White Paper
Fujitsu PRIMERGY Servers
512e HDDs: Technology, Performance, Configurations

The new technology of physical 4k sectors for HDDs increases storage density, performance and the capability to correct errors. The first of these HDDs are currently being introduced for PRIMERGY servers. Outwardly, they also emulate logical 512-byte sectors. Suitable versions of the involved software components should be used for optimal performance.

Version
1.2a
2014-01-29

Contents

Document history

Version 1.0 (2013-06-06)

Initial version

Version 1.0a (2013-06-12)

Minor corrections

Version 1.0b (2013-06-18)

Corrected versions of software components

Version 1.1 (2013-12-18)

New chapter about operating system installations
Minor corrections

Version 1.2 (2014-01-28)

Corrected version of SVIM

Version 1.2a (2014-01-29)

Corrected version of SVIM

Technology and examples

For a long time now HDDs have been physically structured in sectors that each has 512 bytes of user data and a certain amount of overhead (including an "error correction code"). However, this sector format has in recent years reached its limits. It is no longer compatible with today's requirements for storage density, performance and the capability to correct errors. To overcome these limits the format for physical sectors of 4096 bytes was defined in 2009 on a cross-company basis ("Advanced Format", 4k sectors). The use of this format presumes support for all affected HW and SW components:

HDDs, disk controllers, operating systems, drivers, applications and other tools.

Since the conversion of all these components by the manufacturers will by its nature take some time, the technology of the 512-byte sector emulation has been introduced for the interim period. This means that although HDDs already have internal physical 4k sectors, from an external viewpoint they still emulate logical 512-byte sectors. This ensures that these new hard disks can also be run in existing environments. Such HDDs sometimes also have the short name "512e" ("e" for "emulation").

Among the hard disks currently on offer for PRIMERGY servers 4k technology is realized solely in the form of these 512e HDDs. In this case, the emulation of the 512-byte sectors is consistently enabled, a switching mode to 4k native does not exist.

The following 512e HDDs are currently available for PRIMERGY servers (only SATA):

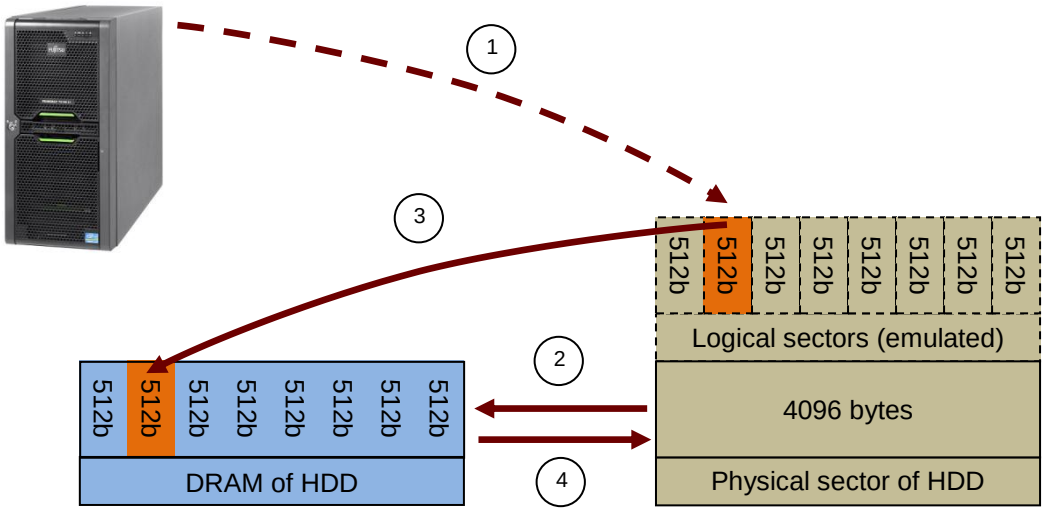
Order code	Description	Deliverable as of
S26361-F3797-E100/L100 *)	HD SATA 6G 1TB 5.4K HOT PL 2.5" ECO	17.06.2013
S26361-F3798-E100/L100	HD SATA 6G 1TB 5.4K HOT PL 2.5" ECO	17.06.2013
S26361-F3700-E250/L250	HD SATA 6G 250GB 7.2K HOT PL 3.5" ECO	08/2012
S26361-F3700-E500/L500	HD SATA 6G 500GB 7.2K HOT PL 3.5" ECO	08/2012
S26361-F3701-E250/L250	HD SATA 6G 250GB 7.2K HOT PL 3.5" ECO	08/2012
S26361-F3701-E500/L500	HD SATA 6G 500GB 7.2K HOT PL 3.5" ECO	08/2012
S26361-F3574-E250/L250 **)	HD SATA 3G 250GB 7.2K NO HOT PL 3.5"	08/2012
S26361-F3293-E250/L250	HD SATA 3G 250GB 7.2K HOT PL 3.5" ECO	08/2012

*) *Special release for TX140 S1 with S26361-F3554-E8, RAID Ctrl SAS 6G 0/1 (D2607)*

**) *S26361-F3574-E250 - EOL*

Some components may not be available in all countries / sales regions.

Accesses, which only process a part of a physical 4k sector, work correctly as a result of the emulation with such HDDs. There are only performance implications with the write processing of parts of a 4k sector. To better understand this you should realize that the rewriting of the "error correction code" for a physical sector can only be done on the basis of the entire 4k user data of the sector; in other words, the HDD must in such cases read the entire content of the 4k sector beforehand. In a nutshell, it must perform a "Read-Modify-Write" cycle:



1) 512-byte sector write request by host
2) HDD reads 4kB sector into DRAM
3) HDD updates requested 512-byte sector in DRAM
4) HDD overwrites 4kB sector on media

Performance

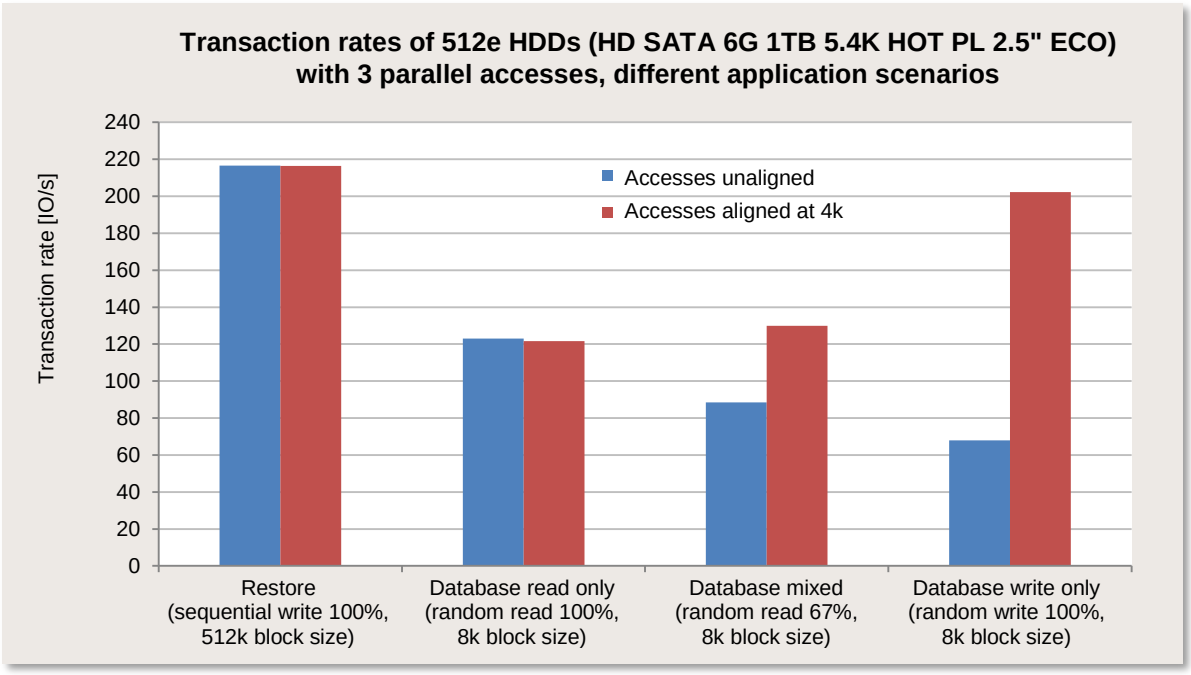
Since "Read-Modify-Write" cycles cause additional head movements on the HDD, they reduce performance. Thus, in the interest of best possible performance on the part of the application and operating system, accesses that only write a part of a physical 4k sector should be avoided. There are two cases of such accesses: On the one hand when write access takes place with a block size that is not a full multiple of 4k. On the other hand, when write access does take place with a block size that is a full multiple of 4k, but this access is not aligned to the boundaries of the physical sectors. The first case is hardly a problem in today's applications and operating systems, because the size of the written blocks is almost always a multiple of 4k. Thus, the essential optimization potential is to be found in the second case, i.e. in the alignment of write accesses.

In order to ensure the alignment of write accesses from a HDD viewpoint it is necessary to meet two requirements in particular:

- On the one hand, the partition on which the write is to take place (or in the case of RAID arrays their parts on every HDD) should be aligned to the physical 4k sectors. This prerequisite is for example given with effect from Windows Server 2008. This partition alignment is the basic prerequisite for subsequent accesses, which are 4k-aligned with regard to the start of the partition, to also actually cause 4k-aligned access from a HDD viewpoint.
- On the other hand, all write accesses must be 4k-aligned with regard to the start of the partition. This is the more comprehensive requirement for all the software components involved. It means that the physical sector size of the partition has to be correctly determined and passed on so that it can ultimately be taken as a basis for the write operations. This applies both for the write accesses that are initiated by the operating system onto the partition (and thus also for "buffered writes" from applications that write in this way) and always for "unbuffered writes".

A readable description of the various aspects of HDDs with physical 4k sectors can be found in the Microsoft article "Advanced format (4K) disk compatibility update". The article also deals with the aspects of software adaptation to physical 4k sectors.

The following diagram shows which differences in the performance of 512e HDDs exist between unaligned and aligned accesses.



There is no performance difference for sequential accesses and random accesses with 100% read. Aligned accesses only have a discernible advantage in the case of random accesses with a write part. In the diagram a write part of 33% is shown as an example in the third column group (corresponds to the frequent access profile typical of databases). Consequently, larger or smaller write parts have a greater or smaller advantage with aligned accesses. In the case of random load profiles with 100% write the performance advantage of aligned over unaligned accesses can be as much as three-fold. This is illustrated by the fourth group of columns in the diagram.

Supplement:

Some hard disks have internal additional mechanisms to soften the impact of unaligned write accesses (e.g. "SmartAlign" technology from Seagate).

Recommended configurations for high-performance support of 512e hard disks

For optimal performance the information about the physical sector size must be known and correctly interpreted in all components. This is guaranteed in the case of the "buffered writes" that are usually implemented by the operating system and applications if the hard disk controller, the driver and the operating system meet the prerequisites described in the following three tables.

Disk controller			
Type	**Order code**	**Description**	**Firmware package versions**
RAID, PCIe slot	S26361-F3554-E512/L512 S26361-F4481-E1/L1 S26361-F3593-E201/L201	RAID Ctrl SAS 6G 5/6 512 MB (D2616) PY SAS RAID Mezz Card 6Gb RAID Ctrl SAS 6G 8Port ex 512MB LP LSI	≥ 12.12.0-0129
	S26361-F3554-E8/L8	RAID Ctrl SAS 6G 0/1 (D2607)	≥ 20.10.1-0120
	S26361-F3669-E1/L1 S26361-F3669-E3/L3 S26361-F4531-E512/L512 S26361-F4531-E513/L513 S26361-F4531-E100/L100 S26361-F4531-E300/L300 S26361-F3713-E201/L201 S26361-F3713-E203/L203	RAID Ctrl SAS 6G 5/6 1GB (D3116) RAID Ctrl SAS 6G 5/6 1GB (D3116C) PY SAS RAID HDD Module PY SAS RAID HDD Module 3.0 PY SAS RAID HDD Module w/o cache PY SAS RAID HDD Module w/o cache 3.0 RAID Ctrl SAS 6G 8Port ex 1GB LP LSI RAID Ctrl SAS 6G 8Port ex 1GB LP LSI V3	Controller not released for 512e HDDs currently on offer
RAID, onboard	(contained in base units)	Onboard controllers, RAID mode	Firmware doesn't have any effect
NON-RAID	S26361-F4480-E1 S26361-F3554-E118	PY SAS HBA Mezz Card 6Gb PSAS CP200i	Firmware doesn't have any effect

Operating system	Driver versions
Windows variants	With controllers in PCIe slots, RAID mode: Megasas ≥ 6.505.5 With onboard controllers in RAID mode: MegaSR ≥ 15.02.2013.0425 Controller in NON-RAID mode: All
Linux variants	With controllers in PCIe slots in RAID mode and with controllers in NON-RAID mode: Native support With onboard controllers in RAID mode: MegaSR ≥ 15.02.2013.0425

Operating system	
Name	**Additional requirements**
Windows Server 2008	KB2553708
Windows Server 2008 R2	KB982018 or Service Pack 1
Windows Server 2012	-
RHEL	Version ≥ 5
SLES	Version ≥ 11 SP2

Some components may not be available in all countries / sales regions.

Comments:

- All the configurations not named in the tables support 512e HDDs, namely functionally, but not with optimal performance.
- Please see the system configurators for the released combinations of controllers and hard disks.
- To already ensure optimal performance during an operating system installation it is necessary for the above mentioned software components to be available in good time in the installation procedure. The following section shows how that can be achieved for example in the standard Windows versions for PRIMERGY servers.
- If work is done outside normal operating system environments, thus e.g. using imaging tools to create partitions or to clone entire hard disks, you should for optimal performance also bear the 512e support of these tools in mind.
- The following applies for all applications: If the application works with "unbuffered writes", this must take the physical sector size into account for optimal write performance in the case of accesses to 512e HDDs.

Windows operating system installations

The above mentioned hotfixes and drivers are normally only installed after the operating system installation. If keeping the duration of an operating system installation onto 512e-HDDs to a minimum is required, the relevant, additional software components should be made available as early as possible in the installation procedure. The following table lists by way of example which options you have in the standard Windows operating system versions for PRIMERGY servers so as to meet this requirement of "earliest possible availability".

This table also takes the "ServerView Installation Manager" (abbreviated to SVIM) into account. It is supplied with PRIMERGY servers and offers additional hotfixes and drivers (including ones to support 512e-HDDs), which are also available in good time in the installation procedure. The name "installation source" is used below as a generic term to mean the various options of providing an operating system installation medium locally or via a network.

<table>
<tr><th rowspan="2">Operating system version</th><th colspan="2">Software components to be used and information about high-performance installation (in addition to optimal controller FW)</th></tr>
<tr><th>for the OS</th><th>for the controller driver</th></tr>
<tr><td>Windows Server 2008</td><td>SVIM version ≥11.13.08 together with every installation source
OR
Installation source that has incorporated hotfix KB2553708 for all phases of the installation *)</td><td rowspan="4">SVIM version ≥11.13.08 together with every installation source
OR
Recommended controller driver according to the table of driver versions above. This driver should be added during the installation when you see the prompt "Where do you want to install?" by selecting "Load Driver" (alternatively, by pressing F6)</td></tr>
<tr><td>Windows Server 2008 R2</td><td>SVIM version ≥11.13.08 together with every installation source
OR
Installation source that has incorporated hotfix KB982018 for all phases of the installation *)
OR
Installation source that has incorporated Service Pack 1 *)</td></tr>
<tr><td>Windows Server 2012</td><td>-</td></tr>
<tr><td>Windows Server 2012 R2</td><td>-</td></tr>
</table>

*) *Through suitable use of administrative tools ("Deployment Image Servicing and Management" (DISM) from the "Windows Automated Installation Kit" (Windows AIK or WAIK) or the "Windows Assessment and Deployment Kit" (Windows ADK)) it is possible for an installation image of a Windows operating system version to be extended, compared with the original, to include selected packages.*

The following example illustrates the time savings that can be achieved with an operating system installation onto 512e-HDDs, taking as an example a "HD SATA 6G 500GB 7.2K HOT PL 3.5" ECO" hard disk that is connected to an onboard controller.

Software components	Duration of the installation
Windows Server 2008 R2; Installation source without incorporated hotfix KB982018 and/or Controller driver < 15.02.2013.0425	approx. 110 min
Windows Server 2008 R2; Installation source with incorporated hotfix KB982018 and Controller driver ≥ 15.02.2013.0425	approx. 40 min

This clearly shows that both the prerequisites as regards the operating system as well as the controller driver must be met for the shortest possible installation times.

Literature

PRIMERGY Systems

http://primergy.com/

Basics of Disk I/O Performance

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

Technology "Advanced Format" for HDDs

Seagate Technology Paper:
SmartAlign™ Technology for Advanced Format Hard Drives
http://www.seagate.com/docs/pdf/whitepaper/tp615_smartalign_for_af_4k.pdf

Compatibility of Operating Systems and Applications with "Advanced Format" HDDs

Advanced format (4K) disk compatibility update
http://msdn.microsoft.com/en-us/library/windows/desktop/hh848035(v=vs.85).aspx

Hotfixes and Service Packs for Windows Server Operating Systems

Hotfix KB2553708 for Windows Server 2008
http://support.microsoft.com/kb/2553708/en-us

Hotfix KB982018 for Windows Server 2008 R2
http://support.microsoft.com/kb/982018/en-us

Service Pack 1 for Windows Server 2008 R2
http://support.microsoft.com/kb/976932/en-us

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

2014-01-29 WW EN