4K Sector HDD FAQ

Dell Enterprise Disk Engineering
January 3, 2014

Terminology

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

Logical Block Address (LBA) – An atomic unit of 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.

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 ECC burden per physical sector is also reduced. The larger physical sector not only improve format efficiency but also improve media defect correction ability and 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 HDD:

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 System.
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 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 System.

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)

Note: 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 unknown scenarios of functionality issues, data loss, or sub-optimal performance. Thus, Dell strongly cautions against using 512e media with Windows such as Windows Server 2003.

Common Names	Reported Logical Sector Size	Reported Physical Sector Size	Windows 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

*Red Hat Enterprise Linux 6 supports 4k-sector devices as data disks. 4K-sector boot disks are supported in UEFI mode only

**SUSE Linux Enterprise fully supports 4 KB/sector drives in all conditions and architectures with one exception. The 4KB/sector hard disk drives are not supported as a boot drive on x86_64 systems booting with a legacy BIOS.

3. Why do 512 emulation HDD has performance issue and potential data integrity risk?

The 512e HDD has 4k bytes physical sector, the internal HDD read and write functions are performed one physical sector (4k bytes) at a time OR a group of eight logical sectors (512 bytes) at a time. Since 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 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 READ-MODIFY-WRITE operation could corrupt the physical sector causing data loss/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. What are the mitigations to 512e performance and data risk?

As stated before, the main reasons for 512e performance is mis-alignment 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 paper, most versions of Client, Cloud, and Cold drives do not have power-loss-protection during READ-MODIFY-WRITE operation. 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 version of this non-volatile cache system:

1. Solid state non-volatile cache (NVC NOR flash or NVC NAND flash) using built-in spindle back-EMF to power the flash.
2. Disk Media non- volatile cache aka. Media Based Cache (MBC or MC) using reserved area of the drive media to buffer the non-aligned writes from the host
3. Combination of step 1 and step 2.

At the time of this paper, we are seeing various non-volatile cache implementations:

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

6. When is 4k sector HDD happening?

4k sector HDD started on Notebook HDD in Q3 '2010, on Desktop HDD in 2011, selective Cloud Enterprise HDD in 2013 and will be on mainstream Enterprise HDD 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 (750GB) new drive families in late 2010. The mainstream 2.5" capacities (250GB-500GB) for notebooks transitioned to 4k sector (512e) 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 lagging the 2.5" HDD. The transition occurred with the 4TB product introduction in 2012.

Enterprise: New products in 2014 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, with very few options for 512 native (such as with Helium drives).

7. How will customer be affected?

Customer who buys early version of 512e HDD and uses them with legacy OS and software will have performance degradation and data integrity risk. See section 2 above for Known legacy OSes that are not 4k aware. Most home grown cloud operating systems have been designed to operate with 4k native or 512e format.

8. How does the customer avoid performance degradation?

Early adopter Client customer can minimize the performance degradation of 512e HDD by converting to new OS and software that are "4k aware". There is also 3rd party alignment software that reduces the data misalignment on 512e HDD in conjunction with legacy OS & software device driver.

Currently, client systems are shipped with new OS that is 4k aware so the misaligned incidents are reduced significantly. Benchmark showed no performance concerns with Client systems using latest OS and later generation 512e drives.

Enterprise cloud customer have their own home grown OS that converts the atomic operation to 4k bytes sector so 512e and 4k native drives are ready.

Traditional mission critical enterprise customer running traditional OS (Microsoft, Linux) and database (Oracle, SQL) will have to align their disk partition and to apply best practice (see Microsoft knowledge base and our next paper).

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

Customer can also choose 4k native HDD if he/she is only running with newer version 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 newer BIOS/OS/Software.

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

If volatile write cache is disabled:

- 512n and 4kn HDD, there is no data loss during sudden power off.
- Enterprise 512e HDD with NVC and/or MBC/MC feature (see section 5) also does not have data integrity risk.
- For 512e without NVC or MBC/MC (Client grade HDD), 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.