White Paper
Fujitsu PRIMERGY Servers
RAID Controller Performance 2013

This technical documentation is aimed at the persons responsible for the disk I/O performance of Fujitsu PRIMERGY servers. The document is intended to help you become acquainted - from a performance viewpoint - with the options and application areas of various RAID controllers for internal disk subsystems. Depending on the requirements for data security and performance as well as planned or existing server configuration, specific recommendations arise for the selection and parameterization of controllers. Controllers that are available for PRIMERGY systems in 2013 are to be considered here.

Version
2.0a
2013-08-01

Contents

Document history

Version 1.0a

Version 2.0

- Adaptation of the document title to 2013
- Addition of controllers with the support of PCIe Gen3, which were introduced in 2013
- Addition of the MegaRAID Advanced Software options
- New measurement results with the controller comparisons for RAID 1 based on modified recommendations for the cache settings
- Minor corrections

Version 2.0a

- Minor corrections

Introduction

Hard disks are a security factor as well as critical performance components in the server environment. It is thus important to bundle the performance of such components via intelligent organization so that they do not cause a system bottleneck. They should simultaneously compensate for any failure of an individual component. Methods exist for arranging several hard disks in an array so that any hard disk failure can be compensated. This is known as a “Redundant Array of Independent Disks” or in short RAID. Special RAID controllers are normally used.

The PRIMERGY servers are available in a wide range of internal configuration versions with different RAID controller and hard disk configurations. The “Modular RAID” concept that is offered as a standard for all servers in the PRIMERGY family consists of a modular controller family and standardized management via the Fujitsu RAID Manager software known as “ServerView RAID”. The comprehensive offer of RAID solutions enables the user to select the appropriate controller for a particular application scenario. The performance of the disk subsystem is defined by the controller, the selected hard disks and the features of the RAID level.

Several documents have been created in the PRIMERGY white paper series which illustrate all aspects of “Modular RAID” regarding performance:

- We recommend - as a comprehensive introduction to disk I/O performance - the White Paper “Basics of Disk I/O Performance”.
- The document “Single Disk Performance” presents the hard disks currently available for PRIMERGY and their performance in different application scenarios.
- This document “RAID Controller Performance 2013” covers all the RAID controllers currently on offer for PRIMERGY and their performance.

When sizing internal disk subsystems for PRIMERGY servers you can proceed in such a way that a suitable hard disk type is selected and the necessary number of hard disks for the required RAID level is estimated using rules of thumb. Due to the number and technology of the hard disks that are to be connected as well as the required RAID level the RAID controller is self-evident. This may be adequate for years in order to accurately size a disk subsystem.

However, the technology of storage media (for example Solid State Drives, or in short SSDs) or in the internal interfaces of the server has progressed over the years and the new disk subsystem no longer meets the increased requirements. Or, in a productive server configuration the application scenario changes and the achieved disk I/O performance is - despite an adequate number of hard disks - not as desired. In both these cases it can be worthwhile to look at the influence of the RAID controller on performance more closely. Sometimes the right controller, or even simply the correctly configured controller, is prerequisite for the best possible performance.

That outlines the objective of this document. First, there will be an overview of the internal RAID controllers that are available for the PRIMERGY systems. The throughput limits of the involved controller interfaces will then be presented under the aspects of performance. After a brief introduction into the measurement context, the different RAID controllers will be compared at various RAID levels and in different application scenarios, which will be substantiated by the measurement results.

In the past the terms “Hard Disk” and also “Hard Disk Drive” (HDD) were used for a hard magnetic-coated, rotating, digital, non-volatile storage medium that could be directly addressed. Technical development has now seen new “hard disk” versions introduced as storage media; they use the same interface to the server and are accordingly handled as hard disks by the server. An SSD, which as an electronic storage medium does not contain any moving parts, can be stated as a typical example, but which nevertheless is also colloquially referred to as a hard disk. Throughout this document the term “hard disk” is used as a generic term, with the names “SSD” and “HDD” being used as a means of differentiation.

This document specifies hard disk capacities on a basis of 10 (1 TB = 10^{12} bytes) while all other capacities, file sizes, block sizes and throughputs are specified on a basis of 2 (1 MB/s = 2^{20} bytes/s).

RAID controllers for PRIMERGY: Basics

Firstly, the RAID controllers that are available for PRIMERGY servers are to be presented with their essential functions in this section Then the throughput limits, which result for the individual controllers due to their interfaces in the server, will be looked at more closely. We will subsequently deal with the possible settings of the controllers, and ultimately discuss the characteristics of onboard controllers.

Presentation of the RAID controllers

The following table summarizes the most important data with regard to the functionality of the available RAID controllers.

To simplify the naming of these controllers this white paper will for the most part only use the short name from the column “Alias”, thus for example LSI2008. These names come from the basic chip name of the controller and, optionally, the cache size.

Controller name	Alias	FF	Cache	Fre-quency	Supported interfaces		Max. # disks	RAID levels	BBU/ FBU
LSI SW RAID on Intel C200 (Onboard SATA)	Cougar Point	I	-	3G	SATA 3G		4	JBOD, 0, 1, 10	-/-
LSI SW RAID on Intel C600 (Onboard SATA)	Patsburg A	I	-	3G	SATA 3G		4	0, 1, 10	-/-
LSI SW RAID on Intel C600 (Onboard SAS)	Patsburg B	I	-	3G	SATA 3G SAS 3G		4	0, 1, 10	-/-
RAID 0/1 SAS based on LSI MegaRAID 4Port	LSI1064	P	-	3G	SATA 3G SAS 3G	PCIe 1.0 x4	4	0, 1, 1E	-/-
RAID 0/1 SAS based on LSI MegaRAID 8Port	LSI1068	P	-	3G	SATA 3G SAS 3G	PCIe 1.0 x4	8	0, 1, 1E	-/-
RAID Ctrl SAS 6G 0/1 (D2607)	LSI2008	P	-	6G	SATA 3G/6G SAS 3G/6G	PCIe 2.0 x8	8	0, 1, 1E, 10	-/-
RAID Ctrl SAS 6G 5/6 512 MB (D2616)	LSI2108	P	512 MB	6G	SATA 3G/6G SAS 3G/6G	PCIe 2.0 x8	8	0, 1, 5, 6, 10, 50, 60	✓/-
PY SAS RAID Mezz Card 6Gb		M							
RAID Ctrl SAS 6G 5/6 1GB (D3116)	LSI2208-1G-2.0	P	1 GB	6G	SATA 3G/6G SAS 3G/6G	PCIe 2.0 x8	8	0, 1, 1E, 5, 6, 10, 50, 60	-/✓
RAID Ctrl SAS 6G 5/6 1GB (D3116C)	LSI2208-1G-3.0	P	1 GB	6G	SATA 3G/6G SAS 3G/6G	PCIe 3.0 x8	8	0, 1, 1E, 5, 6, 10, 50, 60	-/✓
PY SAS RAID HDD Module [1)2)]	LSI2208-512-2.0 [1)2)]	M	512 MB	6G	SATA 3G/6G SAS 3G/6G	PCIe 2.0 x8	8	0, 1, 1E, 5, 6, 10, 50, 60 [2)]	-/✓
PY SAS RAID HDD Module 3.0 [1)2)]	LSI2208-512-3.0 [1)2)]	M	512 MB	6G	SATA 3G/6G SAS 3G/6G	PCIe 3.0 x8	8	0, 1, 1E, 5, 6, 10, 50, 60 [2)]	-/✓
PY SAS RAID HDD Module w/o cache [1)]	LSI2208-Lite-2.0 *	M	-	6G	SATA 3G/6G SAS 3G/6G	PCIe 2.0 x8	8	0, 1	-/-
PY SAS RAID HDD Module w/o cache 3.0 [1)]	LSI2208-Lite-3.0 *	M	-	6G	SATA 3G/6G SAS 3G/6G	PCIe 3.0 x8	8	0, 1	-/-

1) *These controllers are only available for PRIMERGY BX920 S3.*

2) *All RAID levels except for 0 and 1 can only be used for the LSI2208-512 controller in connection with more disk slots in PRIMERGY SX980 Storage Blades.*

The column “FF” expresses the form factor; “I” means “integrated”, “P” means “PCIe slot” and “M” means “mezzanine card”. The column “Max. # disks” specifies the maximum number of hard disks that can be directly run on the controller within the context of the RAID Management concept of PRIMERGY servers. This information can be of help in detecting whether the controller could be a theoretical bottleneck. In some PRIMERGY models so-called “expanders” (special components defined in the SAS standard) are used – in connection with specific controller models – in order to further increase the maximum number of hard disks. In so doing, the expander cannot increase the bandwidth of the existing ports, but makes it available in total to all connected hard disks.

In the case of “RAID 0/1” SAS-RAID controllers the LSI2008 controller is the 6G successor of the LSI1068 controller, whereas the LSI2208 controllers are the 6G successors of the LSI2108 controller, and together both form the “RAID 5/6” SAS-RAID controller family.

The LSI2208 controller exists with two different cache sizes, shortened to the two alias names LSI2208-1G and LSI2208-512. If only LSI2208 is mentioned below, any statements made apply for both cache sizes.

A new chip, which supports PCIe Gen3, was introduced in 2013 for the group of three LSI2208 controllers. In order to make a better distinction in this document the alias names have been extended by the endings 2.0 and 3.0 respectively. If alias names are used without these endings, the statement applies for both controller versions.

In the evaluation of the performance of disk subsystems, processor performance and memory configuration do not for the most part play a significant role in today's systems - a possible bottleneck usually affects the hard disks and the RAID controller, and not CPU or memory of the server system. Thus the various RAID controllers can be compared independently of the PRIMERGY models in which they are used - even if all the configurations are not possible in all PRIMERGYs due to their expandability with hard disks.

The following table is a compilation of which RAID controllers are released or will be released in the near future (or were released in the past) in the individual PRIMERGY systems for the connection of hard disks at the time this white paper was written and how many hard disks the models support at most. Please see the configurators of the systems for the possible combinations of PRIMERGY configuration versions and controllers.

PRIMERGY	Max. # disks	Expander	Onboard controller			Controller with PCIe interface						
			CougarPoint	Patsburg A	Patsburg B	LSI1064	LSI1068	LSI2008	LSI2108	LSI2208-1GB	LSI2208-512	LSI2208-Lite
BX920 S3	2			✓	✓						✓	✓
BX924 S3	2			✓	✓				✓			
CX250 S1	8	-/✓		✓					✓			
CX270 S1	24	-/✓		✓				✓	✓			
RX100 S7	4		✓			✓		✓	✓			
RX100 S7p	4		✓			✓		✓	✓	✓		
RX200 S7	8			✓	✓			✓	✓	✓		
RX300 S7	16	-/✓		✓	✓			✓	✓	✓		
RX350 S7	24	-/✓		✓	✓			✓	✓	✓		
RX500 S7	8			✓	✓			✓		✓		
RX600 S6	8							✓	✓	✓		
RX900 S2	8								✓			
SX940 S1	4					✓		✓	✓			
SX960 S1	10	✓							✓	✓		
TX100 S3	4		✓						✓			
TX100 S3p	4		✓						✓	✓		
TX120 S3	4		✓			✓		✓	✓			
TX120 S3p	4		✓			✓		✓	✓	✓		
TX140 S1	8		✓				✓	✓	✓			
TX140 S1p	8		✓				✓	✓	✓	✓		
TX300 S7	24	-/✓		✓	✓			✓	✓	✓		

The LSI2108 controller is available as a mezzanine card in the server blade PRIMERGY BX924 S3. The same applies for the LSI2208 controllers in the case of the server blade PRIMERGY BX920 S3.

This white paper only examines the previously mentioned mezzanine cards in connection with internal hard disks in the same server blade. The performance of PRIMERGY SX980 storage blades, which are connected to such mezzanine cards by means of SAS switches, is dealt with in the performance report of the PRIMERGY SX980.

Controller interfaces and their throughput limits

A RAID controller needs an interface on the one hand to the hard disks and on the other hand to the chip set of the motherboard. The first one is typically SAS or SATA, the second one is typically PCIe or, in the event of integrated onboard controllers, an internal bus interface of the motherboard. The upper limits for the throughputs of SAS, SATA and PCIe have been put together below.

SAS and SATA

“Serial Attached SCSI” (SAS) and “Serial Advanced Technology Attachment” (SATA) are serial interfaces, whose data throughput depends on the frequency. These interfaces are used to connect non-volatile storage media, such as hard disks, optical drives and tape drives.

Type	Frequency	Theoretical throughput	Practical throughput (85%)
SAS 3G / SATA 3G	3000 MHz	286 MB/s	243 MB/s
SAS 6G / SATA 6G	6000 MHz	572 MB/s	486 MB/s

The frequencies are to be found in the abbreviation 3G or 6G as part of the controller or hard disk name. Alternatively, a version number is also used with SAS - 1.0 for 3G and 2.0 for 6G.

The theoretically achievable throughput is calculated as follows: 1 bit per 1 Hz, minus 20% redundancy of the serial transfer due to the so-called 8b/10b coding. The throughput that can be achieved in practice can be estimated by multiplying this with 0.85. This 85% is a mean empirical value taken from the values that have been observed over the years for various components.

All the components of a connection between end devices must use the same version of the SAS or SATA protocol. In addition to the hard disks, these also include the controllers and any expanders that are possibly used. If different components come together here, the most high-performance standard that is jointly supported by all components is automatically used, i.e. a lower frequency is possible. In this respect, the higher protocols are downwards compatible.

Whereas each port with SATA is often individually connected to a hard disk, four SAS connections and cables are frequently put together and referred to as an “x4 SAS” or “x4 wide port”. This makes it possible to directly connect a maximum of four SAS hard disks via a backplane. The throughput of x4 SAS is four times that of the corresponding individual SAS connection; this also applies similarly for SATA.

Interface	Connection	Frequency	Theoretical throughput	Practical throughput (85%)
SAS 3G / SATA 3G	1 × x4	3000 MHz	1144 MB/s	973 MB/s
SAS 3G / SATA 3G	2 × x4	3000 MHz	2289 MB/s	1945 MB/s
SAS 6G / SATA 6G	1 × x4	6000 MHz	2289 MB/s	1945 MB/s
SAS 6G / SATA 6G	2 × x4	6000 MHz	4578 MB/s	3890 MB/s

Some PRIMERGY models can be expanded with a larger number of hard disks than the controller has hard disk connections. In this case, the number of connectable hard disks is increased by means of an expander. As already mentioned, an expander can only distribute the data flow, not increase the throughput.

The SAS protocol is defined in such a way that it can also transport the SATA protocols of the same or a lower frequency (tunneling). This enables the controllers of both SAS versions to communicate with SATA hard disks. Conversely, it is not possible to connect SAS hard disks via a SATA interface.

PCIe

PCIe is also a serial interface between the controller and the motherboard. The connectors are designed with a different width and number of lanes. x4 (four lanes) and x8 (eight lanes) are normal, whereby the actual number of electrically used lanes is the important thing (here referred to below as the "functional PCIe width"). The throughput of a lane is also determined by the frequency.

Interface	Connection	Frequency	Theoretical throughput	Practical throughput (90%)
PCIe.1.0, PCIe Gen1	x4	2500 MHz	954 MB/s	858 MB/s
PCIe.1.0, PCIe Gen1	x8	2500 MHz	1907 MB/s	1716 MB/s
PCIe.2.0, PCIe Gen2	x4	5000 MHz	1907 MB/s	1716 MB/s
PCIe.2.0, PCIe Gen2	x8	5000 MHz	3815 MB/s	3433 MB/s
PCIe 3.0, PCIe Gen3	x4	8000 MHz	3756 MB/s	3380 MB/s
PCIe 3.0, PCIe Gen3	x8	8000 MHz	7512 MB/s	6761 MB/s

PCIe 1.0 is also often referred to as "PCIe Gen1", PCIe 2.0 as "PCIe Gen2" and PCIe 3.0 as "PCIe Gen3".

The theoretically achievable throughput is calculated as follows: 1 bit per 1 Hz multiplied by the number of connections (x4 or x8), minus 20% redundancy of the serial transfer due to the so-called 8b/10b coding for PCIe 1.0 and 2.0 and minus 1.54% redundancy due to the 128b/130b coding for PCIe 3.0. The throughput that can be achieved in practice can be estimated by multiplying this with 0.90. This 90% value is a mean empirical value taken from the values for various components that have been observed over the years.

All PRIMERGY servers, beginning with the generation introduced in 2010 (e. g. PRIMERGY RX300 S5), support PCIe 2.0 and from the generation introduced in 2012 (e. g. PRIMERGY RX300 S7) PCIe 3.0. If different components come together here, the highest frequency jointly supported by all components is used.

Application to the RAID controllers

The next table presents the performance-determining key data for all RAID controllers. The throughput limits listed here are obtained with the help of the two previous subsections "SAS and SATA" and "PCIe". The significant throughput limit in each case is highlighted in the table in bold print.

Controller alias	# CPU cores	Cache memory type	# Disk connections	Limit for throughput of disk interface	PCIe version	Effective PCIe width	Limit for throughput of PCIe interface
Cougar Point			4 × SATA 3G	**973 MB/s**			
Patsburg A			4 × SATA 3G	**973 MB/s**			
Patsburg B			4 × SAS 3G	**973 MB/s**			
LSI1064	1		4 × SAS 3G	973 MB/s	1.0	x4	**858 MB/s**
LSI1068	1		8 × SAS 3G	1945 MB/s	1.0	x4	**858 MB/s**
LSI2008	1		8 × SAS 6G	3890 MB/s	2.0	x4	**1716 MB/s**
	1					x8	**3433 MB/s**
LSI2108	1	DDR2 / 800 MHz	8 × SAS 6G	3890 MB/s	2.0	x4	**1716 MB/s**
	1					x8	**3433 MB/s**
LSI2208-2.0	2	DDR3 / 1333 MHz	8 × SAS 6G	3890 MB/s	2.0	x4	**1716 MB/s**
						x8	**3433 MB/s**
LSI2208-3.0	2	DDR3 / 1333 MHz	8 × SAS 6G	3890 MB/s	3.0	x4	**3380 MB/s**
				3890 MB/s		x8	6761 MB/s

In the majority of cases the throughput limits do not represent a bottleneck. In practice, the application scenarios with random access to conventional hard disks prevail in particular, in which no high throughputs are achieved.

The limitation for application situations with pure sequential access and the frequent block size 64 kB only occurs in the event of PCIe 1.0, x4, with five conventional hard disks; and in the event of PCIe 2.0, x8, it only occurs with 17 conventional hard disks. In the case of sequential or random accesses to modern SAS-SSDs with the same block size the limitation in the event of PCIe 2.0, x8, occurs with nine to 15 SSDs, depending on the write share.

The throughput values in the column "Limit for throughput of disk interface" apply for the connections between the controller and the hard disks in their entirety. The throughputs via this SAS/SATA interface are only in the case of RAID 0 identical with the throughputs from the viewpoint of the application. With other RAID levels the throughput via the SAS/SATA interface is from the viewpoint of the application multiplied by a specific factor compared with the throughput. This factor is always ≥ 1 and depends on the RAID level and several characteristics of the access pattern. The real throughput limits are therefore always lower than the values in the column "Limit for throughput of disk interface" by the mentioned specific factor.

Safeguarding the controller cache against power failure

In order to back up data in the cache in the event of a power failure two options are currently on offer for PRIMERGY servers:

Battery Backup Unit (BBU)

The conventional method works with a battery backup unit (BBU). In this case, the power supply of the volatile cache memory is backed up during a power failure by means of a rechargeable battery (accumulator). Since the battery has a limited capacity, which also decreases with time due to physical and chemical processes, it can only ensure supply to the cache memory for a limited amount of time. This time is not constant, but depends on several influencing factors, such as the age of the BBU, charging status, temperature, etc. Thus, the BBU is subject to a limited warranty.

As soon as the power is available again and the server has been powered up, the RAID controller can continue to work with the content of the cache memory and the data remains consistent.

Flash Backup Unit (FBU)

With a flash backup unit (FBU), which is based on more recent technology, the data is not kept in the cache memory in case of a power failure; the content of the cache memory is in contrast copied to a non-volatile flash memory. The data can remain in this flash memory for almost as long as you like, which means that the retention time known from the BBU is no longer a problem.

The energy required for the copying process from the cache memory to the flash memory in case of a power failure comes from a super capacitor.

As soon as the power is available again and the server has been powered up, the cache content is written from the flash memory back to the cache memory. The RAID controller can now continue to work again and the data remains consistent.

Performance-relevant settings

Accurate parameter setting for the respective RAID array is essential to the optimal performance of the RAID controller. Depending on the controller there is a varying number of parameters that can be set. For the purpose of easy and reliable handling of the settings for RAID controllers and hard disks it is advisable to use the RAID-Manager software “ServerView RAID”, which is supplied for PRIMERGY servers. All the cache settings for controllers and hard disks can usually be made en bloc – specifically for the application – by using the pre-defined modi “Performance” or “Data Protection”. The “Performance” mode ensures the best possible performance settings for the majority of the application scenarios.

Any existing controller and hard disk caches are enabled by the “Performance” mode, which is why the cache of the RAID controller should be protected in this mode against data loss in the event of a power failure by means of a battery backup unit (BBU) or flash back unit (FBU). Furthermore, the hard disk caches should also be safeguarded by the use of an uninterruptible power supply (UPS).

In special cases, a parameter setting that deviates from the standard setting of the “Performance” mode can make sense. Reference is made to whether it makes sense in the appropriate part of the section “Controller comparison”.

The cache settings of the “ServerView RAID” software contain – depending on the controller – all or part of the following setting options of the RAID controller and the hard disks. The first three setting options control the RAID controller, and the last one controls the hard disks of the RAID array. All the parameters can be specifically set for each RAID array.

Read mode

The “Read mode” parameter can be used to influence cache behavior for read. Two options “No read-ahead” and “Read-ahead” are available. No caching during read takes place in the case of “No read-ahead”. On the other hand, the controller uses caching during read in the case of “Read-ahead”.

In the case of the “Read-ahead” option the onboard controllers (e. g. Patsburg A) generally read blocks in advance. The PCIe controllers with a cache work in a more differentiated way for this option: The requested blocks are continuously analyzed to see whether there is sequential read access. If the controller detects such an access, it starts to also read the sequentially following blocks – in addition to the requested block – in the cache in order to have them available for the expected, next requests. The current option “Read-ahead” is in other words adaptive. This is a merger of the two previous options “Read-ahead” and “Adaptive”.

Write mode

The setting options of the controller cache that control the handling of write requests are summarized under the term “Write mode”. There are three options for setting the write cache: “Write-through”, “Write-back” and “Always Write-back (independent of BBU state)”. The “Write-through” option ensures that each write request from the controller is only reported back as completed when it has been acknowledged by the hard disk. With the “Write-back” and “Always Write-back” options the requests are cached in the controller cache, immediately acknowledged to the application as completed and only transferred to the hard disk later. This procedure enables optimal utilization of controller resources, faster succession of the write requests and therefore higher throughput. Any power failures can be bridged by means of an optional BBU/FBU, thus guaranteeing the integrity of the data in the controller cache. The “Always Write-back” option enables the write cache on a permanent basis; it is also used if the BBU/FBU is not operational. On the other hand, the “Write-back” option automatically switches to “Write-through” as long as the controller cache is not safeguarded by the BBU/FBU.

Cache mode

The “Cache mode” parameter, sometimes also referred to as “I/O Cache” influences the behavior of the controller cache during read. The “Direct” option defines that the data to be read is read from the hard disk directly and is also not stored in the controller cache. The alternative “Cached” causes an attempt to first be made to find the data in the controller cache and to satisfy the read request before the hard disk is accessed. All the data here is written to the controller cache so as to be available for subsequent read requests.

Disk cache mode

The possible values here are “enabled” and “disabled”. In most cases, the enabling of the hard disk cache entails an increase in throughput for write access. If the system is safeguarded by a UPS, the enabling of the hard disk cache is recommended for performance reasons.

The next table shows which of these setting options exist for the individual controllers.

Controller alias	Read mode	Write mode	Cache mode
Onboard controller Cougar Point	✓		
Onboard controller Patsburg A	✓		
Onboard controller Patsburg B	✓		
LSI1064 / LSI 1068			
LSI2008			
LSI2108	✓	✓	✓
LSI2208-1GB	✓	✓	✓
LSI2208-512	✓	✓	✓
LSI2208-Lite			

To complete matters the following table also provides a compilation of the settings that are currently implemented in the modi “Data Protection” and “Performance” in ServerView RAID. It should be noted that the settings for the controllers with a controller cache also depend on the existence of a BBU/FBU, but are independent of the selected RAID level.

Controller alias	BBU/ FBU?	Data Protection				Performance			
		Read mode	Write mode	Cache mode	Disk cache	Read mode	Write mode	Cache mode	Disk cache
Onboard controller Cougar Point		Read-ahead			off	Read-ahead			on
Onboard controller Patsburg A		Read-ahead			off	Read-ahead			on
Onboard controller Patsburg B		Read-ahead			off	Read-ahead			on
LSI1064 / LSI1068					off				on
LSI2008					off				on
LSI2108		Read-ahead	Write-through	Direct	off	Read-ahead	Always Write-back	Direct	on
	✓	Read-ahead	Write-back	Direct	off	Read-ahead	Write-back	Direct	on
LSI2208-1GB / LSI2208-512		Read-ahead	Write-through	Direct	off	Read-ahead	Always Write-back	Direct	on
	✓	Read-ahead	Write-back	Direct	off	Read-ahead	Write-back	Direct	on
LSI2208-Lite					off				on

MegaRAID Advanced Software Options

The MegaRAID Advanced Software options are CacheCade 2.0 and FastPath. Two different approaches to increasing performance, which need to be considered separately, are pursued here.

The MegaRAID Advanced Software options are available for the LSI2208-1GB and LSI2208-512 controllers. A special license key is required to activate them. Although the smallest configuration for the MegaRAID Advanced Software options uses two hard disks, more hard disks will normally be used so that in the event of blade servers an LSI2208-512 controller has to be used together with PRIMERGY SX980 storage blades.

CacheCade 2.0

CacheCade 2.0 is read and read/write caching software for hot spot data, which significantly reduces the access times for the use of HDD arrays by using SSDs as a non-volatile dynamic high-performance controller cache pool. This simply combines the two advantages of the hard disk types HDD and SSD – large capacities and short access times.

Prerequisite and limitation

One important prerequisite is a so-called application hot spot for IO accesses. The data is on a RAID array consisting of HDDs, to which access is random. To improve access times you need SSDs which are configured as a CacheCade 2.0 cache and assigned to the HDD-RAID array. These HDDs and SSDs must be connected to the same controller. Possible configurations of the CacheCade 2.0 cache are RAID 0, RAID 1 and RAID 1E up to a maximum usable cache capacity of 512 GB. The parameter settings "write through" and "write back" define whether the cache is only used as a read-cache or write/read-cache.

Example configuration

A typical configuration for CacheCade 2.0 would be e. g. a HDD array of type RAID 0 consisting of four SATA-6G-HDDs and an SSD array of type RAID 1 consisting of two SAS-6G-SSDs as a CacheCade 2.0 cache. These configurations are considered in more detail in this document both separately and in combination as CacheCade 2.0.

FastPath

FastPath is a high-performance IO accelerator for RAID arrays that are made up of SSDs. This optimized version of LSI MegaRAID technology permits a clear-cut increase in the performance of applications with a high IO load for random access if SATA-6G-SSDs or SAS-6G-SSDs are used.

Prerequisite and limitation

The transaction rate of RAID arrays consisting of high-performance 6G-SSDs can be limited in the case of random access with small blocks, although the throughput limit of SAS-/SATA-6G and PCIe has not been reached yet and an adequate number of SSDs is available. In this case, it is possible to significantly increase the transaction rate with the use of FastPath.

The IO range, as of which FastPath can be successfully used, begins from about 200000 IO/s per controller. The section "Measurement environment" lists a maximum transaction rate for a SAS-6G-SSD of 14184 IO/s for random access with 8 kB block size and a read share of 67%. It can therefore be concluded that the limit starts at approximately fourteen SSDs.

Characteristics of the onboard controllers

Some PRIMERGY models offer with an onboard controller – at present CougarPoint, Patsburg A or Patsburg B – a simple, low-priced entry-level solution for operation with up to four SATA hard disks. With the support of RAID 0, RAID 1 and RAID 10 such an onboard controller offers a range of common RAID levels and does not occupy a PCIe slot.

All the controllers that are connected via the PCIe interface are not treated as onboard controllers here, even if they are integrated on the motherboard (see table in the section "Presentation of the RAID controllers").

The onboard controller is implemented as a firmware/driver-based software RAID solution. It is typically integrated in the "Southbridge" chip, which belongs to the motherboard chip set. During the boot phase accesses to the RAID array are implemented by the firmware. As soon as the operating system is active, suitable drivers take on this task.

The onboard controller does not have a processor of its own, but uses the CPU of the server system for RAID functionality. The pro rata consumption of the server's processor performance is increasingly less important in newer servers.

CougarPoint

The onboard controller CougarPoint is a pure SATA controller that can be set in various modes via the BIOS. Although only the "RAID" mode is suitable for the effective use of this type of controller, all the modes of these SATA controllers are presented for the sake of completeness. The pro rata consumption of the server's processor performance is increasingly less important in newer servers. There are three modes:

RAID	Recommended mode on account of its flexibility. A trouble-free migration of a SATA-HDD from a non-RAID to a RAID configuration is only possible here. All the functionalities of SATA are supported, i.e. also NCQ and "hot swapping". A firmware named "LSI Logic Embedded MegaRAID" is integrated in the controller BIOS in PRIMERGY servers for the supported RAID levels. Only in this mode are RAID arrays already possible during the boot phase, and only in this mode can controllers and hard disks be seen in the RAID-Manager software "ServerView RAID", where they can be administered. Special drivers are needed.
AHCI	AHCI stands for "Advanced Host Controller Interface" and is a cross-manufacturer interface standard for SATA controllers. Support is provided for NCQ and "hot swapping". Special drivers in the operating system are also necessary for AHCI.
IDE	In this operating mode the SATA ports as such are made visible to the operating system. NCQ is not supported. Appropriate SATA drivers are required, which are supplied for various operating systems on the "ServerStart DVD".

Patsburg A and Patsburg B

The C600 chipset of the server generation released in 2012 with two and more processor sockets (e. g. PRIMERGY RX300 S7) has a completely new architecture, which integrates an SAS/SATA controller for the first time. As standard, this controller works in the SATA mode (Patsburg A). You can upgrade to the higher value SAS/SATA version (Patsburg B) with the "SAS Enabling Key", which can be ordered separately.

Although provision is made in the BIOS of a number of servers with Patsburg A/B for SATA configuration options (incl. "AHCI"), they generally refer however solely to the SATA connections for optical drives and are not used in the RAID configurations covered here.

Measurement context

Now that the various controllers have been presented and their technical features explained, it is our intention in the following section "Controller comparison" to discuss the controllers in various application scenarios and to back this up on the basis of measurement results. Hence, a brief introduction to begin with of the measurement method and the measurement environment.

All the details of the measurement method and the basics of disk I/O performance are described in the white paper "Basics of Disk I/O Performance".

Measurement method

As standard, performance measurements of disk subsystems in PRIMERGY servers are carried out with a defined measurement method, which models the hard disk accesses of real application scenarios on the basis of specifications.

The essential specifications are:

- Share of random accesses / sequential accesses
- Share of read / write access types
- Block size (kB)
- Number of parallel accesses (# of outstanding I/Os)

A given value combination of these specifications is known as "load profile". The following five standard load profiles can be allocated to typical application scenarios:

Standard load profile	Access	Type of access		Block size [kB]	Application
		read	write		
File copy	random	50%	50%	64	Copying of files
File server	random	67%	33%	64	File server
Database	random	67%	33%	8	Database (data transfer) Mail server
Streaming	sequential	100%	0%	64	Database (log file), Data backup; Video streaming (partial)
Restore	sequential	0%	100%	64	Restoring of files

In order to model applications that access in parallel with a different load intensity, the "# of Outstanding I/Os" is increased, starting with 1, 3, 8 and going up to 512 (from 8 onwards in increments to the power of two).

The measurements of this document are based on these standard load profiles.

The main results of a measurement are:

- Throughput [MB/s] Throughput in megabytes per second
- Transactions [IO/s] Transaction rate in I/O operations per second
- Latency [ms] Average response time in ms

The data throughput has established itself as the normal measurement variable for sequential load profiles, whereas the measurement variable "transaction rate" is mostly used for random load profiles with their small block sizes. Data throughput and transaction rate are directly proportional to each other and can be transferred to each other according to the formula

Data throughput [MB/s]	*= Transaction rate [IO/s] × Block size [MB]*
Transaction rate [IO/s]	*= Data throughput [MB/s] / Block size [MB]*

Measurement environment

All the measurement results discussed in this document were determined using the hardware and software components listed below:

System under Test (SUT)	
Models	PRIMERGY RX300 S7 PRIMERGY TX300 S7 PRIMERGY TX120 S3p PRIMERGY BX920 S3
Operating system	Microsoft Windows Server 2008 Enterprise x64 Edition SP2
RAID Manager software	ServerView RAID Manager 5.5.2 ServerView RAID Manager 5.7.3
Initialization of RAID arrays	RAID arrays are initialized before the measurement with an elementary block size of 64 kB (“stripe size”)
File system	NTFS
Measuring tool	Iometer 2006.07.27
Measurement data	Measurement files of 32 GB with 1 – 8 hard disks; 64 GB with 9 – 16 hard disks; 128 GB with 17 or more hard disks
Onboard SATA Controller “Cougar Point”	Intel BD82C202 PCH, Cougar Point (in PRIMERGY TX120 S3p) Driver name: megasr1.sys, Driver version: 15:00.0329.2012 BIOS version: A. 10.03031333R SATA RAID mode
Onboard SATA Controller “Patsburg A”	Intel C600 chipset (in PRIMERGY RX300 S7) Driver name: megasr1.sys, Driver version: 15:00.0329.2012 BIOS version: SCU.11.03201000R
Onboard SAS Controller “Patsburg B”	Intel C600 chipset (in PRIMERGY RX300 S7) Driver name: megasr1.sys, Driver version: 15:00.0329.2012 BIOS version: SCU.11.03201000R SAS Enabling Key
Controller “RAID 0/1 SAS based on LSI MegaRAID 4Port” (LSI MegaRAID SAS 1064)	Driver name: lsi_sas.sys, Driver version: 1.34.03.00 Firmware version: 1.33.00.00 BIOS version: 6.36.00.00
Controller “RAID 0/1 SAS based on LSI MegaRAID 8Port” (LSI MegaRAID SAS 1068)	Driver name: lsi_sas.sys, Driver version: 1.34.03.00 Firmware version: 1.33.00.00 BIOS version: 6.36.00.00
Controller “RAID Ctrl SAS 6G 0/1” (LSI MegaRAID SAS 2008)	Driver name: megasas2.sys, Driver version: 5.2.103.64 Firmware package: 20.10.1-0075 Firmware version: 2.120.224-1451
Controller “RAID Ctrl SAS 5/6 512MB (D2616)” (LSI MegaRAID SAS 2108)	Driver name: megasas2.sys, Driver version: 5.2.116.64 Firmware package: 12.12.0-0087, Firmware version: 2.120.233-1471 Controller cache: 512 MB
Controller “RAID Ctrl SAS 6G 5/6 1GB (D3116)”	Driver name: megasas2.sys, Driver version: 5.2.116.64 Firmware package: 23.2.1-0048, Firmware version: 3.152.65-1626 Controller cache: 1 GB
Controller „RAID Ctrl SAS 6G 5/6 1GB (D3116C)”	Driver name: megasas2.sys, Driver version: 6.505.5 Firmware package: 23.9.0-0023, Firmware version: 3.220.75-2196 Controller cache: 1 GB
Controller “PY SAS RAID HDD Module”	Driver name: megasas2.sys, Driver version: 5.2.116.64 Firmware package: 23.2.1-0049, Firmware version: 3.152.75-1658 Controller cache: 512 MB
Controller “PY SAS RAID HDD Module w/o cache”	Driver name: megasas2.sys, Driver version: 5.2.116.64 Firmware package: 23.2.1-0049, Firmware version: 3.152.74-1659
SATA-6G-HDD	Seagate ST91000640NS, 2.5″, 1000 GB, 7200 rpm, 6 Gb/s
SAS-6G-HDD	Toshiba MK1401GRRB, 2.5″, 146 GB, 15000 rpm, 6 Gb/s
SATA-6G-SSD	Seagate ST200FM0002, 2.5″, 200 GB, 6 Gb/s

The hard disk models used for the controller comparison are summarized again below in detail together with their fundamental performance data, because this is important for your understanding of the performance values achieved with the controllers. A high-performance SATA-6G and SAS-6G hard disk were chosen in each case for the classic hard disks (HDDs), and a 200 GB SAS-6G-SSD represents the SSD class.

The table depicts the maximum values measured with a single hard disk for the five standard load profiles that were shown in the previous subsection "Measurement method". The hard disk cache is enabled in all cases, because this almost always ensures optimal performance.

Hard disk type	Short name (alias)	Sequential maximum throughput [MB/s] 64 kB block size		Maximum transaction rates for random accesses [IO/s]		
				8 kB block size	64 kB block size	
		Read	Write	Read share:	Read share:	
				67%	67%	50%
HDD SATA, 6 Gb/s, 2.5″ 1000 GB, 7200 rpm hot-pluggable category: Business-Critical (BC)	SATA-6G-HDD	108 MB/s	107 MB/s	372 IO/s	317 IO/s	314 IO/s
HDD SAS, 6 Gb/s, 2.5″ 146 GB, 15000 rpm, category: Enterprise (EP)	SAS-6G-HDD	207 MB/s	204 MB/s	594 IO/s	506 IO/s	492 IO/s
SSD SAS, 6 Gb/s, 2.5″ 200 GB, hot-pluggable, category: Enterprise (EP)	SAS-6G-SSD	339 MB/s	196 MB/s	14184 IO/s	3006 IO/s	2985 IO/s

This information and more about the hard disks released for the PRIMERGY servers is available in the white paper "Single Disk Performance".

Controller comparison

All the important preliminary information about controllers has been provided in the previous sections. This information will in many cases already narrow down the choice of controller for a given application. If further customer information about the planned use of the controller is added, a great deal more can be said about the performance to be expected with the individual controllers. Thus this section is to compare the controllers differentiated for various RAID levels, application scenarios, load intensities, numbers of hard disks as well as hard disk technologies. The statements are illustrated with the help of measurement results. The comparisons are divided into the following subsections, which can be read independently of each other:

- RAID 1 (two SATA hard disks)
- RAID 1 (two SAS hard disks)
- RAID 0 and 10 (four SATA hard disks)
- RAID 0 and 10 (five to eight hard disks)
- RAID 0 and 10 (more than eight hard disks)
- RAID 5

General preliminary remarks about the comparisons:

- The five load profiles described in the section “Measurement method”, “File copy”, “Database”, “File server”, “Streaming” and “Restore”, are mostly used in the comparisons, thus enabling the random and sequential application scenarios to be reasonably covered. If the customer load profile significantly differs from this, the statements made here no longer apply without restrictions.
- As the benchmark for the performance of a disk subsystem the transaction rate is specified - as is common practice - in IO/s for random load profiles, and throughput in MB/s for sequential load profiles.
- All the controllers that support the RAID level and hard disk type that have just been considered are to be discussed.
- To make things more easily understandable, the diagrams in this section are for the most part restricted to the maximum values achievable. These are usually only achieved with a high load intensity of the disk subsystem.
- The three hard disks that are dealt with more closely in the section “Measurement environment” (SATA-6G-HDD, SAS-6G-HDD and SAS-6G-SSD) are used as example for hard disk technologies. Their key performance data is also presented there. In some places of the following comparisons the achieved performance values are explained on the basis of the performance data of these hard disk types.
- Since the maximum performance achievable with the controllers is the subject of this document, the unchanged cache and hard disk settings of “ServerView RAID” in “Performance” mode, and not those of the “Data Protection” mode, were taken as a basis for the measurements. This setting makes sense if the customer provides an adequate safeguard against power failures. If a change to these settings is beneficial, it will be mentioned in that given instance.
- Conventional hard disks (in contrast to SSDs) are now only referred to in short as “HDDs” in the following controller comparisons.
- In the latest generation the group of LSI2208 controllers supports PCIe 3.0. This basically means a higher throughput limit, which is achieved in configurations with numerous hard disks. Thus, statements about configurations with up to eight hard disks are still equally valid for the generations PCIe 2.0 and 3.0. Therefore, the two generations are not presented separately in the appropriate sections. It only makes sense to handle the PCIe 3.0 generation separately for the larger configurations RAID 0 and RAID 10 with more than eight hard disks, as well as RAID 5 with 24 hard disks, and namely on the basis of the LSI2208-1G-3.0 controller.

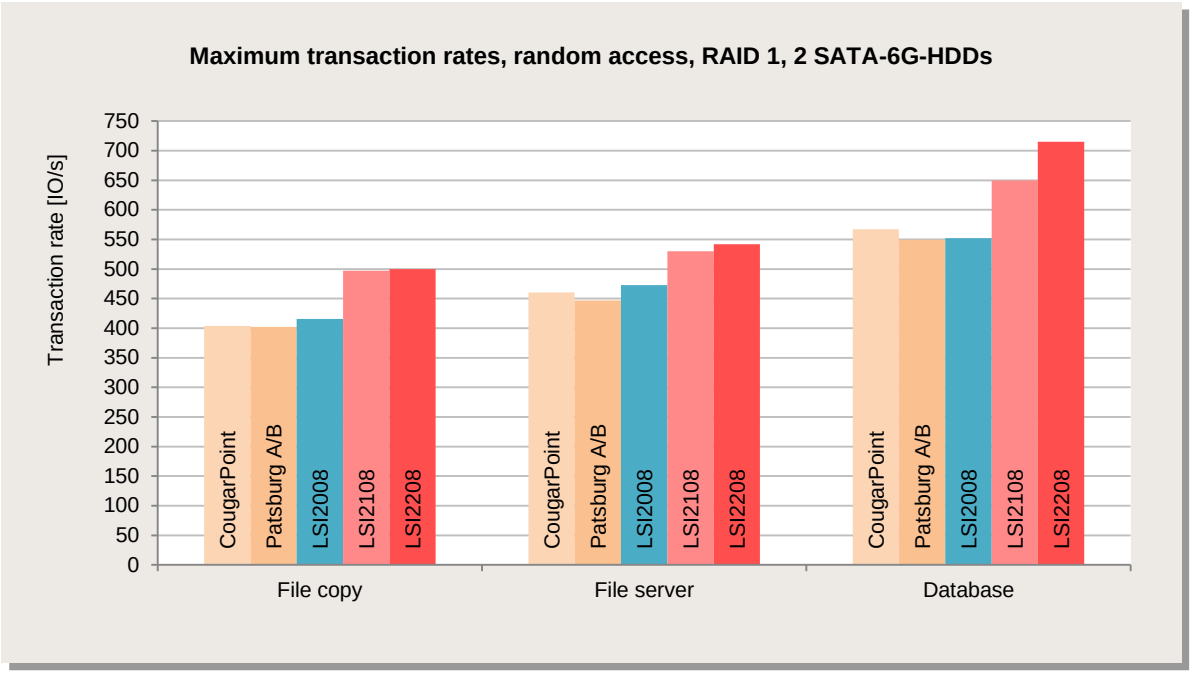
RAID 1 (two SATA hard disks)

It is worth looking at this case separately, because all onboard controllers can only be compared in this way with most of the PCIe controllers. This is done using the SATA-6G-HDD, which was described more closely in the section "Measurement environment". The performance values for the individual load profiles are mostly determined by the HDDs, because with this number of HDDs the controllers are not yet working at their limits.

Random accesses

RAID 1 with two SATA-6G-HDDs

The diagram shows a controller comparison for two SATA-6G-HDDs configured as RAID 1. The three groups of columns in the diagram represent the transaction rates for the standard load profiles "File copy" (random access, 50% read, 64 kB block size), "File server" (random access, 67% read, 64 kB block size) and "Database" (random access, 67% read, 8 kB block size).

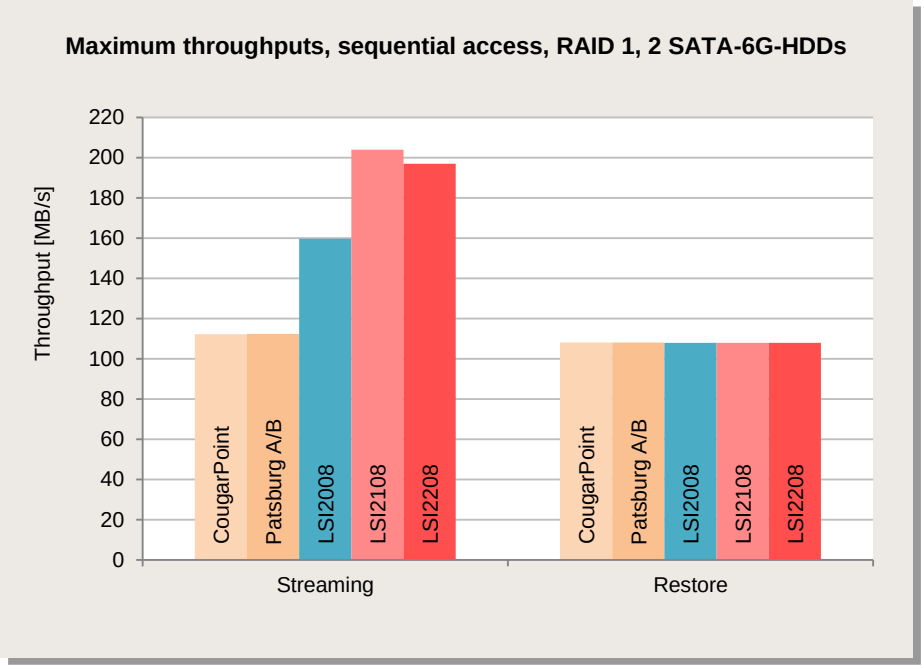


The LSI2108 and the LSI2208-1G controllers provide the highest overall transaction rates here.

Sequential accesses

RAID 1 with two SATA-6G-HDDs

The next diagram shows a controller comparison for two SATA-6G-HDDs configured as RAID 1. The two groups of columns in the diagram represent the throughputs for the standard load profiles “Streaming” (sequential access, 100% read, 64 kB block size) and “Restore” (sequential access, 100% write, 64 kB block size).



All in all, in the majority of cases for this RAID level the maximum throughputs are close to the maximum values of a single hard disk.

For read the LSI2008, the LSI2108 and the LSI2208 controllers use both hard disks for high load intensities and consequently show a higher maximum throughput.

RAID 1 (two SAS hard disks)

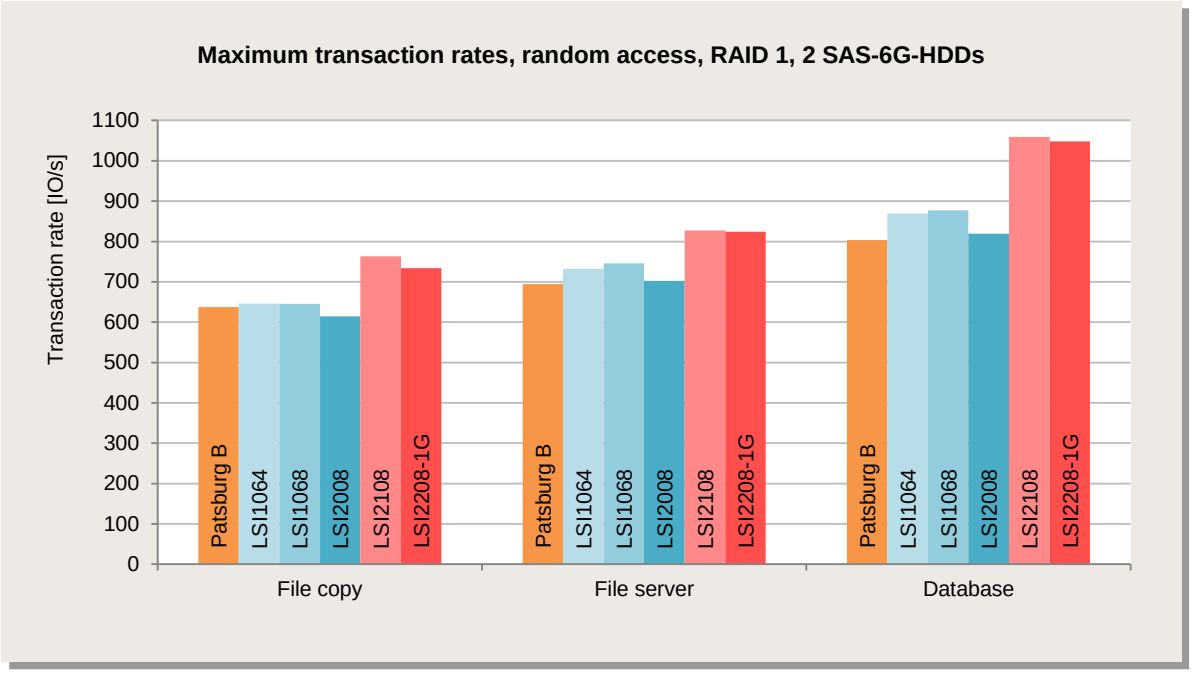
It is also worth looking at this range of hard disk numbers separately, because all PCIe controllers and at least one onboard controller can only be compared here. This is done using the SAS-6G-HDD and the SAS-6G-SSD, which were described more closely in the section "Measurement environment".

In the case of RAID 1 the controllers with a cache usually perform better if you disable the controller cache.

Random accesses

RAID 1 with two SAS-6G-HDDs

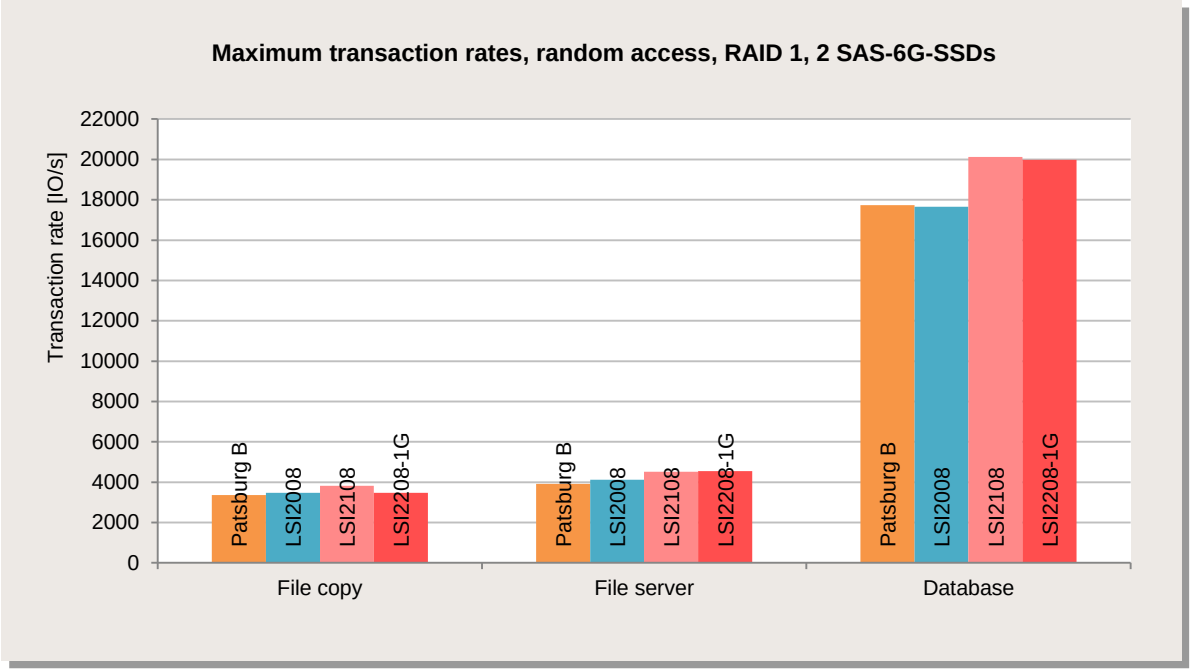
The diagram shows a controller comparison for two SAS-6G-HDDs configured as RAID 1. The three groups of columns in the diagram represent the transaction rates for the standard load profiles "File copy" (random access, 50% read, 64 kB block size), "File server" (random access, 67% read, 64 kB block size) and "Database" (random access, 67% read, 8 kB block size).



The LSI2108 and the LSI2208-1G controllers provide the highest overall transaction rates here.

RAID 1 with two SAS-6G-SSDs

The diagram shows a controller comparison for two SAS-6G-SSDs configured as RAID 1. The three groups of columns in the diagram represent the transaction rates for the standard load profiles “File copy” (random access, 50% read, 64 kB block size), “File server” (random access, 67% read, 64 kB block size) and “Database” (random access, 67% read, 8 kB block size).



The LSI2108 and the LSI2208-1G controllers provide the highest overall transaction rates here.

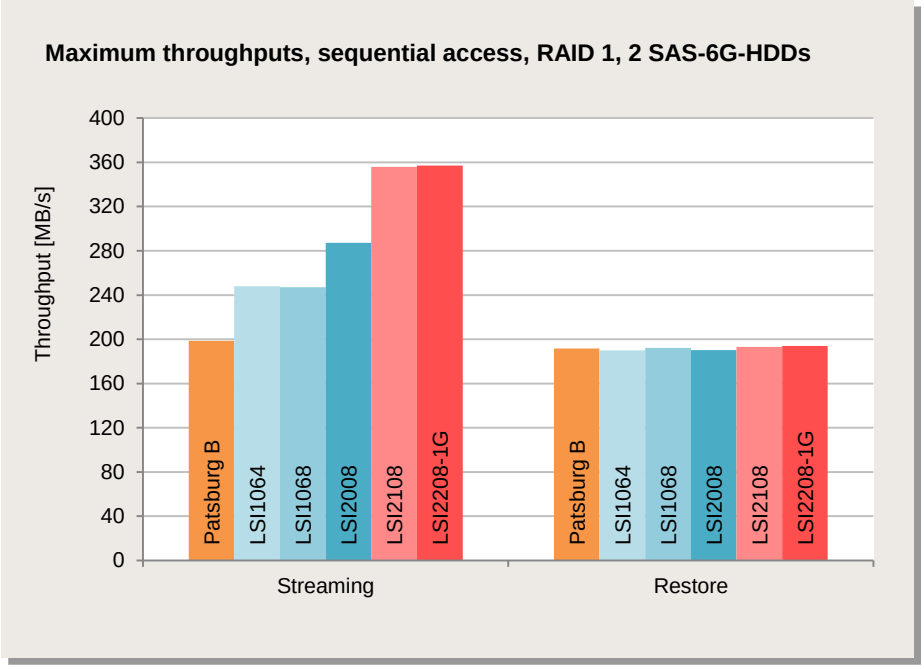
Sequential accesses

RAID 1 with two SAS-6G-HDDs

The next diagram shows a controller comparison for two SAS-6G-HDDs configured as RAID 1. The two groups of columns in the diagram represent the throughputs for the standard load profiles “Streaming” (sequential access, 100% read, 64 kB block size) and “Restore” (sequential access, 100% write, 64 kB block size).

For write the maximum throughputs in all cases for this RAID level are close to the maximum values of a single hard disk.

For read all controllers apart from Patsburg B use both HDDs for higher load intensities and consequently achieve higher read throughputs than a single HDD would provide.

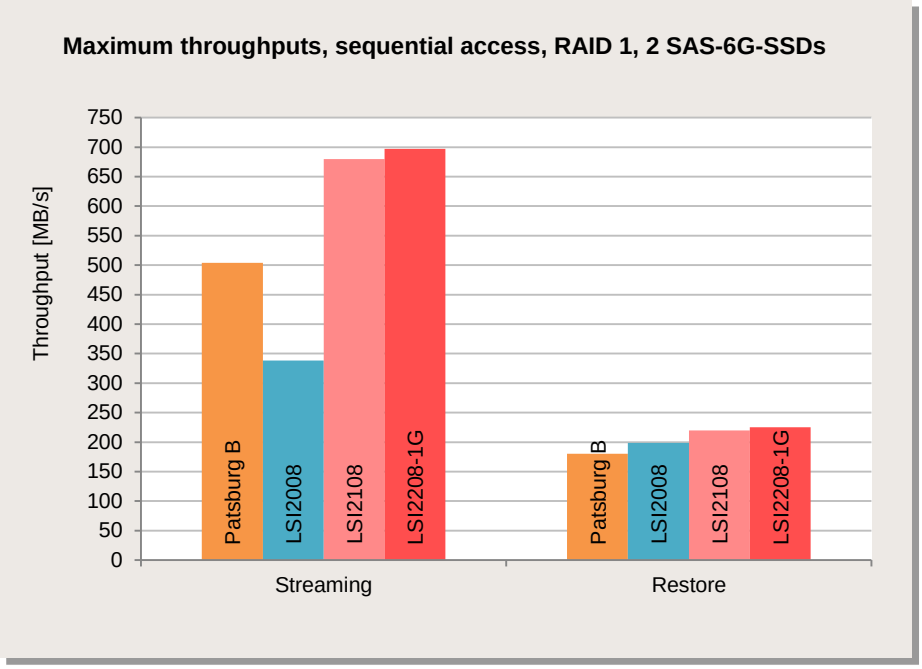


RAID 1 with two SAS-6G-SSDs

The next diagram shows a controller comparison for two SAS-6G-SSDs configured as RAID 1. The two groups of columns in the diagram represent the throughputs for the standard load profiles “Streaming” (sequential access, 100% read, 64 kB block size) and “Restore” (sequential access, 100% write, 64 kB block size).

For write the maximum throughputs in all cases for this RAID level are close to the maximum values of a single hard disk.

For read most controllers use both SSDs for higher load intensities and consequently achieve higher read throughputs than a single SSD would provide.



The Patsburg B controller achieves about 500 MB/s for the standard load profile “Streaming”, which is equivalent to about twice the practical throughput limit of an SAS-3G connection. This means that the two SAS-3G connections are used to full capacity. The maximum sequential throughputs are achieved here with the LSI2108 and the LSI2208 controller. Both achieve throughputs of considerably more than 500 MB/s for the standard load profile “Streaming”.

RAID 0 and 10 (four SATA hard disks)

This subsection continues the controller comparisons made in RAID 1 (two SATA hard disks) for four hard disks. Clear differences between the controllers can be seen with this higher number of hard disks.

In the case of RAID 0 the controllers with a cache usually have a higher maximum performance if you disable the controller cache. Thus, the transaction rates of the LSI2108 and the LSI2208-1G controllers in this subsection are based on the following changes compared with the standard settings of the "Performance" mode of ServerView RAID:

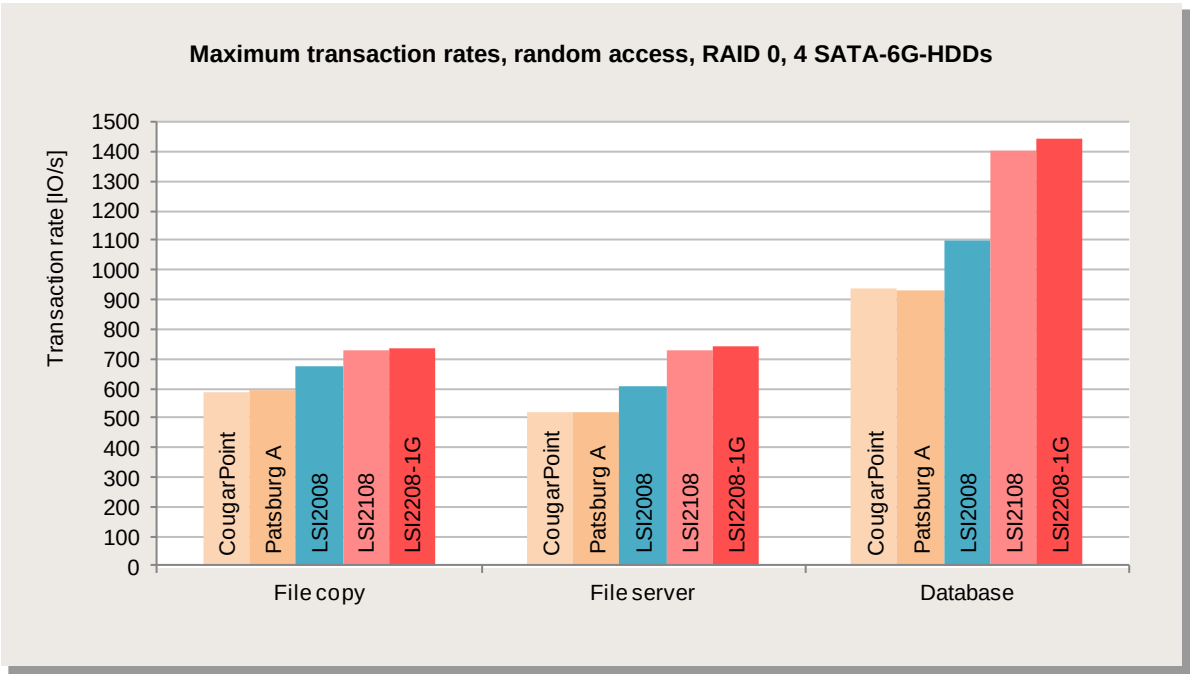
- Read mode from "Read-ahead" to "No read-ahead"
- Write mode from "Always write-back" to "Write-through"

The standard settings were used for RAID 10.

Random accesses

RAID 0 with four SATA-6G-HDDs

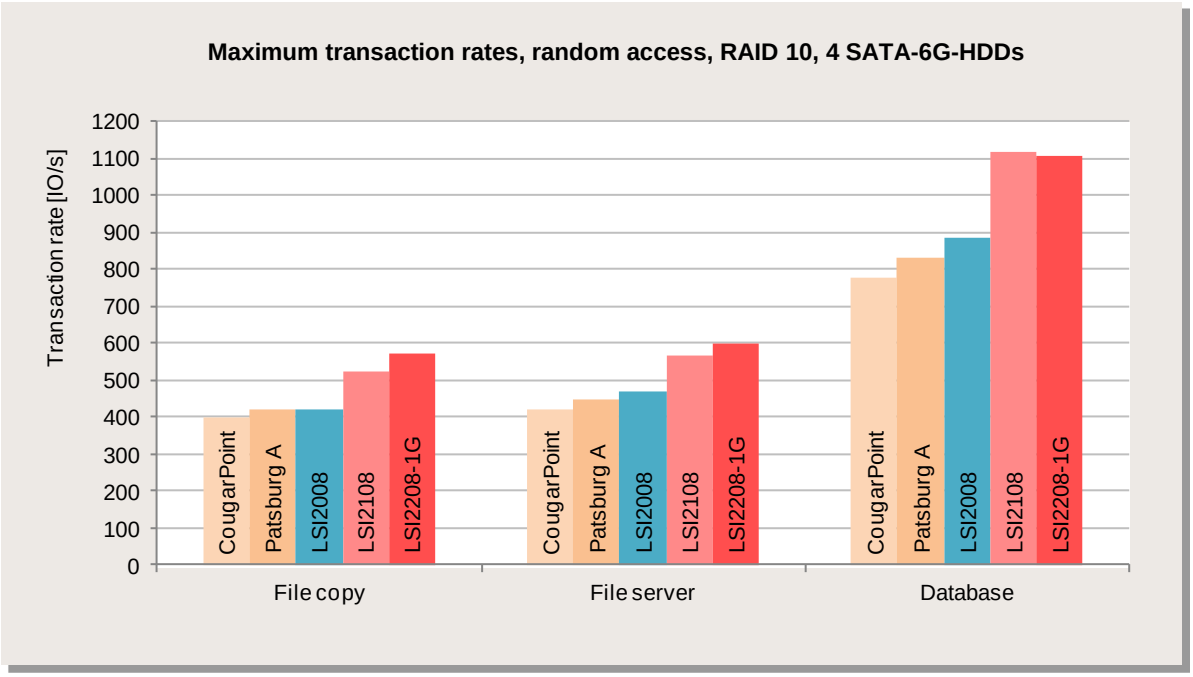
The next diagram shows the transaction rates of the hard disk array for RAID 0 for random load profiles that can be achieved with various controllers. The three groups of columns show the transaction rates for the standard load profiles "File copy" (random access, 50% read, 64 kB block size), "File server" (random access, 67% read, 64 kB block size) and "Database" (random access, 67% read, 8 kB block size).



It can be clearly seen that the transaction rates are higher if the quality of the controller is higher.

RAID 10 with four SATA-6G-HDDs

The next diagram shows the transaction rates of the hard disk array for RAID 10 for random load profiles that can be achieved with various controllers. The three groups of columns show the transaction rates for the standard load profiles “File copy” (random access, 50% read, 64 kB block size), “File server” (random access, 67% read, 64 kB block size) and “Database” (random access, 67% read, 8 kB block size).



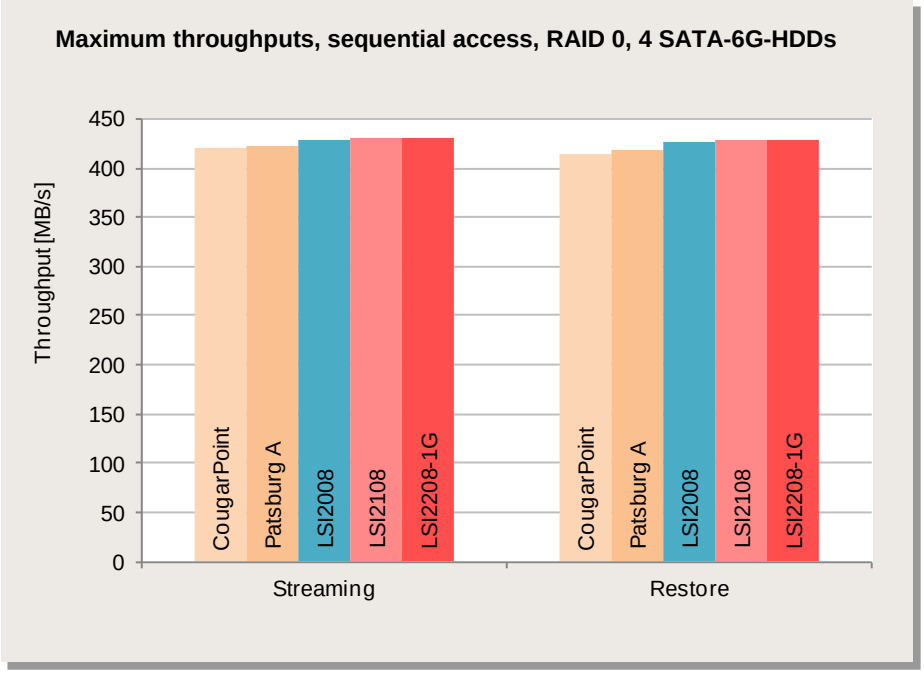
The fact that the transaction rates are higher if the quality of the controller is higher can also be clearly seen here.

Sequential accesses

RAID 0 with four SATA-6G-HDDs

The next diagram shows the maximum throughputs of the hard disk array for RAID 0 for sequential load profiles that can be achieved with various controllers. The two groups of columns in the diagram show the throughputs for the standard load profiles “Streaming” (sequential access, 100% read, 64 kB block size) and “Restore” (sequential access, 100% write, 64 kB block size).

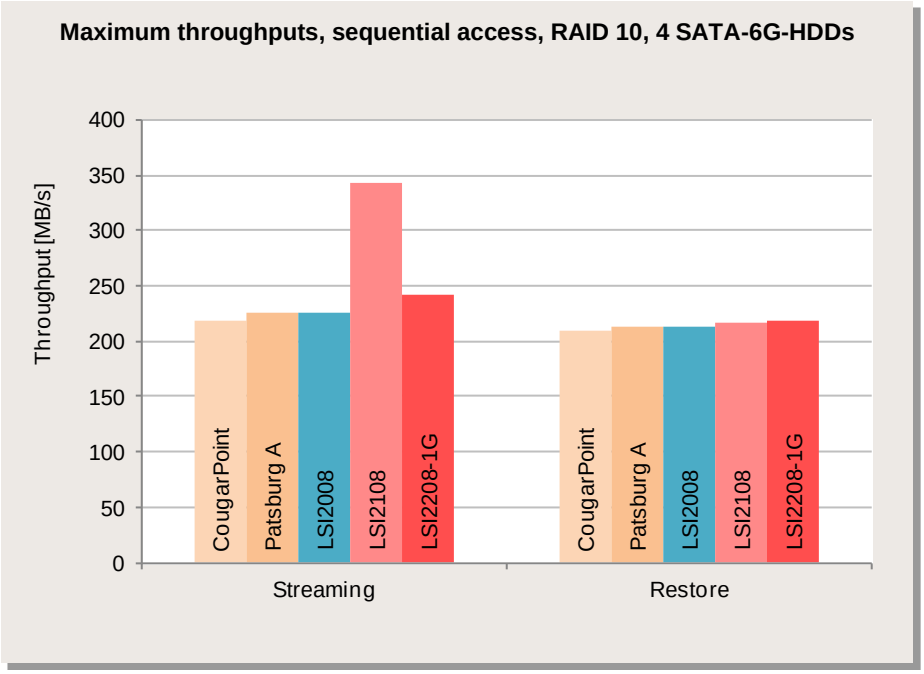
All the controllers deliver approximately the same performance in these cases.



RAID 10 with four SATA-6G-HDDs

The next diagram shows the maximum throughputs of the hard disk array for RAID 10 for sequential load profiles that can be achieved with various controllers. The two groups of columns in the diagram show the throughputs for the standard load profiles “Streaming” (sequential access, 100% read, 64 kB block size) and “Restore” (sequential access, 100% write, 64 kB block size).

All the controllers deliver the same performance in most of these cases. In this case with the standard load profile “Streaming” only the LSI2108 controller works in a range, in which it not only reads from two, but from four HDDs simultaneously and can as a result increase throughput considerably above the double value for a single HDD.



RAID 0 and 10 (five to eight hard disks)

The onboard controllers are no longer sufficient for the operation of more than four hard disks in the current PRIMERGY servers, therefore only the PCIe controllers are compared below. As two of the PCIe controllers are released for a maximum of eight hard disks (LSI1068 and LSI2008), it makes sense to compare all the controllers for the eight connected hard disks. At the same time the performance values presented in this subsection cover the range of medium numbers of hard disks on a representative basis. Since the maximum values for higher performance requirements are of particular interest here, measurements with high-performance SAS-6G-HDDs or SAS-6G-SSDs are used as a means of illustration. These hard disks are described in more detail in the section "Measurement environment".

In contrast to the LSI1068 controller, the LSI1064 controller is no longer taken into consideration here, because the latter only allows the connection of at most four hard disks in all released configurations. The performance-relevant aspects of this controller have been adequately dealt with in the previous subsection "RAID 1 (two SAS hard disks)".

In the case of RAID 0 the controllers with a cache usually have a higher maximum performance if you disable the controller cache. Thus, the transaction rates of the LSI2108 and the LSI2208-1G controllers in this subsection are based on the following changes compared with the standard settings of the "Performance" mode of ServerView RAID:

- Read mode from "Read-ahead" to "No read-ahead"
- Write mode from "Always write-back" to "Write-through"

The standard settings were used for RAID 10.

Random accesses

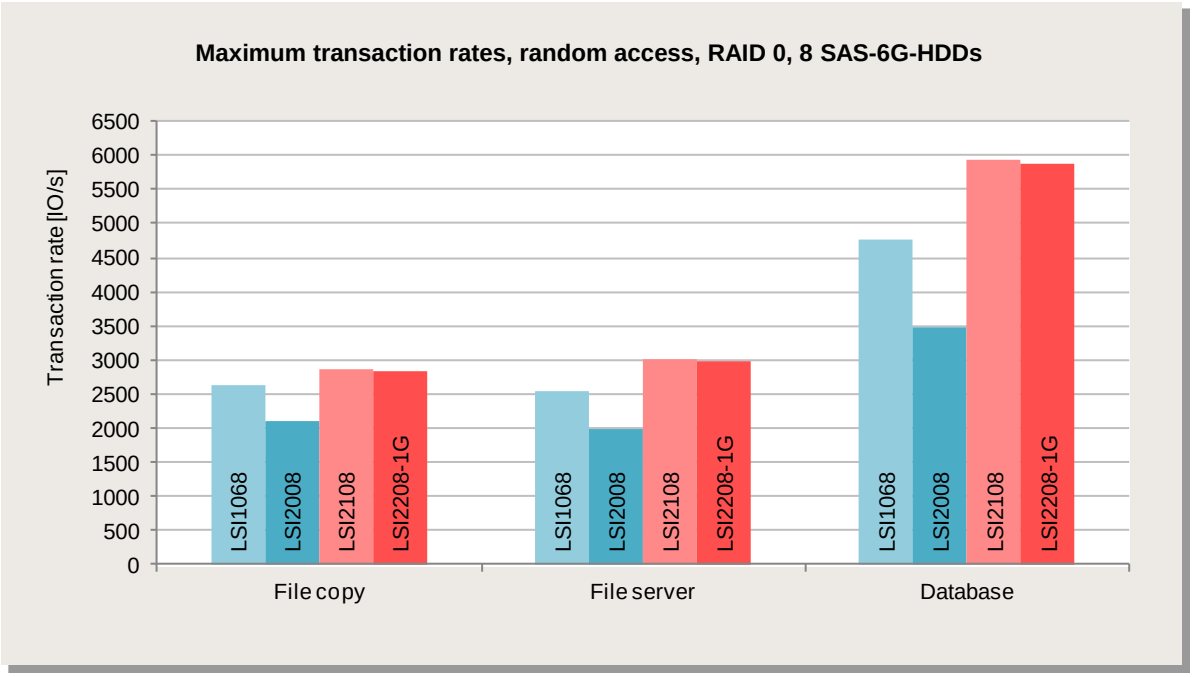
When considering random accesses for larger numbers of hard disks it makes sense to distinguish between HDDs and SSDs, because the maximum values for SSDs are of a quite different magnitude.

HDDs

The controllers are compared below with random accesses to HDDs. The maximum transaction rates of the storage medium for the load profile used are the most important limiting factor here. Nevertheless, performance in such cases is not fully independent of the controller, because there is an additional influence in controllers without a cache (LSI1068 and LSI2008); reference will be made to this in the appropriate places. Although the following results were acquired with eight SAS-6G-HDDs, they can also be used to estimate the maximum transaction rates to be expected for other types and numbers (≤ 8) of hard disks. The throughputs that occur with random accesses to HDDs are so low that any limitations to the PCIe or SAS interface of the controllers do not play a role.

RAID 0 with eight SAS-6G-HDDs

The next diagram shows the transaction rates of the hard disk array for RAID 0 for random load profiles that can be achieved with various controllers. The three groups of columns show the transaction rates for the standard load profiles “File copy” (random access, 50% read, 64 kB block size), “File server” (random access, 67% read, 64 kB block size) and “Database” (random access, 67% read, 8 kB block size).



The two right-hand columns in each of the three groups of columns in this diagram represent the two controllers with a cache (LSI2108 and LSI2208-1G). Both achieve approximately the maximum transaction rate of this RAID array for each of the three load profiles. The two other controllers (LSI1068 and LSI2008) each achieve only about 60% to 90%. A more precise analysis of the transaction rates would show that these two controllers for random load profiles have similar performance values as the two controllers with a cache for low to medium load intensities (up to about 32 parallel accesses with the LSI2008 and 64 respectively with the LSI1068). The LSI1068 and LSI2008 controllers then retain the hitherto achieved transaction rates for higher parallelisms.

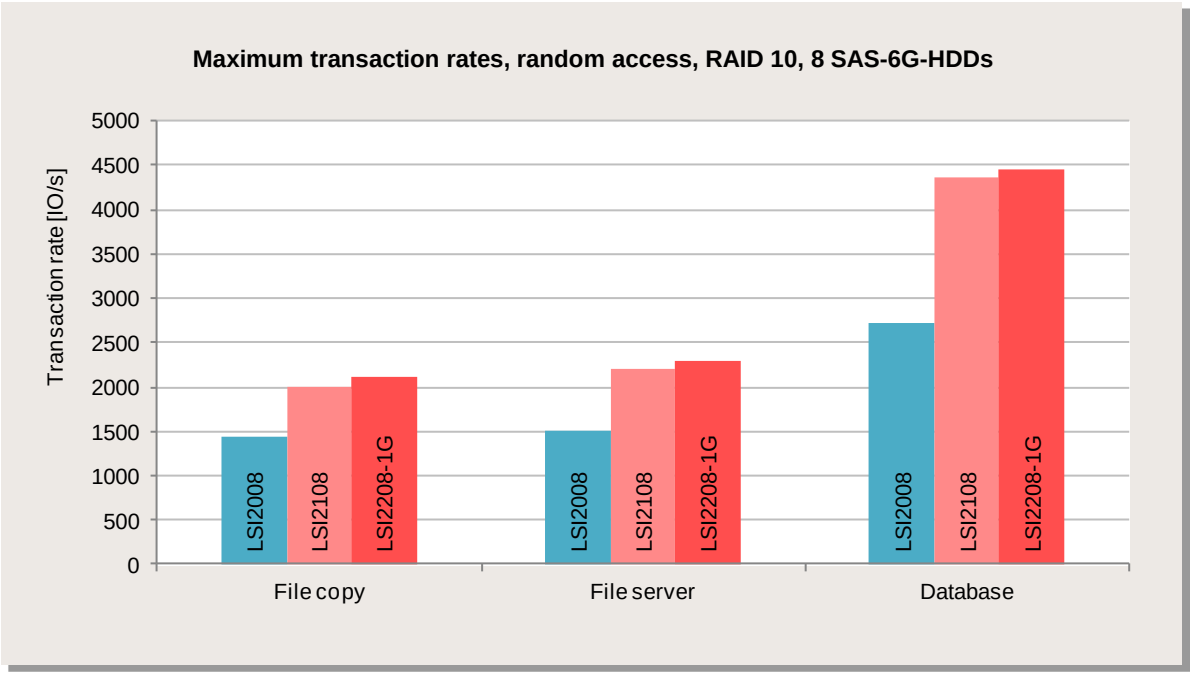
This principal difference between the two groups of controllers for random accesses is shown here by way of an example for eight hard disks. The precise percentage of this difference depends on the size of the RAID array and on the hard disk type. The difference is also relevant for RAID arrays with smaller numbers of hard disks if such RAID arrays can show an increase in performance above 32 parallel accesses. The difference cannot be seen yet for the measurements in subsection “RAID 1 (two SAS hard disks)”, but it is already significant for the measurements in RAID 0 and 10 (four SATA hard disks).

Incidentally, it should be noted for the measurements presented here that only about half the transaction rate is possible for the load profiles with block size 64 kB as with the load profile with the 8 kB block size, although the maximum transaction rates of the HDD type used are relatively similar for all three load profiles. [1]

[1] The reason for this is the initialization of the RAID array with the elementary block size of 64 kB (“stripe size”). The random load profile with block size 8 kB, which is small compared with the “stripe size”, can be estimated directly according to the formula “Number of HDDs × HDD performance”. If the load profile uses the same block size with which the RAID array was initialized, only about half can be achieved. A detailed reason for this here would be taking things too far.

RAID 10 with eight SAS-6G-HDDs

The next diagram shows the transaction rates of the hard disk array for RAID 10 for random load profiles that can be achieved with various controllers. The three groups of columns in the diagram show the transaction rates for the standard load profiles “File copy” (random access, 50% read, 64 kB block size), “File server” (random access, 67% read, 64 kB block size) and “Database” (random access, 67% read, 8 kB block size). The LSI1068 controller is no longer represented here, because it does not support RAID 10.



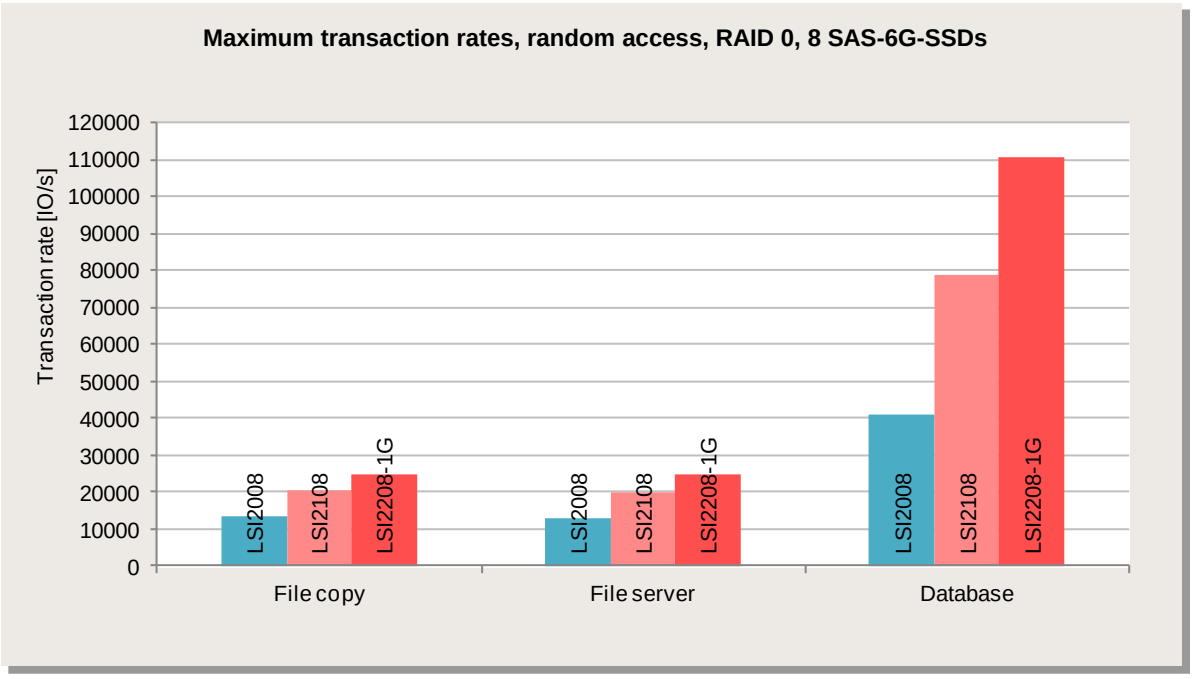
The diagram shows the same principal behavior as with RAID 0. This means that the controller without a cache (LSI2008) achieves at most about 70% of the transaction rate of the other two controllers with a cache for all the three load profiles depicted. Also for this RAID level this is a performance difference that only occurs above 32 parallel accesses to the RAID array.

SSDs

Whereas the maximum possible transaction rate for an enabled hard disk cache is normally < 700 IO/s for random accesses to an individual HDD, it is approximately twenty times larger for a SAS-6G-SSD. As a result of such a high transaction rate it is also possible for arrays consisting of several SAS-6G-SSDs to provide throughputs for random accesses of several hundred MB/s. This in turn means that the resources and interfaces of the controllers are put under a much higher load than with HDDs and that the differences between the controller generations can be seen.

RAID 0 with eight SAS-6G-SSDs

The next diagram shows the transaction rates of the SSD array for RAID 0 for random load profiles that can be achieved with various controllers. The three groups of columns show the transaction rates for the standard load profiles “File copy” (random access, 50% read, 64 kB block size), “File server” (random access, 67% read, 64 kB block size) and “Database” (random access, 67% read, 8 kB block size).



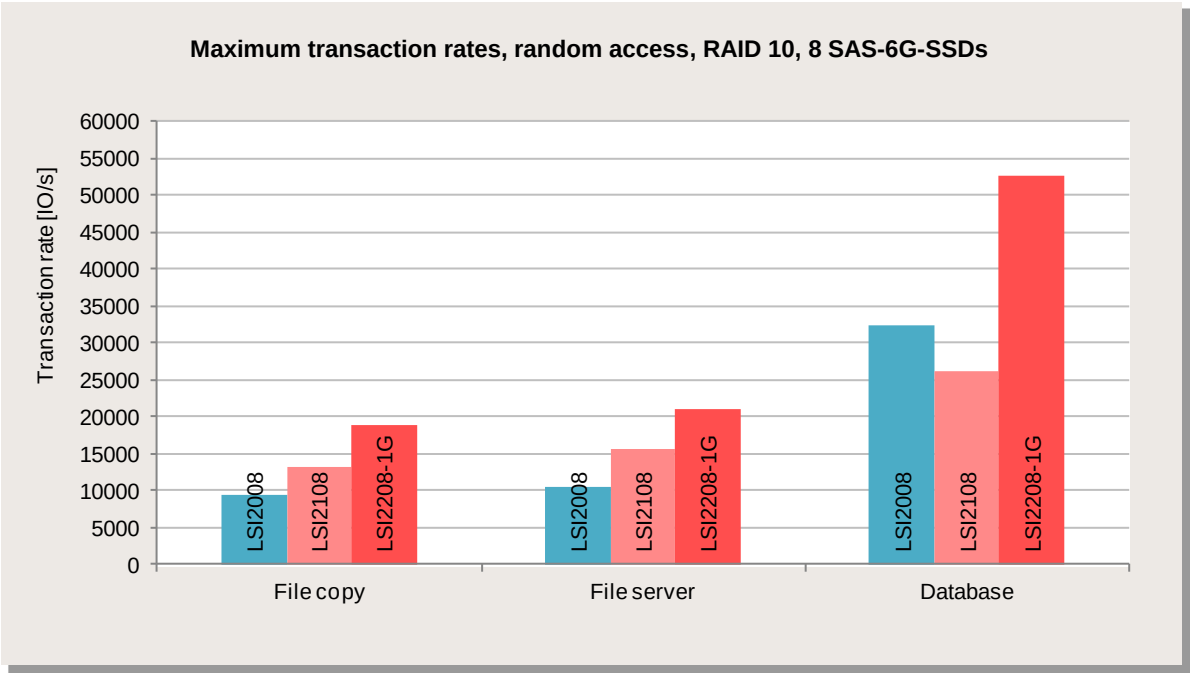
The LSI2208-1G controller is the one with top performance here.

It is also interesting to understand the throughput values that are associated with these transaction rates. Despite the lower transaction rates, the two load profiles with a 64 kB block size have the higher throughputs. For example, the LSI2208-1G controller handles a throughput of about 1563 MB/s with the load profile “File server”.

The next essential question is: To what extent does the effective transaction rate of such a RAID array scale with the number of SSDs? With eight SSDs the LSI2208-1G controller achieves approximately the eight-fold transaction rate of a single SSD for all three random load profiles; in other words it is not yet overloaded. Depending on the random load profile, the LSI2108 controller achieves approximately five to seven times the transaction rate of a single SSD; i.e. it would already be a little in overload.

RAID 10 with eight SAS-6G-SSDs

The next diagram shows the transaction rates of the SSD array for RAID 10 for random load profiles that can be achieved with various controllers. The three groups of columns show the transaction rates for the standard load profiles “File copy” (random access, 50% read, 64 kB block size), “File server” (random access, 67% read, 64 kB block size) and “Database” (random access, 67% read, 8 kB block size).



The LSI2208-1G controller also has a very considerable advantage here, particularly in the case of small blocks.

Sequential accesses

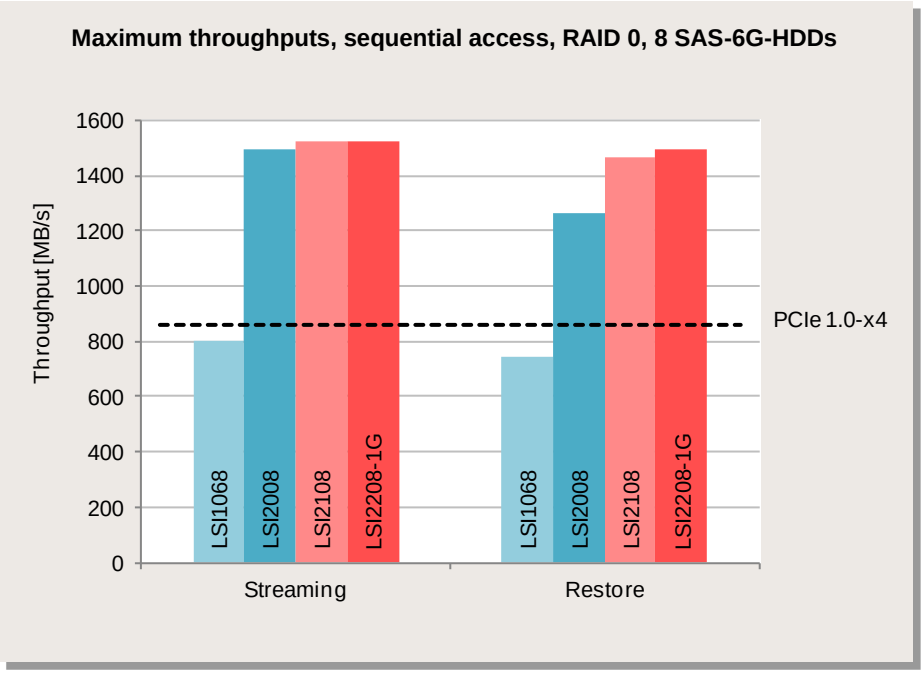
General statements about the controllers for sequential accesses are made below on the basis of measurements with eight hard disks. In this case, it makes sense to distinguish between HDDs and SSDs, because one controller more has been released for the HDDs and the throughput limits of the LSI2008 controller can be illustrated better with the SSDs. The maximum throughputs to be expected for this can be calculated for other types and numbers (≤ 8) of hard disks by means of suitable multiplication. If the throughput calculated in this way exceeds the limiting value of the controller, the controller limiting value becomes effective.

HDDs

RAID 0 with eight SAS-6G-HDDs

The next diagram shows the throughputs of the hard disk array for RAID 0 for sequential load profiles that can be achieved with various controllers. The two groups of columns in the diagram show the throughputs for the standard load profiles “Streaming” (sequential access, 100% read, 64 kB block size) and “Restore” (sequential access, 100% write, 64 kB block size).

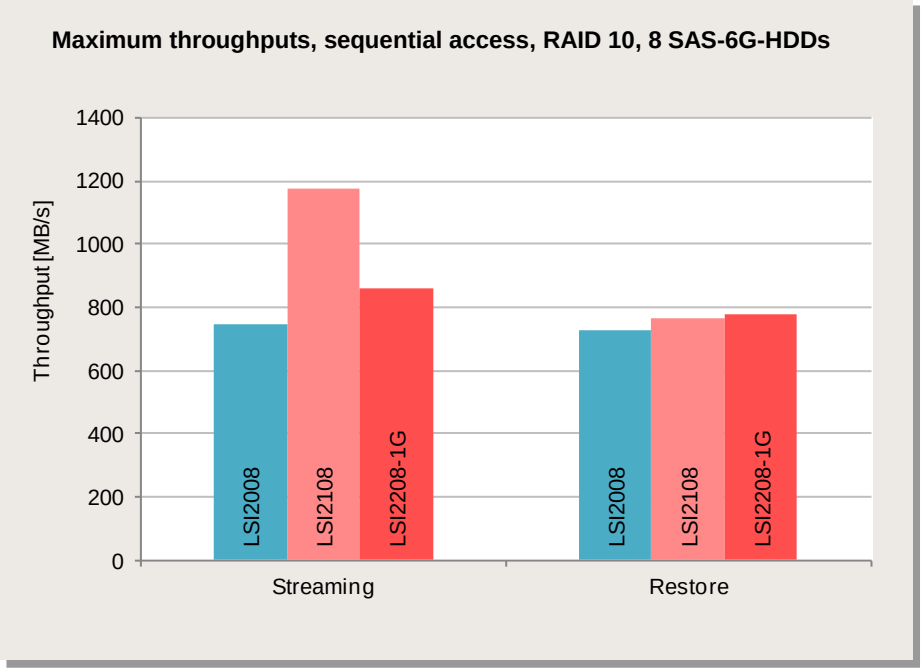
The diagram shows that the LSI1068 controller is no longer sufficient for sequential throughputs of 800 MB/s or more. This is due to the throughput limit of the PCIe interface (PCIe 1.0, x4, black, dotted horizontal line). Therefore, one of the three other controllers should be used for higher throughputs.



RAID 10 with eight SAS-6G-HDDs

The next diagram shows the throughputs of the hard disk array for RAID 10 for sequential load profiles that can be achieved with various controllers. The two groups of columns in the diagram show the throughputs for the standard load profiles “Streaming” (sequential access, 100% read, 64 kB block size) and “Restore” (sequential access, 100% write, 64 kB block size).

For sequential read and write all three of the controllers under consideration for this RAID array achieve or exceed a throughput of approximately four times the maximum throughput of a single HDD (i.e. about 800 MB/s in this case).

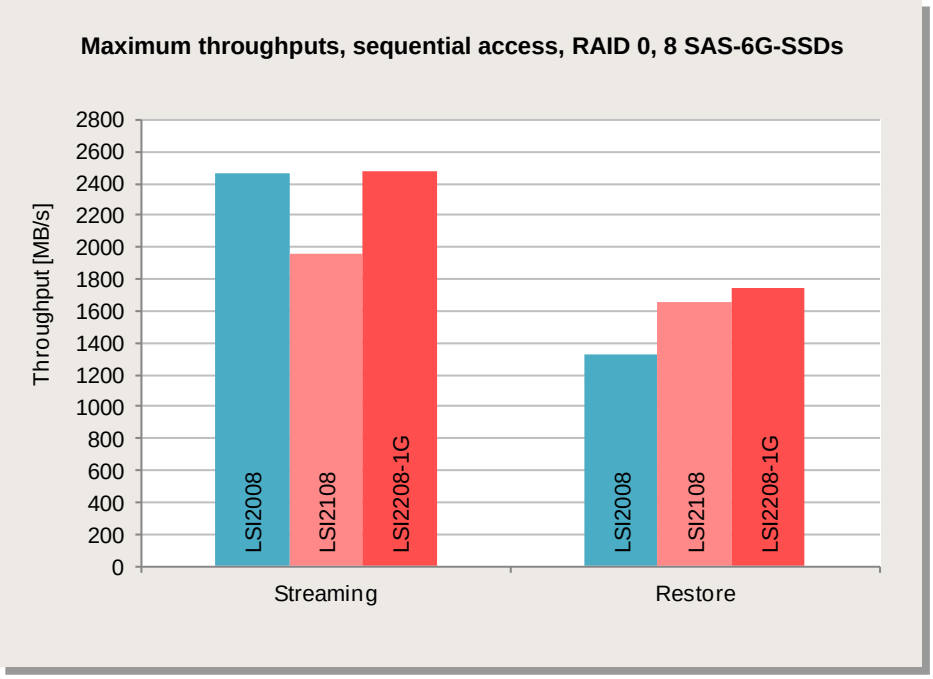


SSDs

RAID 0 with eight SAS-6G-SSDs

The next diagram shows the throughputs of the hard disk array for RAID 0 for sequential load profiles that can be achieved with various controllers. The two groups of columns in the diagram show the throughputs for the standard load profiles “Streaming” (sequential access, 100% read, 64 kB block size) and “Restore” (sequential access, 100% write, 64 kB block size).

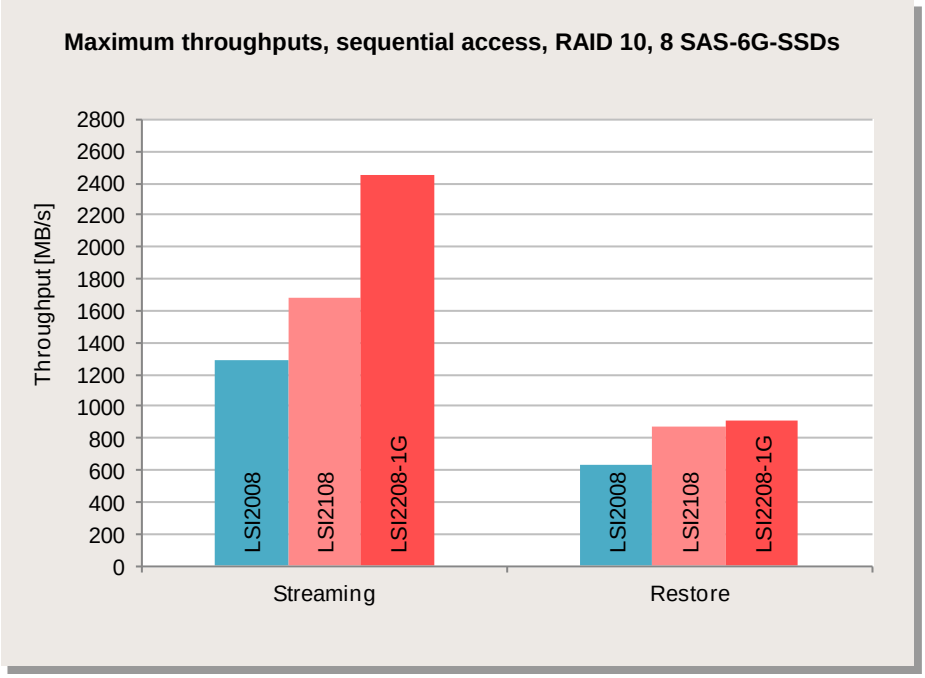
On the whole, the performance with the LSI2208-1G controller is the best in this case.



RAID 10 with eight SAS-6G-SSDs

The next diagram shows the throughputs of the hard disk array for RAID 10 for sequential load profiles that can be achieved with various controllers. The two groups of columns in the diagram show the throughputs for the standard load profiles “Streaming” (sequential access, 100% read, 64 kB block size) and “Restore” (sequential access, 100% write, 64 kB block size).

The performance with the LSI2208-1G controller is also the best in this case.



RAID 0 and 10 (more than eight hard disks)

The three controllers with the top performance are available for the operation of more than eight hard disks in the current PRIMERGY servers: the LSI2108 controller and the LSI2208-1G-2.0 and LSI2208-1G-3.0 controller. They currently enable server configurations with up to 24 internal hard disks (e. g. in the PRIMERGY TX300 S7). The comparisons for RAID 0 are carried out with 24 hard disks; and the comparisons for RAID 10 with 16 hard disks (for RAID 10 the latter is the current upper limit for a single RAID array). As a result, the range of large numbers of hard disks is covered on a representative basis. As was also the case in the previous subsection, measurements with high-performance SAS-6G-HDDs or SAS-6G-SSDs are used as a means of illustration. These hard disks are described in more detail in the section "Measurement environment".

In the case of RAID 0 the controllers with a cache usually have a higher maximum performance if you disable the controller cache. Thus, the transaction rates of the LSI2108 and the LSI2208-1G controllers in this subsection are based on the following changes compared with the standard settings of the "Performance" mode of ServerView RAID:

- Read mode from "Read-ahead" to "No read-ahead"
- Write mode from "Always write-back" to "Write-through"

The standard settings were used for RAID 10.

Random accesses

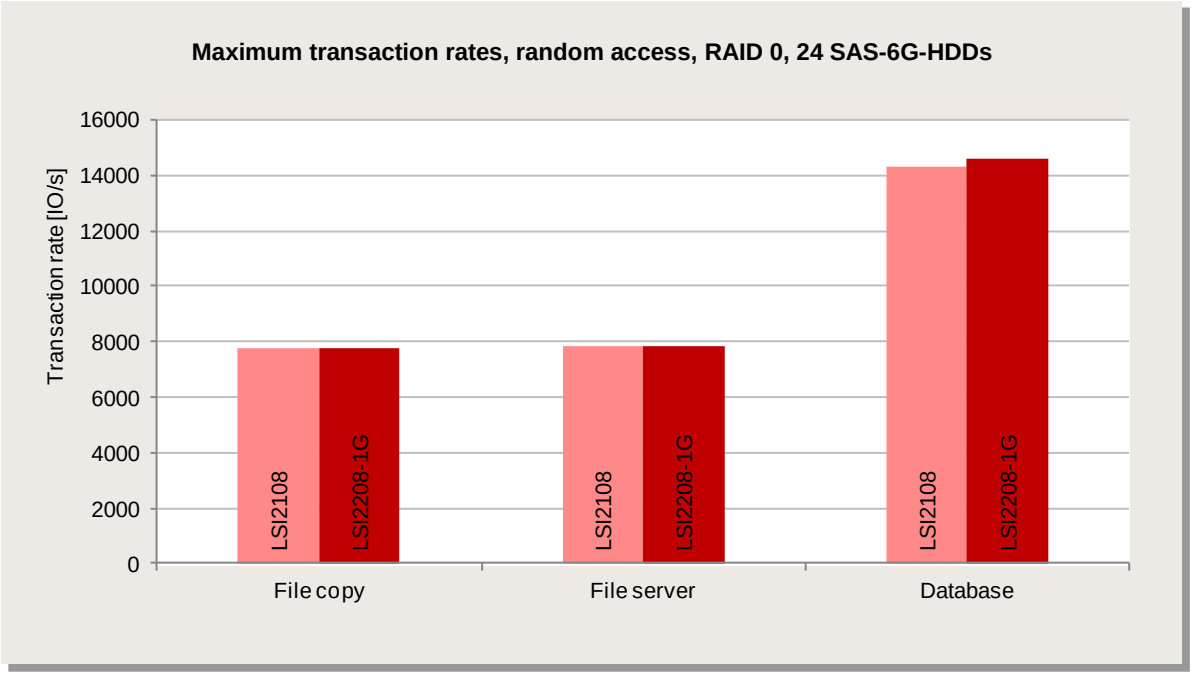
When considering random accesses for larger numbers of hard disks it makes sense to distinguish between HDDs and SSDs, because the maximum values for SSDs are of a quite different magnitude.

HDDs

The controllers are compared below with random accesses to HDDs. The maximum transaction rates of the storage medium for the load profile used are the most important limiting factor here. Although the following results were acquired with 24 SAS-6G-HDDs, they can also be used to estimate the maximum transaction rates to be expected for other types and numbers (≤ 24) of hard disks. The throughputs that occur with random accesses to HDDs are so low that any limitations to the PCIe or SAS interface of the controllers do not play a role.

RAID 0 with 24 SAS-6G-HDDs

The next diagram shows the transaction rates of the hard disk array for RAID 0 for random load profiles that can be achieved with various controllers. The three groups of columns show the transaction rates for the standard load profiles “File copy” (random access, 50% read, 64 kB block size), “File server” (random access, 67% read, 64 kB block size) and “Database” (random access, 67% read, 8 kB block size).



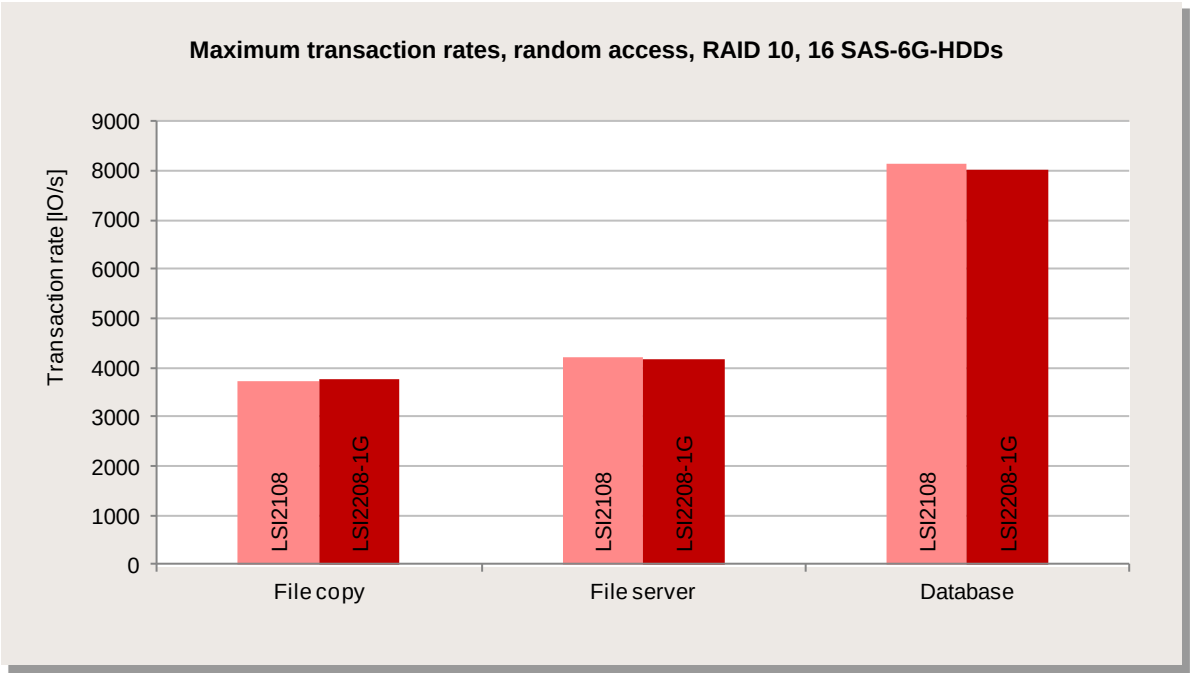
The SAS-6G hard disks used have the maximum transaction rate of 594 IO/s for the load profile “Database”. Theoretically, an array consisting of 24 such hard disks should therefore handle a maximum of 594 IO/s × 24 = 14256 IO/s. The maximum measured transaction rate for the load profile “Database” is 14396 IO/s; thus very much confirming the rough estimate. For load profiles with block size 64 kB only about half the transaction rate of that with the load profile with the 8 kB block size is possible in the measurements presented here, although the maximum transaction rates of the HDD type used for all three load profiles are relatively similar.[2]

All controllers achieve approximately the maximum transaction rate of this RAID array for each of the three load profiles.

[2] The reason for this is the initialization of the RAID array with the elementary block size of 64 kB (“stripe size”). The random load profile with block size 8 kB, which is small compared with the “stripe size”, can be estimated directly according to the formula “Number of HDDs × HDD performance”. If the load profile uses the same block size with which the RAID array was initialized, only about half can be achieved. A detailed reason for this here would be taking things too far.

RAID 10 with 16 SAS-6G-HDDs

The next diagram shows the transaction rates of the hard disk array for RAID 10 for random load profiles that can be achieved with various controllers. The three groups of columns in the diagram show the transaction rates for the standard load profiles “File copy” (random access, 50% read, 64 kB block size), “File server” (random access, 67% read, 64 kB block size) and “Database” (random access, 67% read, 8 kB block size).



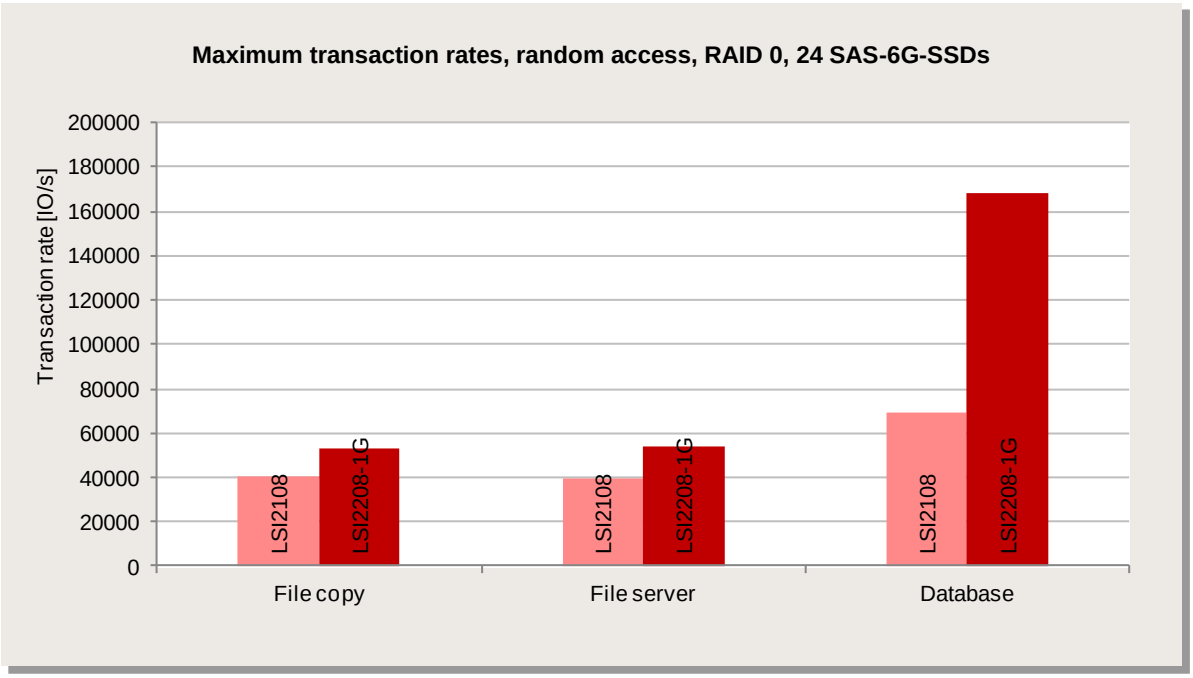
The diagram shows the same principal behavior as with RAID 0.

SSDs

Whereas the maximum possible transaction rate for an enabled hard disk cache is normally < 700 IO/s for random accesses to an individual HDD, it is approximately twenty times larger for a SAS-6G-SSD. As a result of such a high transaction rate it is also possible for arrays consisting of several SAS-6G-SSDs to provide throughputs for random accesses of several hundred MB/s. This in turn means that the resources and interfaces of the controllers are put under a much higher load than with HDDs and that the differences between the controller generations can be seen.

RAID 0 with 24 SAS-6G-SSDs

The next diagram shows the transaction rates of the SSD array for RAID 0 for random load profiles that can be achieved with various controllers. The three groups of columns show the transaction rates for the standard load profiles “File copy” (random access, 50% read, 64 kB block size), “File server” (random access, 67% read, 64 kB block size) and “Database” (random access, 67% read, 8 kB block size).



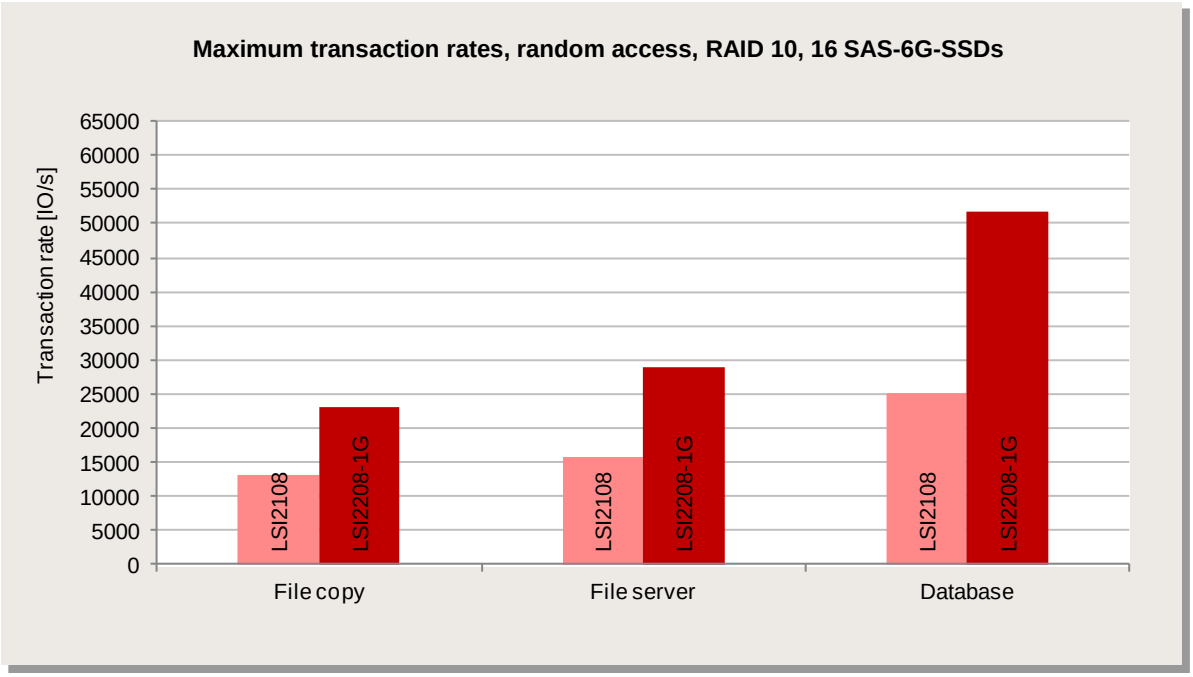
The LSI2208-1G controller is the one with top performance here. For the load profile with the smaller block size ("Database") it achieves approximately two and a half times the transaction rate of the LSI2108 controller.

Expressed in the form of SAS-6G-SSD numbers: you need - depending on the random load profile - between 13 and 17 SAS-6G-SSDs for example to make full use of the possibilities offered by the LSI2208-1G controller in the case of RAID 0. This confirms the statement made in subsection RAID 0 and 10 (five to eight hard disks) that the LSI2208-1G controller with eight SSDs for random load profiles is not yet overloaded.

It is also interesting for us to recognize the throughput values that result through conversion from these transaction rates. Despite the lower transaction rates, the two load profiles with a 64 kB block size have the higher throughputs. For example, the LSI2208-1G controller handles a throughput of about 3200 MB/s with the load profile “File copy”. This value is remarkable, because it is higher than the two sequential maximum throughputs of the controller for 100% read and 100% write with this RAID level. This value would not have been reached without real bidirectional use of the SAS connections.

RAID 10 with 16 SAS-6G-SSDs

The next diagram shows the transaction rates of the SSD array for RAID 10 for random load profiles that can be achieved with various controllers. The three groups of columns show the transaction rates for the standard load profiles “File copy” (random access, 50% read, 64 kB block size), “File server” (random access, 67% read, 64 kB block size) and “Database” (random access, 67% read, 8 kB block size).



The LSI2208-1G controller is the one with top performance here.

A similar picture to that of RAID 10 with eight SAS-6G-SSDs presents itself here. However, the transaction rates can only be slightly increased by doubling the number of SSDs to 16. Hence the recommendation not to connect up more than ten SSDs per controller and to use the LSI2208-1G controller in order to achieve maximum performance for RAID 10 with random load profiles.

Sequential accesses

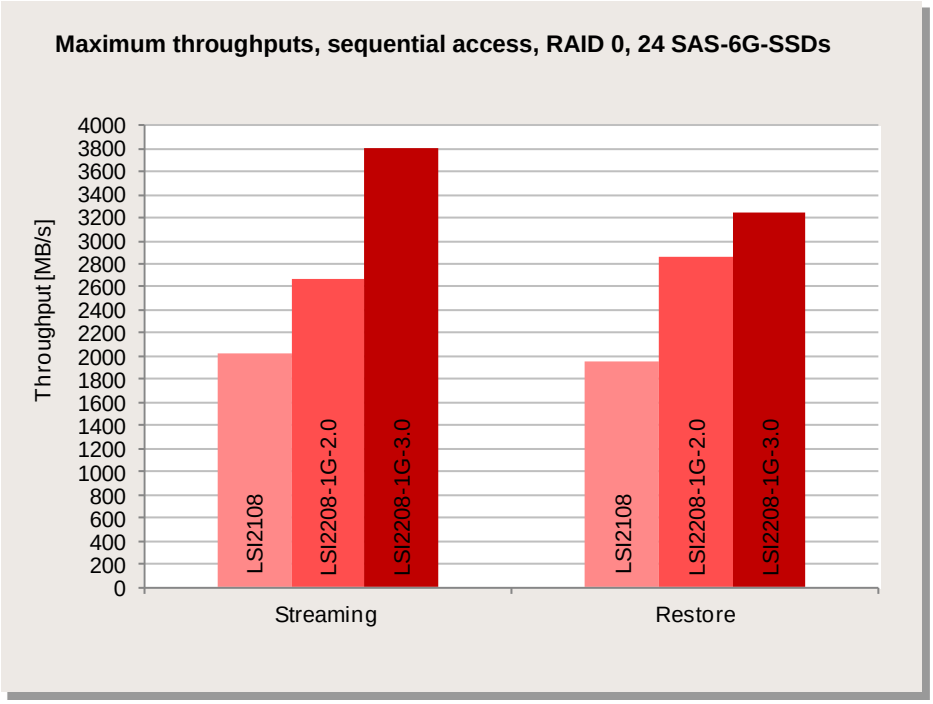
When considering sequential accesses for larger numbers of hard disks it is mostly not necessary to distinguish between HDDs and SSDs, because the maximum throughputs of the two hard disk types are still of a similar magnitude.

RAID 0 with 24 SAS-6G-SSDs

The next diagram shows the throughputs of the hard disk array for RAID 0 for sequential load profiles that can be achieved with various controllers. The two groups of columns in the diagram show the throughputs for the standard load profiles “Streaming” (sequential access, 100% read, 64 kB block size) and “Restore” (sequential access, 100% write, 64 kB block size).

The LSI2208-1G-3.0 controller also provides the best performance here and exceeds the limit of PCIe Gen2 in the case of "Streaming". The SSD array is the limit in the case of "Restore".

Compare this diagram with the corresponding diagram for RAID 0 with eight SAS-6G-SSDs; almost the same values were achieved there for the LSI2108 controller as here.

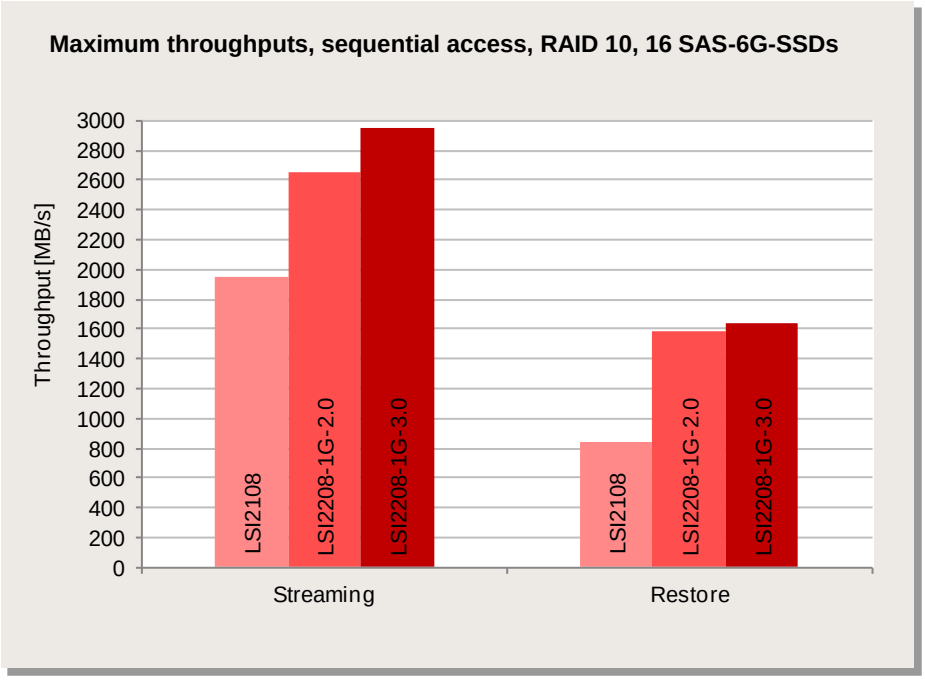


RAID 10 with 16 SAS-6G-SSDs

The next diagram shows the throughputs of the hard disk array for RAID 10 for sequential load profiles that can be achieved with various controllers. The two groups of columns in the diagram show the throughputs for the standard load profiles “Streaming” (sequential access, 100% read, 64 kB block size) and “Restore” (sequential access, 100% write, 64 kB block size).

The LSI2208-1G-3.0 controller also provides the best performance thanks to PCIe Gen3 here. However, the differences to Gen2 are not so large, because 16 SSDs are too few - particular with the load profile "Restore" - to fully utilize the higher throughput limit.

Again compare this diagram with the corresponding diagram for RAID 0 with eight SAS-6G-SSDs; almost the same values were achieved there for the LSI2108 controller as here.



RAID 5

With RAID levels 0, 1 and 10 the controller can restrict itself to assigning the logical position of the data blocks to a physical position in the RAID array (“striping”) and, if necessary, to writing entire blocks in a mirrored way (“mirroring”).

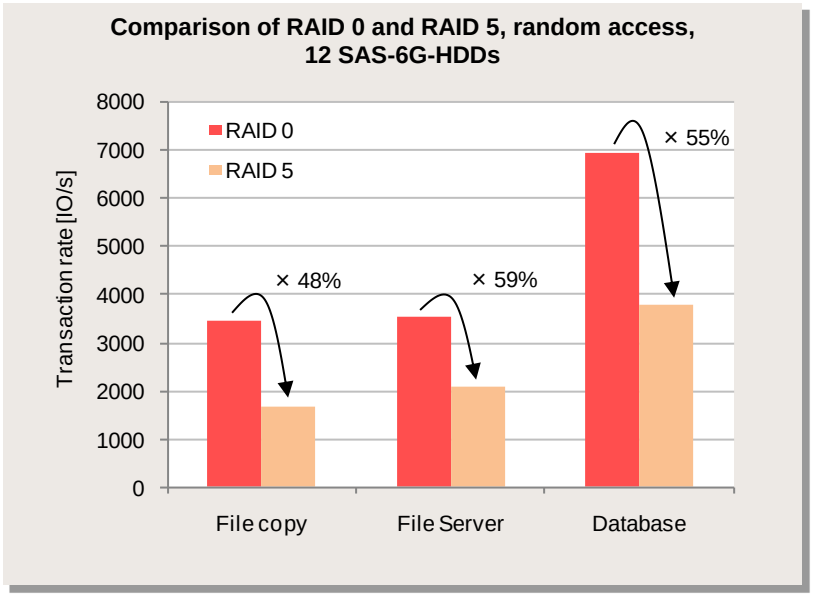
In contrast, a controller must manage considerably higher requirements with RAID levels such as RAID 5, mainly for write. This is due to the fact that such RAID levels require additional blocks with parity data, which first has to be calculated from the actual data.

The measurements of this subsection were made with the LSI2108 and the LSI2208 controllers, because only they support RAID 5 as standard. Since the maximum values for higher performance requirements are of particular interest, measurements with high-performance SAS-6G-HDDs or SAS-6G-SSDs are used as a means of illustration. These hard disks are described in more detail in the section “Measurement environment”.

Random accesses

A specifically increased number of accesses to the RAID array is generated from a write request of the application scenario for RAID 5. This has a particular effect on random accesses. As a result the maximum possible transaction rate drops from the viewpoint of the application to a specific percentage if the same quantity of hard disks is firstly configured as RAID 0, and then as RAID 5. In a first approximation these percentages are specifically for the comparison with RAID 5 and the random load profile used, but they also depend on the controller.

The following diagram represents the maximum measured transaction rates for an array of twelve SAS-6G-HDDs, which was first configured as RAID 0 and then as RAID 5 (by way of example, both times in connection with the LSI2108 controller). The values for RAID 0 are approximately the technical maximum for this HDD array. The three groups of columns in the diagram represent the transaction rates for the standard load profiles “File copy” (random access, 50% read, 64 kB block size), “File server” (random access, 67% read, 64 kB block size) and “Database” (random access, 67% read, 8 kB block size). The value measured for RAID 5 is a certain percentage of the value for RAID 0 for each of these three load profiles. These comparative percentages have been entered in the diagram above the arrows.



The controller enables a comparative percentage of 48% for the load profile with 50% write share (“File copy”); and it makes comparative percentages of 59% and 55% respectively possible for the load profiles with a write share of 33% (“File server” and “Database”). With an estimation based on multiplication factors for write requests it could be deduced that this comparative percentage would theoretically have to contribute 40% in the first case, and theoretically 50% in the two other cases. Thus, the percentages achieved with the controllers are in each case clearly higher. This is only possible on account of the controller cache (enabled as standard). The higher these percentages, the better the controller design.

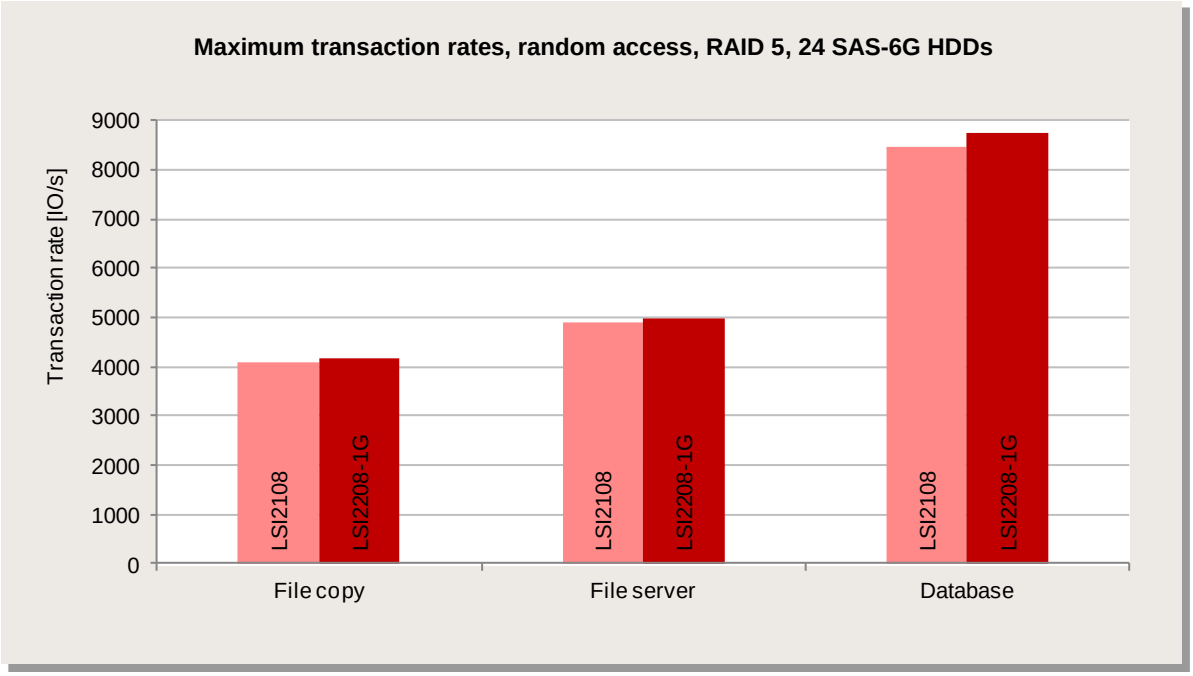
A detailed deduction of the theoretical percentages of 40% and 50% respectively would be taking us too far here.

HDDs

The performance values that can be achieved with RAID 5 for random load profiles are now the first to be illustrated in the case of HDDs. As with RAID-Levels 0, 1 and 10 (without parity calculation) the maximum performance of the storage medium is also dominating for this RAID level with HDDs in contrast to SSDs, which is why it makes sense to deal with the two hard disk technologies separately.

RAID 5 with 24 SAS-6G-HDDs

The next diagram shows the transaction rates of the hard disk array for RAID 5 for random load profiles that can be achieved with various controllers. The three groups of columns in the diagram show the transaction rates for the standard load profiles “File copy” (random access, 50% read, 64 kB block size), “File server” (random access, 67% read, 64 kB block size) and “Database” (random access, 67% read, 8 kB block size).



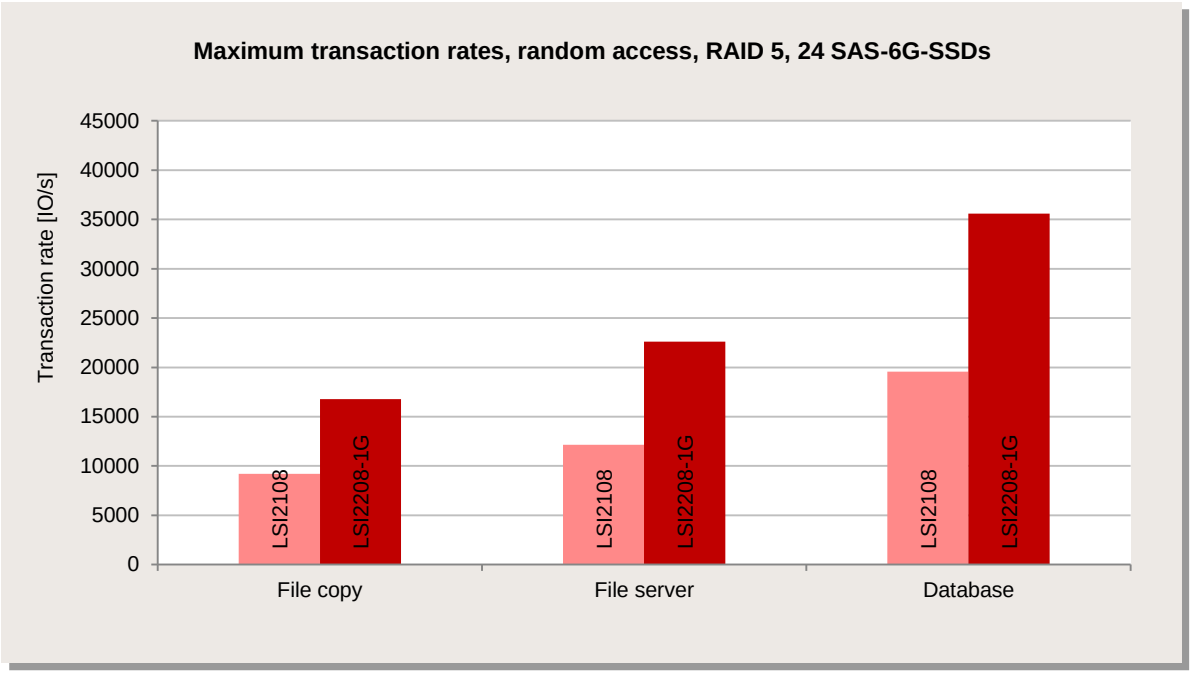
The LSI2208-1G controller, i.e. the newer controller with the larger cache, has a slight advantage.

SSDs

The transaction rates that can be achieved with SAS-6G-SSDs are to be discussed next (still in the case of RAID 5 with random load profiles). In a very similar way to the previously handled RAID levels 0 and 10, the differences between the various controllers also become apparent in RAID 5 due to the use of SAS-6G-SSDs.

RAID 5 with 24 SAS-6G-SSDs

The following diagram puts together the maximum transaction rates for the various controllers using the example of an array consisting of 24 SAS-6G-SSDs. The three groups of columns in the diagram show the transaction rates for the standard load profiles “File copy” (random access, 50% read, 64 kB block size), “File server” (random access, 67% read, 64 kB block size) and “Database” (random access, 67% read, 8 kB block size).



The technological progress from the previous to the current controller generation is as impressive here as in the case of RAID 0 with 24 SSDs: In the case of the standard load profile “Database” the LSI2208-1G controllers provide a transaction rate that is up to two times that of the LSI2108 controller.

Expressed in the form of SAS-6G-SSD numbers: with RAID 5 the LSI2208-1G controller – depending on the random load profile – enables between six and sixteen SAS-6G-SSDs to be fully loaded without a bottleneck arising in the controller. In addition to the SAS-6G-SSD-specific maximum values from the section “Measurement environment”, this estimation is also based on the percentages that are valid for RAID 5 for the reduction in the maximum transaction rate from the start of the current subsection “Random accesses”.

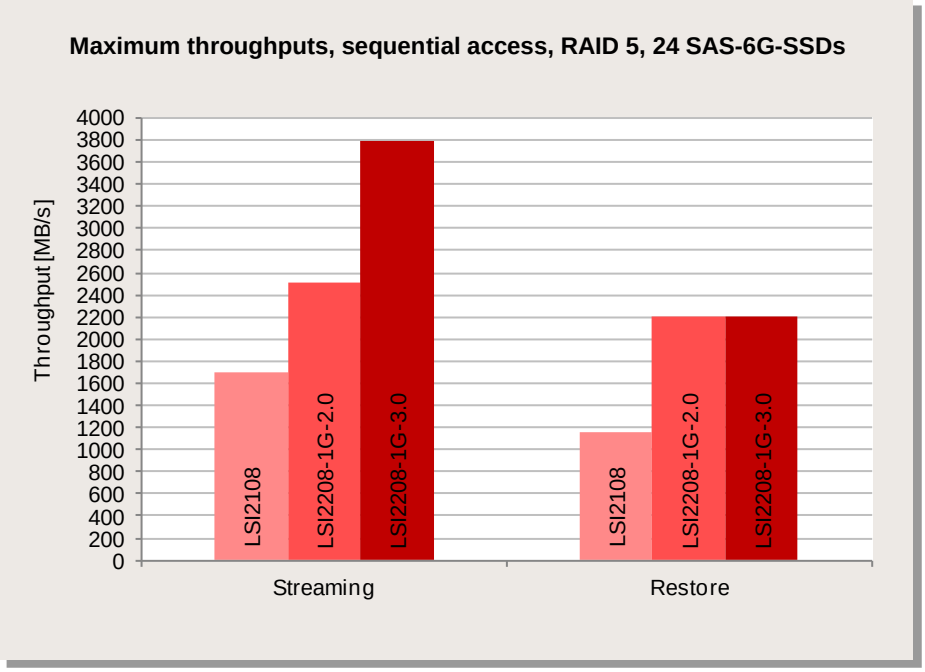
Sequential accesses

When considering sequential accesses with RAID 5 for larger numbers of hard disks it is also not necessary to distinguish between HDDs and SSDs, because the respective maximum values are still of a similar magnitude. General statements about the controllers are made below on the basis of measurements with SAS-6G-HDDs. The maximum throughputs to be expected can be calculated for other types and numbers (≤ 24) of hard disks. If the throughput to be expected exceeds the limiting value of the controller, the controller limiting value becomes effective.

In the case of RAID 5 there are always (N-1) blocks with the actual data and 1 block with parity data to be found in a logically associated group of blocks in an array consisting of N hard disks. Thus in all cases, a share of 1/(N-1) contains parity data when such a group is processed fully sequentially. Consequently, such sequential processing can never provide more than the (N-1)-fold data throughput of a single hard disk for the application.

RAID 5 with 24 SAS-6G-SSDs

The next diagram shows the throughputs of the hard disk array for RAID 5 for sequential load profiles that can be achieved with various controllers. The two groups of columns in the diagram show the throughputs for the standard load profiles “Streaming” (sequential access, 100% read, 64 kB block size) and “Restore” (sequential access, 100% write, 64 kB block size).



The great advantage of the LSI2208-1G-3.0 controller is clear to see here. As with RAID 0, 3800 MB/s are achieved with sequential read, which is clearly above the limit for PCIe Gen2.

The LSI2108 controller achieves a maximum throughput of about 1200 MB/s with sequential write for this RAID level, and the LSI2208-1G controller achieves about 2200 MB/s. In the case of RAID 5 this maximum value for sequential write is a significant indicator for the performance of a RAID controller, as the speed of the controller is reflected in a relatively undistorted way here in the calculation of the parity blocks.

For numbers of hard disks (N) between three and 24 it is possible - as already mentioned - to estimate the throughput values to be expected for the sequential load profiles reviewed here as the (N-1)-fold value of the throughput of a single hard disk. The limiting values of the controllers cannot be exceeded here. From the values of the section Measurement environment the following would e.g. result for the SAS-6G-HDD and SAS-6G-SSD which are taken as the basis for this document: In the case of the load profile “Restore” the LSI2208-1G controller should be used for seven or more HDDs and eight or more SSDs respectively.

Lower load levels

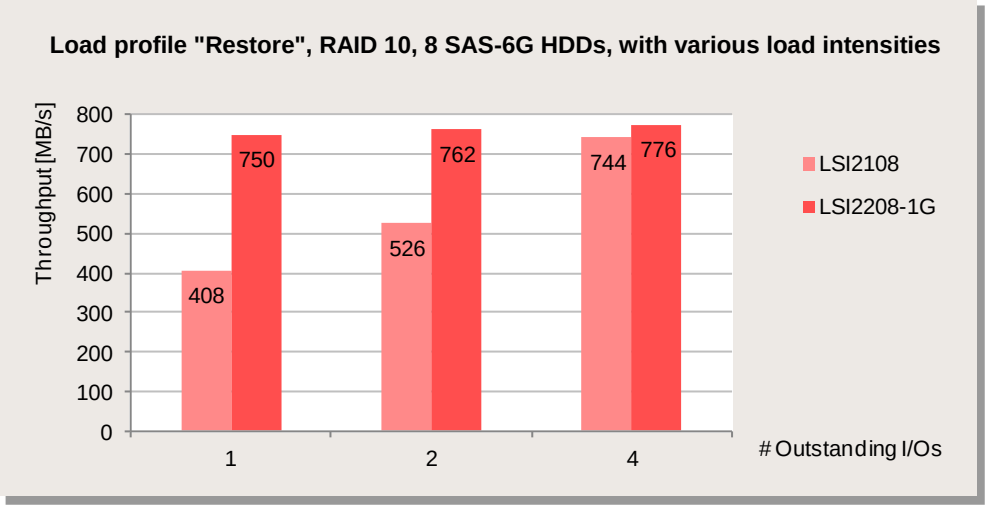
The differences between the controllers can not only be seen in the achievable maximum values under high load, but also under a lower load. Thus, the following table compiles for the individual controllers the maximum values for the sequential throughputs of a single application, which works without parallelism for its accesses (referred to below with “1 outstanding IO”). These maximum values apply for all the RAID levels that are supported by the controllers and cannot be exceeded. These throughputs can only be achieved with RAID 0; and with the other RAID levels the throughputs usually remain below the maximum values.

Controller	Maximum values of sequential throughput, 1 outstanding IO, 64 kB block size	
	100% read (load profile “Streaming”)	100% write (load profile “Restore”)
LSI1068	230 MB/s	180 MB/s
LSI2008	310 MB/s	340 MB/s
LSI2108	670 MB/s	670 MB/s
LSI2208-1G	930 MB/s	890 MB/s

The “Read-ahead” setting is prerequisite to achieving these values for sequential read, as is the setting “Write-back” for sequential write. These maximum throughputs also depend very much on the block size, whereby the interrelations of the table values as regards size are similar for other block sizes.

The differences between the last two controllers in the table become significant at the latest when the RAID array used is large enough to enable more than 900 MB/s of sequential throughput for “1 outstanding IO”. In such cases, an inappropriately selected controller can have the effect of a restriction.

The following example illustrates this on the basis of throughput measurements with a RAID 10 array consisting of eight SAS-6G-HDDs for the load profile “Restore” (sequential access, 100% write, 64 kB block size). The comparison is made between the LSI2108 and the LSI2208-1G controllers with differing numbers of parallel accesses (“# Outstanding IOs”).



You can clearly see that in this case the LSI2108 controller does not achieve the same throughput for 1 and 2 outstanding IOs as the LSI2208-1G controller. The latter already achieves a throughput of 750 MB/s, whereas the LSI2108 controller only achieves a little more than half.

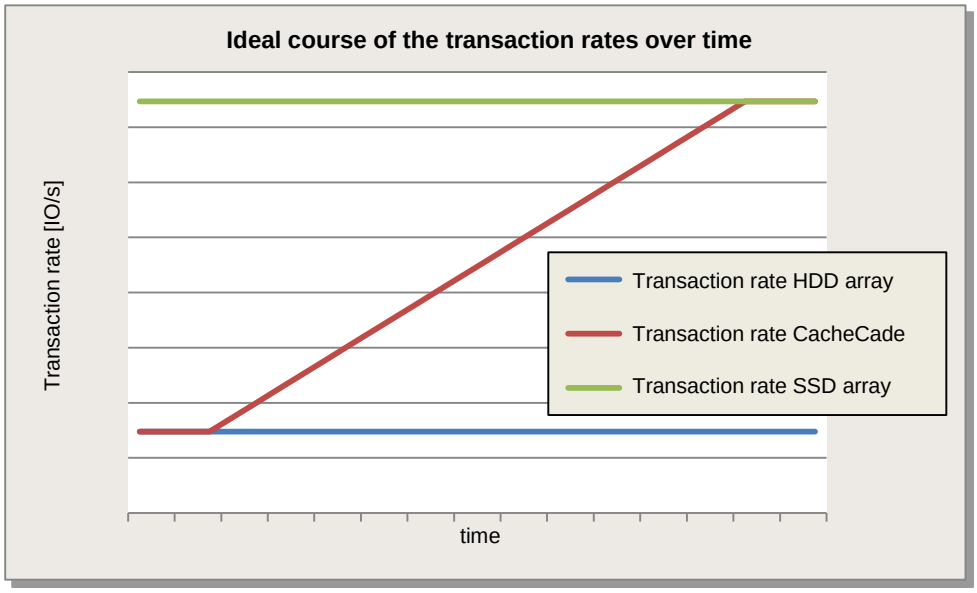
From the viewpoint of response times, this means that it is possible for low load intensities to approximately halve the response times with the LSI2208-1G controller compared with the LSI2108 controller.

MegaRAID Advanced Software Options

CacheCade 2.0

The starting point for the decision to use CacheCade 2.0 is a RAID array consisting of HDDs with predominantly random access. If a RAID level with data security is used or if SATA-HDDs with high capacities are used, then you can assume a not particularly high transaction rate. SSDs and CacheCade 2.0 can be used in order to increase this. However, it is prerequisite for access to the data to have a hot spot, i. e. access to a small part of the data is very intense. The aim with CacheCade 2.0 is to save this hot spot on the SSDs and thus achieve the high transaction rates of the SSDs with the majority of the accesses.

When accessing data, a transaction rate that starts on the level of the pure HDD array will achieve the level of the pure SSD array over time. In an ideal case, you would observe the following:



In relation to the SSDs the HDD array has a low transaction rate. During the data accesses the hot spot is gradually saved in the CacheCade 2.0 cache, which means on the SSDs, and you benefit from the relatively high transaction rate of the SSDs. Ideally, you achieve the transaction rate of a corresponding array consisting of SSDs.

Hot Spot

One essential prerequisite for CacheCade 2.0 is, as already mentioned, a hot spot with random data access. Recognizing this as such and estimating the size is one of the first steps for the planning of the CacheCade 2.0 cache. Beside access to the hot spot, other accesses are possible, but the proportion of the accesses to the hot spot should be 85% or more. Over time a hot spot may relocate, whereby lower transaction rates are set during this transition. A hot spot should not be too large and should suit the size of the CacheCade 2.0 cache. It has been shown that hot spots can be buffered particularly effectively up to 100 GB and that the cache should be at least twice as large as the hot spot.

Transaction rate improvement

A further important aspect with CacheCade 2.0 is the difference between the transaction rates of the HDD array and the SSD array. If it is relatively large, the use of CacheCade 2.0 will be particularly effective. If the HDD array is on the other hand already designed for a high transaction rate, e. g. due to a large number of HDDs, the difference will be minor.

In the section "RAID 0 with 24 SAS-6G-HDDs" you find e. g. a transaction rate of about 14000 IO/s for the standard load profile "Database". That corresponds pretty accurately to the transaction rate of an SAS-6G-SSD, as can be seen in the section "Measurement environment". It can therefore be concluded that such a configuration does not mean an advantage with CacheCade 2.0.

In each case you should assess whether an increase is to be expected and whether a slight increase in the transaction rate possibly meets your requirements.

Time factor

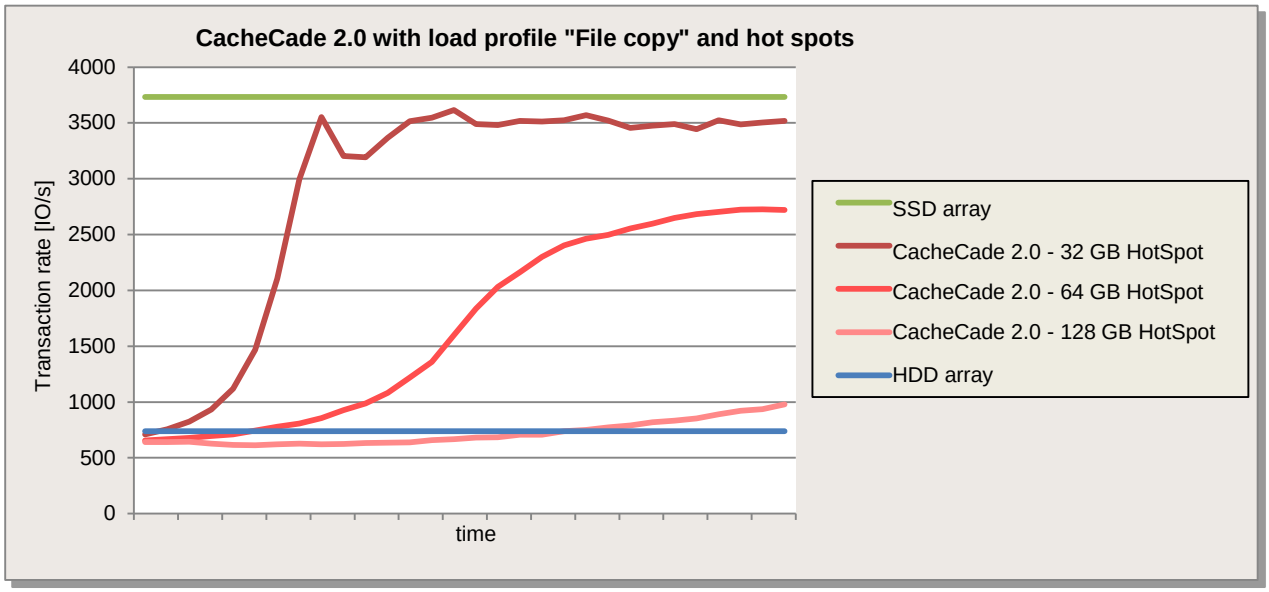
Ultimately, the time factor is of interest, i. e. how long it takes before CacheCade 2.0 provides higher transaction rates. This can be the case after a few minutes for a small hot spot. However, in the case of a large data volume or a HDD array with a low transaction rate it can take several hours.

The duration of the transition in the above diagram is thus influenced by several factors. The larger the hot spot, the longer it takes to transfer it to the CacheCade 2.0 cache. Data can be transferred in a shorter time with a high-performance HDD array. Finally, the intensity with which access takes place also has an impact, i. e. whether the maximum transaction rate of the HDD array is used.

How long it may take should also be considered here in each individual case and with due regard to the requirements.

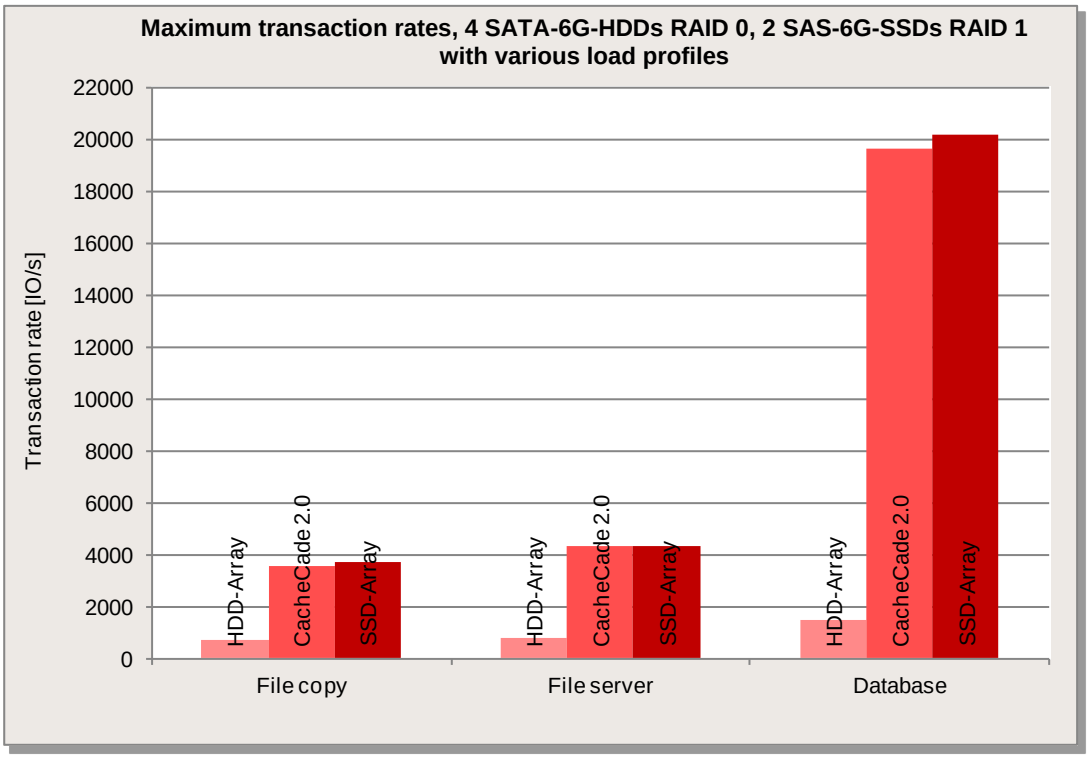
Example: Four SATA-HDDs RAID 0 and two SAS-SSDs RAID 1

A typical configuration for CacheCade 2.0 would be the combination of four SATA-HDDs as RAID 0 and two SAS-SSDs as RAID 1. Both configurations and their transaction rates are described in detail in this document under "RAID 0 and 10 (four SATA hard disks)" and "RAID 1 (two SAS hard disks)". The four SATA-6G-HDDs provide with RAID 0 a total storage space of about 4 TB. The two SAS-6G-SSDs have with RAID 1 approximately 200 GB. The following diagram shows the transaction rates that result over time if access to a hot spot of 32 GB, 64 GB and 128 GB takes place with the load profile "File copy" via CacheCade 2.0.



In practice the maximum transaction rate of the SSDs is not achieved, as a certain overhead for CacheCade 2.0 has to be taken into account. Very good results are achieved with a hot spot of 32 GB. If you have a hot spot of 64 GB, the transaction rates are clearly higher than with the HDD array, but no longer achieve the level of the SSDs. In the case of a 128 GB hot spot you can see that the transaction rate only increases very slowly and that a considerable amount of time is required until the transaction rate of the HDD array is exceeded.

In the previous measurement the "File copy" load profile was selected as a representative. Based on the simplified assumption that access is made to precisely 32 GB of the data, the maximum transaction rates for the three standard load profiles "File copy", "File server" and "Database" are shown as follows.



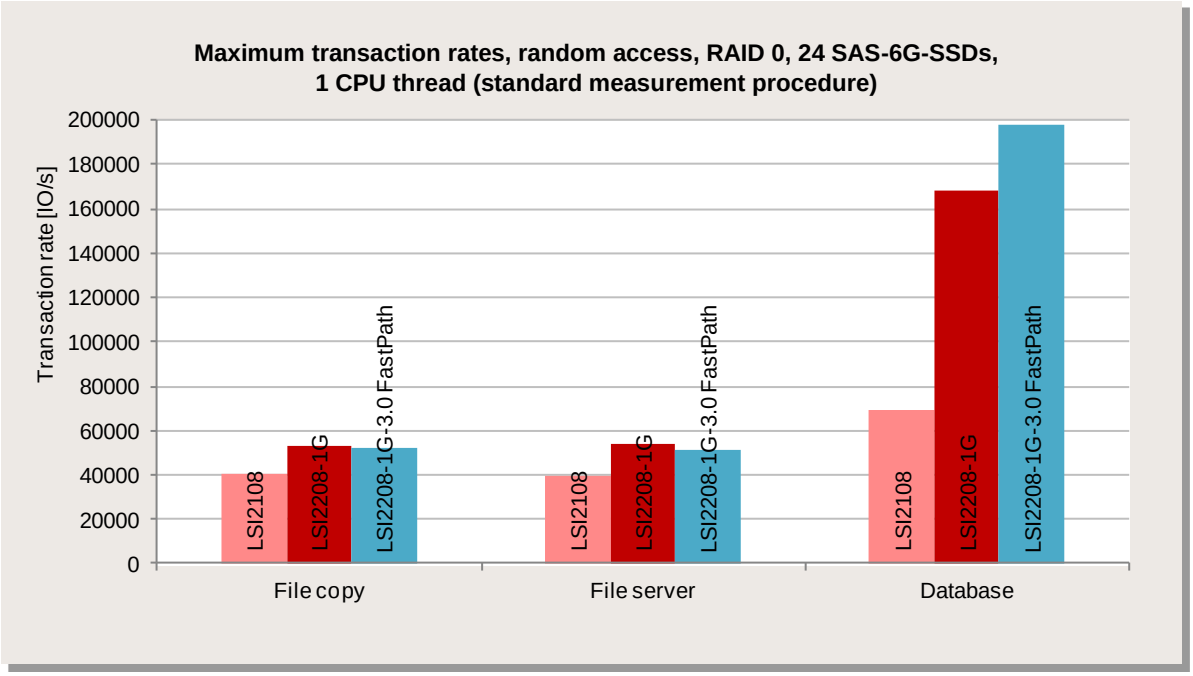
The measurement values refer to the LSI2208-1G controller. The HDD array consists of four SATA-6G-HDDs as RAID 0 and the CacheCade 2.0 cache consists of two SAS-6G-SSDs as RAID 1. The CacheCade 2.0 cache is configured for write and read. The maximum achievable transaction rates are shown in each case. A hot spot of 32 GB was assumed for access to the data.

FastPath

On the topic of large numbers of hard disks the configuration RAID 0 with 24 SAS-6G-SSDs was looked at more closely and presented in the section "RAID 0 and 10 (more than eight hard disks)". The highest transaction rates for random accesses are achieved here and a comparison with FastPath is appropriate.

RAID 0 with 24 SAS-6G-SSDs and FastPath

The next diagram shows the transaction rates of the SSD array for RAID 0 for random load profiles that can be achieved with various controllers. The three groups of columns show the transaction rates for the standard load profiles "File copy" (random access, 50% read, 64 kB block size), "File server" (random access, 67% read, 64 kB block size) and "Database" (random access, 67% read, 8 kB block size). Based on the measurement results already presented in the section "Controller comparison" the results for the LSI2208-1G-3.0 controller in combination with FastPath have been added here.



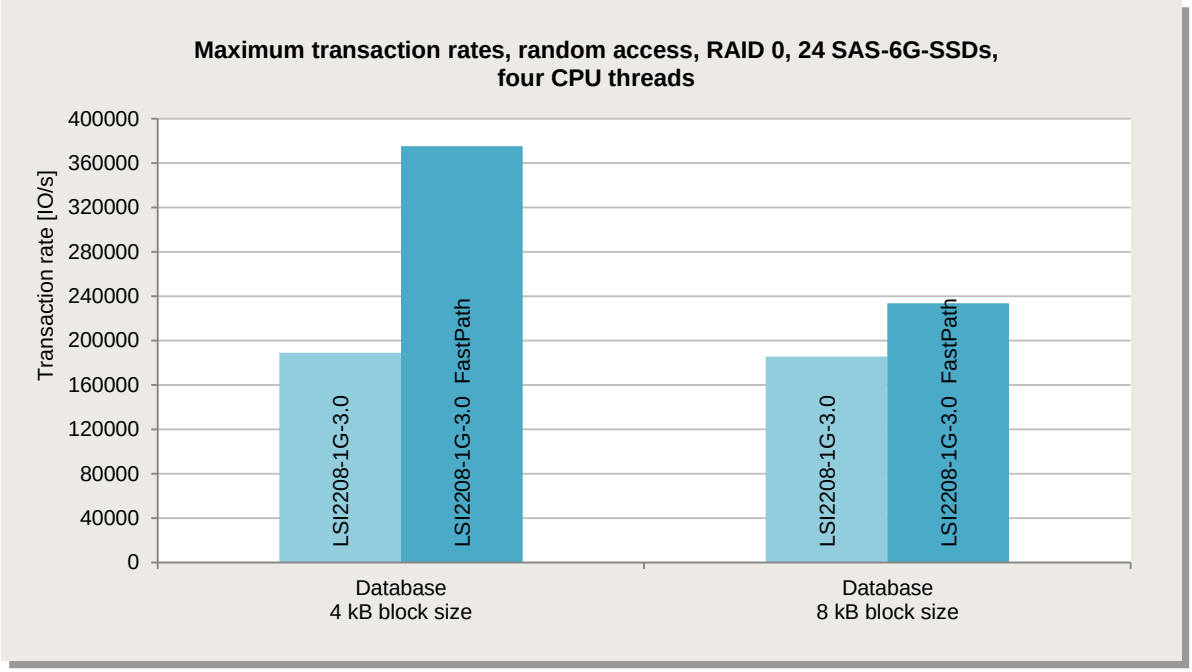
The two load profiles "File copy" and "File server" represent applications with a block size of 64 kB. As already mentioned, despite relatively low transaction rates from the LSI2208-1G controller high data throughput rates of more than 3000 MB/s are achieved, which represent a limitation to performance. Thus, no increase is achieved with FastPath. The load profile "Database" represents applications with a block size of 8 kB. At about 168000 IO/s the throughput amounts to approximately 1310 MB/s and is not close to the limit. With the standard measurement procedure an increase in performance can already be seen for the "Database" load profile thanks to FastPath. As described in the section "Measurement method", the "# of outstanding I/Os" is used in order to model applications that access in parallel. During the standard measurement procedure with Iometer this number varies per RAID array within just one load-generating CPU thread. With approximately 200000 I/Os the situation arises that one CPU thread of the server is at 100% utilization. If real applications behave in a similar way, i. e. handling the entire IO load with a CPU thread, a maximum of 200000 IO/s is also achieved and FastPath will not bring about an increase.

However, in the case of applications with high transaction rates, e. g. with standard databases, the access activities are typically spread over several CPU threads. The modeling of the "Database" application is adapted accordingly and spread over four CPU threads within Iometer. As a supplement to the block size of 8 kB the block size of 4 kB is also taken into consideration. A higher transaction rate could be achieved with even smaller block sizes. However, block sizes smaller than 4 kB are rather untypical.

Compared to the standard settings of the “Performance” mode of ServerView RAID, the ServerView RAID mode “Fast Path optimum” is used, as specified by ServerView RAID. This currently has the following settings:

- Read Mode “No read-ahead”
- Write Mode “Write-through”
- Cache Mode “Direct”
- Disk Cache “Enabled”

The following diagram shows the transaction rates of the LSI2208-1G-3.0 controller for the load profile “Database” with 4 kB and 8 kB block size if several CPU threads are used.



The transaction rate of the LSI2208-1G-3.0 controller without FastPath is close to 190000 IO/s with 4 kB and 8 kB. It can therefore be concluded that the controller, and not the CPU of the system, has a limiting impact, because approximately 170000 IO/s were already achieved in the previous diagrams with only one CPU thread. In contrast, transaction rates of more than 360000 IO/s are achieved with FastPath if block sizes of 4 kB are used; applications with 8 kB block size also show an increase of up to 240000 IO/s with FastPath, but are then limited on the SAS side with a throughput of about 1900 MB/s.

The "Database" load profile shown here uses a write share of 33%. It is quite conceivable for applications to use a smaller write share or for access to take place with a 1 kB or 2 kB block size. In such cases, the absolute transaction rates can also be significantly higher than 400000 IO/s.

Conclusion

The PRIMERGY servers use the “Modular RAID” concept to offer a plethora of opportunities to meet the requirements of various application scenarios.

An onboard controller is a low-priced entry-level alternative for the RAID levels 0, 1 and 10, which saves one PCIe slot but is restricted to four hard disks. The pro rata consumption of the server's processor performance is increasingly less important in newer servers.

On the SAS/SATA side the current onboard controllers support the standards up to frequency 3G. In the case of modern SAS-6G-SSDs a controller with an SAS-6G interface is also required to achieve the best possible throughputs. This would mean e. g. for a PRIMERGY BX920 S3 that you should not use the “Onboard Intel C600 SAS RAID” controller, but e. g. the “PY SAS RAID HDD Module” controller (LSI2208-512).

The “SAS 0/1” RAID controllers without a controller cache support the basic RAID solutions RAID 0, RAID 1 as well as RAID 1E and RAID 10 respectively, and provide a good performance for these RAID levels. Solely the “RAID 0/1 SAS controller based on the LSI MegaRAID 8Port” controller (LSI1068) from the previous generation supports sequential throughputs up to about 800 MB/s only. If higher throughputs are needed, a suitable, current controller should be used.

The “SAS 5/6” RAID controllers offer all the current standard RAID solutions RAID 0, RAID 1, RAID 5, RAID 6, RAID 10, RAID 50 and RAID 60 in the High-End sector. These controllers have a controller cache and can as an option be backed up using a BBU/FBU. Manifold options to set the use of the cache make it possible to flexibly adapt the controller performance to suit the RAID level used. In many application scenarios, for example if random accesses take place on conventional hard disks with a high load intensity, these controllers enable a 70% higher transaction rate than the “SAS 0/1”-RAID controllers.

According to technological progress the current "SAS 5/6" RAID controllers can be divided into three groups: The first group is based on the LSI SAS2108 chip, the second group is based on the LSI SAS2208 chip in the version for PCIe 2.0, and the third group is based on the LSI SAS2208 chip in the version for PCIe 3.0. The first group contains the "RAID Ctrl SAS 6G 5/6 512 MB (D2616)" controller and the "PY SAS RAID Mezz Card 6Gb" controller. The second group contains the "RAID Ctrl SAS 6G 5/6 1GB (D3116)" controller, the "PY SAS RAID HDD Module" controller and the "PY SAS RAID HDD Module w/o cache" controller. The third group contains the "RAID Ctrl SAS 6G 5/6 1GB (D3116C)" controller, the "PY SAS RAID HDD Module 3.0" controller and the "PY SAS RAID HDD Module w/o cache 3.0" controller.

The majority of the application scenarios that put a load on the disk subsystem come along with a random read / write access. If SSDs are used to manage very high IO rates, the controller has considerable influence on the maximum transaction rate. Thus, for example the “RAID Ctrl SAS 6G 5/6 1GB (D3116)” controller (LSI2208-1G) partially enables almost a tripling of the maximum transaction rate compared with the first group. Therefore, a check should be made as to whether the “RAID Ctrl SAS 6G 5/6 512 MB (D2616)” controller (LSI2108) is still sufficient, or whether a newer controller should be used for every planned usage with more than four SSDs per controller for the respective load profile and the planned RAID array using the diagrams in the section “Controller comparison”.

Regardless of the hard disk type, the various controllers each have maximum sequential throughputs that are specific to the RAID level and the load profile. These maximum values have in part increased substantially between the three groups of the “SAS 5/6” controllers, for example in the case of RAID 5 and sequential read from 1700 MB/s in the first group and 2500 MB/s in the second group to ultimately 3800 MB/s in the third group.

If a higher transaction rate or higher throughput is required for the planned application scenario than a single controller can provide, two controllers can be used. A number of PRIMERGY servers provide this option (e. g. PRIMERGY TX300 S7).

A further aspect of faster controllers with sequential access profiles is the increased throughput that is already achieved with low access parallelism. If the RAID array is efficient enough, it means that more than 900 MB/s is possible for read and write with the “RAID Ctrl SAS 6G 5/6 1GB (D3116)” controller in this special application. Compared with controllers of the previous generation this also means a significant increase in the maximum throughput for these special cases.

Although the MegaRAID Advanced Software options comprise the two functions CacheCade 2.0 and FastPath, they will not be used simultaneously on the same controller as they follow different optimization approaches.

CacheCade 2.0 accelerates random accesses to HDD arrays, which should have a certain minimum size. Configurations with one HDD and one SSD make little sense. It should on the other hand be noted that large HDD arrays can already achieve a high transaction rate and that no gain is possibly to be expected through the CacheCade 2.0 cache. CacheCade 2.0 is especially effective if access is made to a small subarea (hot spot) of data between 32 GB and 64 GB.

FastPath is in contrast to CacheCade 2.0 designed for use with high transaction rates, as occur in SSD arrays. If a maximum transaction rate of 200000 IO/s is not exceeded in a configuration, but the SSDs still have adequate reserves and the SAS resources have not yet been exhausted, FastPath can then be used to increase the transaction rate.

The RAID-Manager software “ServerView RAID” that is supplied for PRIMERGY servers is recommended for the configuration of controllers and hard disks. This utility program makes it possible to conveniently adapt controller and hard disk settings to meet customer requirements regarding performance and data security in a controller-independent way for the majority of the application scenarios. If BBUs/FBUs and UPSs are used as buffers in the case of power failures, maximum performance can be reconciled with data security.

Literature

PRIMERGY Systems

http://primergy.com/

PRIMERGY Performance

http://www.fujitsu.com/fts/products/computing/servers/primergy/benchmarks/

This White Paper:


http://docs.ts.fujitsu.com/dl.aspx?id=e2489893-cab7-44f6-bff2-7aeea97c5aef

http://docs.ts.fujitsu.com/dl.aspx?id=71fac54b-7ec3-4b3f-b13d-f80fbb42d583

http://docs.ts.fujitsu.com/dl.aspx?id=e34159fa-0196-4a01-99ff-8792b5f644eb

RAID Controller Performance (previous white paper, valid until 2011)

http://docs.ts.fujitsu.com/dl.aspx?id=ada7c1bf-74e3-4953-b783-839cdeec790b

Single Disk Performance

http://docs.ts.fujitsu.com/dl.aspx?id=0e30cb69-44db-4cd5-92a7-d38bacec6a99

Basics of Disk I/O Performance

http://docs.ts.fujitsu.com/dl.aspx?id=65781a00-556f-4a98-90a7-7022feacc602

Information about Iometer

http://www.iometer.org

Contact

FUJITSU

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

PRIMERGY Product Marketing

mailto:Primergy-PM@ts.fujitsu.com

PRIMERGY Performance and Benchmarks

mailto:primergy.benchmark@ts.fujitsu.com

2013-08-01 WW EN