# NAS4Free Forums

Welcome to the NAS4Free Forums!
http://forums.nas4free.org/

## WD advanced format 4k drives - definitive answer?

http://forums.nas4free.org/viewtopic.php?f=59&t=1494

### WD advanced format 4k drives - definitive answer?



by **striken4f** Posted: **October 21st, 2012, 6:11 pm**

Hi all

I have done a search on this issue and the FAQ refers me to "NAS4Free with Western Digital Advanced Format Drive" before posting questions, but for the life of me neither me nor Mr. Google can track that FAQ down, so apologies if this has been asked before. I'm a long-time knowledgeable Windows technician but my expertise of unix-type OSes doesn't go much beyond what I learned during my CompSci degree at college many years ago.

I am about to build my first ZFS NAS and after much research have settled on Nas4Free over all the alternatives (FreeNAS, Napp-It, Nexenta etc.) mostly because I'm not convinced on the open-source future of the Solaris OSs

I have 7x WD Red 2TB drives WD20EFRX which I intend to use in a RAIDZ2 pool of 6 drives with a hot spare. I have searched on this issue of Advanced Format drives and get numerous conflicting answers across the web, none of which give a definitive answer. I'd really appreciate a clear answer on this issue from the community if possible...

So my questions are:

1) As I understand it, the 2tb Reds are 512E drives - i.e. they are 4k sector drives which report 512 byte sectors to the OS. Is this correct?
2) If I select Advanced Format drive option in Nas4Free, is that all I have to do to make them work at peak performance? Should/must I format them first using Western Digital's align tool under Windows? Use the pin 8/9 jumper? Something else? Or is it just hook 'em up and go? At the moment they are raw, unformatted and unpartioned drives except for I've run extended drive fitness tests on each of them (which BTW was well worth it as despite getting them from three different stores to ensure they weren't from the same batch I had bad sectors on two of the drives and had to RMA them).
3) Is there any way to force the drives to report 4k sectors and would it be advisable to do so if possible?

I'd very much appreciate any advice - when it's built I'm going to be trusting my personal data to this NAS and obviously I'd like to set it up from the get-go for maximum compatibility and performance.

My NAS box will consist of:

Supermicro X9SCA-F-O mobo
E3-1220V2 Xeon cpu
Super Talent DDR3-1333 16GB ECC ram
6x Western Digital Red WD20EFRX 2TB drives in RAIDZ2
OCZ Agility 4 64GB SSD to run the OS and serve as L2ARC
Intel 313 Series Hawley Creek 20GB SLC SSD for ZIL
Additional SATA card to run the SSDs of (haven't settled on which one yet but this forum has been very helpful in narrowing down my choices)
All running over a gigabit wired home network.

I know this is all way overkill as Nas4Free can run on low end components but I really wanted a proper server-level NAS with ECC to absolutely ensure reliability and data integrity.

Thanks in advance for the help...

## Re: WD advanced format 4k drives - definitive answer?

by **kernow** Posted: **October 21st, 2012, 6:43 pm**

I'm not sure there is any way to get their firmware to report 4K sectors. I have WD20EARs and just ticked the advanced sector box when setting them up.

Apart from using WDIDLE to set the idle timer to maximum, I haven't done anything else. Performance was never my main priority though.

## Re: WD advanced format 4k drives - definitive answer?

by **fsbruva** Posted: **October 22nd, 2012, 8:12 am**

Here's the deal (just went through this):
You need to trick ZFS into looking past the 512 sector size reported by the drive when creating the pool.
1. Add the disks
2. Format as ZFS
3. Create virtual devices, choosing advanced format (this will create the pool, but with ad#.nop as the members)
If you run

```
zdb | grep ashift
```

at this point, you should get ashift=12. However, zfs is still working through GEOM to access the drives (unnecessary overhead). Now, before you actually write any data to the drive, do the following:

4. From a command prompt, export the zpool

```
zpool export {poolname}
```

If the above command doesn't work because something is accessing the device, then force it with the -f flag.

5. From a command prompt

```
gnop destroy /dev/ad0.nop /dev/ad1.nop
```
(assuming ad0 and ad1 are your devices. Mine are ad4 and ad6)

6.
```
zpool import {poolname}
```

7. When you type zpool status, you should see your pools have the physical devices listed as members, not the GEOM logical ones.

The GEOM devices will be created at boot time still, but it won't matter, because zfs is dealing directly with the drive, rather than the gnop layer. You can also verify that this whole shebang will work by interrogating the drives for the features. When you add the disks to the management interface, ensure you activate SMART monitoring. Then, you ask about the drive by issuing the following command:
```
smartctl -i /dev/ad0
```
(or whatever ad # you have)

You should see something similar to this:
```
nas4free:~/# smartctl -i /dev/ad4
smartctl 5.43 2012-06-30 r3573 [FreeBSD 9.1-RC2 amd64] (local build)
Copyright (C) 2002-12 by Bruce Allen,
http://smartmontools.sourceforge.net

=== START OF INFORMATION SECTION ===
Model Family:     Toshiba 2.5" HDD MK..59GSM (Adv. Format)
Device Model:     TOSHIBA MK1059GSM
Serial Number:    914GPE74T
LU WWN Device Id: 5 000039 3829010a4
Firmware Version: GU001U
User Capacity:    1,000,204,886,016 bytes [1.00 TB]
Sector Sizes:     512 bytes logical, 4096 bytes physical
Device is:        In smartctl database [for details use: -P show]
ATA Version is:   8
ATA Standard is:  Exact ATA specification draft version not indicated
Local Time is:    Mon Oct 22 05:18:45 2012 PDT
SMART support is: Available - device has SMART capability.
SMART support is: Enabled
```

The crucial part is: Sector Sizes: 512 bytes logical, 4096 bytes physical. That's a clue that this method will actually do some good. Following the pool's export and subsequent import, it would be wise to carry out a scrub before you put data on it, and re-issue the zdb command to double check the alignment.

## Re: WD advanced format 4k drives - definitive answer?

by **shaitan667** Posted: **October 23rd, 2012, 6:02 am**

So I have 4x3TB Red Drives in RAIDZ - what benefits will I see in doing this?

## Re: WD advanced format 4k drives - definitive answer?

by **fsbruva** Posted: **October 23rd, 2012, 7:33 am**

> ***shaitan667 wrote:***
> So I have 4x3TB Red Drives in RAIDZ - what benefits will I see in doing this?

You should see higher read/write speeds. You could do some tests with large file transfers (over SMB or just within the system), and then make the change and try it again. Odds are, if you chose "Advanced Formatting" when you created the vdevs, then you already have an ashift of 12. The method I am posting has two uses:
1. If creating the zpool only lets it use strip size of 512 (because that's what the drive reports), and the pool has an ashift of 9. This method allows zfs to use the correct ashift, and play ball with the physical stripe size of the drive.
2. If creating the zpool was done with a gnop device to allow for 4096 byte stripes, even though the drive has 4096 physical stripe size. This method removes the middleman of the gnop logical device.

> ***wiki.illumos.org wrote:***
> There is no functional or reliability problem with 4KB physical sectors being represented as 512 byte logical sectors. This technique has been used for decades in computer systems to allow expansion of device or address sizes. The general name for the technique is read-modify-write: when you need to write 512 bytes (or less than the physical sector size) then the device reads 4KB (the physical sector), modifies the data, and writes 4KB (because it can't write anything smaller). For HDDs, the cost can be a whole revolution, or 8.33 ms for a 7,200 rpm disk. Thus the performance impact for read-modify-write can be severe, and even worse for slower, consumer-grade, 5,400 rpm or variable speed "green" drives. Bottom line: for best performance, the HDD needs to properly communicate the physical block size via the inquiry commands for best performance.

Some empirical data: http://digitaldj.net/2010/11/03/zfs-zpool-v28-openindiana-b147-4k-drives-and-you/
Note that you are using "sub-optimal" drive count for RAIDZ. I am not sure where that number comes from, or how big an impact it may or may not have.

## Re: WD advanced format 4k drives - definitive answer?

by **shaitan667** Posted: **October 23rd, 2012, 10:06 am**

Ok cool, thanks for the info. I had selected the Advanced Formatting option when creating the pool.

I think I know where the Drive count recommendations came from. There a large thread somewhere (I think it was maybe OCAU or Hardforum,) by the sub.mesa guy where he benchmarked a whole heap of different setups and posted results. I saw it earlier today when searching for something else -should have bookmarked it

Basically it showed that RAIDZ works best when using an odd number of drives so that there are an even number of data drives plus the parity drive. That way the amount of data being written to each drive is equal to a power or multiple of 2. I really can't remember the exact theory behind it, but it sounded logical. I think I will ditch the SSD and get another Red.

I will hunt through my history and find the thread - although I think I was at work when I saw it. Will post it when I find it.

EDIT - Found it! Yay for shared history across all signed in installations of Chrome

http://hardforum.com/showpost.php?p=103 ... ostcount=2

Here is what he said

sub.mesa 2[H]4U, 2.7 Years

Status:
The theory behind RAID-Z performance and 4K sector drives:

Quote:
RAID-Z is somewhat odd; it is more like RAID3 than RAID5 really. To avoid confusion, let me explain on how i understand this to work:

Traditional RAID
In traditional RAIDs we know stripesize; normally 128KiB. Depending on the stripe width (number of actual striped data disks) the 'full stripe block' would be <data_disks> * <stripesize> = full stripe block. In RAID5 the value of this full stripe block is very important:

1) if we write exactly the amount of data of this full stripe block, the RAID5 engine can do this at very high speeds, theoretically the same as RAID0 minus the parity disks.

2) if we write any other value that is not a multiple of the full stripe block, then we have to would have to do a slow read+xor+write procedure which is very slow.

Traditional RAID5 engines with write-back essentially build up a queue (buffer) of I/O requests and scan for full stripe blocks which can be written efficiently; and will use slower read+xor+write for any smaller or leftover I/O.

RAID-Z
RAID-Z is vastly different. It will do ALL I/O in ONE phase; thus no read+xor+write will ever happen. How it does this? It changes the stripe size so that each write request will fit in a

full stripe block. The 'recordsize' in ZFS is like this full stripe block. As far as i know, you cannot set it higher than 128KiB which is a shame really.

So what happens? For sequential I/O the request sizes will be 128KiB (maximum) and thus 128KiB will be written to the vdev. The 128KiB then gets spread over all the disks. 128 / 3 for a 4-disk RAID-Z would produce an odd value; 42.5/43.0KiB. Both are misaligned at the end offset on 4K sector disks; requiring THEM to do do a read whole sector+calc new ECC+write whole sector. Thus this behavior is devasting on performance on 4K sector drives with 512-byte emulation; each single write request issues to the vdev will cause it to perform 512-byte sector emulation.

## Re: WD advanced format 4k drives - definitive answer?

by **striken4f** Posted: **October 24th, 2012, 4:59 pm**

Fsbruva -

Thanks so much for the detailed response. Helpful and much appreciated.

## Re: WD advanced format 4k drives - definitive answer?

by **fsbruva** Posted: **October 24th, 2012, 5:32 pm**

> ***striken4f wrote:***
> Fsbruva -
>
> Thanks so much for the detailed response. Helpful and much appreciated.

What are ~~friends~~ forums for? 
Another thing (before you get too far down the path...)
Make sure you enable smart monitoring, and then set the power management to 254 (max performance). This will keep the drive heads from parking all the d@mn time, which can lead to premature failure. This method uses ataidle, and works for my Toshiba drives. Check out some posts/howtos on wdidle, which does the same thing, I'm lead to believe.

## Re: WD advanced format 4k drives - definitive answer?

by **fsbruva** Posted: **October 25th, 2012, 7:56 am**

Another thing (before you get too far down the path...)
Make sure you enable smart monitoring, and then set the power management to 254 (max performance). This will keep the drive heads from parking all the d@mn time, which can lead to premature failure. This method uses ataidle, and works for my Toshiba drives. Check out some posts/howtos on wdidle, which does the same thing, I'm lead to believe.

## Re: WD advanced format 4k drives - definitive answer?

by **dboy** Posted: **October 25th, 2012, 5:21 pm**

I installed FreeNAS 0.72 (8191) about 18 months ago, I think I started with an earlier version and upgraded about a year ago...
4*2TB WesternDigital Advanced Format Drive disks in RAIDZ ZFS pool.

Did a fair bit of research and tried my best to get it right... but think I am still working through the GNOP layer.

I typically never get above 7 MB per second.
I seem to be about the same performance via the network as if I transfer files locally between the zfs pool and an USB2 drive I use for backup.
(I use rsync for backups)

I think this is pretty bad for a disk to disk transfer, but probably acceptable for a standard 100Mbps network with a theoretical max speed of 100 Mbps/8=12.5 MB/s.
Is this likelly to be due to the Advanced Formad Drives?
I dont hit the roof CPU and RAM wise, 4 GB RAM, fast enough CPU.

## Re: WD advanced format 4k drives - definitive answer?

by **fsbruva** Posted: **October 26th, 2012, 7:41 am**

> ***dboy wrote:***
> I installed FreeNAS 0.72 (8191) about 18 months ago, I think I started with an earlier version and upgraded about a year ago...
>
> Did a fair bit of research and tried my best to get it right... but think I am still working through the GNOP layer.

Wait - are you using 0.7? Or are you using NAS4Free?

If you use N4F, the gui can be used to check, otherwise you have to use a commandline.

From a commandline:

```
zpool status
```

You should see something akin to:

```
  pool: zpool
 state: ONLINE
  scan: scrub repaired 0 in 0h42m with 0 errors on Tue Oct 23
14:00:09 2012
config:

   NAME        STATE     READ WRITE CKSUM
```

```
   zpool       ONLINE       0     0     0
     mirror-0  ONLINE       0     0     0
       ada2    ONLINE       0     0     0
       ada1    ONLINE       0     0     0

errors: No known data errors
```

See how the members of the pool are ada2 & ada1? If you were working through the gnop layer, you would see ada2.nop and ada1.nop. (I realize this is a simple mirror, and you're using RAIDZ1 - we're only concerned with the members of the vdev)

> ***dboy wrote:***
> I typically never get above 7 MB per second.
>
> Is this likelly to be due to the Advanced Formad Drives?
> I dont hit the roof CPU and RAM wise, 4 GB RAM, fast enough CPU.

I doubt it has anything to do with the drives. 75% of theoretical network capacity is pretty darn good - and just because it's a USB2 disk, that doesn't necessarily imply that your motherboard is USB 2.0.

Also, other possible issues (without getting too far afield):
1. Even though you have 4GB of RAM, there are kernel tuning parameters that can improve the performance/stability of ZFS.
2. Are you running a 64 bit system? If not, the 32 bit kernel will only allow for 512 MB (12.5%) of ram to be allocated to ZFS. Most guides I have read recommend more than that.

## Re: WD advanced format 4k drives - definitive answer?

by **al562** Posted: **December 14th, 2012, 11:57 pm**

Looks like I just found that FAQ the OP was looking for . Topic trimmed, moved, locked & added to FAQs.
Thanks everyone.

Regards,
Al