Block alignment for best performance on Nimble Storage

(Version 20130326-A)

The purpose of this KB article is to describe various types of block alignment pathologies which may cause degradations in performance if not corrected. These include:

Partition misalignment: Occurs if the starting location of the partition (on Windows, Linux, or VMFS) is not aligned with the volume block size.

IO/block size mismatch: Occurs if the volume block size is larger than the IO size.

VMware linked clones: Occurs with VMware linked clones because of tight packing of written chunks in the clone's delta file (aka redo log).

From a performance perspective, partition misalignment has a larger impact than IO size mismatches and therefore should always be avoided and resolved if possible. However, to gain the best possible performance efficiencies out of a Nimble array it is advisable to ensure IO sizes are also matched appropriately. Outlined below are the conditions where each type of block alignment issue can be observed and steps to take to resolve them.

Partition Alignment

VMFS partitions that will need to be aligned:

1. VMFS partitions in ESX 3.x
2. VMFS partitions in vSphere 4.x AND were created by command line "vmkfstools" with starting block that is NOT a multiple of 4KB. If the GUI was used in vSphere 4.x to create the VMFS file system then the partition is aligned properly.

How to fix

1. For ESX 3.x VMFS partition alignment, follow: http://www.vmware.com/pdf/esx3_partition_align.pdf to use FDISK to align a partition manually.
2. For ESX 4.x VMFS partition alignment, use vSphere client to create VMFS partition. If "vmkfstools" is required for VMFS partition creation, follow http://www.vmware.com/pdf/esx3_partition_align.pdf to use FDISK to align a partition manually.

Guest and physical partition that will need to be aligned:

1. Windows server 2003/2000/XP data disks
2. Windows server 2008/2008-R2/Vista data disks that are *in-place* upgraded from Windows server 2003/2000/XP
3. Linux OS

How to fix

1. For Windows, follow the Microsoft KB article - http://support.microsoft.com/kb/929491
    a) ***Please note:*** The strong recommendation is to use a starting offset of 2,048 sectors (1MB).
2. For Linux OS, use fdisk to align a partition manually:
    1. `Enter fdisk -u /dev/sd<x> where <x> is the device suffix`
    2. `Type n to create a new partition`
    3. `Type p to create a primary partition`
    4. `Type 1 to create partition No. 1`
    5. `Type 2048 for the first sector`
    6. `Use default value for the last sector`
    7. `Type w to write label and partition information to disk`

IO/Block Size Matching

There are essentially 3 layers of IO/block sizes that need to be matched:
1.) Application IO size
2.) Host file system block size (e.g., NTFS cluster size)
3.) Nimble volume block size

To avoid misalignment with Windows-based applications, the **Nimble Volume Block Size** should be no larger than the **Application IO Size** and also no larger than the **NTFS Cluster Size**:

NIMBLE_VOL_BLOCK_SIZE <= APP_IO_SIZE **AND** NIMBLE_VOL_BLOCK_SIZE <= NTFS_CLUSTER_SIZE

> *Note: with LINUX-based applications you can use the following constraints instead:*
> NIMBLE_VOL_BLOCK_SIZE <= FILESYS_BLOCK_SIZE <= APP_IO_SIZE

For optimal performance, the Nimble volume block size should be set to the highest value that satisfies the above constraints. But this optimality is secondary to meeting the constraints: if the

Nimble volume block size is set so large as to violate the constraints, there will be a severe penalty because of misaligned IO.

The NTFS cluster size should be set based on Microsoft best practices. In general Microsoft recommends setting the NTFS cluster size to 4KB for small files (e.g., file shares), and to a larger value such as 64KB for large files (e.g., SQL Server).

For example; if the application is doing 8KB IO, and NTFS cluster size is 8KB, then the Nimble volume block size can be either 4KB (aligned but sub-optimal) or 8KB (optimal), but not 16KB (misaligned). Otherwise, if the NTFS Cluster size is 4KB, the Nimble Volume block size can be 4KB (optimal), but not 8KB (misaligned).

For SQL Server, the default IO size is 8KB, and Microsoft recommends setting the NTFS cluster size at 64KB. Given these parameters, the Nimble volume block size can be 4KB (aligned but sub-optimal) or 8KB (optimal), but not larger (misaligned). Please consult the table at the end of this document for guidelines on optimal settings.

The basic steps to ensure proper matching include:
1.) Determining the application IO size.
2.) Setting the Nimble volume block size via the Nimble.
3.) Setting the NTFS cluster size (during formatting) based on Microsoft best practices using *Computer Manager->Disk Management*

Determining the APP_IO SIZE

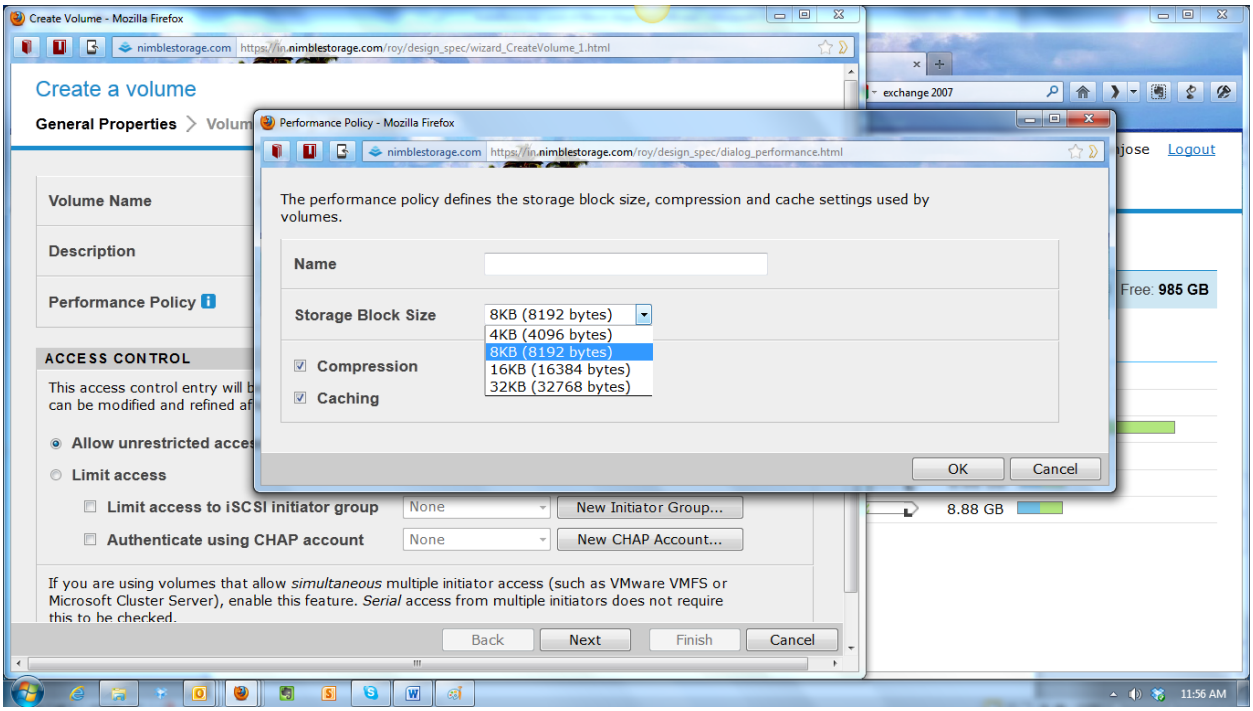
Using the table below one can identify what the proper APP_IO_SIZE should be used for optimal performance. The table below is not an exhaustive list, but includes the commonly used ones.

Application (perf)	App IO Size	NTFS cluster size	Nimble volume block size
Default (Best)	**4K**	**4K**	**4K**
SQL (Best)	**8K**	**64K**	**8K**
SQL (Good)	8K	8K	8K
SQL (Ok)	8K	4K	4K
SQL (Really BAD!)	8K	4K	8K
SQL (Really BAD!)	8K	64K	32K
Exchange 2010 (Best)	**32K**	**64K**	**32K**
Exchange 2010 (Good)	32K	32K	32K
Exchange 2010 (Ok)	32K	32K	4K
Exchange 2010 (Ok)	32K	4K	4K
Exchange 2010 (Really BAD!)	32K	4K	32K
Exchange 2007 (Best)	**8K**	**64K**	**8K**
Exchange 2007 (Good)	8K	8K	8K
Exchange 2007 (Ok)	8K	8K	4K

Exchange 2007 (Ok)	8K	4K	4K
Exchange 2007 (Really BAD!)	8K	4K	8K

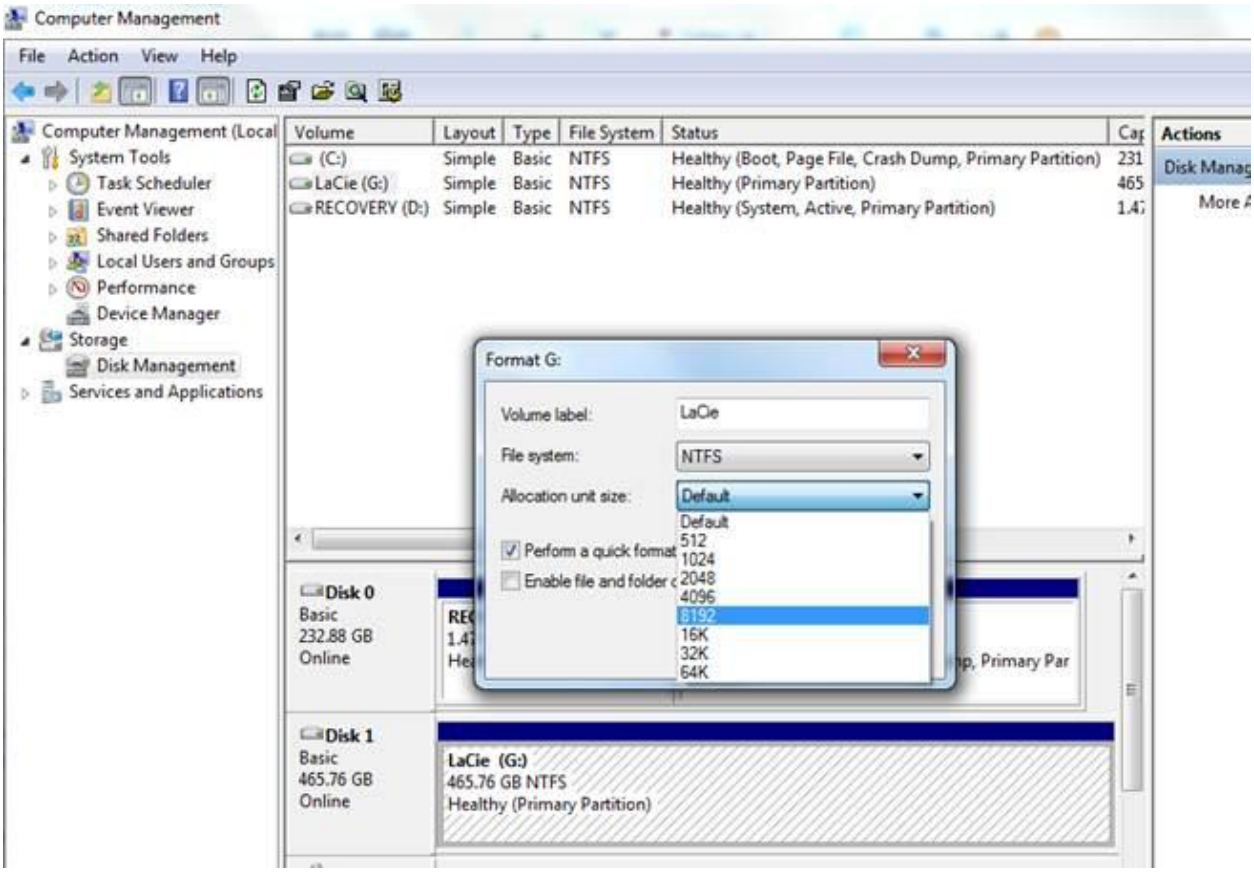
Setting the Nimble volume block size

For optimal performance, set the Nimble volume block size to the minimum of the application IO size and the recommended NTFS cluster size. If in doubt, set the volume block size to a lower value such as 4KB, which might be suboptimal but will avoid misalignment. The Nimble volume block size can be configured (via the Nimble GUI) by specifying an existing performance policy or a custom one with that block size (as shown below.)



Setting the NTFS cluster size

The last step requires setting the NTFS cluster size based on Microsoft best practices. To do this you will need to use *Computer Manager->Disk Management* and ensure that the proper cluster size is chose ***before*** formatting (this is a one-way operation and cannot be undone.) In *Disk Management* (refer to image below) right click the volume; select format and under "*Allocation unit size*" drop-down select the matching block size. From here you can proceed with actual formatting. See image below for how to set the cluster size on an NTFS volume.



VMware Linked Clones

vSphere 5.1 and older versions implement a linked clone using a delta file that acts as a "redo log" on top of a base or snapshot file. Writes to the linked clone are packed tightly in this redo log, aligning to 512 byte boundaries, but not necessarily to 4KB. Note that all storage systems are prone to this misalignment. VMware is aware of the problem and is planning to resolve the problem in the upcoming vSphere 5.5 release using what they call "space-efficient sparse disks".