4K Sector HDD FAQ

Dell Engineering
May 2015

Authors:

Frank Widjaja and Chetan Kumar, Dell Enterprise Disk Engineering

Revisions

Date	Description
Jan 2014	Initial release
May 2015	Updated the template

Contents

Terminology

Sector – An atomic unit of data transfer size from/to a hard Disk Drive

LBA (Logical Block Address) – An atomic unit of hard disk drive (HDD) sector address (location).

Physical Sector – Sector size at the HDD media level, normally is 512 bytes

Physical LBA – LBA layout on HDD media level, each LBA has the Physical Sector size

Logical Sector – Sector size defined at the host–to-disk drive interface. Normally the same size as Physical Sector unless the HDD is emulating.

Logical LBA – LBA layout at host-to-disk drive interface, each LBA has the Logical Sector size.

Frequently Asked Questions

1. What is a 4K sector HDD? How is it different from standard 512 bytes sector HDD?

4K sector HDD is a new generation HDD whose physical sector is 4K bytes. As HDD areal density increases, the footprint of each physical sector shrinks. However, since the natural contamination (dust, particulate, etc.) remains about the same size, the ratio between the footprint of the media defect and the physical sector is increasingly larger. It requires more embedded ECC per each physical sector to maintain the published sector error rate. By increasing the physical sector size to 4K bytes, the ratio is reduced therefore the error correction code (ECC) burden per physical sector is also reduced. The larger physical sector not only improves format efficiency but also improves media defect correction ability and signal-to-noise (S/N) design margin. Current shipping HDDs are based on 512 byte physical sector and logical sector.

2. What is 4K native HDD and what is 512 emulation (512e) HDD?

There are two models of 4K sector HDDs:

1. 4K native HDD is a 4K sector HDD whose logical sector is the same as physical sector. Therefore the logical sector size is 4K bytes instead of the legacy 512 bytes. This is incompatible with legacy BIOS and Operating Systems.
2. 512 emulation (512e) HDD is a 4K sector HDD whose logical sector is 512 bytes (not matching with physical sector). It no longer has a one-to-one relationship between physical sector and logical sector. Instead, one 4K physical sector consists of eight logical 512 bytes sectors (4K bytes = 8 * 512 bytes). This is done to make the 512 emulation HDD compatible to legacy BIOS and Operating Systems.

Common Names	Reported Logical Sector Size	Reported Physical Sector Size	Windows Version with Support
512-byte Native, 512n	512 bytes	512 bytes	All Windows versions
Advanced Format, AF, 512e, 512E, 512-byte Emulation	512 bytes	4096 bytes	Windows Server 2012 Windows Server 2008 R2 w/ MS KB 982018 Windows Server 2008 R2 SP1 Windows Server 2008 w/ MS KB 2553708
Advanced Format native, AFn, 4K Native, 4Kn*	4096 bytes	4096 bytes	Windows Server 2012 (4K data disks are supported and as boot disks in UEFI mode)

While not stressed in the preceding table, Windows Server 2003, and Windows Server 2003 R2 do not support 512e or 4Kn media. While the system may boot up and be able to operate minimally, there may be functionality issues, data loss, or sub-optimal performance. Thus, Dell strongly cautions against using 512e media with legacy Windows operating systems such as Windows Server 2003.

Common Names	Reported Logical Sector Size	Reported Physical Sector Size	Linux Version with Support
512-byte Native, 512n	512 bytes	512 bytes	All Linux versions
Advanced Format, AF, 512e, 512E, 512-byte Emulation	512 bytes	4096 bytes	RHEL 6.1 * SLES 11 SP2 ** Ubuntu 13.10 Ubuntu 12.04.4
Advance Format native, AFn, 4K Native, 4Kn	4096 bytes	4096 bytes	RHEL 6.1 * SLES 11 SP2 ** Ubuntu 13.10 Ubuntu 12.04.4

3. What are 512 emulation HDD performance issues and potential data integrity risks?

As the 512e HDD has a 4K bytes physical sector, the internal HDD read and write functions are performed on one physical sector (4K bytes) at a time or as a group of eight logical sectors (512 bytes) at a time. As the legacy host performs data transfer at 512 bytes boundary, any of the write data could start and end at the beginning, in the middle or at the end of the 4K physical sector. When the data starts or ends in the middle of the physical sector, it is called misaligned data. On misaligned data, the 512e HDD must perform READ-MODIFY-WRITE (RMW) functions to complete the write operations. Therefore, 512e HDD suffers significant performance loss (50%) in the random writes, misaligned data operations. In addition, a sudden power loss during RMW operation could corrupt the physical sector causing data loss or corruption on adjacent logical sectors within the affected physical sector. The host will not be aware of these corruptions on the adjacent logical sectors since they were not part of the data transfer during the emergency power loss condition.

4. What is Advanced Format?

Advanced Format is the term for 4K sector HDD or 512e HDD implementation.

5. How can performance issues and data risk be mitigated?

As stated before, the main reason for 512e performance issues is misalignment during writes.

Newer operating systems and applications are 512e-disk aware and minimize the incident of READ-MODIFY-WRITE. Writing data on an aligned 4K boundary and in multiple of 4K bytes eliminates the READ-MODIFY-WRITE operation.

As of the date of this document, most versions of Client and Enterprise Entry, Cloud and Archive drives do not have power-loss-protection during READ-MODIFY-WRITE operations. An emergency power loss during READ-MODIFY-WRITE operation of a physical sector (4K) could corrupt the adjacent logical sectors (512 bytes) within the affected physical sector.

Enterprise drives incorporate an advanced non-volatile cache system to buffer the READ-MODIFY-WRITE operation.

There are three versions of this non-volatile cache system:

a. Solid state non-volatile cache (NVC NOR flash or NVC NAND flash) using built-in spindle back-EMF to power the flash in event of a power loss
b. Disk Media non-volatile cache which is also known as Media Based Cache (MBC or MC) using a reserved area of the drive media to buffer the non-aligned writes from the host
c. Combination of 1 and 2

As of now, there are various non-volatile cache implementations:

RMW Cache System	**Implemented in**
None	Client, Entry, Cloud, and Archive HDDs
NVC solid state	Vendor A&B* for Enterprise drives
MBC/MC	Vendor C for Enterprise drives
NVC + MC	Available as proto for enterprise 512e drives but no plan for production

*HDD vendors' caching deployments are unique.

6. When is 4K sector HDD launching?

4K sector HDD started on notebook HDDs in Q3 '2010, on Desktop HDDs in 2011, selective Cloud Enterprise HDDs in 2013 and mainstream Enterprise HDDs in 2014. Dell is leading this transition by adopting these drives in PowerEdge, Power Vault, EqualLogic and Compellent systems.

Notebooks: The transition started with 2.5" notebook drives, starting with the largest capacity (750 GB) new drive families in late 2010. The mainstream 2.5" capacities (250GB-500GB) for notebooks transitioned to 4K sector (512e) in mid-2011.

Desktops: 3.5" HDD has ~3x the capacity point of 2.5" HDD so the demand of high capacity 3.5" HDD was less than the 2.5" HDD. The transition occurred with the 4TB product introduction in 2012.

Enterprise: New products will have options for 512 native, 512e and 4K native models. This is to provide seamless transition. By 2016, all new products will either be 512e or 4K native models. Legacy capacities will be supported with 512n models until 2020.

7. How will 512e-disks affect our customers?

Customers who are using client 512e HDDs and use them with legacy OS and software will have performance degradation and data integrity risks. See question 2 above for known legacy operating systems that are not 4K aware. Most home-grown cloud operating systems have been designed to operate with 4K native or 512e format.

8. How to mitigate performance degradation?

Early adopter client customers minimized the performance degradation of 512e HDDs by converting to new OS and software that are "4K aware". There is also third-party alignment software that reduces the data misalignment on 512e HDDs in conjunction with legacy OS and software device drivers.

Currently, client systems are shipped with new OS that is 4K aware so the misaligned incidents are reduced significantly. Some of the benchmarks (Jetstress) showed no performance concerns with client systems using latest OS and later generation 512e drives.

Enterprise cloud customers have their own home-grown OS that converts the atomic operation to 4K bytes sector so 512e and 4K native drives can be easily deployed in these cases.

Traditional mission critical enterprise customers running traditional OS (Microsoft, Linux) and database (Oracle, SQL) will have to align their disk partitions to avoid performance and data loss possibilities.

In addition, customers can choose to use the Enterprise version of 512e HDD with non-volatile cache system to further minimize the performance degradation impact.

Customers can also choose 4K native HDD if their servers are using newer versions of BIOS/OS/Software. This solution stack will have the optimal performance since the physical and logical sectors are matched. However, it is limited to only the newer BIOS/OS/Software.

9. Is there data integrity risk during sudden power off?

If volatile write cache is disabled:

- For 512n and 4Kn HDD, there is no data loss during sudden power off.
- Enterprise 512e HDD with NVC and/or MBC/MC feature (see question 5) also does not have data integrity risk.

- For"512e"without"NVC"or"MBC/MC"(Client"grade"HDD"for"example),"there"is"a"risk"of"data"integrity" during"sudden"power"loss.

If volatile write cache is enabled, then sudden power loss will result in data cache loss regardless of the HDD format types.