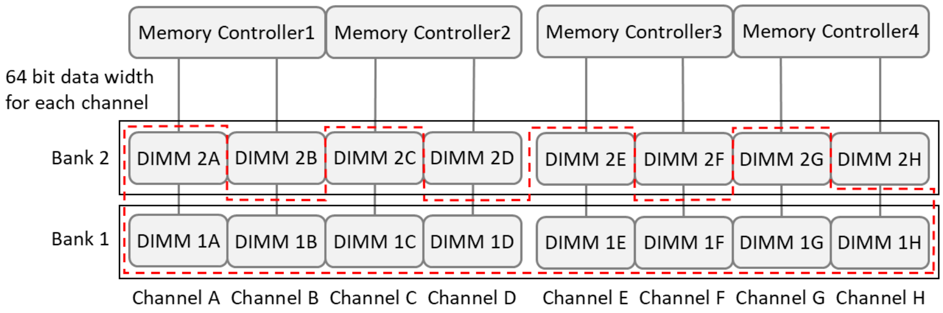
The Sapphire Rapids based PRIMERGY servers have 16 DIMM slots per processor. The data path width is 64 bits, as in the DDR4, but in DDR5, it operates independently as two 32 bit sub-channels. This greatly improves parallel access performance over DDR4.

As with the previous generation of the processor, there are four memory controllers and eight memory channels in one processor. However, with the adoption of 4800 MT/s DDR5 memory, its memory bandwidth is increased by 50% in theory.

In the Sapphire Rapids generation, changing the value of DPC (the term is used hereinafter), which is the number of DIMMs per channel, may cause changes in the memory transfer rate and affect memory performance, depending on the processor model. This is important to note because the previous Xeon Scalable Processor based PRIMERGY servers did not vary in memory transfer rate depending on DPC.

Although the speed of the UPI link in the previous generations (Ice Lake) was 11.2 GT/s, it has been improved to the maximum of 16 GT/s in the Sapphire Rapids generation. Moreover, while the number of UPI links between processors was maximum of three for the 2-socket RX server in the Ice Lake generation, it has been improved to the maximum four for the Sapphire Rapids generation. Thanks to these improvement, it is expected to improve the performance of the applications which have frequent memory accesses between processors, such as the database processing.

We also use the term “memory bank” in the following. In the figure below, a group of eight 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.



Number of DIMM slots per CPU

RX2530/2540 M7:	16 (Bank 1 and 2)
CX2550 M7, TX2550 M7:	8 (Bank 1 only)
CX2560 M7:	12 (Red dashed line)

For a 64-bit bandwidth of the data, the individual DRAM chips on the DIMM are responsible for 4 bits or 8 bits each (see code x4 or x8 for type name). Such a chip group is called a rank. There are DIMM types of one, two, four, or eight ranks.
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.

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

Processors									
Processor	Type	Cores	Threads	L3 Cache [MB]	UPI Links & Speed [GT/s]	Nominal Frequency [GHz]	Max. Turbo Frequency [GHz]	Max. Memory Transfer rate [MT/s]	TDP [Watt]
Xeon Platinum 8490H	XCC	60	120	112.5	4x 16	1.9	3.5	4800	350
Xeon Platinum 8480+	XCC	56	112	105.0	4x 16	2.0	3.8	4800	350
Xeon Platinum 8470	XCC	52	104	105.0	4x 16	2.0	3.8	4800	350
Xeon Platinum 8468V	XCC	48	96	97.5	3x 16	2.4	3.8	4800	330
Xeon Platinum 8468	XCC	48	96	105.0	4x 16	2.1	3.8	4800	350
Xeon Platinum 8462Y+	MCC	32	64	60.0	3x 16	2.8	4.1	4800	300
Xeon Platinum 8460Y+	XCC	40	80	105.0	4x 16	2.0	3.7	4800	300
Xeon Platinum 8458P	XCC	44	88	82.5	3x 16	2.7	3.8	4800	350
Xeon Platinum 8452Y	XCC	36	72	67.5	4x 16	2.0	3.2	4800	300
Xeon Platinum 8450H	XCC	28	56	75.0	4x 16	2.0	2.6	4800	250
Xeon Platinum 8444H	XCC	16	32	45.0	4x 16	2.9	3.2	4800	270
Xeon Gold 6454S	XCC	32	64	60.0	4x 16	2.1	3.4	4800	270
Xeon Gold 6448Y	MCC	32	64	60.0	3x 16	2.1	4.1	4800	225

Xeon Gold 6444Y	MCC	16	32	45.0	3x 16	3.6	4.1	4800	270
Xeon Gold 6442Y	MCC	24	48	60.0	3x 16	2.6	4.0	4800	225
Xeon Gold 6438Y+	MCC	32	64	60.0	3x 16	2.0	4.0	4800	205
Xeon Gold 6438N	MCC	32	64	60.0	3x 16	2.0	3.6	4800	205
Xeon Gold 6438M	MCC	32	64	60.0	3x 16	2.2	3.9	4800	205
Xeon Gold 6430	XCC	32	64	60.0	3x 16	1.9	3.4	4400	270
Xeon Gold 6428N	MCC	32	64	60.0	3x 16	1.8	3.8	4000	185
Xeon Gold 6426Y	MCC	16	32	37.5	3x 16	2.5	4.1	4800	185
Xeon Gold 5420+	MCC	28	56	52.5	3x 16	2.0	4.1	4400	205
Xeon Gold 5418Y	MCC	24	48	45.0	3x 16	2.0	3.8	4400	185
Xeon Gold 5418N	MCC	24	48	45.0	3x 16	1.8	3.8	4000	165
Xeon Gold 5416S	MCC	16	32	60.0	3x 16	2.0	4.0	4400	150
Xeon Gold 5415+	MCC	8	16	22.5	3x 16	2.9	4.1	4400	150
Xeon Silver 4416+	MCC	20	40	37.5	2x 16	2.0	3.9	4000	165
Xeon Silver 4410Y	MCC	12	24	30.0	2x 16	2.0	3.9	4000	150
Xeon Silver 4410T	MCC	10	20	26.25	2x 16	2.7	4.0	4000	150
Xeon Gold 6414U	XCC	32	64	60.0	-	2.0	3.4	4800	250
Xeon Gold 5412U	MCC	24	48	45.0	-	2.1	3.9	4400	185
Xeon Bronze 3408U	MCC	8	8	22.5	-	1.8	1.9	4000	125

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.

DDR5 topics and available DIMM types

The Sapphire Rapids based PRIMERGY servers use the DDR5 SDRAM memory module unlike the previous Xeon Scalable Processor based PRIMERGY servers use. The Sapphire Rapids based systems have the following improvement.

- DDR5 supports a memory frequency up to 4800 MT/s. The previous generation systems with Ice Lake supported up to 3200 MHz using DDR4 SDRAM.
- Like the previous generation systems, the new Sapphire Rapids based system can be equipped with up to 4 TB of DRAM per socket with 256 GB 3DS RDIMMs.

The following table shows the DIMMs supported by the Sapphire Rapids based PRIMERGY servers . In DIMM, there are Registered DIMM (RDIMM) and 3DS Registered DIMM (3DS RDIMM) types. RDIMM x4, RDIMM x8, and 3DS RDIMM cannot be mixed.

DIMM type	Control	Max. Transfer rate (MT/s)	Volt (V)	# of Ranks	Capacity
16GB (1x16GB) 1Rx8 DDR5-4800 R ECC	Registered	4800	1.1	1	16 GB
32GB (1x32GB) 2Rx8 DDR5-4800 R ECC	Registered	4800	1.1	2	32 GB
32GB (1x32GB) 1Rx4 DDR5-4800 R ECC	Registered	4800	1.1	1	32 GB
64GB (1x64GB) 2Rx4 DDR5-4800 R ECC	Registered	4800	1.1	2	64 GB
128GB (1x128GB) 4Rx4 DDR5-4800 3DS R ECC	3DS Registered	4800	1.1	4	128 GB
256GB (1x256GB) 8Rx4 DDR5-4800 3DS R ECC	3DS Registered	4800	1.1	8	256 GB

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

- RDIMM: The control commands of the memory controller are buffered in the register (that gave the name), which is in its own component on the DIMM. This relief for the memory channel enables configurations with up to 2DPC (DIMMs per channel).
- 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 or 3DS RDIMM is desirable is usually determined by the memory capacity required. But 3DS RDIMM have a little overhead in performance.
Depending on the sales area, there are DIMM types that cannot be used.

Definition of the memory transfer rate

There are four types of memory transfer rate on Sapphire Rapids based PRIMERGY servers: 4800, 4400, 4000, and 3200 MT/s. The tranfer rate is defined by the BIOS when the system is switched on and applies per system, not per processor.

The memory transfer rate is affected by the maximum memory transfer rate of the processor model, the DPC value of the memory configuration, and the BIOS setting. The maximum memory transfer rate of processors is 4800, 4400, or 4000 MT/s, depending on the model, as shown in the table in DIMM slots and memory controllers. For models with a maximum memory transfer rate of 4800 MT/s, the memory transfer rate in a 2DPC configuration is reduced to 4400 MT/s. This cannot be disabled in the BIOS.

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 optimezed”, the effective memory tranfer rate is as shown in the following table. This is the default BIOS setting.

DDR Performance = Performance optimized				
Processor type	RDIMM		3DS RDIMM	
	1DPC	2DPC	1DPC	2DPC
DDR5-4800	4800	4400	4800	4400
DDR5-4400	4400	4400	4400	4400
DDR5-4000	4000	4000	4000	4000

When “Energy optimezed” is selected, the effective memory transfer rate is as shown in the following table. As mentioned earlier, DDR5 memory modules do not currently have a low voltage version. The DDR5 module always operates at a voltage of 1.1 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				
Processor type	RDIMM		3DS RDIMM	
	1DPC	2DPC	1DPC	2DPC
DDR5-4800	3200	3200	3200	3200
DDR5-4400	3200	3200	3200	3200
DDR5-4000	3200	3200	3200	3200

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 has 10 parameters. The default is underlined each time.

- Memory Mode : Independent / Mirroring / Address Range Mirroring
- ADDDC Sparing : Disabled / Enabled
- DDR5 ECS : Disabled / Enabled
- NUMA : Disabled / Enabled
- Virtual NUMA : Disabled / Enabled
- DDR Performance : Performance optimized / Energy optimized
- PPR Type : Hard PPR / Soft PPR / PPR Disabled
- Patrol Scrub : Disabled / Enabled
- SNC(Sub NUMA) : Disabled / Enable SNC2 / Enable SNC4
- UMA-Based Clustering : Hemisphere (2-clusters) / Quadrant (4-clusters)

The first three parameters, "Memory Mode", "ADDDC (Adaptive Double Device Data Correction) Sparing", and "DDR5 ECS (Error Check and Scrub)" 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. It enables error correction of failures on two DRAM devices. When mirroring is enabled, ADDDC Sparing is disabled.
"DDR5 Error Check and Scrub (ECS)" is a feature of DDR5 that improves reliability and error correction. DDR5 enables error correction inside the device when reading data (On-Die ECC function). ECS uses this function to read data in DRAM and to correct and write back data in case of errors. When enabled, checks are performed periodically.
There are restrictions on the configuration available for "Memory Mode" or "ADDDC Sparing". Please refer to the respective configurator for these.
If these functions are requested, 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 fourth 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 fifth parameter "Virtual NUMA" is used when Windows runs on a processor with more than 64 logical CPUs. Because the maximum number of logical CPUs in a processor group that Windows

uses to manage logical CPUs is 64, logical CPUs that exceed the limit are managed as a separate processor group. As a result, the size of the processor group is unbalanced, resulting in a disadvantage in performance. With “Virtual NUMA” enabled, a processor is used by dividing it into two equally sized virtual NUMA nodes. “Virtual NUMA” is similar to SNC described below but it doesn’t have the same effect on the performance improvement as SNC has.
The sixth parameter “DDR Performance” concerns memory transfer rate and was dealt with in the last section in detail.
The seventh parameter “PPR type” treats Post Package Repair (PPR), which is the feature of DDR5. 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 eighth 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 last two parameters are the settings for the clustering in the processor.
“SNC(Sub NUMA)” is a parameter for dividing the processor cores, the L3 cache and the memory controllers into clusters. Three options are available for XCC type processors: “Enable SNC4”, “Enable SNC2”, and “Disabled”. For MCC type processors, two options are available: “Enable SNC2” and “Disabled”. The default setting is “Disabled”.
When set to “Enable SNC4”, these resources are assigned to either of the clusters in the processor divided into four. When set to “Enable SNC2”, they are assigned to either of the two clusters in the processor. The cluster is treated as one NUMA domain from the operating system. If disabled, the processor is treated as a single cluster which is UMA (Uniform Memory Access).
SNC improves the latency of the access to the L3 cache and memory in NUMA node. SNC is particularly recommended for NUMA optimized applications because it can minimize local memory latency and maximize local memory bandwidth.
The “UMA-Based Clustering” parameter is only available for XCC type processors. It changes the behavior of cache coherency in a UMA configuration. With default “Quadrant (4-clusters)” setting, it divides the L3 cache and the memory controllers to four clusters based on their proximity to each other. Cores are not split. With “Hemisphere (2-clusters)” setting, they are divided into two clusters. The smaller the divided area, the shorter the distance between the L3 cache and memory, and the better the latency.
There are limitations to the configuration available for SNC or UMA-Based Clustering. When DIMMs are populated as shown in the Upgrade and Maintenance Manual, SNC2 is available when the number of DIMMs is a multiple of two, SNC4 and Quadrant are available when the number of DIMMs is a multiple of four, and Hemisphere is available when the number of DIMMs is a multiple of two (except six).

Performant memory configurations

The memory transfer rate and the number of memory channels used greatly affect memory performance. Since the memory transfer rate depends on the type of processor installed and DPC. In addition, Xeon Scalable Processor has eight 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 (a 16 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 eight memory channels per processor are the same. In each bank configuration, the same type of eight DIMMs set is used. This ensures that memory accesses are evenly distributed among these memory system resources. Technically speaking, the optimum 8-way interleaving is realized via the memory channel. In this document, this is called Performance Mode configuration.

Xeon Scalable Processor (Sapphire Rapids) 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	Max. memory transfer rate MT/s	Comment
128 GB	256 GB	DDR5-4800 R	16		4800	
192 GB	384 GB	DDR5-4800 R	16	8	4400	Mixed configuration
256 GB	512 GB	DDR5-4800 R	16	16	4400	
256 GB	512 GB	DDR5-4800 R	32		4800	
384 GB	768 GB	DDR5-4800 R	32	16	4400	Mixed configuration
512 GB	1024 GB	DDR5-4800 R	32	32	4400	
512 GB	1024 GB	DDR5-4800 R	64		4800	
768 GB	1536 GB	DDR5-4800 R	64	32	4400	Mixed configuration
1024 GB	2048 GB	DDR5-4800 R	64	64	4400	
1024 GB	2048 GB	DDR5-4800 3DS R	128		4800	
2048 GB	4096 GB	DDR5-4800 3DS R	128	128	4400	

2048 GB	4096 GB	DDR5-4800 3DS R	256		4800	Maximum configuration at memory transfer rate of 4800 MT/s
4096 GB	8192 GB	DDR5-4800 3DS R	256	256	4400	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. 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 8 per bank.
The smallest configuration in the table has 128 GB for one processor because the eight 16 GB DIMMs (i.e., 128 GB) must be counted for each processor.
The Performance Mode configuration requires an identical DIMM group of eight per bank, but it does not forbid different DIMM sizes in different banks if the following restrictions are observed:

- RDIMMs 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 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 but please refer to the respective configurator for details.

- RDIMMs 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, two, four, six, eight, twelve, or sixteen.

You also need to pay attention to configurations where the number of DIMMs per processor does not become a multiple of eight, 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, operation with one or two DIMMs 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.

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 16 DIMMs in the first processor and eight DIMMs in the second processor satisfies the Performance Mode criteria. This is because the memory channels per processor are handled in the same way. 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 RX2540 M7 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 RX2540 M7
Processor	2x Xeon Platinum 8480+ (56 cores, 2.0 GHz, Max. memory transfer rate 4800 MT/s) 2x Xeon Gold 6430 (32 cores, 2.2 GHz, Max. memory transfer rate 4400 MT/s) 2x Xeon Silver 4416+ (20 cores, 2.0 GHz, Max. memory transfer rate 4000 MT/s)
Memory types	16GB (1x16GB) 1Rx8 DDR5-4800 R ECC 32GB (1x16GB) 2Rx8 DDR5-4800 R ECC 32GB (1x32GB) 1Rx4 DDR5-4800 R ECC 64GB (1x64GB) 2Rx4 DDR5-4800 R ECC 128GB (1x128GB) 4Rx4 DDR5-4800 3DS R ECC 256GB (1x256GB) 8Rx4 DDR5-4800 3DS R ECC
Disk subsystem	1x SATA 6G SSD (via onboard SATA controller)
Software	
BIOS	R1.6.0
Operating system	SUSE Linux Enterprise Server 15 SP4

The 64 GB 2Rx4 RDIMM was usually used for the test set described below. All the other DIMMs listed in the table were used only in the test set for the impact of the DIMM type except for the evaluation of the impact of interleaving across the memory channels.

The following table shows relative performance. The absolute measurement values for the STREAM and SPECrate2017 Integer 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 Integer.

STREAM Benchmark

The STREAM benchmark (Developer: Mr. John McCalpin) 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 Integer Benchmark

The SPECrate2017 Integer benchmark has been added as a model for commercial application performance. This is part of the Standard Performance Evaluation Corporation (SPEC) SPEC CPU2017. SPEC CPU2017 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.

SPEC CPU2017 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.

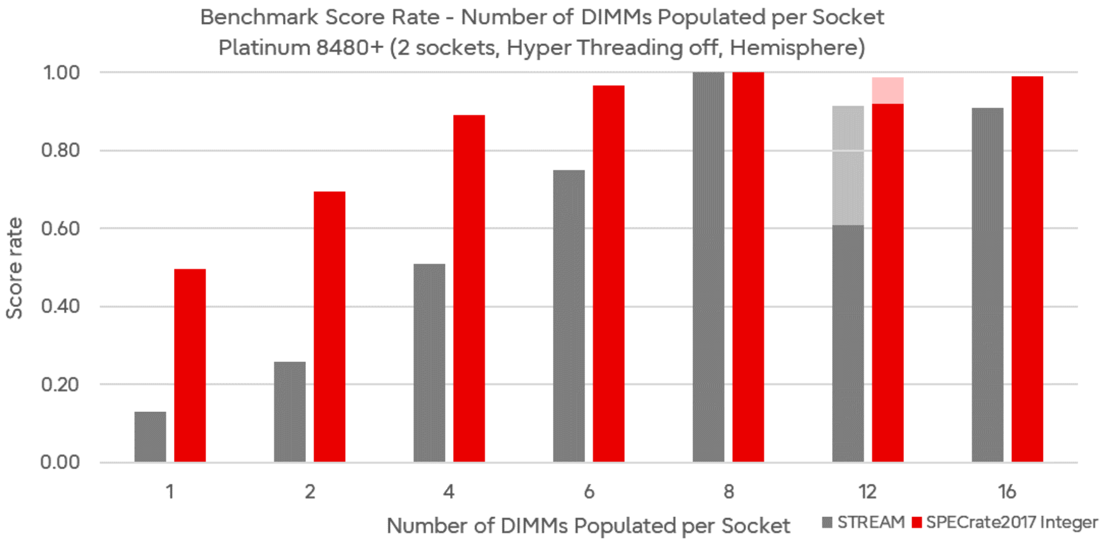
A distinction is also made in the suites between the speed run with only one process and the rate run with a configurable number of processes working in parallel. The second version is evidently more interesting for servers with their large number of processor cores and hardware threads.

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 SPEC CPU2017. The SPECrate2017 Integer suite was selected, because commercial applications predominate in the use of PRIMERGY servers.

Interleaving across the memory channels

Interleaving is a method of setting a physical address area so that eight 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 following figure shows the ratio of the performance, when DIMMs are not mounted in a set of eight pieces per processor and the ideal 8-way interleave is not performed; the value is considered as 1 when the number of DIMMs is eight. The number of DIMMs populated per one processor is limited to one, two, four, six, eight, twelve, or sixteen for the Sapphire Rapids based PRIMERGY servers. Only the results for the settings with “SNC(Sub NUMA) = Disabled” and “UMA-Based Clustering = Hemisphere” are shown here. The results for the settings with “SNC(Sub NUMA) = Enable SNC2” or “Enable SNC4”, and for the settings with “SNC(Sub NUMA) = Disabled” and “UMA-Based Clustering = Quadrant” are omitted since configurable DIMM patterns are limited for them and the performance ratios of them are almost same as that shown here.
The processor model used for this test is a Xeon Platinum 8480+. The DIMM type used is 128 GB 4Rx4 3DS RDIMM. This was chosen to ensure that there is enough memory to satisfy the working set.



In particular, marked declines are seen in the STREAM index that measures memory throughput. When the number of DIMMs is equal to or less than eight, the performance is improved according to the increase in the number of DIMMs. With 12 DIMMs or more, the 2DPC configuration reduces the memory transfer rate, resulting in about 10% drop in the performance of STREAM.

You need to pay attention for the configuration with 12 DIMMs per one processor. In the configuration where 8 DIMMs are populated in Bank 1 and 4 DIMMs are in Bank 2, the physical address area configured by the former DIMMs is eight-channel interleaved and that configured by the latter DIMMs is four-channel interleaved. Because of it, the performance of the application depends on the area where it runs. On 12 DIMMs in the figure above, two types of results are shown in two bar charts: light and dark.

Evaluation on SPECrate2017 Integer 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 Integer 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 or 6-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.
In addition, as shown by the results of 12 DIMM configuration, the imbalanced DIMM configuration is not recommended from the view of maximizing performance with stability.

Memory transfer rate

The memory transfer rate of the Xeon Scalable Processor based PRIMERGY servers may vary depending on DPC. The type of the processors and the BIOS parameter also affect the memory transfer rate. Power-saving (managed via the BIOS parameter DDR Performance) can be the reasons why the effective memory tranfer rate 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 table are based on an ideal case, in other words, the maximum speed in the processor class.

Benchmark	Processor type	Maximum memory transfer rate	3200 MT/s	4000 MT/s	4400 MT/s	4800 MT/s
STREAM	Platinum 8480+	4800 MT/s	0.72			1.00
	Gold 6430	4400 MT/s	0.78		1.00	
	Silver 4416+	4000 MT/s	0.89	1.00		
SPECrate2017 Integer	Platinum 8480+	4800 MT/s	0.92			1.00
	Gold 6430	4400 MT/s	0.96		1.00	
	Silver 4416+	4000 MT/s	0.98	1.00		

The processor models used in this test are the Xeon Platinum 8480+ (Maximum memory transfer rate 4800 MT/s), Xeon Gold 6430 (Maximum memory transfer rate 4400 MT/s), and Xeon Silver 4416+ (Maximum memory transfer rate 4000 MT/s). The DIMM type used is 64GB 2Rx4 RDIMM. It is used with a 1DPC configuration.
If you set "DDR Performance = Energy optimized" in the BIOS, the frequency will always be 3200 MT/s.

Influence of the DIMM types

Maximum six types of DIMMs are planned when the Sapphire Rapids 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 8480+ (Maximum memory transfer rate 4800 MT/s) and Xeon Gold 6430 (Maximum memory tranfser rate 4400 MT/s).
- 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 16 for 1DPC measurement and 32 for 2DPC measurement.
- All the measurements were carried out with the maximum memory transfer rate of each processor type. That is, the DIMMs were running at 4800 MT/s for 1DPC configration with Xeon Platinum 8480+, and 4400 MT/s for 2DPC configration with it. In the case of Xeon Gold 6430, they were running at 4400 MT/s regardless of DPC.
- Hyper Threading setting were disabled for the measurement with Xeon Platinum 8480+ due to the limited memory capacity required for measurement.
- The table is standardized to the 1DPC configuration with the 64 GB 2Rx4 RDIMM (highlighted in bold print), which will provide the best memory performance. This DIMM is preferred in benchmarking as long as the memory capacity that can be achieved with it is sufficient.

			Platinum 8480+ (Max. 4800 MT/s)		Gold 6430 (Max. 4400 MT/s)	
DIMM type	Configuration	# of ranks per channel	STREAM	SPECrate 2017 Integer	STREAM	SPECrate 2017 Integer
16GB (1x16GB) 1Rx8 DDR5-4800 R ECC	1DPC	1	0.86	0.96	0.87	0.98
	2DPC	2	0.91	0.99	0.97	1.00
32GB (1x32GB) 2Rx8 DDR5-4800 R ECC	1DPC	2	1.00	1.00	1.00	1.00
	2DPC	4	0.85	0.98	0.93	1.00
32GB (1x32GB) 1Rx4 DDR5-4800 R ECC	1DPC	1	0.85	0.97	0.87	0.98
	2DPC	2	0.91	0.99	0.97	1.00
64GB (1x64GB) 2Rx4 DDR5-4800 R ECC	**1DPC**	**2**	**1.00**	**1.00**	**1.00**	**1.00**
	2DPC	4	0.85	0.98	0.92	1.00
128GB (1x128GB) 4Rx4 DDR5-4800 3DS R ECC	1DPC	4	0.94	0.99	1.00	0.99
	2DPC	8	0.86	0.97	0.91	0.98
256GB (1x256GB) 8Rx4 DDR5-4800 3DS R ECC	1DPC	8	0.99	0.99	0.99	0.99
	2DPC	16	0.73	0.96	0.81	0.98

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.

As the table above shows, you can see that the performance is better when the number of the ranks per memory channel is two than that of one rank. On the other hand, a 1DPC configuration using 16 GB 1Rx8 RDIMM or 32 GB 1Rx4 RDIMM, i.e., 1-way rank interleaving results in the noticeable performance degradation.

Optimization of the clustering in the processor

The Sapphire Rapids processor has the setting called "UMA-Based Clustering" in addition to "SNC(Sub NUMA)" for the settings of the clustering in the processor. The Sapphire Rapids based PRIMERGY servers can select the four clustering modes, SNC4, SNC2, Quadrant, and Hemisphere by these settings. 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 1DPC configurations with 64 GB 2Rx4 RDIMMs.

The table shows that performance is affected in the range of one to four percentage point. 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 Integer for commercial application performance therefore only applies at this stage in a restricted manner.

Benchmark	Processor type	SNC4	SNC2	Quadrant	Hemisphere
STREAM	Platinum 8480+	1.03	1.01	1.00	1.00
	Gold 6430	1.04	1.02	1.00	1.00
	Silver 4416+	-[1]	1.03	-[1]	1.00
SPECrate2017 Integer	Platinum 8480+	1.03	1.02	1.01	1.00
	Gold 6430	1.03	1.01	1.00	1.00
	Silver 4416+	-[1]	1.01	-[1]	1.00

The values of the BIOS settings for each clustering mode are the following.

Clustering mode	SNC(Sub NUMA)	UMA-Based Clustering
SNC4	Enable SNC4	-[2]
SNC2	Enable SNC2	-[2]
Quadrant	Disabled	Quadrant (4-clusters)
Hemisphere	Disabled	Hemisphere (2-clusters)

[1] This setting is not configurable if the processor is MCC.

[2] This setting is not configurable when SNC(Sub NUMA) is set to Enable SNC4 or Enable SNC2.

Access to remote memory

For the tests using the STREAM and SPECrate2017 Integer 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 8 way rank-interleaved Performance Mode configuration with 64 GB 2Rx4 RDIMMs operating at the highest possible memory transfer rate 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. Especially, Silver 4416+ processor which has two UPI links shows the noticeable performance degradation in the bandwidth intensive benchmark such as STREAM.

Benchmark	Processor type	# of UPI links	NUMA = Enabled	NUMA = Disabled
STREAM	Platinum 8480+	4	1.00	0.68
	Gold 6430	3	1.00	0.66
	Silver 4416+	2	1.00	0.58
SPECrate2017 Integer	Platinum 8480+	4	1.00	0.88
	Gold 6430	3	1.00	0.90
	Silver 4416+	2	1.00	0.92

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

We evaluate the impact of two redundancy and reliability options on performance here.
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 replaces 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 1DPC configuration with 64 GB 2Rx4 RDIMM. The columns in the table correspond to the default settings and the options of the BIOS parameter “Memory Mode” and “ADDDC 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, loss of performance isn’t observed.

Benchmark	Processor type	Default	Mirroring	ADDDC Sparing
STREAM	Platinum 8480+	1.00	0.72	1.00
	Gold 6430	1.00	0.73	1.00
	Silver 4416+	1.00	0.82	1.00
SPECrate2017 Integer	Platinum 8480+	1.00	0.95	1.00
	Gold 6430	1.00	0.97	1.00
	Silver 4416+	1.00	0.98	1.00

Literature

PRIMERGY Servers

https://www.fujitsu.com/global/products/computing/servers/primergy/

Memory performance

This Whitepaper

🇺🇸 https://docs.ts.fujitsu.com/dl.aspx?id=fec08359-b897-435c-96ff-b2bd0daabbfc

🇯🇵 https://docs.ts.fujitsu.com/dl.aspx?id=c2f30fb8-a486-4934-a773-b76b18c5d407

Past Issue of White Paper

Memory performance of Xeon Scalable Processor (Ice Lake) based Systems

https://docs.ts.fujitsu.com/dl.aspx?id=1930d389-7521-4c85-bcf9-86a71a14a7c3

Benchmarks

STREAM

https://www.cs.virginia.edu/stream/

SPECcpu2017

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

BIOS settings

BIOS optimizations for 4th Generation Xeon Scalable Processor-based systems

https://docs.ts.fujitsu.com/dl.aspx?id=d9d38fdc-87de-4b78-9c2f-5ad25ceb32ae

PRIMERGY Performance

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

Document change history

Version	Date	Description
1.0	2023-07-04	Initial version

Contact

Fujitsu

Web site: https://www.fujitsu.com

PRIMERGY Performance and Benchmarks

mailto:fj-benchmark@dl.jp.fujitsu.com