White Paper
Fujitsu PRIMERGY Servers
Solid State Drives - FAQ

For some years now Solid State Drives (SSDs), based on flash memory, have existed as an alternative to conventional hard disks. Deciding to deploy them depends on the requirements made of storage media reliability, system reaction speed, system efficiency and last but not least of course, it also depends on the costs. This document is dedicated to frequently asked questions (FAQ) about the topic SSD. It deals with technical aspects of conventional SSDs, PCIe-SSDs and DOM (Disk On Module), paying particular attention to lifespan, but also provides information about the optimal use of SSDs in the PRIMERGY server family.

Version
1.1
2014-03-31

Contents

Document history

Version 1.0

Original version

Version 1.0a

- Minor corrections
- Regarding the question "In which cases should a RAID controller with cache be used?":
  The exception with regard to RAID 1 no longer applies and has therefore been removed.

Version 1.1

- Drive classes updated
- Endurance classes updated
- New SSD models added
- DOM added
- Performance values updated

General

As is commonly the case, this white paper uses decimal prefixes according to the SI standard for capacity specifications (1 kB = 10^3 bytes, 1 MB = 10^6 bytes, …, 1 EB = 10^{18} bytes).

As regards the specification of block sizes and throughputs – and this is also usual – binary prefixes are used. To distinguish decimal prefixes, binary prefixes are specified according to the IEC standard (1 KiB = 2^{10} bytes, 1 MiB = 2^{20} bytes, …, 1 TiB = 10^{40} bytes).

Please note that the difference between figures with a binary and a decimal prefix is greater, the higher the unit:

1 KiB = 1.024 kB = 1 kB + 2.4%

1.TiB ≈ 1.1 TB = 1 TB + 10%

FAQs about internal SSD topics

What are the fundamental differences between SSDs and HDDs?

Contrary to conventional hard disks, SSDs do not have any moving parts, but are based on flash memory. The external profile and interface of conventional SSDs correspond to those of conventional hard disks. PCIe-SSDs on the other hand are, as the name suggests, directly operated via the PCIe interface. In comparison with conventional hard disks, SSDs stand out for very low access times, mechanical robustness, noiselessness and very low energy consumption, but still have a lower capacity today and a higher price.

SSDs save information in flash memory cells, of which there are various types:

- Single-level cells (SLC) save one bit per cell. About 100,000 write operations per cell are possible with SLCs in 50-nm technology. SSDs with these cells usually only have low capacity, but very high write performance and write endurance.
- Multi-level cells (MLC) save two bits per cell. The number of possible write operations varies depending on the structure size. About 3,000 write operations per cell are possible with MLCs in 25-nm technology. Compared with SLC-based SSDs, MLC-based SSDs are lower priced, have a higher capacity, but in return also lower write performance and write endurance.
- Triple-level cells (TLC) save three bits per cell. About 1,000 write operations per cell are possible with TLCs in 25-nm technology. Thus, they offer even higher capacity with lower write performance and write endurance.

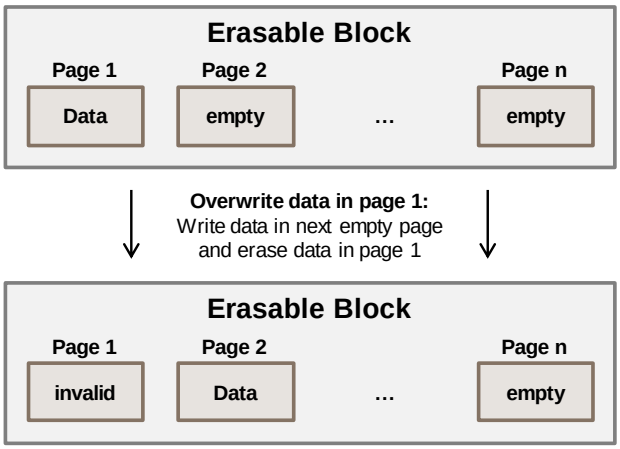
Particularly the low access times, compared with HDDs, make SSDs especially desirable for use in servers. SSDs are indeed not writable to an unlimited extent. Internal intelligent controllers in SSDs nevertheless ensure even distribution over all flash memory cells for write operations and thus also uniform wearing. Furthermore, a range of SSD models has an internal storage capacity that goes beyond their nominal capacity. This is why certain SSD models are not only suited for use in servers, but are genuine competition for HDDs.

How is flash memory accessed?

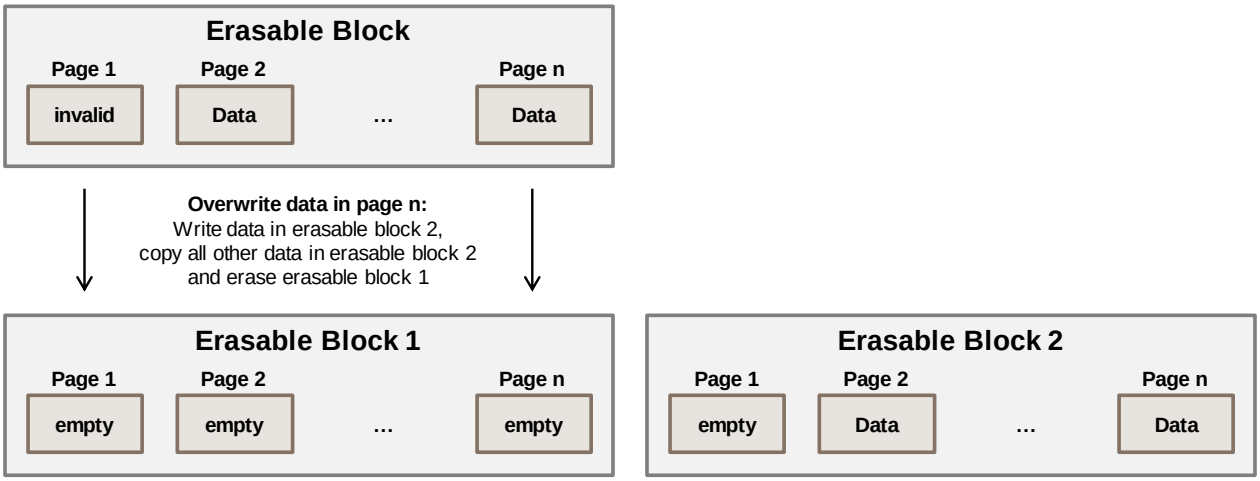
On an SSD the smallest unit in the case of read and write operations is a **page** with a size of 4 KiB. Regardless of the required data volume at least 4 KiB are always read. And write jobs, independent of their block size, also cause write operations in 4 KiB units.

It is only possible to write on flash memory cells if they are empty. Re-use therefore requires the prior erasure of the flash memory cell. This operation is known as the **program/erase cycle**. The number of possible program/erase cycles per flash memory cell is limited. Thus, if a file had a set allocation to a physical memory frequent file updating would result in the flash cells, on which they are saved, wearing out faster than other cells. This could very quickly result in some memory areas no longer being updatable, while others are still completely unused. To counteract this there is a procedure known as **wear-leveling**, with which the controller of an SSD distributes write blocks to the memory cells in such a way that the wearing of the SSDs is uniform. Memory cells of the internal reserve are also used in this case (see Overprovisioning). The uniform nature of the wearing, as caused by wear-leveling, is bought in return for an increase in the write effort (see Write Amplification). This cannot be influenced externally.

In an SSD wear-leveling presumes the mapping of logical block addresses (LBAs) on physical addresses. If data is to be updated, the original physical address is marked as invalid and a new physical address is used for the updating. Pages, whose addresses are invalid, must be erased by the SSD before re-use. The details of this are as follows: In an SSD pages are consolidated to form a so-called **erasable block**. If the content of a page is to be changed, the latter is marked as invalid and the new contents are written on the one that is least used of the free pages of an erasable block.



If there is no longer a free page available in the erasable block, its valid pages are copied to a free erasable block and the old erasable block is erased.



In this way, previously invalid pages are made available again for write jobs. This operation is called **garbage collection**. This ensures uniform wearing of the flash memory cells of pages that are updated (**dynamic wear-leveling**). To also enable the inclusion of flash memory cells of erasable blocks that are not updated in this wearing process, their data is also automatically moved at specific intervals (**static wear-leveling**).

What does write amplification mean?

The actual write effort within an SSD can be distinctly higher than the size of a write job might suggest. It can amount to a multiple of this, particularly in the case of random access with small block sizes. This effect is referred to as write amplification. Theoretically, it can be calculated as follows:

$$\text{Write Amplification} = \frac{\text{Data volume that is physically written to an SSD}}{\text{Data volume of the logical write job}}$$

In practice, however, the level of the write amplification depends on so many factors that an estimation can at best be made.

What influence does write amplification have on performance and write endurance?

High write amplification has a negative influence on the performance of write jobs and on the write endurance of SSDs. Write amplification is quite high in the case of random write for small data blocks in particular, the result of which is a high write effort and fast wearing of the flash cells. This is counteracted by so-called overprovisioning and intelligent data management on the part of the internal SSD controller. In particular the wear-leveling procedures, which ensure equal distribution of the data and thus uniform wearing of the flash cells, should be mentioned here. This helps an SSD to a lifespan that is generally comparable with the conventional hard disks of the Enterprise class (see Drive classes).

The specifications of SSD manufacturers on write endurance and their KPIs (TBW, PBW, DWPD) take typical write amplification for a standardized write load with random access into account. Thus, a different write load to this also entails different write endurance as a consequence. In most different scenarios this is higher.

May SSDs be defragmented?

SSDs can be defragmented, but should be avoided at all costs.

Defragmentations are intended for HDDs and also only make sense there. Files on HDDs are initially continuous and not in fragments. Fragmentations may occur there in the course of time, which go hand in hand with a gradual reduction in performance. Recovering the original performance level can by all means justify the effort of defragmentation.

In the case of an SSD it does not matter for performance whether the data is continuous or fragmented. This is why defragmentation provides no benefits. The time invested is unnecessary. And the increase in write amplification associated with defragmentation is detrimental to the write endurance of the SSD.

Therefore, if Windows Server 2012 recognizes that an SSD would be affected by the defragmentation of a drive, no defragmentation is carried out. If PCIe-SSDs are in use, this recognition is a given, but not with conventional SSDs via a RAID controller. Therefore, a system administrator should in the latter case ensure that no defragmentation is performed, whether it be manually or as a scheduled task.

What is the function of the command TRIM?

The command TRIM is intended to also erase data, which was erased at a logical level, physically on the data media.

Background:

The erasing of files and also high-level formatting, such as the Quick format under Windows, mean that previously saved data can no longer be referenced. However, the data itself is not necessarily erased at data medium level in this case. In the case of HDDs this complex erase operation would also be completely superfluous. It is a different matter with an SSD: The latter now sees itself as "fuller" than is actually the case. The affected memory areas are thus not available as empty blocks for dynamic wear-leveling. Unlike empty blocks, they are also still subject to static wear-leveling, which results in additional write amplification. If the memory areas are used by the server again, the indispensable erasing of the flash memory cells can, contrary to what is otherwise the case, only take place immediately before a new write job. And writing on such flash memory cells is associated with higher response times.

This is why newer operating system versions use the command TRIM for physical data erasure when erasing files and for high-level formatting if they recognize an SSD as a part of a logical drive.

When does the command TRIM work?

Conventional SSDs:

As RAID controllers do not support the TRIM command, this function is not available in the case of conventional SSDs.

PCIe-SSDs:

Support of the TRIM command is usually provided under Windows and Linux.

Windows:

- RAID volumes (mirrored, spanned or striped) apart from RAID 5.
- Simple volumes.
- Each combination of the above RAID levels across several data media, as long as at least one PCIe-SSD is affected.
- PCIe-SSD in a RAID array.
- Several partitions on a PCIe-SSD.
- NTFS and FAT32 file system.
- Volumes with mount points.
- Compressed volumes.
- Different cluster sizes, package sizes and sector sizes.
- So-called "extended" and "shrunk" volumes.

Linux:

- For all operating systems that support "Discard".
- If "Discard" is enabled.
- Not all Linux commands, with which drives can be created, issue "Discards" (see documentation of the respective Linux distribution).

What effects are to be expected if TRIM does not function?

A drop in performance should be expected with higher write amplification and when accessing previously erased memory areas.

What does overprovisioning mean?

SSDs typically have an internal reserve of flash memory cells. This is used as a buffer during write operations and as a replacement if memory cells can no longer be written. This means it has a positive influence on the write performance and the write endurance of SSDs. The reserve is normally specified as a percentage of the SSD capacity.

$$\text{Overprovision as a percentage of the SSD capacity} = \frac{\text{Physical capacity - Logical capacity}}{\text{Logical capacity}} \times 100$$

The size of the overprovisioning area differs from SSD model to SSD model. It is normally not specified by any SSD manufacturer, but is considered in their information about DWPD and PBW (or TBW).

Several SSD manufacturers offer SSDs completely without an overprovisioning area for SSDs in the consumer sector in particular. They assume that the customer will set up individual overprovisioning for himself, by only using for example 70% of the SSD capacity.

FAQs about drive classes

What drive classes does Fujitsu have to offer?

For PRIMERGY servers Fujitsu offers a large number of storage media (HDDs and SSDs) with different performance and availability features for all conceivable scenarios - from the small department server that is only needed during working hours up to the high-availability database server. To facilitate making a selection from this wide range Fujitsu has introduced five drive classes:

Drive class	Description/Suitability	Type	Capacity [1])
Economic (ECO)	Characterized by a low unit price. However, their performance and reliability levels mean that they are only suitable for entry-level applications. They should be used in non-critical areas with low I/O traffic and moderate speed requirements as higher workloads can impair their reliability. ECO drives have rotating speeds of 5400/7200 rpm and are equipped with a SATA interface.	2.5" SATA 5.4k HDD	1 TB
		3.5" SATA 7.2k HDD	250 GB – 500 GB
Business-Critical (BC)	Provide maximum capacity with the lowest costs per GB. They are designed for good performance and corresponding reliability. Depending on the server implementation, BC drives can be equipped with a SAS or a SATA interface and offer a rotating speed of 7200 rpm. If top I/O throughput rates are required, then Enterprise HDDs or SSDs should be used as an alternative.	2.5" SATA 7.2k HDD	250 GB – 1 TB
		2.5" SAS 7.2k HDD	500 GB – 1 TB
		3.5" SATA 7.2k HDD	500 GB – 4 TB
		3.5" SAS 7.2k HDD	1 TB – 4 TB
Enterprise (EP) [2])	Offer maximum performance and reliability. They are designed for high throughputs and low latencies.	2.5" SAS 10k HDD	300 GB – 1.2 TB
		2.5" SAS 15k HDD	146 GB – 300 GB
		2.5" SSD	100 GB – 1.6 TB
		3.5" SAS 15k HDD	300 GB – 600 GB
		PCIe-SSD	365 GB – 1.2 TB
		DOM	64 GB

[1]) Date: 2014-04-01

[2]) formerly: Enterprise (EP), Enterprise Mainstream (EP MAIN) and Enterprise Performance (EP PERF)

Which SSDs are offered for PRIMERGY servers?

Name	Interface	Capacity [GB]	Write Endurance [PBW]	DWPD (rounded down)	MiB/s until PBW is reached	Unrecoverable read errors (Bit error rate)	MTBF [million hrs.]	Form factor
PCIe-SSD 1.2TB MLC	PCIe Gen2 x4	1200	17	7	103	$1/10^{20}$	1.5	LP [1])
PCIe-SSD 785GB MLC	PCIe Gen2 x4	785	11	7	67	$1/10^{20}$	2	LP [1])
PCIe-SSD 640GB MLC [2])	PCIe Gen1 x4	640	10	8	60	$1/10^{20}$	1.24	LP [1])
PCIe-SSD 365GB MLC	PCIe Gen2 x4	365	4	6	24	$1/10^{20}$	2	LP [1])
PCIe-SSD 320GB MLC [2])	PCIe Gen1 x4	320	4	6	24	$1/10^{20}$	1.24	LP [1])
SSD SAS 12G 1.6TB Main 2.5" H-P EP	SAS 12G	1600	29.2	10	176	$1/10^{17}$	2	2.5"
SSD SAS 12G 800GB Main 2.5" H-P EP	SAS 12G	800	14.6	10	88	$1/10^{17}$	2	2.5"
SSD SAS 12G 400GB Main 2.5" H-P EP	SAS 12G	400	7.3	10	44	$1/10^{17}$	2	2.5"
SSD SAS 12G 200GB Main 2.5" H-P EP	SAS 12G	200	3.65	10	22	$1/10^{17}$	2	2.5"
SSD SAS 6G 400GB SLC HOT P 2.5" EP PERF [2])	SAS 6G	400	35	47	212	$1/10^{16}$	2	2.5"
SSD SAS 6G 400GB MLC HOT PL 2.5" EP PERF [2])	SAS 6G	400	7.5	10	45	$1/10^{16}$	2	2.5"
SSD SAS 6G 200GB SLC HOT P 2.5" EP PERF [2])	SAS 6G	200	18	49	109	$1/10^{16}$	2	2.5"
SSD SAS 6G 200GB MLC HOT PL 2.5" EP PERF [2])	SAS 6G	200	3.75	10	23	$1/10^{16}$	2	2.5"
SSD SAS 6G 100GB SLC HOT P 2.5" EP PERF [2])	SAS 6G	100	9	49	54	$1/10^{16}$	2	2.5"
SSD SAS 6G 100GB MLC HOT PL 2.5" EP PERF [2])	SAS 6G	100	1.875	10	11	$1/10^{16}$	2	2.5"
SSD SATA 6G 800GB Main 2.5" H-P EP	SATA 6G	800	14.6	10	88	$1/10^{17}$	2	2.5"
SSD SATA 6G 400GB Main 2.5" H-P EP	SATA 6G	400	7.3	10	44	$1/10^{17}$	2	2.5"
SSD SATA 6G 400GB MLC HOT P 2.5“ EP MAIN [2])	SATA 6G	400	7.5	10	45	$1/10^{16}$	2	2.5“
SSD SATA 6G 200GB Main 2.5" H-P EP	SATA 6G	200	3.65	10	22	$1/10^{17}$	2	2.5"
SSD SATA 6G 200GB MLC HOT P 2.5" EP MAIN [2])	SATA 6G	200	3.75	10	23	$1/10^{16}$	2	2.5"
SSD SATA 6G 100GB Main 2.5" H-P EP	SATA 6G	100	1.825	10	11	$1/10^{17}$	2	2.5"
SSD SATA 6G 100GB MLC HOT P 2.5" EP MAIN [2])	SATA 6G	100	1.875	10	11	$1/10^{16}$	2	2.5"
SSD SATA 3G 64GB SLC HOT P 2.5“ EP MAIN [2])	SATA 3G	64	2	17	12	$1/10^{15}$	2	2.5“
SSD SATA 3G 32GB SLC HOT P 2.5“ EP MAIN [2])	SATA 3G	32	1	17	6	$1/10^{15}$	2	2.5“
DOM SATA 3G 64GB Main N H-P	SATA 3G	64	1.6425	14	9	n/a	3	DOM [3])

[1]) LP: Low profile

[2]) EOL (end of life)

[3]) DOM: Disk on module

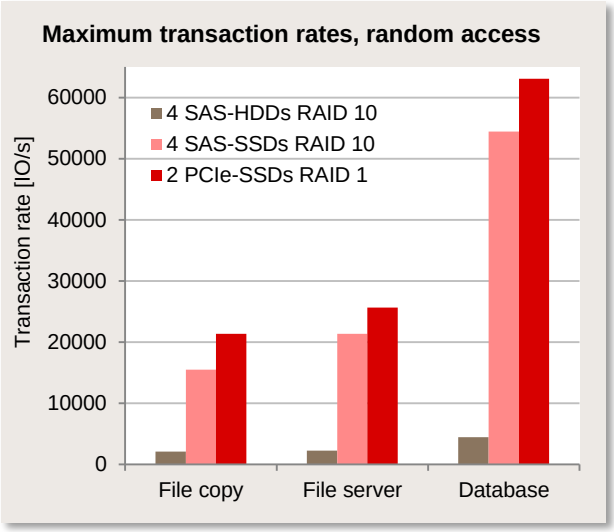
What is a DOM?

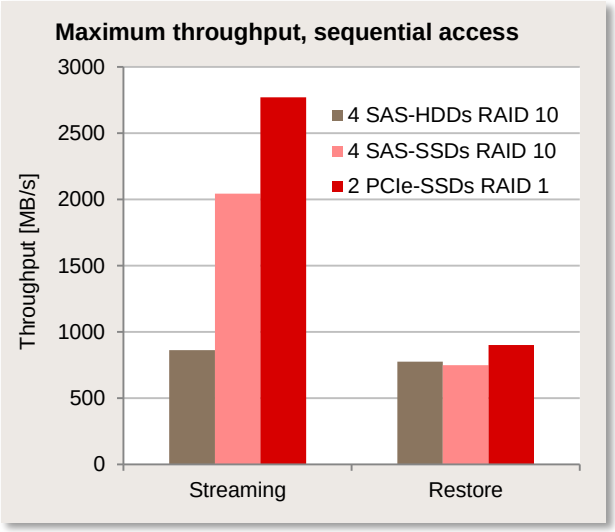
DOM stands for "Disk on module". It is an extremely space and energy-saving flash memory that is particularly used as a boot drive in servers. The memory technology is equivalent to that of SSDs. For a range of PRIMERGY servers Fujitsu offers a DOM with a SATA 3G interface, which can be inserted directly into the SATA port of the system board.

How do PCIe-SSDs differ from conventional SSDs (SATA/SAS)?

Conventional SSDs are operated via host bus adapters, usually RAID controllers, with a SATA or SAS interface. The interface of the RAID controller to the chipset of the systemboard is typically PCIe or, in the case of the integrated onboard controllers, an internal bus interface of the systemboard. PCIe-SSDs on the other hand are operated directly and only via the PCIe interface.

If an individual PCIe-SSD is compared with a single conventional SSD, then with regard to performance the former is vastly superior to the latter. PCIe-SSDs can therefore be a worthwhile alternative to a RAID configuration with conventional SSDs. For example, two PCIe-SSDs in a RAID 1 can surpass the performance of four SAS-SSDs in a RAID 10. More information is available in the papers "RAID Controller Performance" and "Performance Report PCIe-SSDs ioDrive®2".





Why are there such large price differences between the various SSDs?

Similar to HDDs, capacity, interface and performance play a special role in the pricing for SSDs. Another important and SSD-specific criterion is their write endurance. In this case, SSDs can be roughly divided into three categories, which meet completely different requirements:

Endurance Class	DWPD	Description/Suitability
Value Endurance (Read-intensive)	< approx. 5 usually < 0.3	SSDs of the lower price category, predominantly with MLC flash memory. Over a period of 5 years these usually manage an average write load of below 3 MiB/s. They are suited to load situations that are characterized by a low write intensity. Examples: System drive, streaming services
Mainstream Endurance	approx. 5 – approx. 15 usually approx. 10	SSDs of the medium price category, PCIe-SSDs with their high prices due to exceptional performance and DOM, predominantly with MLC flash memory. Over a period of 5 years these usually manage an average write load in the double-digit MiB/s range. They are suited to load situations that are characterized by a moderate write intensity. Examples (SSDs): File servers, web servers Examples (PCIe-SSDs): Virtual servers, databases, mail servers Examples (DOM): System drive
High Endurance (Write-intensive)	> approx. 15 often approx. 50	SSDs of the upper price category, predominantly with SLC flash memory. Over a period of 5 years these usually manage an average write load in the two to three-digit MiB/s range. They are suited to load situations that are characterized by a high write intensity. Examples: Virtualization servers, databases, mail servers

The commonly used name "Enterprise-SSD" provides no information as to which Endurance class an SSD can be assigned. Many manufacturers use the term for SSDs of the Mainstream and High Endurance class, some use it for SSDs of the Value Endurance class, and some use it for SSDs in general.

FAQs about fail-safety and write endurance

Are SSDs less secure than HDDs?

The objection that SSDs are less secure than HDDs is frequently used as an argument against the use of SSDs in server environments. On the whole, this can neither be confirmed nor denied. It is necessary for the individual security aspects to be viewed separately from each other:

- Fail-safety (Unrecoverable Read Errors, MTBF, AFR):
  With the exception of three PCIe-SSD models there are no differences between SSDs and HDDs.
- Write endurance:
  This is an SSD-specific aspect; the SSD models offered for PRIMERGY servers cover the entire range of requirements.
- Archiving
  SSDs are not a medium for long-term archiving; short-term archiving is possible.

The questions handled in this section explain this topic in more detail.

What are Unrecoverable Read Errors?

Data media failures are often due to so-called unrecoverable read errors. The frequency with which this phenomenon occurs is an important indicator as regards the quality of HDDs and SSDs.

Unrecoverable read errors are caused by a fault on a storage medium. Such fatal data losses are counteracted by redundantly storing data in RAID arrays. However, unrecoverable read errors can also play a significant role during a RAID array rebuild. Rebuild times are especially on the increase because of the high capacity per data medium. Since an unrecoverable read error would have fatal effects during the rebuild of a RAID 5 array, RAID 6 or RAID 10 solutions are increasingly being used. Whereas RAID 6 solutions are only slightly more expensive, but also less high-performance than RAID 5 solutions, RAID 10 solutions offer higher performance and lower rebuild times, but also at a price that is notably higher.

"Patrol Read" is a RAID controller function that has the goal of tracking down defective blocks in a RAID array before an application accesses them for the first time. These blocks can then be marked as faulty and withdrawn from the data storage pool. This also occurs with low-level formatting. Consequently, the probability of an unrecoverable read error occurring can be considerably reduced.

What do the KPIs MTBF and AFR mean?

In order to estimate the approximate lifespan of a storage medium, either HDD or SDD, their MTBF value is commonly used. **MTBF** stands for **M**ean **T**ime **B**etween **F**ailures and specifies in hours the average operating time between two failures. As this is simply a statistic, there is no guarantee that the MTBF value is achieved in a particular situation.

Now and again the indicator **AFR** (**A**nnualized **F**ailure **R**ate) is also used. This is derived from the MTBF value

$$AFR = 100 \times \frac{24 \times 365}{MTBF}$$

and provides a percentage of the devices - in this case HDDs or SSDs - which will probably fail within a year.

Conversely

$$1 - AFR$$

is an estimate for the percentage of devices that will run error-free throughout the entire year.

Example: For MTBF = 2 million hours the following is valid: AFR = 0.44% and 1 – AFR = 99.56%.

How failsafe are SSDs in comparison to HDDs?

Not only the capacity, performance and price play a role in the decision for or against the procurement of SSDs, fail-safety also plays an equally important role. The failure rates for Enterprise HDDs and SSDs are more or less the same. This is why the same security criteria must be applied when deciding in favor of a specific RAID level for the configuration of RAID systems.

Name	MTBF [million hrs.]
PCIe-SSD 1.2TB MLC	1.5
PCIe-SSD 785GB MLC	2
PCIe-SSD 640GB MLC *)	1.24
PCIe-SSD 365GB MLC	2
PCIe-SSD 320GB MLC *)	1.24
SSD SAS 12G 1.6TB Main 2.5" H-P EP	2
SSD SAS 12G 800GB Main 2.5" H-P EP	2
SSD SAS 12G 400GB Main 2.5" H-P EP	2
SSD SAS 12G 200GB Main 2.5" H-P EP	2
SSD SAS 6G 400GB SLC HOT P 2.5" EP PERF *)	2
SSD SAS 6G 400GB MLC HOT PL 2.5" EP PERF *)	2
SSD SAS 6G 200GB SLC HOT P 2.5" EP PERF *)	2
SSD SAS 6G 200GB MLC HOT PL 2.5" EP PERF *)	2
SSD SAS 6G 100GB SLC HOT P 2.5" EP PERF *)	2
SSD SAS 6G 100GB MLC HOT PL 2.5" EP PERF *)	2
SSD SATA 6G 800GB Main 2.5" H-P EP	2
SSD SATA 6G 400GB Main 2.5" H-P EP	2
SSD SATA 6G 400GB MLC HOT P 2.5“ EP MAIN *)	2
SSD SATA 6G 200GB Main 2.5" H-P EP	2
SSD SATA 6G 200GB MLC HOT P 2.5" EP MAIN *)	2
SSD SATA 6G 100GB Main 2.5" H-P EP	2
SSD SATA 6G 100GB MLC HOT P 2.5" EP MAIN *)	2
SSD SATA 3G 64GB SLC HOT P 2.5“ EP MAIN *)	2
SSD SATA 3G 32GB SLC HOT P 2.5“ EP MAIN *)	2
DOM SATA 3G 64GB Main N H-P	3

*) EOL (end of life)

RAID solutions with data redundancy as well as hot-spare and hot-plug should be regarded as technical reactions, which have been available for a long time in order to come to terms with the problem of data media failures and thus ensure high availability.

Another indicator, known as the **B**it **E**rror **R**ate (BER), is used to determine the statistical frequency of unrecoverable read errors. This is an error quotient, which specifies how many bit errors can be expected in relation to a specific storage capacity.

The following table indicates the bit error rate for the current data media for PRIMERGY servers:

Drive class	Data medium type	Bit Error Rate
Economic (ECO)	HDD	1 LBA / 10^{15} Bit [1])
Business Critical (BC)	HDD	1 LBA / 10^{16} Bit
Enterprise (EP)	HDD	
	SSD	1 LBA / 10^{16} Bit
	PCIe-SSD	1 LBA / 10^{20} Bit
	DOM	n/a

[1]) LBA = **L**ogical **B**lock **A**ddress

10^{15} bits = 125 TB, 10^{16} bits = 1.25 PB, 10^{20} bits = 12.5 EB

As regards the probability of unrecoverable read errors, the quality of SSDs according to the table corresponds to that of the Business Critical HDDs and Enterprise HDDs. This is why the same security criteria must be applied when deciding in favor of a specific RAID level for the configuration of RAID systems.

PCIe-SSDs offer a quality here that is several orders of magnitude higher. However, this is not reflected in generally better fail-safety.

What does write endurance mean?

As flash memory cells are wearing parts, an SSD can only tolerate a limited number of write jobs. In contrast to HDDs, not only is there therefore a general, average lifespan, but also one related to write jobs, which is called write endurance.

SSD manufacturers normally specify the write endurance of an SSD in **PBW** (**P**eta**b**ytes **W**ritten) – and in the past also in **TBW** (**T**era**b**ytes **W**ritten), but today only for SSDs with a low capacity.

PBW specifies the maximum number of petabytes that can be written to an SSD:

$$\text{PBW} = \frac{\text{Capacity[GB]} \times (1 + \text{Overprovisioning area[in \% of capacity]}) \times \text{Flash endurance (number of program/erase-cycles)}}{\text{Write amplification factor} \times 10^6}$$

The higher the value, the better. This already takes into account the SSD-typical write amplification for a standardized write load with random access. In other words, this practically concerns petabytes from a user point of view. You should bear in mind that write amplification is lower with sequential access patterns, which means that write endurance is higher in these cases than specified by PBW. Conversely, access patterns that permit higher write amplification and therefore as a consequence lower write endurance are also possible. PBW should therefore be viewed as a benchmark.

DWPD (**D**rive **W**rites **P**er **D**ay) is an indicator which is derived from the PBW value. Some SSD manufacturers specify the write endurance of an SSD in DWPD instead of PBW. When calculating this indicator, the capacity of an SSD and a previously defined writability lifespan are also taken into account:

$$\text{DWPD} = \frac{\text{PBW} \times 10^6}{\text{Capacity in GB} \times \text{Defined lifespan in years} \times 365}$$

The indicator specifies which daily write load an SSD can on average tolerate over its previously defined lifespan. A lifespan of five years, i.e. 1825 days, is generally used. The higher the DWPD value, the better. Example for an SSD with 400 GB capacity and 7.5 PBW:

$$\frac{7.5 \times 10^6}{400 \times 5 \times 365} \approx 10 \text{ DWPD}$$

The SSD tolerates a daily write load, which corresponds to ten times its own capacity, nonstop for the entire lifespan. It should be noted here that the SSD capacity is used as a unit of measure. This means that of two SSDs with 10 DWPD the one with the greater capacity can also manage the higher write load.

SSD capacity of 400 GB and 10 DWPD: 10 DWPD × 400 GB ≈ 3.6 TiB/day ≈ 44 MiB/s

SSD capacity of 200 GB and 10 DWPD: 10 DWPD × 200 GB ≈ 1.8 TiB/day ≈ 22 MiB/s

The first one not only has double the capacity, but also tolerates twice as much write load. It can be sustained for twice as long at the same write load.

It is important to keep an eye on the length of the defined lifespan. Not all competitors use this for every SSD model with 5 years, they possibly use less. A shorter lifespan drives up the DWPD value: the SSD then tolerates a higher write load, but only for a shorter period of time. If we consider an SSD with 400 GB capacity and 7.5 PBW with a lifespan of only 3 years, instead of the above five years, the result is then:

$$\frac{7.5 \times 10^6}{400 \times 3 \times 365} \approx 17 \text{ DWPD}$$

$$17 \text{ DWPD} \times 400 \text{ GB} \approx 6.2 \text{ TiB/day} \approx 75 \text{ MiB/s}$$

When comparing different SSDs the PBW indicator is ultimately of more significance. A comparison of the write endurance of different SSDs on the basis of their DWPD values is only possible when their capacities are taken into consideration and if they have the same lifespan. If this is not considered, an SSD with 10 DWPD and the same write load can by all means last longer than an SSD with 15 DWPD.

How does the DWPD value behave in relation to the actual daily write load?

The DWPD value represents an upper limit for the average daily write load, which an SSD tolerates over a specific period. If the actual daily write load is lower, the write endurance of the SSD increases accordingly.

Example of an actual write load of 500 GiB/day:

If an "SSD SAS 6G 400GB MLC HOT PL 2.5" EP PERF" is used with 10 DWPD, the write endurance of the SSD is:

$$\frac{\text{DWPD} \times \text{Capacity in GB} \times \text{Defined lifespan}}{\text{Write load in GiB per day}}$$

$$\frac{10 \times 400 \text{ GB} \times 5 \text{ years}}{500 \text{ GiB}} \approx 37 \text{ years}$$

In this case, – and probably in most cases in general – write endurance should not mean a real restriction in practice.

I would like to replace the SSDs I am currently using with new ones, which have at least the same write endurance. How do I find out which ones are suitable?

All the SSDs whose PBW value is at least as high as that of your current SSDs, including those with a lower DWPD value, are suitable. See the SSD table.

Does the choice of RAID configuration influence the write endurance of SSDs?

In principle, yes.

Firstly, the write load on an individual data medium depends on the type of RAID array into which the SSD is integrated:

RAID level	Write load per SSD
Single SSD / JBOD	Write load of SSD and JBOD
RAID 0	Write load of RAID array / Number of drives
RAID 1	Write load of RAID array
RAID 1E	Write load of RAID array × 2 / Number of drives
RAID 5	Write load of RAID array × 2 / Number of drives
RAID 6	Write load of RAID array × 3 / Number of drives
RAID 10	Write load of RAID array × 2 / Number of drives
RAID 50	Write load of RAID array × 2 / Number of drives
RAID 60	Write load of RAID array × 3 / Number of drives

By designing a RAID array with a higher number of SSDs the write load per SSD can be reduced, enabling a lower write endurance – and thus under certain circumstances a more economical SSD type – to suffice.

Example:

A RAID 5 with seven "*HD SAS 6G 146GB 15K HOT PL 2.5" EP"* is replaced by a RAID 5 with SSDs. The average write load for the logical drive will probably not exceed 6 TiB/day in the next 5 years. The total capacity should at least correspond to the current configuration. No more than 5 drive bays are available.

Solution 1: RAID 5 with three SSDs of 400 GB capacity
Write load per SSD: 6 TiB/day × 2 / 3 = 4 TiB/day ~ required 11 DWPD
→ "*SSD SAS 6G 400GB SLC HOT P 2.5" EP PERF"* with 47 DWPD

Solution 2: RAID 5 with four SSDs of 400 GB capacity
Write load per SSD: 6 TiB/day × 2 / 4 = 3 TiB/day ~ required 9 DWPD
→ "*SSD SAS 6G 400GB MLC HOT P 2.5" PL EP"* with 10 DWPD

Solution 2 is clearly the more economical, because mainstream SSDs can be used instead of high endurance SSDs. A third solution would be to use three SSDs with 10 DWPD and a capacity of at least 440 GB.

What do you do when write endurance ends?

The SSDs and PCIe-SSDs sold for PRIMERGY servers increasingly reduce the write operation speed when 95% of the PBW value is reached. In practice, the PBW value is as a result hardly reached, whereby the SSD remains operable in principle. The associated increase in response times for write jobs is an indicator for the system administrator to initiate the replacement of the SSDs. This can be checked by monitoring the relevant operating-system-specific performance counters. The respective physical data media can be monitored under Windows operating systems by logging the indicator "Bytes Written/s", and with Linux operating systems by monitoring with the use of the iostat command.

Management software is available for monitoring and reporting the write endurance for PCIe-SSDs.

How long can data be retained without accessing the SSD?

As long as the energy supply of an SSD is ensured, its data can in theory be kept for an unlimited period of time due to an internal SSD procedure called "Static wear-leveling", but in practice for as long as the SSD does not fail (see MTBF and AFR).

How long can data be retained if the SSD is disconnected?

A special feature of SSDs is their restricted data retention capability when switched off. If an SSD is removed from a server and, for example, put in a cabinet as backup, the information stored will remain available for ten years as best-case. Factors such as flash technology (SLC/MLC), the previous intensity of use (PBW) or the ambient temperature have the effect of shortening the storage period. The minimum storage period is 6 months for SSDs with SLC flash storage and 3 months with SSDs with MLC flash storage.

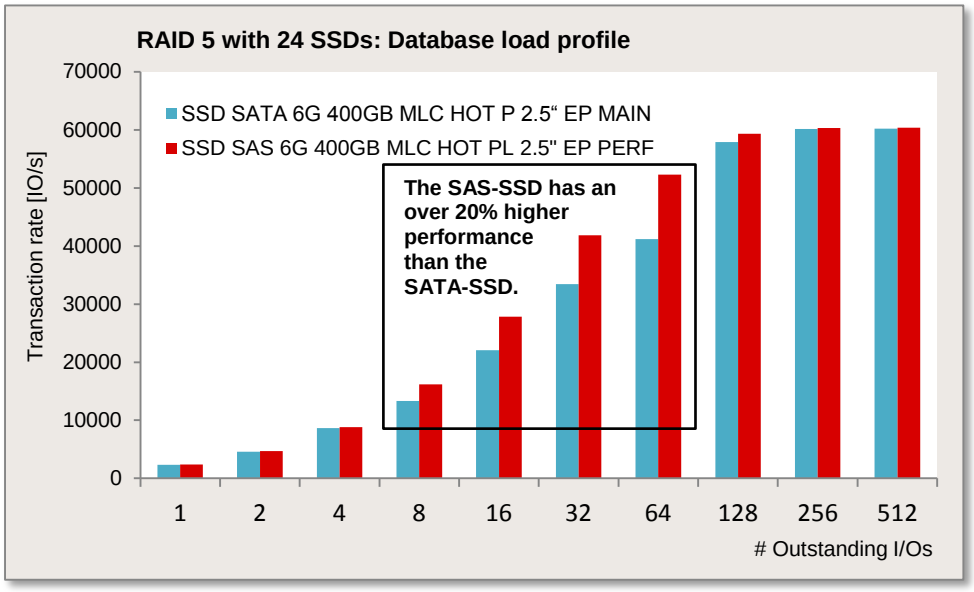
FAQs about interfaces

In which cases should SAS-SSDs be used instead of SATA-SSDs?

Some SSD models merely differ with regard to their external interface (SATA / SAS). If you only consider the top throughput limits of both interfaces, no difference can be found. There is also no difference with regard to the maximum number of SSDs that can be integrated into a PRIMERGY server.

Nevertheless, the SAS interface offers a range of advantages over the SATA interface:

- The enabling of redundant architectures through dual porting:
  A SAS-SSD can be connected to two RAID controllers.
  Example: The PRIMERGY CX420 Cluster-in-a-box solution
- Considerably higher performance of the error-correction algorithm
- Better performance:
  In general, SAS-SSDs and SATA-SSDs hardly reveal any performance differences with a very low load or very high load. However, in the mid-load range and depending on the load profile SSDs with a SAS interface can reach throughputs that are more than 20% above those of SSDs with a SATA interface. This is illustrated here by the example of a configuration with 24 SSDs in a RAID 5 with a database load profile:



Database load profile (transaction processing): random, 67% read, 33% write, 8 kB block size

FAQs about RAID

Is RAID needed at all when using SSDs?

Whether RAID configurations can be omitted or not when using SSDs depends on the reasons that would make a RAID indispensable when HDDs are used:

- Security:
  Fail-safety is just as important with SSDs as it is with HDDs. If a RAID configuration with data redundancy is required in the case of HDDs, then this is also the case with SSDs.
- Capacity:
  A single data medium usually does not suffice to enable the implementation of a logical drive with a high minimum capacity. In this case, it is not possible to do without a RAID configuration – particularly because the capacity range with SSDs is lower than with HDDs.
- Performance:
  If in the case of HDDs only performance aspects speak for the use of a RAID, it is then under certain circumstances possible to do without for SSDs. This should only be considered when replacing HDDs that are configured as a JBOD or a RAID 0.

What should be observed with regard to a RAID configuration when converting from HDDs to SSDs?

If the disk subsystem of a server works with HDDs, it is in the event of a conversion to an SSD-based disk subsystem first necessary to define the SSD configuration that is to replace the existing configuration. A certain number of storage media is required to manage the stipulated write load. This depends on their capacity and performance. A configuration with SSDs will therefore certainly look different than a configuration with HDDs. This should not only apply for the number of storage media, but can possibly also affect the RAID level used or the use of SSDs as a cache for HDD RAID arrays. In this connection, we refer you to the document RAID Controller Performance. Furthermore, the lifespan of SSDs can – other than in the case of HDDs – be extended enormously by increasing the number of SSDs in a RAID array.

Which RAID level should be recommended when using SSDs or PCIe-SSDs?

The decision in favor of a specific RAID level is driven by security aspects. In principle, the same security criteria should be applied for SSDs as for HDDs. MTBF times and failure rates do not fundamentally differ between SSDs and HDDs. The same applies for the rebuild times of a RAID array after the failure of a data medium: In RAID arrays SSDs demonstrate the greatest performance advantages compared to HDDs for random load profiles with small block sizes. These advantages decrease as the block size increases and are only low, if they still exist, with sequential accesses. Depending on the RAID level, number of data media and controllers used, SSDs or Enterprise HDDs provide advantages here. This is why the same RAID levels, if possible, should be used for HDDs and SSDs. However, as SSDs currently do not have such high storage capacities as some HDD types, you could possibly be forced to use a larger number of data media when replacing a HDD RAID array with SSDs. This can for example mean that a RAID 5 or RAID 10 with SSDs should now be used instead of a RAID 1 with HDDs.

HW-RAID vs. SW-RAID

Due to the high performance of SSDs a higher processor load should be taken into account when using a software RAID than with HDDs. Therefore, the alternative use of RAID controllers relieves the load on server processors when using SSDs much greater than when using HDDs. However, it should be noted that RAID controllers do not support TRIM.

Which controller should be recommended when using SSDs?

This depends on the load profile and the requirements for throughput, I/O rate and access times. In addition to the requirements for capacity and data redundancy, these also have a major influence on the minimum number of SSDs required and the choice of controller. The white paper RAID Controller Performance discusses this subject in detail.

In which cases should a RAID controller with cache be used?

In all RAID configurations a controller cache has a positive effect in the lower and middle load range for random accesses with a write share. This also applies for sequential read and write accesses in RAID configurations with 4 or more SSDs.

Is a hot spare required when using SSDs?

In the case of conventional SSDs, the answer is as a matter of principle yes. Indeed, an SSD does not contain any moving parts. Nevertheless, an SSD can still fail. The failure rates (see MTBF) correspond to those of Enterprise HDDs.

FAQs about SSDs vs. HDDs in various application scenarios

How large are the performance differences between HDDs and SSDs?

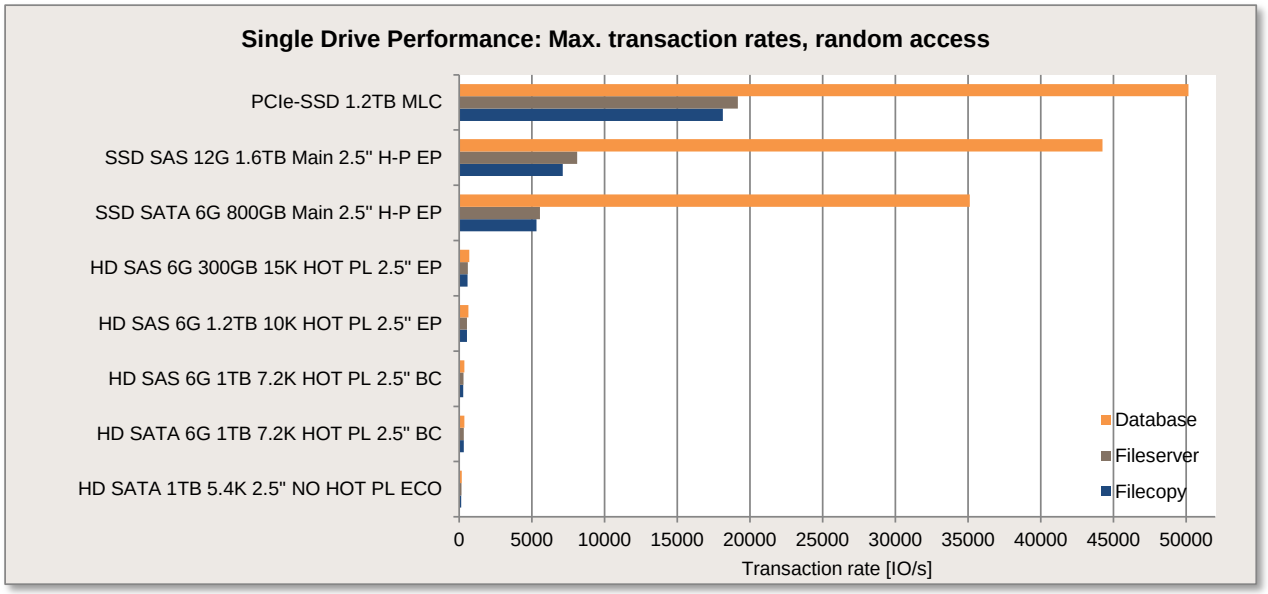
In the case of HDDs the time required to position the read/write head plays a major role. This is high for random accesses, especially with small block sizes. This is why with HDDs the throughput in MB/s for a database-typical access pattern is at most 3% of the throughput for pure streaming.

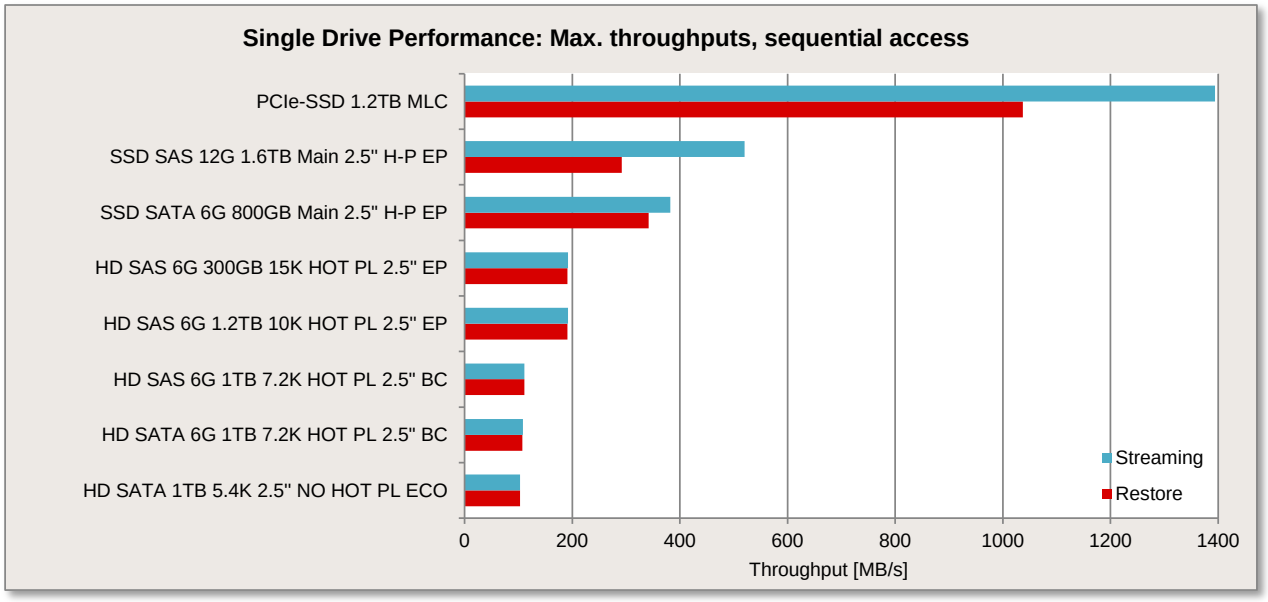
For technological reasons this is fundamentally different for SSDs. With the same block size the throughput for random accesses is not much lower than with sequential accesses. And even with a database-typical access pattern an SSD still achieves about ⅔ of the streaming throughput.

This is why the performance advantage of SSDs stands out particularly with random access patterns – and the smaller the block size in this case, the more it stands out. In comparison, performance differences between Economic HDDs and Enterprise HDDs appear to be almost negligibly small.

The two following diagrams illustrate these differences for a selection of data media from all drive categories. The standard load profiles shown here in table form were used.

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





The topic of performance is discussed in detail in the papers RAID Controller Performance, Performance Report PCIe-SSDs ioDrive®2 and the disk-I/O sections of the PRIMERGY Performance Reports.

If you have any further questions, please contact PRIMERGY Performance and Benchmarks (mailto:primergy.benchmark@ts.fujitsu.com).

Why is my question of all things not answered in this paper?

Perhaps the author has indeed already asked himself your question, and is revising the paper to that effect right now.

Simply contact:

PRIMERGY Product Marketing (mailto:Primergy-PM@ts.fujitsu.com) or

PRIMERGY Performance and Benchmarks (mailto:primergy.benchmark@ts.fujitsu.com).

Literature

PRIMERGY Systems

http://primergy.com/

PRIMERGY Components

This White Paper:

 http://docs.ts.fujitsu.com/dl.aspx?id=78858d6c-4c0f-479a-8ceb-705fe1938f4e

 http://docs.ts.fujitsu.com/dl.aspx?id=222ef5d3-0d3b-428a-9472-e8ed1ca1c8b9

 http://docs.ts.fujitsu.com/dl.aspx?id=1d8b7d65-e5f4-4a99-8e7b-f47c74ccc85e

White Paper:
Hard disk drives or solid state disk drives for servers – what is more suitable?
http://docs.ts.fujitsu.com/dl.aspx?id=94cf1265-15d9-4d91-bbba-0345f37eb74b

Solution-specific Evaluation of SSD Write Endurance
http://docs.ts.fujitsu.com/dl.aspx?id=3c0db973-3406-476c-a301-1b1ed3b1a9ad

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

Performance-Report PCIe-SSDs ioDrive®2
http://docs.ts.fujitsu.com/dl.aspx?id=2d717c91-8da2-4201-8329-68823ada6ec3

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

PRIMERGY Performance

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

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

2014-03-31 WW EN