

WHITE PAPER

FUJITSU PRIMERGY SERVER

RAID CONTROLLER PERFORMANCE

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.

Version	
1.0b	
2011-09-15	
Content	
Document history	2
Introduction.....	3
RAID controllers for PRIMERGY: Basics.....	4
Presentation of the RAID controllers	4
Controller interfaces and their throughputs	6
Performance-relevant settings.....	8
Characteristics of the onboard controllers.....	10
Measurement context.....	11
Measurement method.....	11
Measurement environment.....	12
Controller comparison	14
RAID 0, 1 and 10 (up to four hard disks).....	15
RAID 0 and 10 (more than four hard disks).....	19
RAID 5	26
Lower load levels	31
Conclusion.....	32
Literature	33
Contact	33

Document history

Version 1.0

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](#)" covers all the RAID controllers currently on offer for PRIMERGY and their performance.
- The document "[RAID Performance](#)" compares the individual RAID levels and provides information about performance and the optimal configuration.

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 Disks, 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 maximum throughputs 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 underlined 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 = 1 billion bytes) while all other capacities as well as throughput are specified on a basis of 2 (1 MB/s = 2²⁰ Byte/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 maximum throughputs, 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 LSI2108. These names come from the basic chip name of the controller and, optionally, the cache size.

Controller name	Alias	Cache	Frequency	Supported interfaces		Max. # disks	RAID levels	BBU
Onboard SATA RAID ICH10R	ICH10R		3G	SATA		4	0, 1, 5*, 10	
Onboard SATA RAID Ibx Peak	Ibx Peak		3G	SATA		4	0, 1, 5*, 10	
Onboard SATA RAID Cougar Point	Cougar Point		3G	SATA		4	0, 1, 5*, 10	
RAID 0/1 SAS based on LSI MegaRAID 4Port	LSI1064		3G	SATA SAS-1.0	PCIe 1.0 x4	4	0, 1, 1E	
RAID 0/1 SAS based on LSI MegaRAID 8Port	LSI1068		3G	SATA SAS-1.0	PCIe 1.0 x4	8	0, 1, 1E	
RAID 5/6 SAS based on LSI MegaRAID 256 MB	LSI1078-256	256 MB	3G	SATA SAS-1.0	PCIe 1.0 x4	8	0, 1, 5, 6, 10, 50, 60	✓
RAID 5/6 SAS based on LSI MegaRAID 512 MB	LSI1078-512	512 MB	3G	SATA SAS-1.0	PCIe 1.0 x4	8	0, 1, 5, 6, 10, 50, 60	✓
RAID Ctrl SAS 6G 0/1 (D2607)	LSI2008		6G	SATA SAS-1.0 SAS-2.0	PCIe 2.0 x8	8	0, 1, 1E, 10	
RAID Ctrl SAS 6G 5/6 512 MB (D2616)	LSI2108	512 MB	6G	SATA SAS-1.0 SAS-2.0	PCIe 2.0 x8	8	0, 1, 5, 6, 10, 50, 60	✓

*) For some PRIMERGY servers it is possible to activate RAID 5 for the onboard controller via the additionally available "iButton".

The column "Max. # disks" specifies the maximum number of hard disks that can be directly run on the controller. 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.

With "RAID 0/1" SAS-RAID controllers the LSI2008 controller is the 6G successor of the LSI1068 controller, whereas the LSI2108 controller is the 6G successor of the LSI1078 controller, and together both form the "RAID 5/6" SAS-RAID controller family.

The LSI1078 controller exists in two different cache sizes, LSI1078-256 and LSI1078-512. If only LSI1078 is mentioned below, any statements made apply 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 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	LSI1064	LSI1068	LSI2008	LSI1078-256	LSI1078-512	LSI2108
BX620 S5	2			✓	✓				
BX620 S6	2			✓	✓				
BX920 S1	2			✓					
BX920 S2	2			✓					✓
BX922 S2	2		ICH10R						
BX924 S2	2		ICH10R						
BX960 S1	2		ICH10R						
CX120 S1	2		ICH10R						
CX122 S1	2		ICH10R						
RX100 S6	4		IbexPeak	✓		✓			✓
RX100 S7	4		CougarPoint	✓		✓			✓
RX200 S5	8		ICH10R	✓	✓		✓	✓	
RX200 S6	8		ICH10R	✓		✓			✓
RX300 S5	12	✓			✓		✓	✓	✓
RX300 S6	12	✓			✓	✓			✓
RX600 S4	8							✓	
RX600 S5	8					✓			✓
RX600 S6	8					✓			✓
RX900 S1	8								✓
RX900 S2	8								✓
SX940 S1	4			✓		✓		✓	✓
SX960 S1	10	✓							✓
TX100 S2	4		IbexPeak						✓
TX120 S2	4		ICH9R	✓		✓	✓		✓
TX120 S3	4		CougarPoint	✓		✓			✓
TX140 S1	8		CougarPoint		✓	✓			✓
TX150 S7	8		IbexPeak		✓	✓			✓
TX200 S5	16		ICH10R		✓	✓	✓	✓	✓
TX200 S6	16		ICH10R		✓	✓			✓
TX300 S5	20	✓			✓		✓	✓	
TX300 S6	20	✓				✓			✓

The LSI1064 controller is already integrated in the server blades BX920 S1 and BX920 S2 on the motherboard. The LSI2108 controller is available as a mezzanine card in the server blade BX920 S2, whereas the LSI1064/LSI1068 controller is implemented in the server blades BX620 S5 and BX620 S6 as a SAS storage module and SAS/RAID storage module respectively.

However, these versions are only another technical implementation of the comparable PCIe plug-in card controller, which has no influence on the level of performance.

Controller interfaces and their throughputs

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 maximum possible 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 for the connection of hard disks, whose data throughput depends on the frequency.

Type	Frequency	Theoretical throughput	Practical throughput (85%)
SAS-1.0, SAS 3G	3000 MHz	286 MB/s	243 MB/s
SAS-2.0, SAS 6G	6000 MHz	572 MB/s	486 MB/s
SATA, SATA 3G	3000 MHz	286 MB/s	243 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 divided by 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-1.0	1 x x4	3000 MHz	1144 MB/s	973 MB/s
SAS-1.0	2 x x4	3000 MHz	2289 MB/s	1945 MB/s
SAS-2.0	1 x x4	6000 MHz	2289 MB/s	1945 MB/s
SAS-2.0	2 x x4	6000 MHz	4578 MB/s	3890 MB/s
SATA	1 x x4	3000 MHz	1144 MB/s	973 MB/s
SATA	2 x x4	3000 MHz	2289 MB/s	1945 MB/s

Some PRIMERGY models can be expanded with a larger number of hard disks than the controller has the hard disk channels. 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 hard disks of the standards SATA 1.5 Gbit/s and SATA 3.0 Gbit/s. 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 1.0 is also often referred to as "PCIe Gen1", whereas PCIe 2.0 is also called "PCIe Gen2".

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 divided by the so-called 8b/10b coding. 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. 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 maximum throughput values listed here are obtained with the help of the two previous subsections "[SAS and SATA](#)" and "[PCIe](#)". The significant maximal throughput in each case is highlighted in the table in bold print.

Controller alias	Cache memory type	# Disk channels	Maximum throughput of disk interface	PCIe version	Effective PCIe width	Maximum throughput of PCIe interface
ICH10R		4 x SATA	973 MB/s			
Ibex Peak		4 x SATA	973 MB/s			
Cougar Point		4 x SATA	973 MB/s			
LSI1064		4 x SAS-1.0	973 MB/s	1.0	x4	858 MB/s
LSI1068		8 x SAS-1.0	1945 MB/s	1.0	x4	858 MB/s
LSI1078	DDR2 / 667 MHz	8 x SAS-1.0	1945 MB/s	1.0	x4	858 MB/s
LSI2008		8 x SAS-2.0	3890 MB/s	2.0	x4	1716 MB/s
					x8	3433 MB/s
LSI2108	DDR2 / 800 MHz	8 x SAS-2.0	3890 MB/s	2.0	x4	1716 MB/s
					x8	3433 MB/s

In the majority of cases the maximum possible throughputs do not represent a bottleneck. First, the application scenarios with random access, in which high transaction rates but no high throughputs are achieved, prevail in practice.

The limit for PCIe 1.0 only arises with six to seven conventional hard disks for applications with purely sequential access. The controllers that support PCIe 2.0 and the PRIMERGY servers are so attuned to each other that no notable limitation to throughput arises in the controllers, even in the PCIe-x4 slot with maximum server configuration and with high-performance SAS-2.0-HDDs for purely sequential accesses.

Performance-relevant settings

Accurate parameter setting 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 supplied RAID-Manager software "ServerView RAID" for the current servers. All the cache settings for controllers and hard disks can usually be made at one go – 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). 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.

Read mode

The "Read mode" parameter can be used to influence cache behavior for read. A maximum of three options "No read ahead", "Read ahead" and "Adaptive" are available. No caching during read takes place in the case of "No read ahead". If a data block is requested, "Read ahead" causes further sequentially following data blocks to be already read from the hard disk to the cache of the controller just in case, and in the hope that they will be called for in subsequent requests. In the case of the "Adaptive" setting the controller itself attempts to determine whether a read-ahead makes sense or not.

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 and they are 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, 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 battery of the BBU is empty or if no BBU is installed, whereas the "Write-back" option automatically switches to "Write-through" as long as the controller cache is not battery-buffered.

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 in such a way that the hard disk is accessed beforehand. 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 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
	Read ahead / No read ahead	Adaptive		
Onboard controller ICH10R	✓			
Onboard controller Ibex Peak	✓			
Onboard controller Cougar Point	✓			
LSI1064				
LSI1068				
LSI1078	✓	✓	✓	✓
LSI2008				
LSI2108	✓	✓	✓	✓

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 and are set independent of the RAID level.

Controller alias	BBU?	Data Protection				Performance			
		Read mode	Write mode	Cache mode	Disk cache	Read mode	Write mode	Cache mode	Disk cache
Onboard controller ICH10R		Read ahead			off	Read ahead			on
Onboard controller Ibex Peak		Read ahead			off	Read ahead			on
Onboard controller Cougar Point		Read ahead			off	Read ahead			on
LSI1064					off				on
LSI1068					off				on
LSI1078		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
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

Characteristics of the onboard controllers

Some PRIMERGY models offer with an onboard controller – ICH10R, IbexPeak or CougarPoint – 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. In all the cases dealt with here it has four SATA ports (3 Gbit/s) to offer. 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. One example of this is a PRIMERGY TX200 S6 with 2 × Xeon E5506 (2.13 GHz), i.e. with a CPU from 2009, and an ICH10R chip. Even if you fully load the maximum possible four SATA-SSDs on this system with hard disk requests, no more than 10% of the CPU performance is used to support the onboard controller.

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 can be up to four 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.
Compatible	The SATA ports are emulated as PATA ports (PATA = parallel ATA, predecessor of the SATA standard). As a consequence of this emulation NCQ is not supported. This operating mode should only be used if the operating system used does not provide any SATA drivers for the "Southbridge" chip used or these should not be used for other reasons.
Native	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".

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 [I/O/s] Transaction rate in I/O operations per second
- Latency [ms] Average response time in ms

The throughput rate 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

<i>Data throughput [MB/s]</i>	$= \text{Transaction rate [I/O/s]} \times \text{Block size [MB]}$
<i>Transaction rate [I/O/s]</i>	$= \text{Data throughput [MB/s]} / \text{Block size [MB]}$

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 S6 PRIMERGY TX120 S3 PRIMERGY TX150 S7 PRIMERGY TX200 S6
Operating system	Windows Server 2008, Enterprise Edition Version: 6.0.6001 Service Pack 1 Build 6001
RAID Manager software	ServerView RAID Manager 5.0.2
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	lometer 2006.07.27
Measurement data	Measurement files of 32 GB with 1 – 8 hard disks; 64 GB with 9 – 16 hard disks
Onboard SATA Controller "ICH10R"	Intel 82801JR I/O Controller Hub, ICH10R (in PRIMERGY TX200 S6) Driver Name: megasr1.sys, Driver Version: 13.2.0614 BIOS: A 09.04151432R SATA RAID mode
Onboard SATA Controller "Ibex Peak"	Intel BD3420 PCH, Ibex Peak (in PRIMERGY TX150 S7) Driver Name: megasr1.sys, Driver Version: 13.2.0614 BIOS: A 09.07211059R SATA RAID mode
Onboard SATA Controller "Cougar Point"	Intel BD82C202 PCH, Cougar Point (in PRIMERGY TX120 S3) Driver Name: megasr1.sys, Driver Version: 14.04.0322 BIOS: A. 10.03031333R SATA RAID mode
Controller "RAID 0/1 SAS based on LSI MegaRAID 4Port" (LSI MegaRAID SAS 1064)	Driver Name: lsi_sas.sys, Driver Version: 1.32.00.04 Firmware Version: 1.30.00.00 BIOS Version: 2D.35
Controller "RAID 0/1 SAS based on LSI MegaRAID 8Port" (LSI MegaRAID SAS 1068)	Driver Name: lsi_sas.sys, Driver Version: 1.25.06.22 Firmware Version: 1.30.00.00 BIOS Version: 2D.35
Controller "RAID Ctrl SAS 6G 0/1" (LSI MegaRAID SAS 2008)	Driver Name: megasas2.sys, Driver Version: 4.23.0.64 Firmware Version: 20.7.1-0025 BIOS Version: 4.18.00
Controller "RAID 5/6 SAS based on LSI MegaRAID 256 MB" (LSI MegaRAID SAS 1078)	Driver Name: megasas2.sys, Driver Version: 4.23.0.64 Firmware package: 11.0.1-0028 Firmware Version: 1.40.152-0827 BIOS Version: 2.07.00 Controller Cache: 256 MB
Controller "RAID 5/6 SAS based on LSI MegaRAID 512 MB" (LSI MegaRAID SAS 1078)	Driver Name: megasas2.sys, Driver Version: 4.23.0.64 Firmware package: 11.0.1-0028 Firmware Version: 1.40.152-0827 BIOS Version: 2.07.00 Controller Cache: 512 MB
Controller "RAID Ctrl SAS 5/6 512MB (D2616)" (LSI MegaRAID SAS 2108)	Driver Name: megasas2.sys, Driver Version: 4.23.0.64 Firmware package: 12.4.0-0031 Firmware Version: 2.40.33-0894 BIOS Version: 2.07.00 Controller Cache: 512 MB
SATA-HDD	Seagate ST9500530NS, 2.5", 500 GB, 7200 rpm, 3 Gb/s
SAS-2.0-HDD	Seagate ST3146356SS, 2.5", 146 GB, 15000 rpm, 6 Gb/s
SATA-SSD	Intel SSDSA2SH064G1GC, 2.5", 64 GB, 3 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 and SAS-2.0 hard disk were chosen in each case for the classic hard disks (HDDs), and a 64 GB SATA-SSD represents the SATA-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]		
		Read	Write	8 kB block size	64 kB block size	
				Read share:	Read share:	
				67%	67%	50%
HDD SATA, 3 Gb/s, 2.5" 500 GB, 7200 rpm hot-pluggable category: Business-Critical (BC)	SATA-HDD	96 MB/s	93 MB/s	181 IO/s	180 IO/s	171 IO/s
HDD SAS, 6 Gb/s, 2.5" 146 GB, 15000 rpm, category: Enterprise (EP)	SAS-2.0-HDD	155 MB/s	155 MB/s	589 IO/s	470 IO/s	454 IO/s
SSD SATA, 3 Gb/s, 2.5" 64 GB, hot-pluggable, category: Enterprise (EP)	SATA-SSD	252 MB/s	187 MB/s	8083 IO/s	2023 IO/s	1580 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 three subsections "[RAID 0, 1 and 10 \(up to four hard disks\)](#)", "[RAID 0 and 10 \(more than four hard disks\)](#)" and "[RAID 5](#)", which can be read independently of each other.

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-HDD, SAS-2.0-HDD and SATA-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.

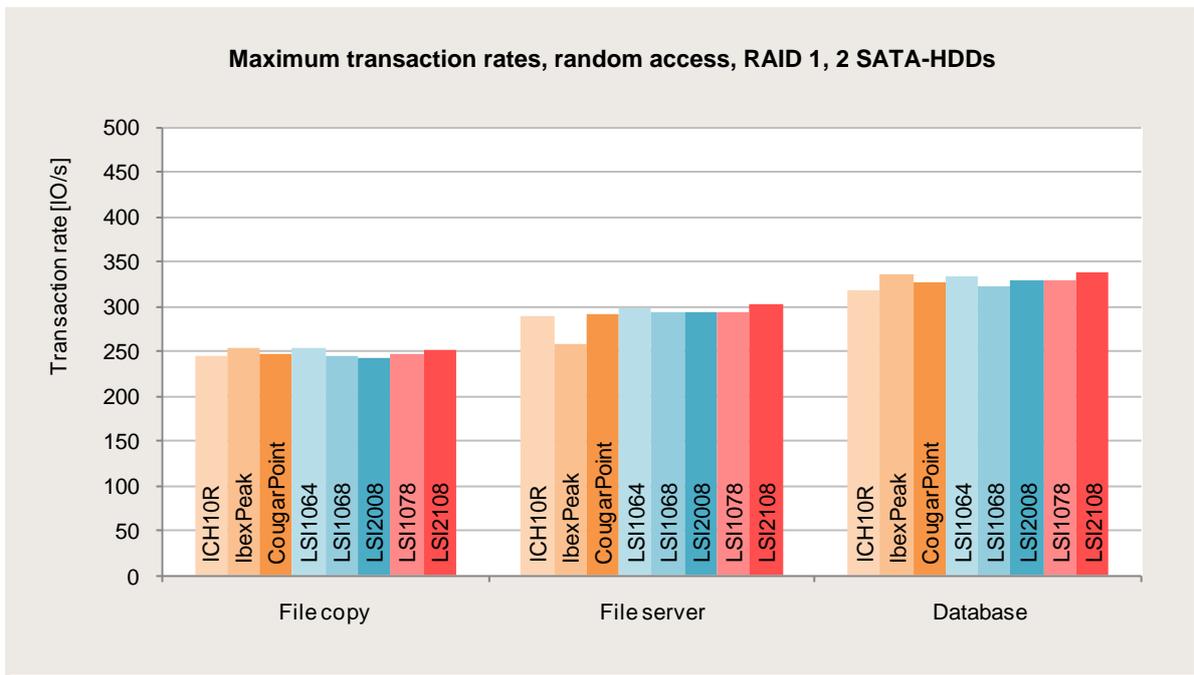
RAID 0, 1 and 10 (up to four hard disks)

It is worth looking at this range of hard disk numbers separately, because the onboard controllers can only be compared with the PCIe controllers here. This is done using the SATA-HDD, which was described more closely in the section "[Measurement environment](#)", because all the controllers to be reviewed support SATA. 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-HDDs

The diagram shows a controller comparison for two SATA-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 performance achieved only slightly depends on the controller here.



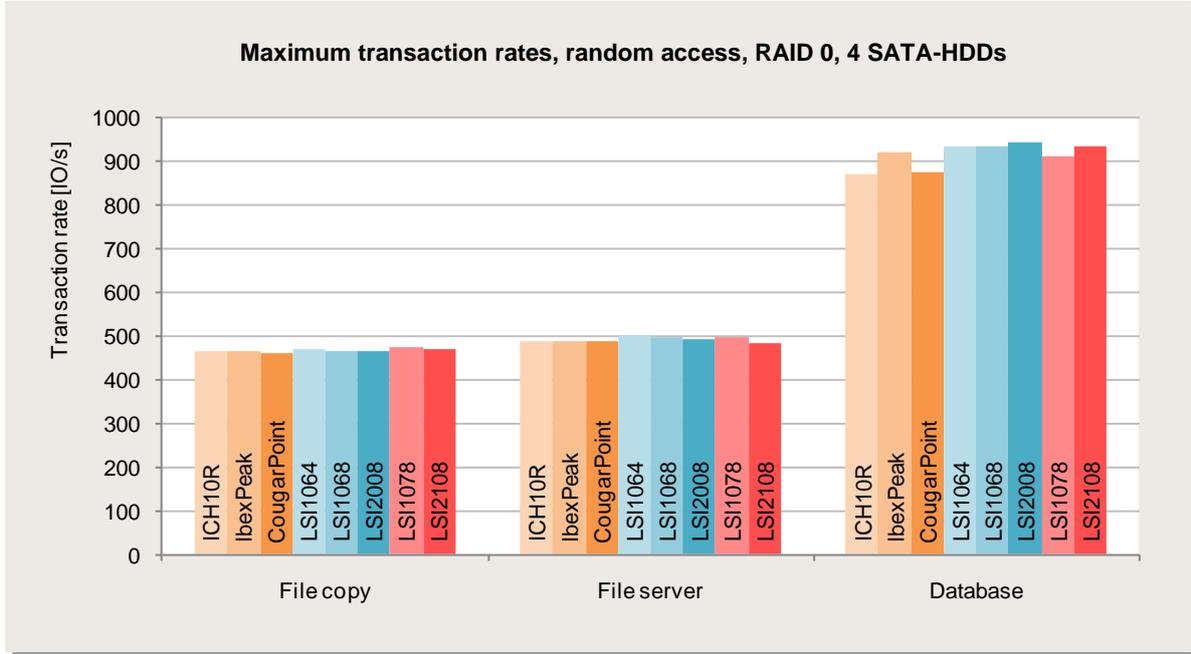
In the case of RAID 1 the controllers with a cache usually perform better if you disable the controller cache. Thus, the transaction rates of the LSI1078 and the LSI2108 controllers in this diagram are based on the following changes compared with the [standard settings](#) of the "Performance" mode of ServerView RAID:

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

These changes are generally recommended for RAID 1.

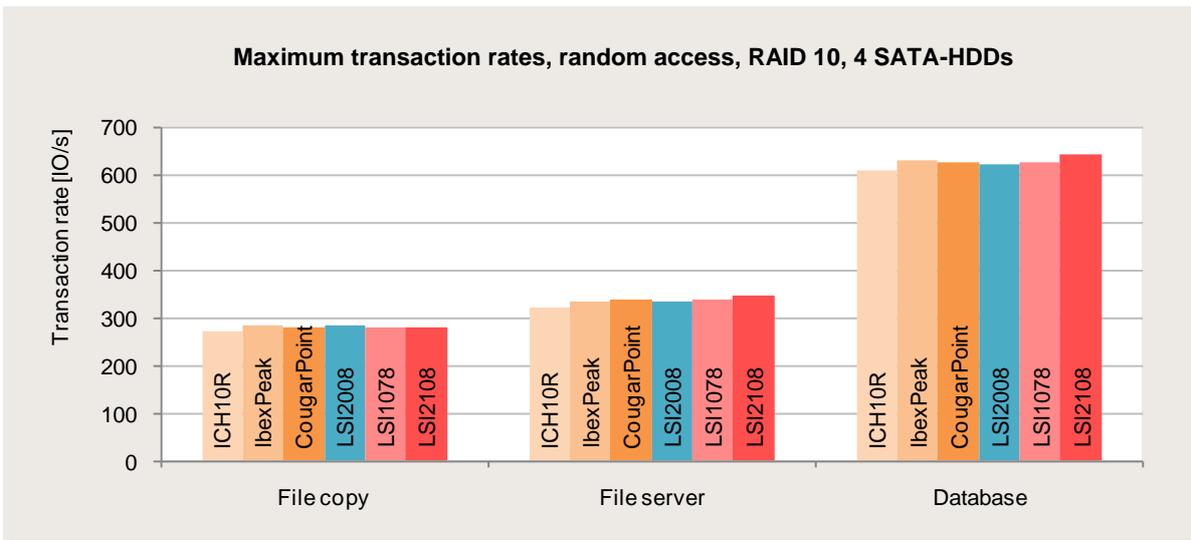
RAID 0 with four SATA-HDDs

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



RAID 10 with four SATA-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). All the controllers deliver approximately the same performance in these cases.

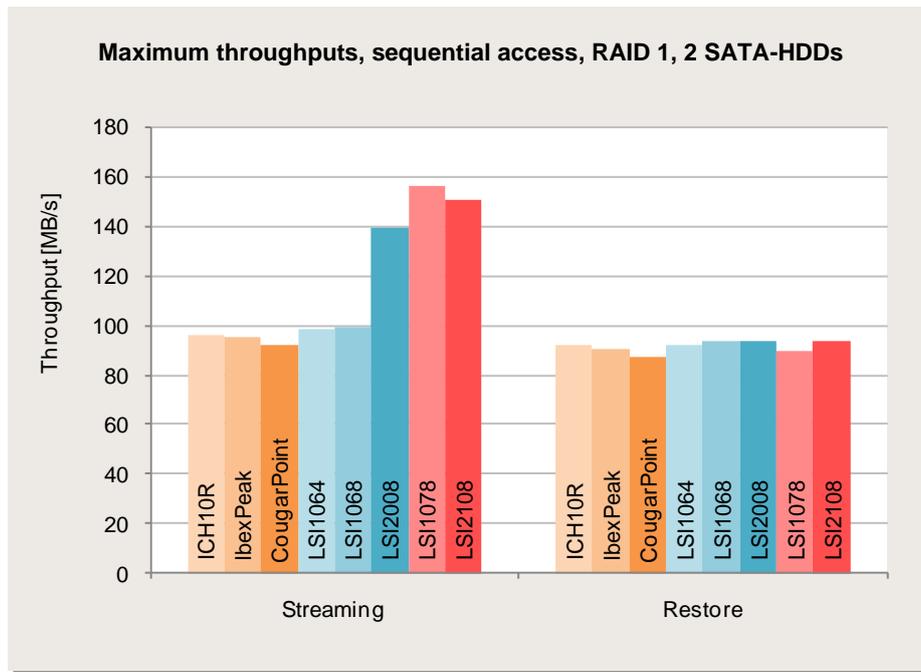


Sequential accesses

RAID 1 with two SATA-HDDs

The next diagram shows a controller comparison for two SATA-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 threshold values of a single hard disk.

For read the LSI2008, the LSI1078 and the LSI2108 controllers use both hard disks for high load intensities and consequently show a higher maximum throughput.



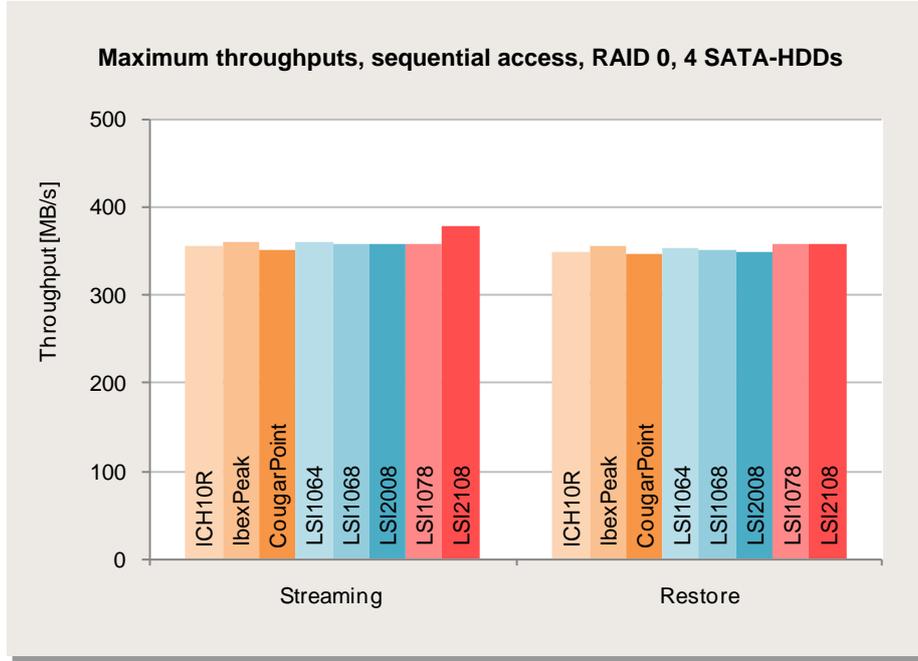
In the case of RAID 1 the controllers with a cache usually perform better if you disable the controller cache. Thus, the transaction rate of the LSI1078 and the LSI2108 controllers in this diagram are based on the following changes compared with the [standard settings](#) of the "Performance" mode of ServerView RAID:

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

These changes are generally recommended for RAID 1.

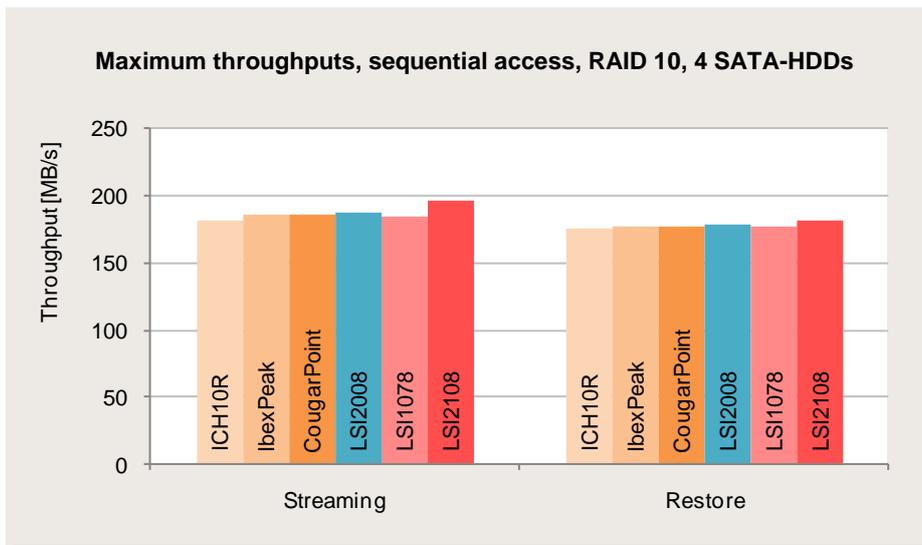
RAID 0 with four SATA-HDDs

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-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 approximately the same performance in these cases.



RAID 0 and 10 (more than four 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. Since the threshold values for higher performance requirements are of particular interest here, measurements usually with twelve high-performance SAS-2.0-HDDs or SATA-SSDs are used as a means of illustration. These hard disks are described in more detail in the section "[Measurement environment](#)".

The performance statements about the LSI1068 controller are based on measurements with eight instead of twelve hard disks, because eight is the maximum number of hard disks of the controller in connection with the currently available PRIMERGY models.

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 0, 1 and 10 \(up to four hard disks\)](#)".

Random accesses

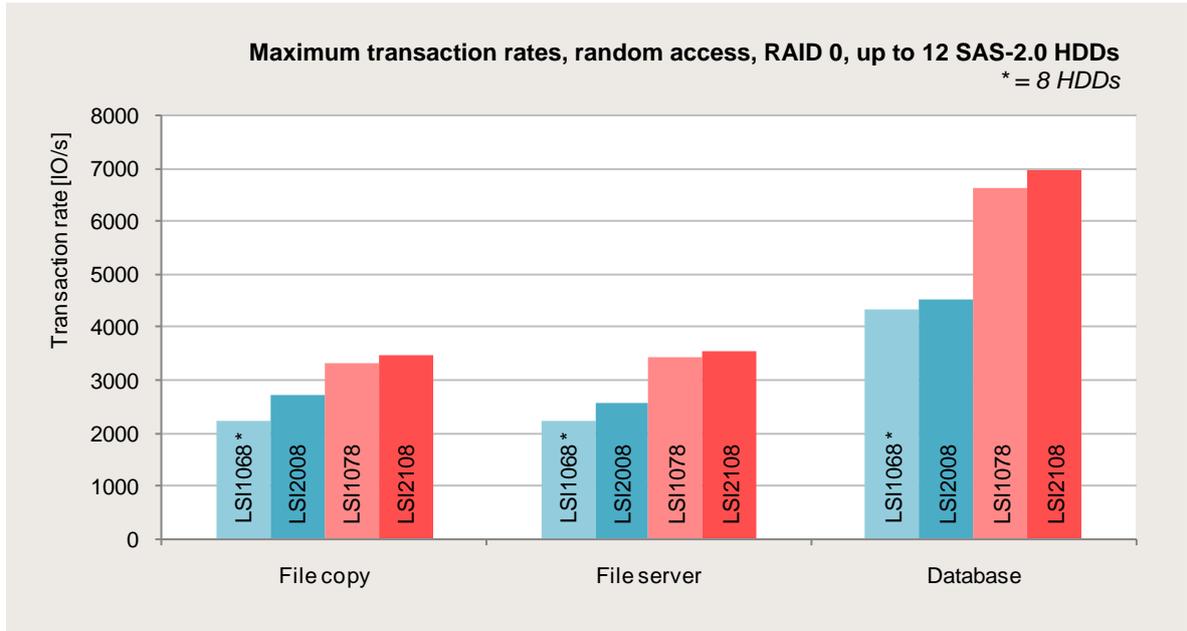
When considering random accesses for larger numbers of hard disks it makes sense to distinguish between HDDs and SSDs, because the threshold values for SSDs are of a quite different magnitude.

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 up to twelve SAS-2.0 HDDs, they can also be used to estimate the maximum transaction rates to be expected for other types and numbers 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 up to twelve SAS-2.0 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. As was already previously mentioned, the LSI1068 controller was operated with eight hard disks. The SAS-2.0 hard disks used have the maximum transaction rate of 589 IO/s for the load profile "Database" (random access, 67% read, 8 kB block size). Theoretically, an array consisting of twelve such hard disks should therefore handle a maximum of 589 IO/s × 12 = 7068 IO/s. The maximum measured transaction rate for the load profile "Database" is 6951 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.¹



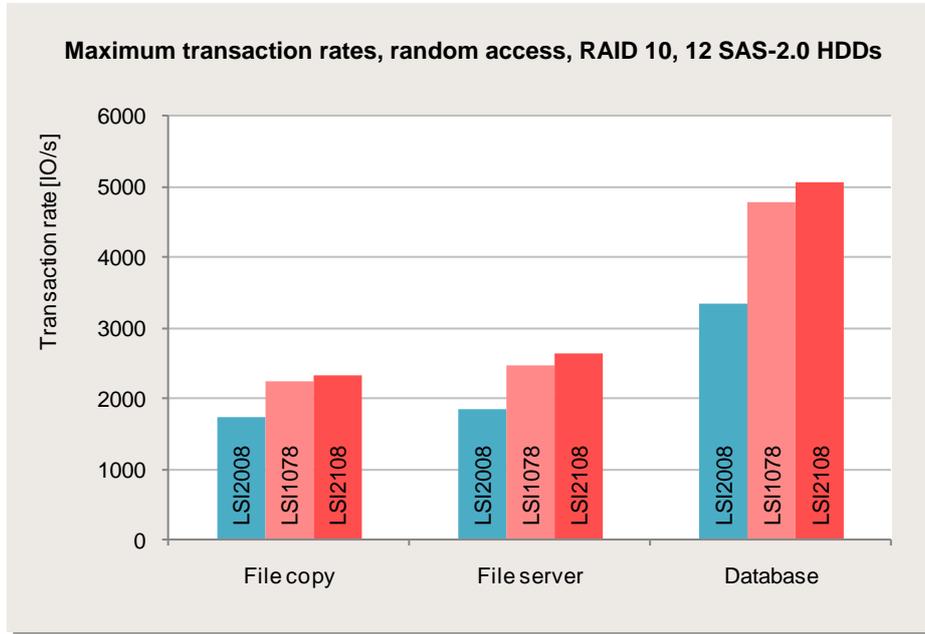
The two right-hand columns in each of the three groups of columns in this diagram represent the two controllers with a cache (LSI1078 and LSI2108). Both achieve approximately the maximum transaction rate of this RAID array for each of the three load profiles. The two other controllers (LSI1068 and LSI2008) each achieve no more than about 70%. A more precise analysis of the transaction rates would show that within the range of up to about 32 parallel accesses these two controllers for random load profiles have similar performance values as the three controllers with a cache. 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 twelve (and eight) hard disks. The difference is also relevant for RAID arrays with smaller numbers of hard disks if the latter can show an increase in performance above 32 parallel accesses. This prerequisite was not met in the measurements in the section "[RAID 0, 1 und 10 \(up to four hard disks\)](#)", because the smaller RAID array had already reached its maximum performance with about 16 parallel accesses there. Thus, there were no significant differences in performance there between the controllers without a cache (LSI1068 and LSI2008) and the controllers with a cache (LSI1078 and LSI2108).

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. The fundamentals of the RAID arrays are described in detail in the document "[RAID Performance](#)".

RAID 10 with twelve SAS-2.0 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.

SATA-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 ten times larger for a SATA-SSD. As a result of such a high transaction rate it is also possible for arrays consisting of several SATA-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.

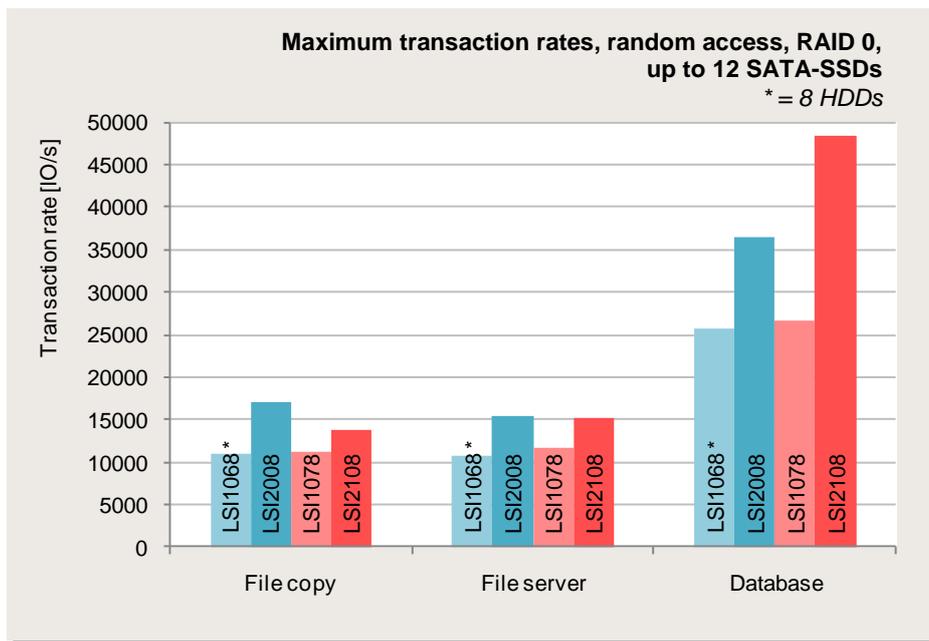
The performance of the controllers with a cache (LSI1078 and LSI2108) essentially depends on the correct choice of cache settings for these cases. Compared with the [standard settings](#) of the "Performance" mode of ServerView RAID, the following should be changed:

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

These changes are generally recommended for RAID levels 0 and 10 with random load profiles in connection with SSDs.

RAID 0 with up to twelve SATA-SSDs

The next diagram shows the transaction rates of the SATA-SSD array for RAID 0 for random load profiles that can be achieved with various controllers. As was already previously mentioned, the LSI1068 controller was operated with eight hard disks. 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 LSI2008 and the LSI2108 controllers are the ones with top performance here. The first one has an advantage with the "File copy" load profile, and the second one a more significant advantage with the "Database" load profile.

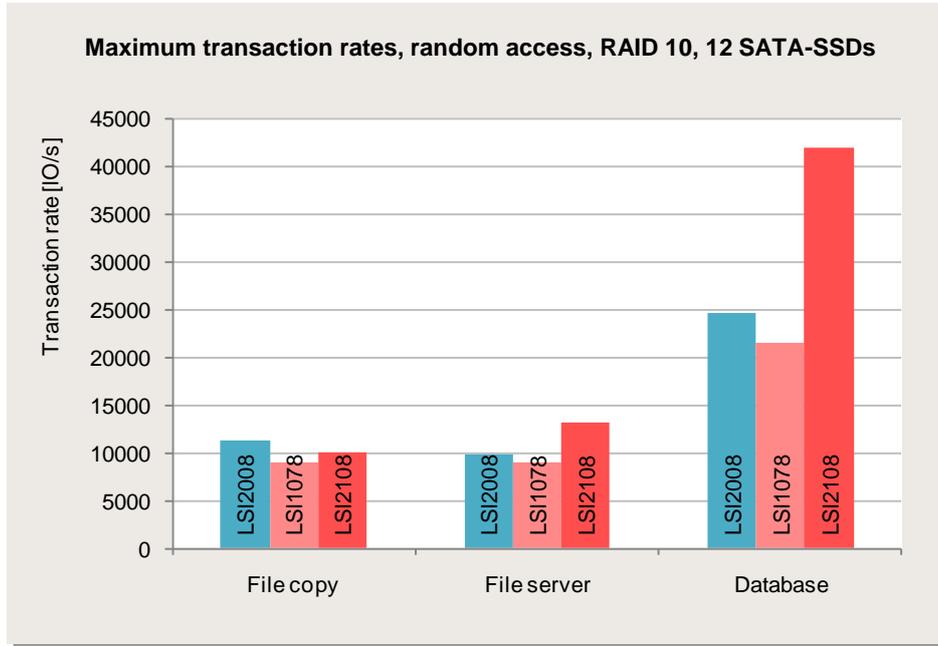


It is also interesting to understand the throughput values that are associated with these transaction rates. Despite the lower transaction rates, the two load profiles with a 64 kB block size have the higher throughputs. For example, the LSI2008 controller handles a throughput of about 1062 MB/s with the load profile "File copy".

If you express the maximum transaction rates in the form of numbers of SATA-SSDs, that means you need - depending on the random load profile - between six and nine SATA-SSDs for example to make full use of the possibilities offered by the LSI2108 controller in the case of RAID 0.

RAID 10 with twelve SATA-SSDs

The next diagram shows the transaction rates of the SATA-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 LSI2008 and the LSI2108 controllers are the ones with top performance here. The first one has a slight advantage with the "File copy" load profile, and the second one has a minor advantage with the "File server" load profile and a great advantage with the "Database" load profile.



Sequential accesses

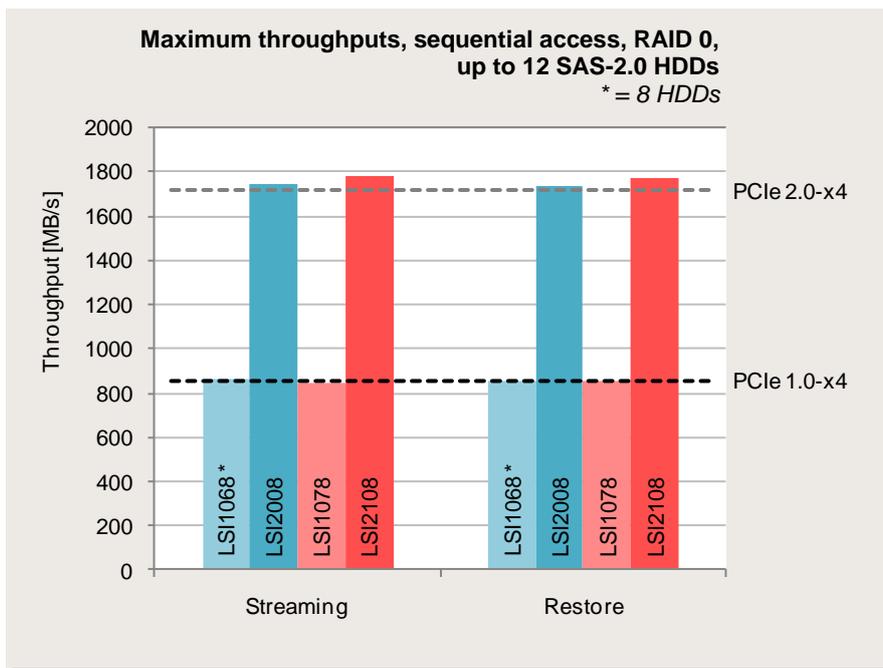
When considering sequential accesses for larger numbers of hard disks it is not necessary to distinguish between HDDs and SATA-SSDs, because the respective threshold values are still of a similar magnitude. General statements about the controllers are made below on the basis of measurements with up to twelve SAS-2.0 HDDs. The maximum throughputs to be expected for this can be calculated for other types and numbers of hard disks by means of suitable multiplication. If the throughput calculated in this way exceeds the threshold value of the controller, the controller threshold value becomes effective.

RAID 0 with up to twelve SAS-2.0 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. As was already previously mentioned, the LSI1068 controller was operated with eight hard disks. 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). In order to make a fair comparison of controllers with and without a cache the values for the controllers with a cache (LSI1078 and LSI2108) in this diagram are based on the following changes to 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"

As a result, a higher throughput is made possible for RAID 0 with sequential write. Through this nothing changes for sequential read. Therefore, these changed settings for RAID 0 arrays with sequential load profiles are generally advisable.

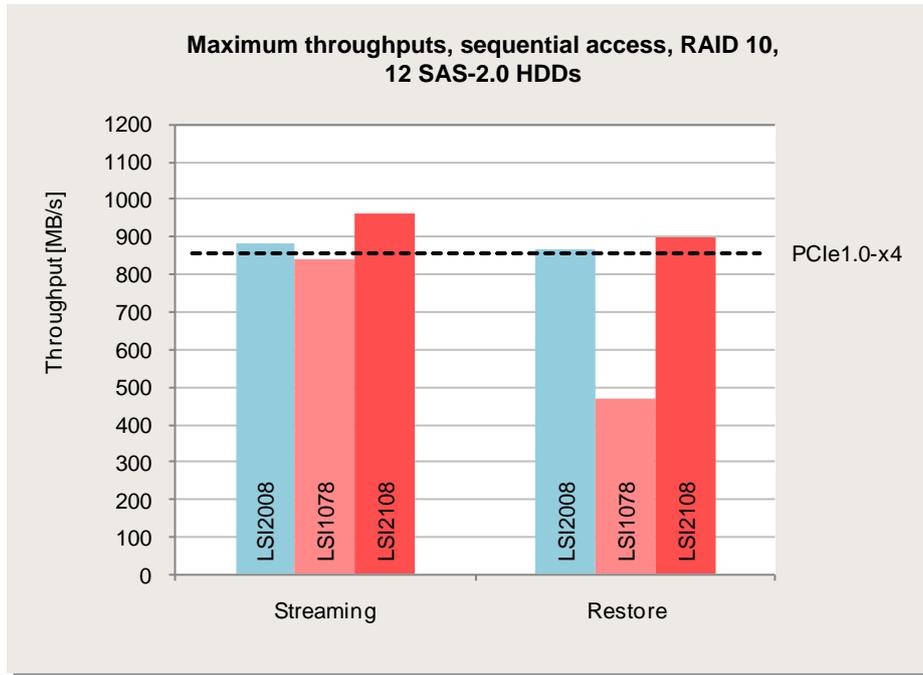


The internal RAID controllers are usually operated in a slot for PCIe 2.0-x4, thus enabling throughputs up to the dotted horizontal line in gray. You can see that the PCIe interface is attuned in such a way that even in this maximum configuration the possibilities offered by the HDDs can be practically used in full. If the two more recent controllers (LSI2008 and LSI208) were to be operated in a slot for PCIe 2.0-x8, they would only achieve a minimally higher throughput (see the length of the columns above the dotted line in gray). Thus, these controllers almost achieve the throughput that is possible from the HDDs (the maximum read/write throughput of such a SAS-2.0 HDD is 155 MB/s, i. e. the RAID array could manage a maximum of 155 MB/s x 12 = 1860 MB/s).

The controllers of the predecessor generation (LSI1068, LSI1078) achieve about 860 MB/s and thus make full use of the possibilities offered by their PCIe interface (PCIe 1.0, x4, dotted horizontal line in black).

RAID 10 with twelve SAS-2.0 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).



The LSI2008 and the LSI2108 controllers achieve or surpass the anticipated throughput of 930 MB/s (= 155 MB/s × 6) for sequential read and write.

The LSI1078 controller achieves 840 MB/s for sequential read; i.e. approximately the throughput that its PCIe interface allows it to achieve (dotted line "PCIe 1.0-x4" in the diagram). Throughput cannot be increased to more than 471 MB/s for sequential write with this controller and RAID 10, not even with a modified cache setting.

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").

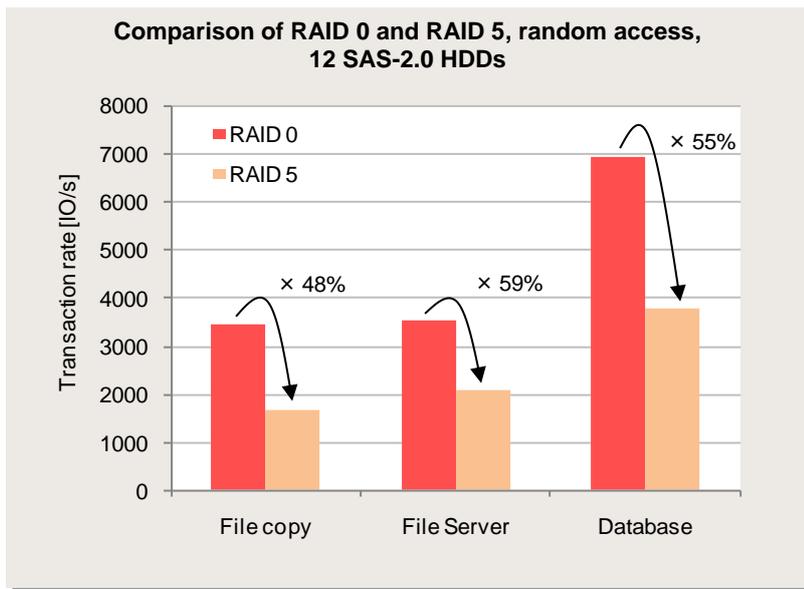
In contrast, a controller must manage considerably higher requirements with RAID levels such as RAID 5, mainly for write. This is due to the fact that such RAID levels require additional blocks with parity data, which first has to be calculated from the actual data.

The measurements of this subsection were made with the LSI1078 and the LSI2108 controllers, because only they support RAID 5 as standard. Since the threshold values for higher performance requirements are of particular interest, measurements with twelve high-performance SAS-2.0-HDDs or SATA-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-2.0 HDDs, which was first configured as RAID 0 and then as RAID 5, 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 LSI2108 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 LSI2108 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.

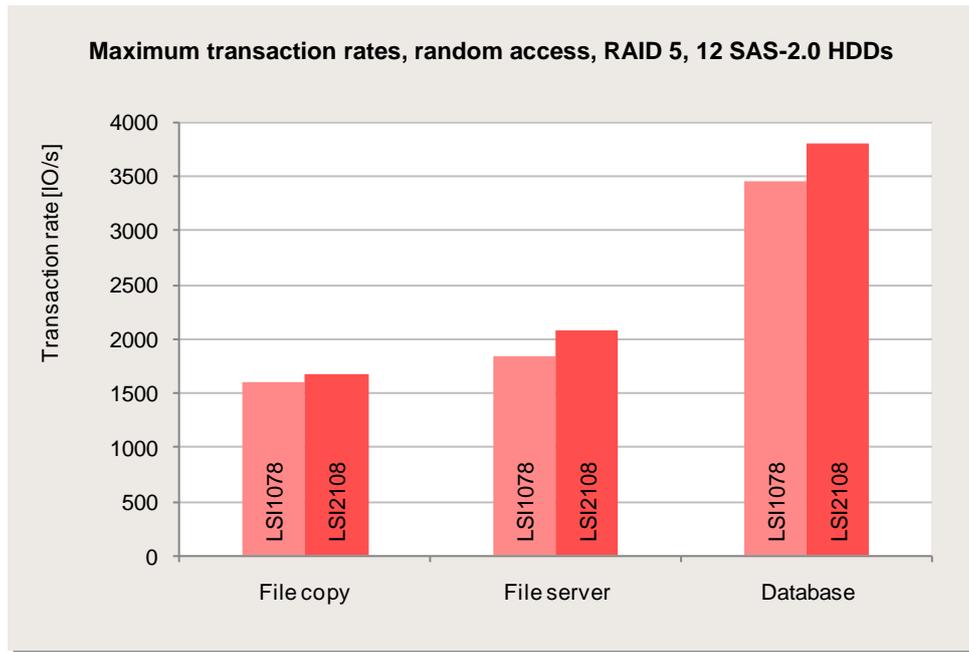
The basics of the individual RAID levels are described in detail in the document "[RAID Performance](#)".

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.

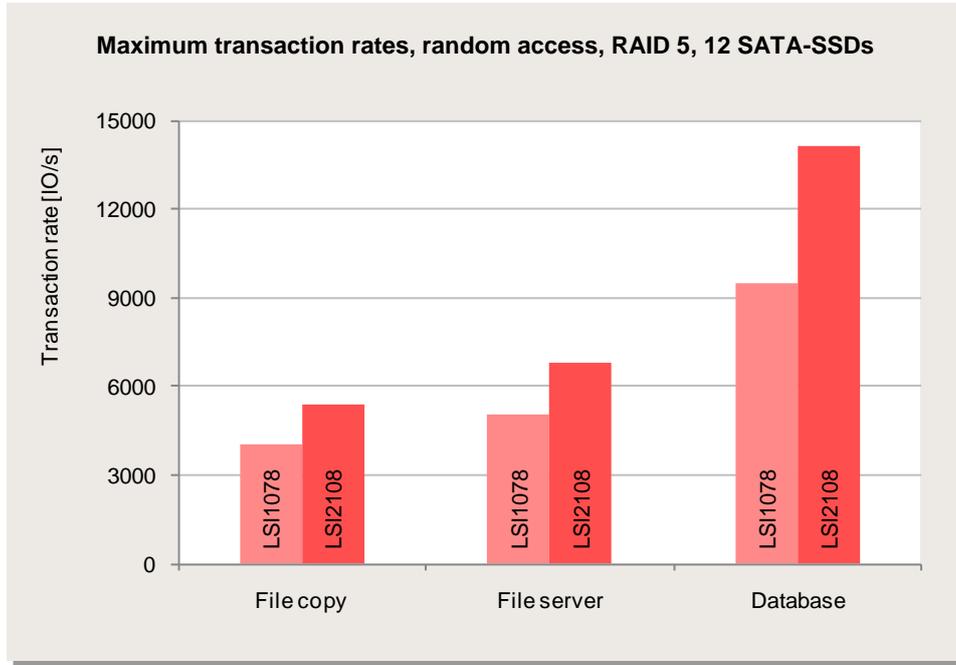
RAID 5 with twelve SAS-2.0 HDDs

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 transaction rates for the LSI2108 controller have already been used above in the comparison of RAID 0 with RAID 5. The LSI2108 controller, i.e. the newer controller with the larger cache, has a slight advantage.



SATA-SSDs

The transaction rates that can be achieved with SATA-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 SATA-SSDs. The following diagram puts together the maximum transaction rates for the various controllers using the example of an array consisting of twelve SATA-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).



If these maximum values are expressed in the form of numbers of SATA-SSDs, it means that with RAID 5 the LSI2108 controller – depending on the random load profile – enables between three and seven SATA-SSDs to be fully loaded without a bottleneck arising in the controller. In addition to the SATA-SSD-specific threshold 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](#)". The differences between the two controller generations would also be clear to see with the three SATA-SSDs, thus in the case under consideration the LSI2108 controller is to be recommended in order to achieve optimal performance.

Sequential accesses

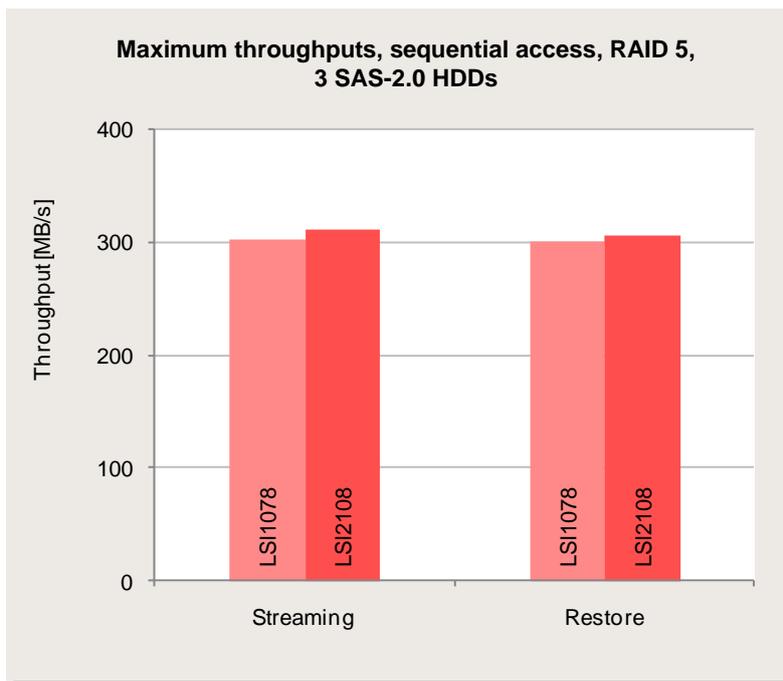
When considering sequential accesses with RAID 5 for larger numbers of hard disks it is also not necessary to distinguish between HDDs and SATA-SSDs, because the respective threshold values are still of a similar magnitude. General statements about the controllers are made below on the basis of measurements with twelve SAS-2.0 HDDs. The maximum throughputs to be expected can be calculated for other types and numbers of hard disks. If the throughput to be expected exceeds the threshold value of the controller, the controller threshold 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.

The next two diagrams show the controllers in sequential application scenarios with RAID 5 arrays of varying size that consist of SAS-2.0 HDDs.

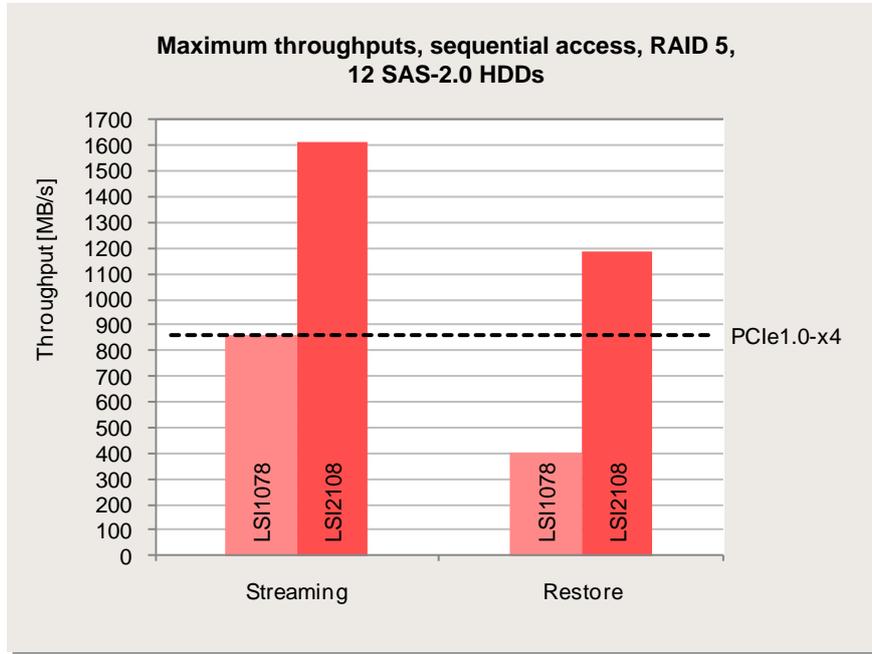
RAID 5 with three SAS-2.0 HDDs

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). In the case of three HDDs almost the expected two-fold maximum throughput of the HDD type used can be achieved in all cases (155 MB/s x 2 = 310 MB/s). There is still no substantial difference between the controllers with this small number of hard disks.



RAID 5 with twelve SAS-2.0 HDDs

The threshold value of the LSI1078 controller is again to be seen in the case of twelve HDDs. The LSI2108 controller achieves 1611 MB/s for read, i.e. almost the maximum possible throughput of 1705 MB/s in this case (= 155 MB/s x 11). The LSI1078 controller allows a maximum of 858 MB/s for read, and thus makes optimal use of the possibilities offered by the PCIe 1.0-x4 interface (dotted line "PCIe 1.0-x4" in the diagram). The LSI1078 controller achieves a maximum throughput of about 400 MB/s with sequential read for this RAID level, and the LSI2108 controller achieves about 1200 MB/s.



For numbers of hard disks (N) between three and twelve 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 threshold values of the controllers cannot be exceeded here.

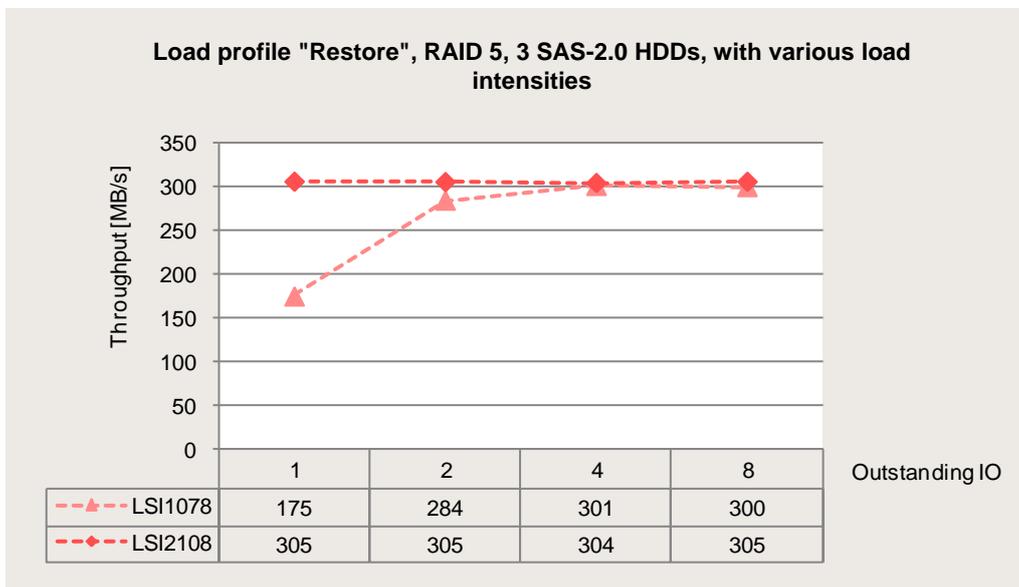
Lower load levels

The differences between the controllers can not only be seen in the maximum achievable values under high load, but also under a lower load. Thus, the following table compiles for the individual controllers the thresholds for the sequential throughputs of a single application, which works without parallelism for its accesses (referred to below with "1 outstanding IO"). These threshold 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 threshold values.

Controller	Threshold 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
LSI1078	290 MB/s	250 MB/s
LSI2108	670 MB/s	670 MB/s

The differences between the controllers become significant if the RAID array used is large enough to enable more than 300 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 5 array consisting of three SAS-2.0 HDDs for the load profile "Restore" (sequential access, 100% write, 64 kB block size). The comparison is made between the LSI1078 and the LSI2108 controllers with differing numbers of parallel accesses ("outstanding IO").



You can clearly see that in this case the LSI1078 controller does not achieve the same throughput for 1 and 2 outstanding IOs as the LSI2108 controller. The latter already achieves the maximum throughput of the RAID 5 array for 1 outstanding IO (in this case, twice the throughput of the SAS-2.0 hard disk used, i.e. 155 MB/s x 2 = 310 MB/s), whereas the LSI1078 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 LSI2108 controller compared with the LSI1078 controller.

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 SATA hard disks. The pro rata consumption of the server's processor performance is increasingly less important in newer servers.

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.

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. Manifold options to set the use of the cache make it possible to flexibly adapt the controller performance to suit the RAID level used. The pro rata consumption of the server's processor performance is increasingly less important in newer servers.

The requirements made of RAID controllers - when PRIMERGY servers are used with internal hard disks - have also grown with the advent of SAS-2.0 and PCIe 2.0 technologies as well as with the increased use of solid state disks (SSDs), which is why a controller of the SAS-2.0 generation (6G version) is advisable in this case.

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 with a RAID 5 array consisting of three SSDs and database-typical accesses, the use of the "RAID Ctrl SAS 6G 5/6 512 MB" (LSI2108) controller instead of a controller of the predecessor generation can cause the maximum transaction rate to be doubled.

Throughputs of 900 MB/s and more are not only possible with SSDs in applications with sequential read access, but also e.g. with six current SAS-2.0 hard disk drives. Without the new generation of PCIe 2.0-compatible controllers the limitation of PCIe 1.0 with an x4 connection would take effect here at about 850 MB/s.

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 600 MB/s is possible for read and write in this special application. For these special cases this means a doubling of the maximum throughput compared with controllers of the predecessor generation.

The RAID-Manager software "ServerView RAID" that is supplied for current 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 and UPSs are used as buffers in the case of power failures, maximum performance can be reconciled with data security.

Literature

PRIMERGY Systems

<http://ts.fujitsu.com/primergy>

PRIMERGY Performance

http://ts.fujitsu.com/products/standard_servers/primergy_bov.html

Basics of Disk I/O Performance

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

Single Disk Performance

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

RAID Performance

<http://docs.ts.fujitsu.com/dl.aspx?id=c55404e9-69c6-4d1e-a556-6a322b27da5b>

Information about Iometer

<http://www.iometer.org>

Contact

FUJITSU Technology Solutions

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

PRIMERGY Product Marketing

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

PRIMERGY Performance and Benchmarks

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

All rights reserved, including industrial property rights. Delivery subject to availability; right of technical modifications reserved. No liability or warranty assumed for completeness, validity and accuracy of the specified data and illustrations. Designations may be trademarks and/or copyrights of the respective manufacturer, the use of which by third parties for their own purposes may infringe the rights of such owners. Further details are available under http://ts.fujitsu.com/terms_of_use.html

2011-09-07 WW EN

Copyright © Fujitsu Technology Solutions GmbH 2011