ZFS Storage Design and other FreeNAS information

Written by Cyberjock on the FreeNAS forums.

This presentation was last updated May 2, 2016.

The latest version of this guide can always be found at http://forums.freenas.org/threads/slideshow-explaining-vdev-zpool-zil-and-l2arc-for-noobs.7775/

Any hardware manufacturers mentioned in this presentation should not be taken as my endorsement of that manufacturer's products over any other manufacturer or product. It is simply to provide information for the reader.

Purposes of this presentation

- Explain the relationship between hard drive, VDev and zpool.
- Explain how to expand a zpool due to increasing storage demands.
- Explain what a ZIL and L2ARC does and how it can be useful.
- Provide some other useful information for newbies trying to get familiar with FreeNAS/FreeBSD.

Preface

Before I start I want to explain where I'm writing this presentation from.

Until March 2012 I had more than 20 years with Windows/DOS exclusively. I'm heavy on both hardware and software usage and optimizations. As such, I'm your typical advanced and experienced Windows user/admin.

When I chose to devote myself to figuring out FreeNAS because I was not happy with Windows file systems. I spent a month of 12 hour days reading forums, experimenting with a VM, and then later a test platform.

I learned a lot about the differences between FreeBSD and Windows. Some are subtle but critical to safely storing your data. Some are obvious but minor.

Much of the information in this presentation is written based on my 18000+ posts on the forum, reading almost every post that has been posted to the FreeNAS forum since I joined the forums in March 2012 and from friendships that I've made on the FreeNAS forum with users, moderators, and developers.

Preface

- This presentation is written for the vast majority of users that are where I was in March 2012. The intent is for me to explain what stuff is important in the FreeBSD world and apply the experiences and knowledge I've gained from reading the forums since I joined.
- Stuff that is often unimportant in the Windows world is extremely critical in FreeBSD(and often Linux as well). One of the worst things you can do is start assuming some warnings and recommendations should be ignored because you've been doing the opposite in Windows and never had a problem. Windows ≠ FreeBSD/FreeNAS. Keep in mind FreeBSD ≠ Linux.
- Please recognize that your Windows hardware knowledge may provide some small insight for selecting hardware but is not equivalent to expertly choosing hardware for a FreeBSD based system. For example, ECC RAM in a desktop isn't too useful. But for ZFS it can be the difference between saving your data and a complete loss of the zpool with no chance for recovery. Realtek NICs are common in the Windows world, but perform extremely poorly (or not at all!) in FreeBSD.
- Take warnings from this presentation seriously. For the most part, they are in this presentation because so many people have had so many problems (many have lost all of their data).
- There's nothing I hate more than seeing threads of people that lost their zpool. So do me a favor(and yourself) and read through the material I link to. ☺

Preface

- You will almost certainly need to learn how to navigate the FreeNAS/FreeBSD command line. Being an expert isn't necessary. But you will need to feel comfortable running commands and changing directories, setting permissions for files and folders from the command line, and checking SMART data on your hard drives.
- If learning the command line isn't for you, you may want to consider an alternative storage solution.

Preface

- FreeBSD has a very steep learning curve. It is not for those looking to learn it in a weekend. I was operating a nuclear reactor before I was old enough to drink alcohol, but I still spent a solid month getting familiar with FreeNAS. Everyone's mileage will vary and your level of "comfort" with FreeNAS will be different than mine. But I haven't lost any data and I have helped many people recover lost data.
- FreeNAS has done some amazing things to simplify access to a file server's most commonly used features while utilizing FreeBSD's amazing power and feature set. Not every feature is available from the WebGUI. In those cases you may need to resort to using the command line to achieving your desired objective. Be careful though, as anything done from the CLI is "behind the WebGUI's back". This can result in serious problems. In short, if you are planning to do something from the CLI seriously consider what you are doing and if you can do it from the WebGUI. If the WebGUI can do it, you better be doing it from the WebGUI. Another way to look at it is if the FreeNAS manual tells you how to do something, don't try to do it any other way.

Preface

- Your dedication to learning how to use the command line and setup FreeNAS with the proper features to monitor your hard drives, your UPS, and your hardware will determine how safe your data is.
- If you choose to do the bare minimum learning, your data is not as safe as someone that has spent a few weeks to become thoroughly familiar with FreeNAS and custom scripts.
- If you choose to spend the necessary time to get familiar with FreeBSD command line and scripting you can have a very safe and reliable storage location for your data.

Preface

Many people choose not to use all of the features of FreeNAS. Features in FreeNAS that I consider absolutely critical to safely storing your data include:

- Set up FreeNAS emails.
- Set up SMART monitoring. This also requires your IDE/SATA controller support SMART. Many "RAID" controllers do not support SMART properly, even in JBOD. You should verify SMART works before using a controller for the long term. The SMART service allows you to include an email address to inform you of problems. Use it! Also don't ignore the emails! Verify you can get all SMART data from your disks before assuming it works.
- Setup the UPS service(and use an UPS). Many people have corrupted their zpool because of an improper shutdown. You should try to minimize the possibility of improper shutdowns. It also supports emailing. Use it!
- Schedule regular scrubs. Typically bi-weekly to monthly is a good place to set your schedule.
- SSH is very handy. SSH is nothing more than a remote command line interface. This allows you to easily log into your server and run commands from the command line. If you have problems the forum will expect you to copy and paste from the SSH window.

What is a hard disk?

- A hard drive stores data. Common sizes are in GB(gigabytes) and TB(terabytes).
- Hard drives use ECC(Error Correction Coding) internally to help protect data from being corrupted due to magnetic domains in the media flipping when media is no longer reliable.
  - Even with ECC, data can become corrupted. This is one area where ZFS can and does help. ZFS can detect and correct for “silent corruption” by using parity data and mirrors.

What is a VDev(Virtual Device)

- A VDev is one or more hard disks that are allocated together and are intended to work together to store data.
- VDevs are allocated in RAID formats such as Mirrors(RAID1 equivalent), RAID-Z1(RAID5 equivalent) and RAID-Z2(RAID-6 equivalent).
- VDevs with single disks are known as “striped” disks. They have no redundancy.
- VDevs can provide redundancy from individual hard disk failure inside the same VDev.
- VDevs cannot operate outside of a zpool.

What is a VDev(Virtual Device)

- You cannot add more hard drives to a VDev once it is created.*
- When a VDev can no longer provide 100% of its data using checksums or mirrors, the VDev will fail.
- If any VDev in a zpool is failed, you will lose the entire zpool with no chance of partial recovery. (Read this again so it sinks in)

* The only exception is converting a VDev from a single disk to a mirror. Consult the FREENAS manual for more information.

What is a zpool?

- A zpool is one or more VDevs allocated together.
- You can add more VDevs to a zpool after it is created.
- If any VDev in a zpool fails, then all data in the zpool is unavailable.
- Zpools are often referred to as volumes.
- You can think of it simply as:
  - Hard drive(s) goes inside VDevs.
  - Vdevs go inside zpools.
  - Zpools store your data.
  - Disk failure isn't the concern with ZFS. Vdev failure is! Keep the VDevs healthy and your data is safe.

Explanation of ZIL

- ZIL stands for ZFS intent log.
- The ZIL stores data that will need to be written to a zpool later and acts as a "non-volatile write cache" for the zpool.
- ZIL drive performance will need to be very fast for writes (read speeds do NOT matter). Typically an SSD is used for this application. An Enterprise class SSD or SSD based on SLC memory is recommended.
- Whatever SSD you plan to use should have it's own battery backup to ensure disk writes are completed even if power is lost suddenly. This often means that consumer SSDs are NOT suitable for slogs.
- FreeNAS does use RAM as a short term write cache for non-sync writes, so a ZIL is not always required for high performance.
- ZILs should have their own redundancy from drive failure in environments where zpools need redundancy. It is highly recommended that ZIL drives be operated in a mirrored mode to prevent data loss for this reason. Otherwise a failure of the ZIL will result in a loss of all data that is in the ZIL and not committed to the zpool.
- ZIL is only useful for sync writes. For all other writes the ZIL is not used. This means if you aren't using NFS and don't have a workload that uses sync writes, a ZIL is pointless.
- For 99.9% of home users, a ZIL will not be useful.
- ZILs are typically very small. Just 1-2GB of ZIL storage for a server with Gb LAN is overkill. This allows the SSD to use its built-in wear leveling mechanism to maximize reliability and performance(Intel only).

Potential ZIL issue

- Failure of a ZIL drive (unless using mirrored ZIL drives) will prevent the zpool from mounting on bootup. This is because part of the mounting process checks the ZIL for uncommitted transactions and commits them before mounting the zpool.
- FreeNAS 9.1.0+ uses ZFS v5000. This is a "fork" of the original ZFS versioning system used by Oracle and is now referred to as OpenZFS.
- If you are creating a new zpool using FreeNAS 9.1.0+ you will not have to worry about this issue since the zpool will be v5000 or higher.
- If you are using an older zpool, consider upgrading to avoid bugs and issues that may exist with older versions.

Explanation of L2ARC

- L2ARC stands for Level 2 Adaptive Replacement Cache.
- The L2ARC is a read cache for the zpool. Note that it is not a read-ahead cache. L2ARC is used for random reads of static data(i.e. databases) and provides no benefit for streaming workloads. It is also only useful when the same data is constantly read over and over.
- L1ARC, or often referred to as simply "ARC" typically uses a significant portion of available RAM on the FreeNAS server(usually around 85%). This is most commonly your read cache.
- The L2ARC stores frequently read data that exceeds the amount of RAM assigned to the ARC.
- SSDs are the primary device used for these functions.
- Failure of the L2ARC will NOT result in a loss of data, but you will lose any performance advantages from using the L2ARC. For this reason, mirroring is generally not recommended.
- Using a L2ARC will consume RAM from the ARC to maintain records of the L2ARC. Because of this, you should be spending money to max out your motherboard's RAM before considering an L2ARC. If you do not have enough RAM, using an L2ARC can result in a decrease in performance. The bottom line is if you don't know if you need/want an L2ARC, you probably don't. Likewise if you have never heard of an L2ARC before this presentation you probably don't need it. Generally, until you hit a minimum of 64GB of RAM, you should not consider an L2ARC. This has to do with how much RAM you have in relation to your L2ARC size.
- Maxing out your system RAM is almost always better than using an L2ARC. Especially since using an L2ARC will consume RAM to index the L2ARC.
- An L2ARC shouldn't be bigger than about 5x your ARC size. Your ARC size cannot exceed $7/8^{th}$ of your system RAM. So for a system with 32GB of RAM, you shouldn't go any bigger than 120GB. This is why maximum system RAM first is a priority!
- Keep in mind that if your L2ARC is too big for your system your performance may actually decrease!

ZIL and L2ARC(more)

- A lot of people in the forum are using a ZIL and L2ARC to increase performance. This is not usually necessary, especially for home users.
- Typical reads from your zpool shouldn't take more than 10 milliseconds, except under EXTREME loading. In this case, using an L2ARC(and assuming the data you are requesting is in the L2ARC) you will save yourself at best 10 milliseconds.
- Typical writes to your zpool are usually cached in RAM unless a sync write is requested(buffer flush to the zpool). Typical write latency should be less than 10 milliseconds plus the time it takes to write the data. So using a ZIL will really save you only 10 milliseconds plus the time it takes to write the data(usually microseconds).
- Because of these reasons, ZILs and L2ARCs are not overly useful and only add cost, complexity, and can add complexity to data recovery if you have issues in the future.
- In general, if you aren't using a zpool to serve iSCSI devices, using an ESXi server to access FreeNAS storage with NFS for the ESXi store, then a ZIL and L2ARC likely provide very little benefit.
- I recommend potential builders (and especially inexperienced builders) never buy or install an SSD for use as a ZIL or L2ARC until they build the server and determine that they actually need a ZIL or L2ARC. FreeNAS includes some tools to help determine if an L2ARC or ZIL could provide any benefit.
- Do not put your ZIL and L2ARC on the same SSD(s). It sounds like a great idea on the surface, but there is plenty of evidence it doesn't work out well and both the ZIL and L2ARC compete for resources causing both to perform poorly.

Keynotes:

- One or more hard disks make up a VDev.
- One or more VDevs make up a zpool.
- A loss of the ZIL drive(s) can result in data loss due to the data that is not committed to the zpool.
- L2ARC drive failure will not cause a loss of data.
- You cannot add more hard drives to a VDev once it is created.
- Once you add a VDev to a zpool it cannot be removed for any reason. Even if you "just" added it because of a typo at the command line or a mistake in the UI.
- Once a zpool has been upgraded your OS must be compatible with the new version to use the pool.

Example 1:

- This zpool provides redundancy against any 2 simultaneous hard disk failures. Any 2 hard disks can fail with no loss of data.

ZPOOL

VDEV (RAID-Z2)

Example 2:

- This zpool provides redundancy against a maximum of 4 simultaneous hard disk failures(2 in each VDev), but not 3 in any one VDev.

ZPOOL

VDEV 1 (RAID-Z2)

VDEV 2 (RAID-Z2)

Example 3:

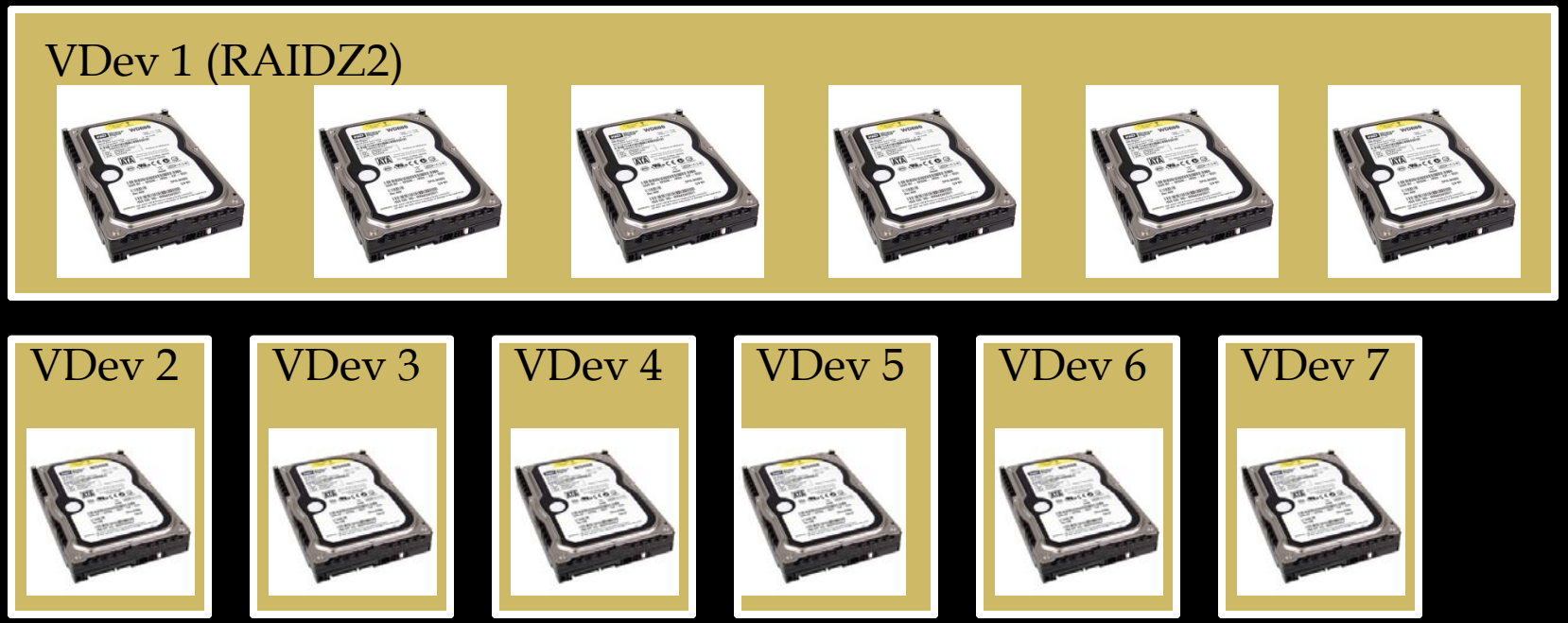
- This zpool provides no redundancy. A failure of any hard disk will result in complete data loss of the zpool.
- A failure of any disk will make its VDev unavailable which will cause a loss of the entire zpool.
- This is a good example where a failure of a VDev results in a loss of the entire pool.

ZPOOL (striped)

Example 4:

- This zpool provides partial redundancy. The zpool can withstand up to 2 hard disk failures in VDev 1, but a failure of any hard disk on VDev 2,3,4,5,6 or 7 will result in a loss of all data.
- This is NOT a recommended configuration because of the single points of failure.

ZPOOL



Expanding a zpool.

- ZFS allows for a zpool to expand in only two ways.
  - Option 1: Replace all of the hard disks in a VDev with larger hard drives (aka autoexpand)
  - Option 2: Add additional VDevs.

Let's discuss each option:

Replace all of the hard disk in a VDev with larger hard drives.

- A VDev can be expanded to include larger drives by replacing the drives one at a time by using the FreeNAS GUI. Consult the manual for instructions on how to replace each drive without a loss of data.

NOTE: Do not unplug a hard disk that is part of a VDev without using the GUI to inform the operating system that a drive will be removed. Failure to properly replace a hard drive may result in a loss of data or system crash.

Example 5

- Replacing all of the drives in a VDev will increase available storage space.
- This method is useful if you want to buy 1 new disk a month. Keep in mind that the zpool will not grow until you have replaced all of the disks in a VDev.

ZPOOL (8TB usable)

VDEV (RAID-Z2)

Add additional VDevs

- More hard disks can be added to a zpool by adding another VDev.
- The new VDev can use a different size hard drive and even a different number of hard drives than the old VDev. It is recommended that they have the same number of disks though.
- The new VDev does not have to have the same characteristics(mirror,Z1,Z2, etc.), however it is highly recommended.
- VDevs cannot be removed from a zpool after they have been added, even if you "just" added it on accident. You should be absolutely 100% sure you are doing what you want to do before adding VDevs. If you aren't sure, test it in a VM to be sure you understand what you are doing. There is no undo!

Example 6

- Click to watch the animation.

ZPOOL (10TB Usable)

VDEV 1 (RAID-Z2)

VDEV 2 (RAID-Z2)

Example 6

- The zpool now provides up to 4 hard disk failures (up to 2 in each VDev) but not 3 in any one VDev.

ZPOOL (10TB Usable)

VDEV 1 (RAID-Z2)

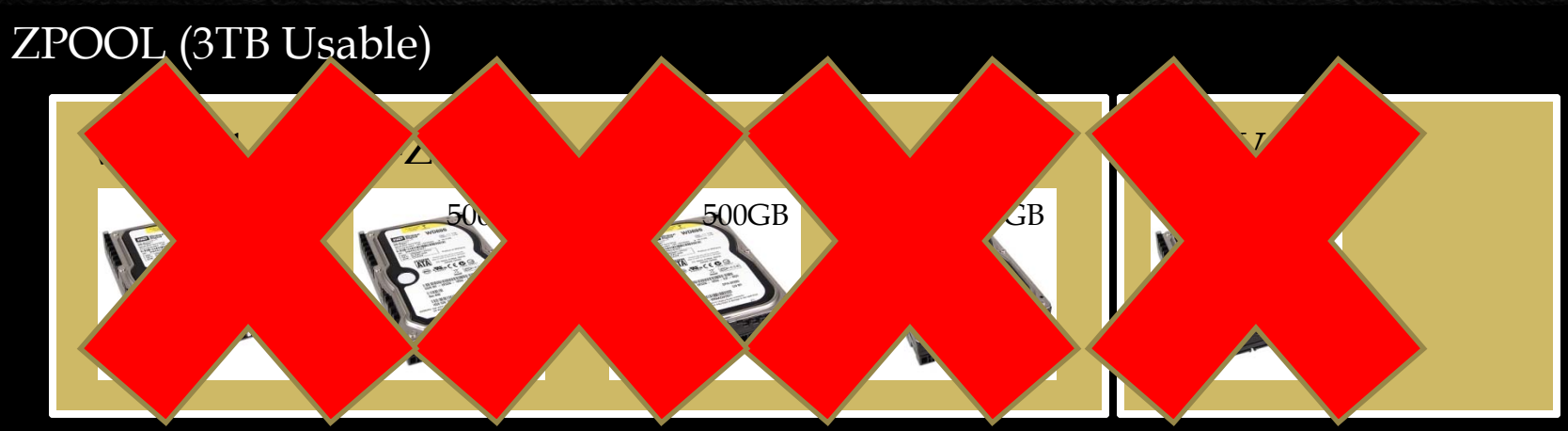
VDEV 2 (RAID-Z2)

Adding VDevs

- You should not add individual hard disks by creating VDevs with 1 hard disk if redundancy is desired. Redundancy will be lost.
- This is a valid configuration for FreeNAS, although it is obviously not recommended.
- Many people have accidentally done this thinking they could add a single disk to a VDev. Later when they find out the mistake they made, it is often because they lost their pool and don't understand why.

Adding VDevs

- In this case, if the 2TB drive(VDev 2) were to fail, then all data in the zpool would be lost.
- FreeNAS does NOT protect you from errors such as this. It is left up to the administrator to properly manage the zpool.



Things to think about

- Failure of a VDev in a zpool will cause the zpool to fail. This means all data in the zpool is lost. There is no chance of recovering data from the remaining VDevs and there are no recovery tools for ZFS.
- FreeNAS will allow configurations that remove redundancy. The administrator is expected to ensure redundancy by proper management of the zpool on his own. Improper management may cost you your data in the future.
- Hard disks put in a VDev cannot be removed- only replaced and only if you have enough redundancy to support the disk replacement.
- VDevs added to a zpool cannot be removed under any circumstances.

Encryption

- Encryption of zpools is supported using geli. Note that the encryption used by FreeNAS is not compatible with Oracle's ZFS v30 implementation of encryption. Oracle's ZFS v30 is not open source and therefore is not expected to be implemented into FreeBSD in the future. Additionally, if you use encrypted drives on FreeNAS you will not be able to easily(if at all) mount them under any other operating system, except FreeBSD.
- It is imperative you maintain a backup of your keys and passwords. Loss of your key or password will result in you not being able to access your data.
- Geli uses AES-XTS 128-bit encryption. FreeNAS recommends using an AES-NI CPU for encryption if you wish to use encryption. CPUs without AES-NI suffer a significant performance penalty. AES-NI supported CPUs will see an unnoticeable performance penalty. In fact, benchmark tests performed were unable to find the performance limit with AES-NI supported CPUs. Only some Intel CPUs manufactured after January 2010 and some AMD CPUs manufactured after Oct 2011 have this feature set. Consult the manufacturer's website to determine if your CPU supports AES-NI. Speeds of less than 20MB/sec are often observed for those that choose to use encryption without a CPU that supports AES-NI.
- Note that any ZIL or L2ARCs used with encrypted zpools are also encrypted.

Installing programs in FreeNAS

- Programs and files should not be installed to the FreeNAS USB stick.
- The available space on the USB stick is also very small(just a few MB). FreeNAS needs this space, so don't assume you can use it.
- If you do want to install programs in FreeNAS/FreeBSD the jail is designed to provide this function. It is a minimal installation of FreeBSD and should allow you to install anything you like. Using the jail will require you be familiar with the FreeBSD command line and you will be responsible for managing updates of the jail appropriately to ensure there are no security risks.
- Consult the manual for instructions on installing and using the jail.

Performance of zpools…

- For maximum performance and reliability, you should never try to use ZFS with less than 8GB of RAM and a 64-bit system under any circumstances, regardless of the size of your pool. We typically see at least 1-2 people every week that break this rule and they lose their pools suddenly. There is no recovery if you damage your pool and ignoring this warning.
- It is not recommended that VDevs contain more than 11 disks under any circumstances.
- FreeNAS' ZFS stripes data in chunks up to 128 kilobytes. If you will be using compression (default is enabled/lz4 with FreeNAS) then the following slide does not apply. Compression adds a layer of complexity because each block of data will be compressed to some arbitrary smaller size, so planning for ideal block allocation is impossible.
- LZ4 compression, which is enabled by default, is considered to be a "free lunch". It is extremely fast and should not impact the performance of your system. For most people, it will result in a performance improvement because less data will have to be read from the relatively slow platter disks. For this reason you should not disable compression.
- For home users in particular, your bottleneck is certainly going to be your Gb NIC and not your pool unless you are using woefully inadequate hardware for FreeNAS.

Performance of zpools…

- If you are not using compression:
  - For performance reasons it is preferred that you use these conventions when creating RAIDZ1, RAIDZ2 and RAIDZ3 VDevs. (n is any whole number you want….1, 2, 3 etc.)
  - RAIDZ1 should have the total number of drives equal to $2^n + 1$. (ie 3, 5, 9, etc. drives for the VDev)
  - RAIDZ2 should have the total number of drives equal to $2^n + 2$. (ie 4, 6, 10, etc drives for the VDev)
  - RAIDZ3 should have the total number of drives equal to $2^n + 3$. (ie 5, 7, 11, etc drives for the VDev)
  - This is to ensure ZFS stripes fall on the 4k sector boundary of newer generation hard drives.
  - If you do not intend to transfer large quantities of data constantly this thumbrule can be totally disregarded. For instance, if you intend to use your server to stream DVD rips of your collection, this thumbrule can be ignored.

Additional thoughts for noobs…

- For maximum performance, more RAM is always better. Maximum allowed motherboard RAM is recommended. The resultant improper shutdowns of your system can result in a corrupted zpool that will no longer mount. Don't be a statistic! Use 8GB+ of RAM with ZFS or use a different NAS solution. I don't even bother responding to posters that can't have the minimum hardware and follow recommendations in the stickies.
  - This is a fairly common occurrence for many forum users that are new to FreeBSD/FreeNAS and think they can "get by" with less RAM because they are using hardware that doesn't support 8GB of RAM, 8GB of RAM costs too much, or they are convinced their workload isn't demanding. There is no exception to the 8GB of RAM minimum rule. Some people get lucky and have no problems, others have lost their zpools. I'd never risk my data by using less than 8GB of RAM and I recommend the same to you.
- If you want to use ZFS with less than FreeNAS' recommended RAM you should consider the NAS4Free project. NAS4Free recommends 1GB per TB of storage without the baseline 8GB. So if you have 5TB of storage you'd need 13GB of RAM for FreeNAS, but only 5GB for NAS4Free. Note that using NAS4Free does not alleviate the risk of improper shutdowns crashing your system. So don't lowball your RAM needs.
- ZILs and L2ARCs have very specific uses. Money spent on SSDs for ZILs or L2ARCs are almost always better spent on RAM.
- For most home users just sharing some files and perhaps some plugins/jails, 16GB of RAM is an excellent place to start. By far most home users will have a stable and fast FreeNAS server with 16GB of RAM.

ZFS versions

- ZFS went closed source with v29 by Oracle after buying Sun (the original creators and maintainers of ZFS). It is not expected that any versions after v28 will ever be open source. The ZFS code has continued as an open source fork by the OpenZFS foundation.
- FreeNAS 9.1.0+ uses ZFS v5000. This is basically v28 with support for flags(features).
- Each newer FreeNAS version can bring new feature flags. An alert will appear in the WebGUI when you can upgrade to new feature flags.

Additional thoughts for noobs…

- Always keep in mind the possible need to expand in the future. Will you want to add more drives later, buy bigger drives to replace your current drives or will you want to build a more modern system in the future?
- Intel Network cards are the NIC of choice. The drivers are well maintained and provide excellent performance(not to mention inexpensive). Other NICs have been known to perform intermittently, poorly, or not at all (especially Realtek). Using "low power" CPUs such as Intel Atoms and AMD C-70s are NOT powerful enough to be used with Realtek and get good performance. If you want a good laugh(or a realization of how bad Realteks are) read the driver notes at http://people.freebsd.org/~wpaul/RealTek/3.0/if_rl.c.
- Some users have complained about slow performance with FreeNAS that was narrowed down to their desktop's NIC. If you want very high speeds(90MB/sec+) you may want to consider Intel NICs for your desktop(s) as well. I run Intel NICs on all of my desktops and servers.
- Wifi was never meant to be fast as it is a convenience technology. So expecting more than 25-30MB/sec through your wifi is generally unrealistic.
- Samba is single threaded on a per-user, per connection basis. Samba provides the "Windows File Sharing" services. If Samba is used it is better to use a higher clocked CPU than one with more cores. There is a balance between more cores and more clock speed however. It is important to choose a CPU that hits that "happy medium" for your exact situation. Generally, any CPU with 2+ cores at 2.5Ghz+ will perform decently with FreeNAS.
- Make sure your on-board SATA BIOS settings are set to AHCI.

Additional thoughts for noobs…

- FreeNAS works great in a virtual machine such as VMWare ESXi, VMWare Workstation and VirtualBox if you want to experiment without dedicating hardware. Do not trust these with real data however!
- Manually optimizing the system settings should not normally be necessary as long as basic system requirements are met. In particular plenty of RAM.
- If you have issues with FreeNAS not operating as expected it is recommended you try upgrading to the latest RELEASE version. Many bugs are fixed with each release. Not all bug fixes are listed in the release notes. You can add yourself to the mailing list if you would like an email when a new version is released.
- Always keep a backup of your FreeNAS configuration (see the FreeNAS manual). Especially before you update FreeNAS.
- RTFM (Read the freaking manual). There are a lot of thumbrules, tips, and tricks in the manual that will help your first build be a smooth process. Generally, questions that are blatantly answered in the manual are often ignored in the forum. A lot of ideas you'll want to try to improve performance will likely be mentioned in the manual with examples. Try to solve it yourself before posting to the forum. The manual can be found at doc.freenas.org.
- Have I mentioned you can never have too much RAM?

Additional thoughts for noobs…

- Virtualization is strongly recommended against for FreeNAS except for experimenting. Many of the self-healing properties of ZFS as well as FreeNAS' hardware monitoring may not function in certain scenarios inside of a virtual machine. Read the thread at http://forums.freenas.org/threads/please-do-not-run-freenas-in-production-as-a-virtual-machine.12484/ . That thread was written because of how many people have lost all of their data due to virtualization. Don't be a statistic!

Additional thoughts for noobs…

- ZFS has very few "recovery tools" unlike many other file systems. For this reason, backups are very important. If the zpool becomes unmountable and cannot be repaired there are no easy software tools or reasonably priced recovery specialists you can use to recover your data. This is because ZFS is enterprise-class software, and no enterprise would waste their time with recovery tools or data recovery specialists. They would simply recover from a known good backup or mirror server.

Additional thoughts for noobs…

- FreeNAS .7 is not the same as FreeNAS 8+. FreeNAS 8+ is a complete rework from scratch while FreeNAS .7 was renamed to NAS4Free. As such the FreeNAS forums provide no support whatsoever for FreeNAS .7. For support with FreeNAS .7 you should visit the NAS4Free forums. Expect to get BBQed in the forum if you post a question about FreeNAS .7.

Additional thoughts for noobs…

- If you aren't an advanced user you should stick to the RELEASE versions of FreeNAS. The Nightly, Alpha, Beta, and Release Candidates are geared towards advanced FreeBSD users that are interested in testing and troubleshooting FreeNAS. I never use anything except RELEASEs on my production servers and only after they've been released for a week or two just to be sure there aren't any major issues based on forum threads. FreeNAS developers are excellent, but every human being makes mistakes. I'd rather not let a mistake cost me my data.
- If you want to test the newer versions do it on a test machine or a virtual machine. (I do!)

Additional thoughts for noobs…

- If you have a problem where your zpool is not mounting:
  - Be very careful what commands you run.
  - Do lots and lots of homework. Many of the commands you'll find on Google are one-way streets and can cause permanent zpool damage if run improperly. Most "guides" for ZFS recovery will NOT warn you of potentially damaging commands.
  - Forcing mounting with the –f or –F is undoable and can cause permanent damage. Do not willy-nilly use the force mount function. If you are having to do this there is probably something wrong, hardware or software.
  - Plenty of users have lost everything because they ran the wrong commands.
- You did make backups, right? You should always backup your irreplaceable data.

Additional thoughts for noobs…

- NTFS is supported in FreeNAS(and FreeBSD). However it is not very reliable and shouldn't be used long term. FreeNAS does not support mounting dynamic NTFS partitions from the GUI.
- FreeNAS is not for everyone. Many people are not able to spend the money on proper hardware. Be 100% sure you are ready to jump into this before starting. If you don't have the time, patience, or money to spend to do a FreeNAS system correctly, you may be better off sticking with what you are familiar with.

Additional thoughts for noobs…

- ZFS is impervious to corruption by design (with sufficient redundancy). Because of this the only "repair" tool is to perform a scrub. It does not fix just file system errors or just file data, it fixes ANY issues that exist using redundancy.
- ZFS relies on redundancy to ensure that corruption is corrected. Because of the design basis for ZFS, it is imperative that you always ensure that sufficient redundancy exists. This is one reason why RAIDZ1 is dangerous. Most users that lose their data had a RAIDZ1 with the remainder being human error or server neglect.
- Naturally, if your RAM is bad you will have corruption of the data stored in the bad RAM locations. And obviously redundancy from bad RAM is not possible, but the next best thing is to have error detection and correction. Enter ECC RAM…

ECC RAM

- ECC RAM(Error Checking and Correcting) RAM is only supported only on some motherboards and some CPUs.
- ECC RAM has typically been left to server class hardware, so most left-over desktops that are repurposed to a FreeNAS server(and virtually all mini-ITX boards) do not support ECC RAM.
- Check with your motherboard and CPU vendor to determine if ECC RAM is supported and that it will provide the ECC function.
- If you choose not to use ECC RAM with ZFS, irreparable damage can occur to the pool despite having redundancy. Because of this using non-ECC RAM is a single point of failure for ZFS.
- Scrubs can completely destroy a zpool that is healthy because bad RAM will cause errors that don't exist to be "corrected" with corrupt data.

ECC RAM (continued)…

- Backups will also be destroyed because the original data itself is damaged.
- This is a limitation of ZFS' design and not FreeNAS. ZFS manuals from Sun (and Oracle) as well as other projects that use ZFS (ZFS on Linux, NAS4Free, etc.) also have this vulnerability and recommend ECC RAM.
- Users in the forums that have used non-ECC RAM have suffered complete pool data loss without recovery if they had bad RAM. The problem is you will not have an indication that your RAM is bad until extensive pool damage has occurred. Most lost all of their backups because the bad data was replicated from the zpool to the backup site. Choosing to use rsync or ZFS replication for backups does not alter the outcome because the original data will be destroyed.
- I would never recommend anyone build a FreeNAS server with non-ECC after seeing the blood, sweat, and tears from the forum users that have lost their data. The risk (total data loss) just isn't worth the reward (saving a few bucks). Don't be a statistic!
- Most people will recognize that RAM doesn't fail that often. So choosing to trust that you won't have bad RAM while using ZFS with non-ECC is a risk you will have to decide on for yourself if choosing to not use ECC RAM and appropriate hardware.

External disks

- Using external disks (USB, Firewire, or eSATA) is not recommended. Many would call this a "recipe for disaster".
- External disks are prone to accidentally being unplugged by bumping the power or data cable. Plenty of users that swore they would be safe with their server in the basement where its locked and in a corner have still made mistakes and lost data.
- Most external disks on a single error of any kind will disconnect and reconnect to the host. This will result in data loss and you may not get any warning that this is going on until it is too late.
- USB and Firewire do not allow for proper read/write error detection.
- USB and Firewire usually do not allow for proper SMART monitoring, reporting or testing. This removes one of your main methods for identifying disk problems early.
- For these reasons, using a laptop is not a good option for FreeNAS because you will have to rely on external disks.
- Plenty of users have lost significant amounts of data by using external disks despite these warnings. Don't be a statistic!
- Laptops also make poor choices because quite often the hardware is customized to the point where there are no FreeBSD drivers for the hardware, and you can't just install a new network card or SATA controller. Even for testing they are pointless as they will not behave properly or effectively to give an accurate assessment of a FreeNAS system.

iSCSI on ZFS

- There are potential performance pitfalls with using iSCSI on ZFS. Excessive fragmentation can result because ZFS is a copy-on-write file system.
- If using iSCSI on ZFS, searching the forums will find some lessons learned by other users for iSCSI on ZFS.
- You can expect that the issue will not be resolved quickly by just making a few changes and a reboot.
- Most users will find that they will spend a month or more of intensive research and testing to resolve performance issues on iSCSI if you have never tuned ZFS before.
- You can expect to have very high server hardware requirements if you use a lot of iSCSI devices.

ESXi datastore on ZFS

- Many users choose to store their ESXi datastore on a zpool.
- Just like with iSCSI, there are unique obstacles to using ESXi's datastore on FreeNAS because of the copy-on-write nature of ZFS.
- See https://support.freenas.org/ticket/1531 for some tips and tricks.
- Just like with iSCSI on ZFS do not expect that the solution is obvious and easy. It will require you to spend significant time tweaking your system to obtain acceptable speed or choosing to use settings that can be hazardous to your data.
- Most users will find that they will spend a month or more of intensive research and testing to resolve performance issues when using ZFS and NFS for an ESXi datastore if you have never tuned ZFS before.
- You can expect to have very high server hardware requirements if you use a lot of VMs on a zpool. Read up in the forum as there are dozens of threads on this topic.

Link Aggregation - LACP

- LACP doesn't work quite how you might think. You do NOT end up with the ability to download files at 2Gb/sec to your desktop… even if you do LACP on your desktop too!
- It also requires your network switches support LACP. Typically only enterprise-class network switches support LACP.
- Generally speaking, unless you have lots of workstations and users that are simultaneously using large quantities of data(aka businesses), LACP will not provide a performance boost.
- If you are new to LACP you shouldn't bother trying to setup LACP. You won't see a performance benefit but you may spend many hours trying to get it to work (assuming your hardware supports LACP)
- You do not need LACP to get excellent performance. I regularly get over 100MB/sec without LACP.

Jumbo Frames

- Jumbo packet setup is commonly attempted by newbies. The promise of free performance improvement entices many newbies.
- All of your network devices must be set to the same MTU(including things such as network printers and routers) in the given network segment.
- Some brands of NICs allow you to set the MTU to 9000, others to 9014. In many cases this also means that those machines are not set to the same size. Because of this most users that haven't bought their network gear specifically to support a particular jumbo frame size will have significant network problems.
- Jumbo frames will not significantly increase the performance under normal conditions and you will find that a well designed FreeNAS machine should be able to achieve excellent speeds without Jumbo Frames.
- For these reasons jumbo frames should not be attempted by newbies. Again, I get over 100MB/sec without jumbo frames and you surely can too.

Long term server maintenance

- For long term reliability and data protection, it is recommended you setup emails from your FreeNAS machine as well as setup SMART monitoring(if supported by your hard disk controller).
- If SMART monitoring is properly setup and your hardware supports SMART the server will email you when a disk is having a problem so you can preemptively replace disks that are failing.
- It is possible to setup scripts to email you nightly (or whatever frequency to desire) the SMART data on all of the drives. See http://forums.freenas.org/showthread.php?6211-Setup-SMART-Reporting-via-email for some good examples.
- Remember that if you change your email password you will need to update the password in FreeNAS or you will not be able to send emails from the FreeNAS server.
- A properly setup FreeNAS server should not require you to log into it regularly or monitor it frequently. The emails should provide you a good indication of any potential problems.

Long term server maintenance(con't)

- ZFS scrubs should be scheduled regularly. I recommend no more frequently than weekly and no less frequently than monthly. Consult the manual for more information.
- ZFS scrubs are CPU and hard drive intensive. They should ideally be scheduled to be performed when the server is not in use and slower server response is acceptable. For example, on Sundays or at night.
- The length of time for a scrub to be performed varies greatly based on many factors. You will have to monitor your server to determine how long you expect a scrub to take so you can schedule it accordingly.
- As more data is stored on the zpool the scrub will take longer.
- Make sure that a scrub and a SMART Long test will NEVER be run at the same time. This can cause scrubs to never end. The exact reason why is not well understood.

Long term server maintenance(con't)

- Check your fans and temperatures regularly.
- CPUs generally are sufficiently cooled with the standard OEM heatsink/fan they come with(especially Intel). Only overclockers and systems with already insufficient cooling really need an expensive aftermarket solution.
- Heat is your hard drive's worst nightmare. Hard drives should stay below 40C at all times for longer lifespan. This generally means that a fan should be used and directed over your hard drives. Google provided some great info on hard drives (http://static.googleusercontent.com/external_content/untrusted_dlcp/research.google.com/en/us/archive/disk_failures.pdf) and is a good read. Page 6 shows a chart of disk failures and lifespan average temperatures. Above 40C and failure rates increase very rapidly. This also means that 7200RPM hard drives require more cooling than lower RPM hard drives(they also used much more electricity).
- Check your hard drive SMART info at least once a month. SMART data is one of the best ways to determine current and future failures before the hard drive fails.
- Keeping a spare hard drive on a shelf is also a good idea. This allows you to restore full redundancy while an RMA is being processed. If 1 disk fails in an array your odds of a second disk failing shortly thereafter go way up.
- Hard drives seem to last longer if you leave them on all of the time and don't put them to sleep. This does mean a higher electricity bill, so you will have to decide if the tradeoff is worth it to you. This has been observed by many experienced server admins and the exact reason why is not known.

Long term server maintenance(con't)

- "Green" drives can and do work well in FreeNAS servers. They are typically lower power and run cooler.
- WD Green drives need a special modification to be used in a non-desktop environment without premature wearing of the drive. See the thread at http://homeservershow.com/forums/index.php?/topic/2235-fix-for-wdeas-green-drives-intellipark-found/ for how to modify the "Intellipark" setting on WD Green drives. If you do not make this change and you have to RMA your hard drive it will be apparent based on the disk lifespan and SMART attribute #193 that you did not use the hard drive in a desktop. Your warranty may not be honored by Western Digital.
- If in doubt, go with WD Reds. They've worked very well for many forum users.

RAIDZ1 versus RAIDZ2

- RAID-5(and its ZFS equivalent RAIDZ1) is "dead" as of 2009. http://www.zdnet.com/blog/storage/why-raid-5-stops-working-in-2009/162
- RAIDZ1 provide redundancy from any single disk failure. However, due to disk sizes increasing faster than hard drive reliability, RAIDZ1 is really not safe. This is expected to continue for the foreseeable future. Eventually even RAIDZ2 will no longer be safe(estimated at around 2019).
- To combat this limitation it is strongly recommended that you use RAIDZ2 or RAIDZ3 over RAIDZ1.
- If your VDev has only 1 disk of redundancy(RAIDZ1 or RAID1) and you have to replace a failed disk and any other disk has even 1 bad sector you are literally running a RAID0(striped pool) until the rebuild is complete. Statistically you can also expect some ZFS corruption and data loss. The data loss can vary from a single file to complete file system corruption resulting in loss of multiple files or metadata. This is not a limitation of ZFS but is a limitation of having only 1 disk of redundancy.
- Remember, ZFS needs to correct for corruption with redundancy. You will lose all redundancy when replacing a disk in RAIDZ1.
- Many people have lost their data on the forums from using RAIDZ1. For this reason it is recommended you fully understand the potential implications if you choose to use only 1 disk for redundancy. Don't be a statistic!

Having problems?

- If you are having problems with FreeNAS, try the following before posting on the forum:
  - Search the forums. Generally if you screwed it up or are having a problem, someone else probably has too!
  - Take screenshots, take a picture of the screen with a digital camera(make sure the picture is legible) or write down the message you got. Save CLI outputs using SSH.
  - Reboot the FreeNAS server.
  - Run a RAM test. When RAM is bad it often causes unpredictable and unexplainable errors. http://www.memtest.org is free and easy to use. Just boot it up and it will automatically start. 3 complete passes with no errors is typically considered "good". It is recommended you start the test and let it run overnight due to the length of the test. ANY errors are unacceptable.
  - Upgrade to the latest RELEASE version of FreeNAS. This protects you from security vulnerabilities and includes a lot of bug fixes and newer ZFS code.
  - Try a different USB stick. USB sticks do wear out and when they start failing they often give unpredictable results. Always use a USB stick that is 4GB or larger and is a name brand such as Sandisk, Kingston or Crucial.
  - If you want fast turnaround of an issue, check us out on Freenode IRC #freenas.
  - Please don't post to the forums if you are doing things that aren't recommended or don't meet the minimum RAM requirements. We have better things to do than to tell you (again) to do what you should have done without adding more noise to the forums.

“How safe is ZFS?”

- This thought provoking question was asked by a forum user and I felt it should be mentioned in this guide.
- ZFS’ “safeness” is directly related to your knowledge and understanding of your hardware choices, software settings and use, and how you setup and maintain your server.
- ZFS and redundant zpools (and redundant hardware RAIDs) are not a substitute for backups!
- If you are extremely knowledgeable you can have a server that is very trustworthy with your data. On the other hand, if you haven’t done your homework and built a system accordingly with emphasis in saving money over design and having sufficient hardware(such as having enough RAM) you can expect your experience with ZFS to be potentially disastrous. Remember, its your data and its only worth as much time, money, and effort as you are willing to put into it. If you put no effort into it, don’t expect the forums to either!

“How safe is ZFS?”

- Always remember that your data is always worth more to you than a stranger on the forums or IRC.
- If your system results in a train wreck don’t expect other forum users or the IRC to spend hours coddling you through trying to restore your data. Very few people have been able to restore a damaged or destroyed zpool.
- The single biggest thing to take away from ZFS is that a properly designed and administered server can bring reliability and performance that exceeds any Windows options. A poorly designed and administered server can cause complete data loss.

Hardware Recommendations

- Read the thread http://forums.freenas.org/threads/so-you-want-some-hardware-suggestions.12276/
- Do not dismiss "server grade" components as being expensive, not energy efficient, or noisy. This is just not true anymore.
- ZFS was designed with the expectation that you would use ECC RAM. Any errors in RAM due to bad ram or "cosmic rays" can be catastrophic to ZFS. If in doubt, go ECC. Remember that your data is stored in RAM and checksummed before being saved to your zpool. If the data in RAM is corrupt, your checksum will be also. Scrubbing a zpool with bad RAM can potentially corrupt good data because of failed RAM. NTFS can recover from errors because it uses chkdsk. There is no chkdsk alternative for ZFS, at all.
- Many of the Supermicro motherboards come with Intel NICs built-in, ECC RAM support, and PCIe slots for expansion. Many users learn the hard way by thinking they can save $50 with desktop parts and end up spending more than a high quality build with server parts because they used a NIC that didn't provide good performance(or wasn't compatible) and had to buy more components after the fact. So choose wisely.

Hardware Recommendations

- My system has 32GB of DDR3-1600 Kingston ECC(KVR16E11K4/32) RAM, Supermicro X9SCM-F-O(IPMI support a plus) with dual Intel Gb NICs and VGA built-in, and an Intel E3-1230V2 CPU and uses only 35 watts idle with no hard drives. The motherboard, CPU, and RAM cost me about $700 USD(cheaper now) and provides amazing performance and reliability. If you want to save some money and still enjoy ECC benefits the Pentium G2020 is an excellent option($70 on Newegg). I can get speeds over 100MB/sec with CIFS on my system. The G2020 is an excellent CPU as long as you don't plan to use encryption, high end compression, or the Plex plugin for transcoding.
- Server grade components are often cheaper than high end consumer products and will have many of the features you will want and none of the stuff you don't. There is no need to have audio, Firewire, etc. on a FreeNAS server, but they will still consume power.

Final Notes

- The intent of this presentation was to become familiar with the relationships between hard disks, VDevs, zpools, ZIL and L2ARCs. Always consult the FreeNAS manual (doc.freenas.org) for proper procedure for any configuration change.
- FreeNAS does require some time to learn how to use it properly. For many people, using FreeNAS is not feasible. Its important for your own data's sake that you know your limitations and determine if this is too much for you. Plenty of people with 10+ years working in the IT industry have lost everything. Others with just a few years have had great success. Remember, your data is on the line with your decision.

- Updates to FreeNAS could make any portion of this presentation incorrect at any time.

9.2.1.x and Samba4

- FreeNAS uses Samba4 (as of 9.2.1). Samba4 adds a lot of new complexity to FreeNAS.
- Samba4 adds SMB3 support(supported in Windows 8+). SMB2 is generally more reliable than using SMB3, but you can determine whether SMB3 is reliable in your environment.
- Do not try to mix Unix and Windows ACLs in CIFS shares. You used to be able to do this with Samba3 with some trickery, but in Samba4 you must dedicate yourself to either Unix permissions or Windows ACLs. Use datasets to separate CIFS shares from other shares.

For more info…

For more info check out:

- FreeNAS FAQ (http://doc.freenas.org/index.php/FAQs)
- FreeNAS Manual(http://doc.freenas.org)
- FreeNAS Forums(http://forums.freenas.org).
- The latest version of this guide can always be found at http://forums.freenas.org/threads/slideshow-explaining-vdev-zpool-zil-and-l2arc-for-noobs.7775/

Thanks for reading!

Hopefully you learned a lot.

This presentation is accurate as of FreeNAS-9.10 RELEASE (March, 2016).

Created by Cyberjock of the FreeNAS forums.