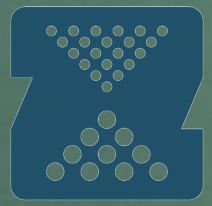
OpenZFS and Linux



OpenZFS

ZFS

ON LINUX

Who is this guy?

Nikolai Lusan

Email: nikolai _@_ lusan.id.au

IRC: Maliuta
on Freenode and OFTC


OpenZFS
Now With Native Encryption!

Licensing

CDDL and GPL are considered incompatible, most distributions will not build installers for containing ZFS support.

Filesystems aren't they fun?

No.

No.

They are not.

Why Not?

Data Loss
Bit Rot
Performance

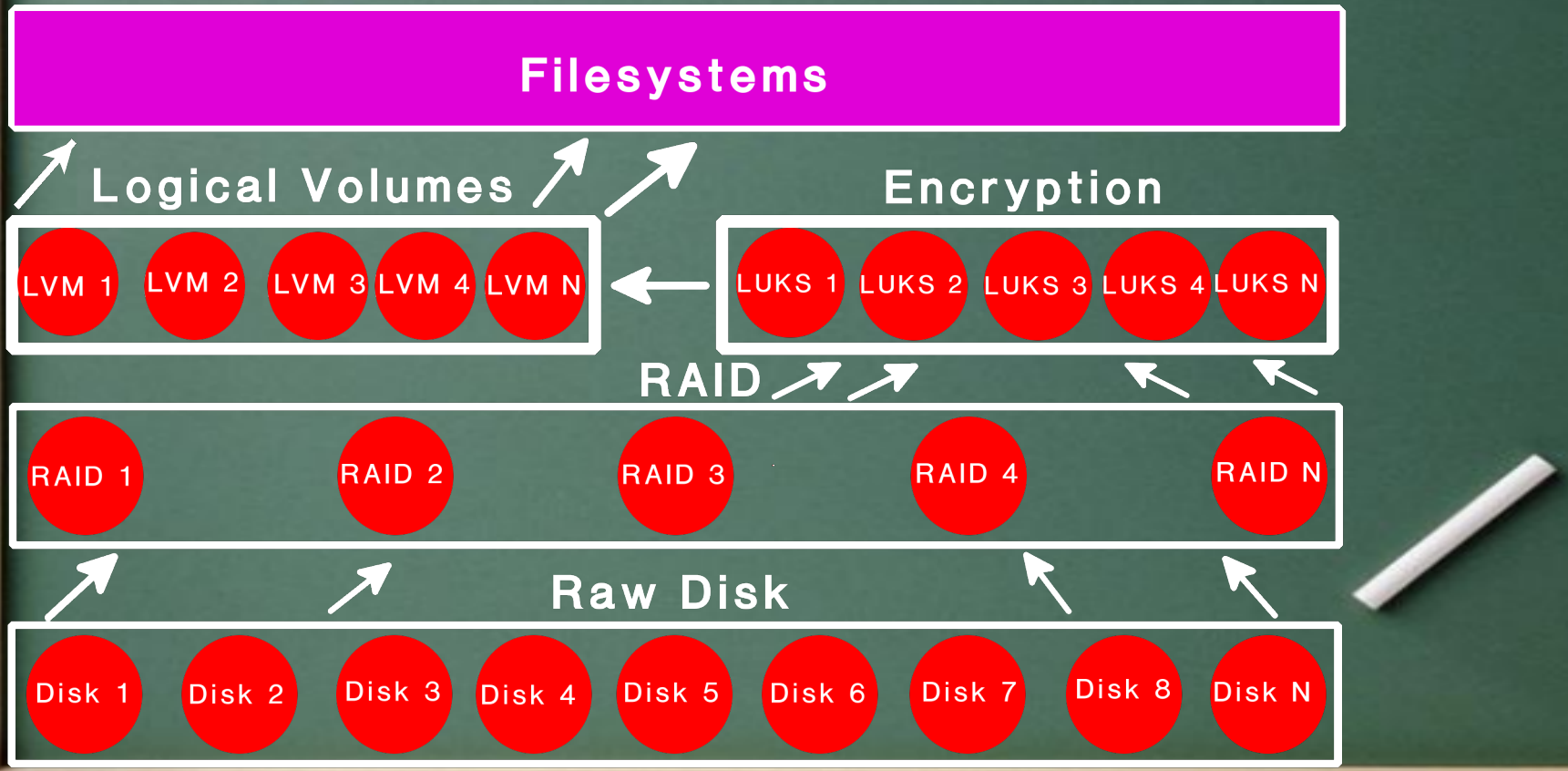
Why ZFS?

- It’s cool
- Stable and established
- Robust
- Good performance - even better with tuning
- Scales up well
- Allows better usage of disk space
- More features than other file systems

Why ZFS?

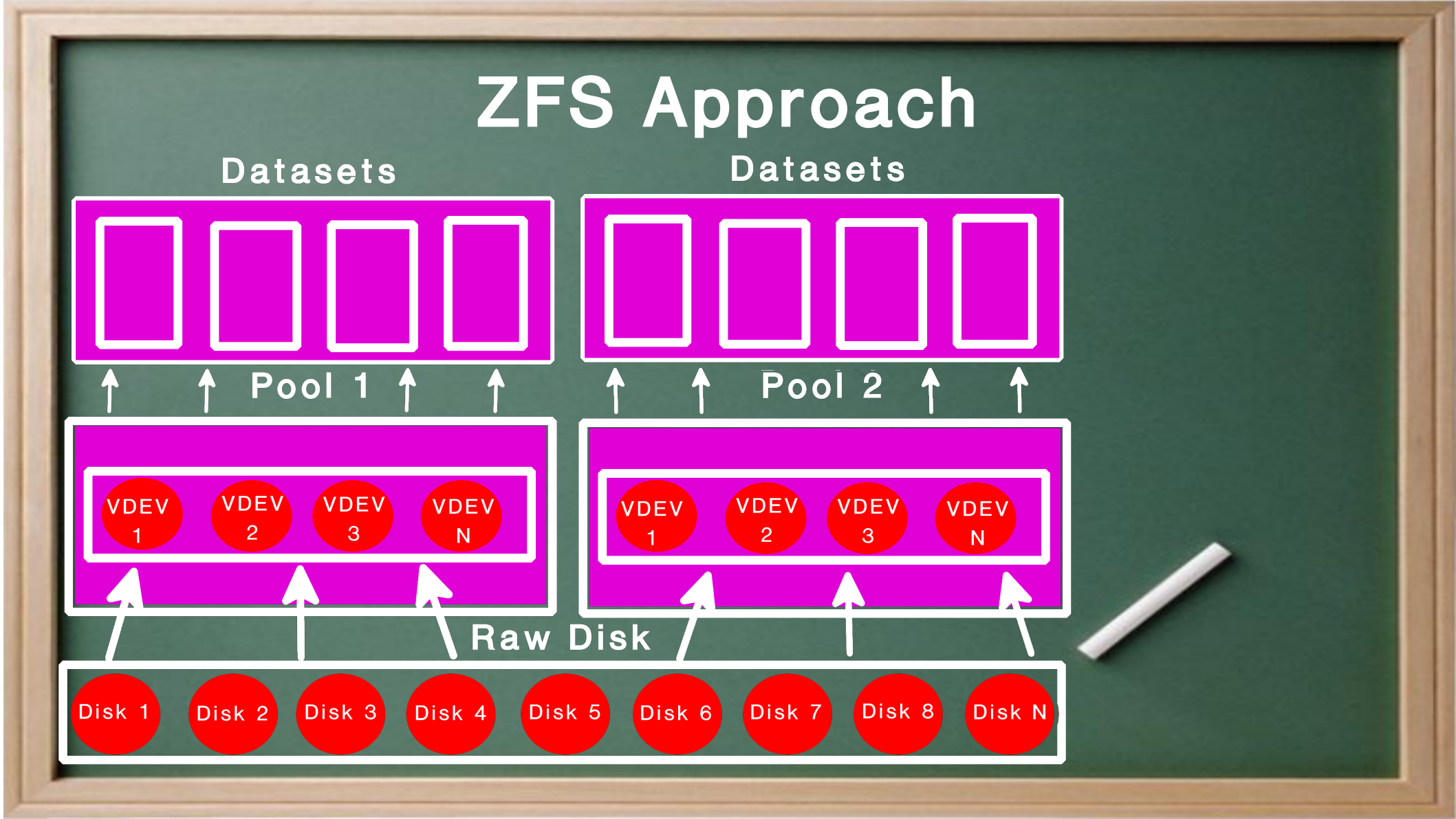
- Designed with systems administrators in mind
- Changes approach to data storage
- Works well in bare metal and virtual environments
- Built in ability to share storage via almost any method available under Linux - eg. NFS, SMB, iSCSI ...

“Traditional” Filesystem Layout



How ZFS Architecture is Different

- Copy on Write (CoW)
- Abstracts storage from disks
- Has internal measures that replace traditional Linux file system access
- Uses pools of virtual devices (VDEV's) which can be of different size and underlying implementation
- Data is stored in datasets, these are similar to LVM logical volumes but far more configurable


ZFS Approach
Datasets
Datasets
Pool 1
Pool 2
VDEV 1
VDEV 2
VDEV 3
VDEV N
VDEV 1
VDEV 2
VDEV 3
VDEV N
Raw Disk
Disk 1
Disk 2
Disk 3
Disk 4
Disk 5
Disk 6
Disk 7
Disk 8
Disk N

VDEVS

- VDEVS are “Virtual Devices”
- The can have different geometries
  - Single disk
  - Mirror of 2 or more disks
  - Multiple types of RAID
- VDEVS are pooled together to create usable storage space
- Writes are striped across VDEVS
- Losing a VDEV means losing data

Pools

- Made up of one or more vdevs
- Writes spread over vdevs
- Mountable filesystem in it's own right
- Many pool level attributes are inherited by datasets
- Pools can be moved from one machine to another with minimal hassle
- When creating pools remember some settings are immutable

Datasets

- Created from ZFS pools
- Each has a set of tuneable attributes
  - Some attribute cannot be changed from inherited, or initial values
- Mountable in arbitrary locations

ZVOLs

- Block devices
- Multiple uses, including swap
- Arbitrary block size
- Not as performant as raw datasets
- Can be exposed to the OS in different ways

ARC/L2ARC/SLOG(ZIL)

- ARC is Adaptive Replacement Cache
- L2ARC is Layer 2 ARC - taken from RAM moved to disk
- SLOG or ZFS Intent Log (ZIL) is an intermediate journal of disk writes that are yet to happen. It allows for a write acknowledgment to be sent to applications/OS faster. ZIL also acts as a kind of journal preventing data loss between boots.

ZFS Tools

- zpool
- zfs
- zed
- zdb

Creating VDEVs

- Use disk/partition names that will remain constant
- Remember not all disks need to be of the same size
- Not all VDEVs need to be of the same type
- There are 3 vdev types
  - Single disk/partition
  - Mirror (with no limit on the number of devices)
  - RAIDZ (with the option for up to triple parity)

```
root@[redacted]:/dev/disk# ls -R
.:
by-id  by-label  by-partlabel  by-partuuid  by-path  by-uuid

./by-id:
ata-Crucial_CT250MX200SSD1_162012[redacted]30          ata-WDC_WD10EFRX-68JCSN0_WD-WCC1U[redacted]34          ata-WDC_WD30EFRX-68EUZN0_WD-WCC4N4[redacted]K3          wwn-0x50014ee[redacted]4c-part9  wwn-0x50014e[redacted]06-part9
ata-Crucial_CT250MX200SSD1_162012[redacted]30-part1    ata-WDC_WD10EFRX-68JCSN0_WD-WCC1U[redacted]34-part1    ata-WDC_WD30EFRX-68EUZN0_WD-WCC4N4[redacted]K3-part1    wwn-0x50014ee[redacted]d9        wwn-0x50014e[redacted]55
ata-Crucial_CT250MX200SSD1_162012[redacted]30-part2    ata-WDC_WD10EFRX-68JCSN0_WD-WCC1U[redacted]34-part9    ata-WDC_WD30EFRX-68EUZN0_WD-WCC4N4[redacted]K3-part9    wwn-0x50014ee[redacted]d9-part1  wwn-0x50014e[redacted]55-part1
ata-Crucial_CT250MX200SSD1_162112[redacted]CD          ata-WDC_WD10EFRX-68JCSN0_WD-WCC1U[redacted]26          ata-WDC_WD30EFRX-68EUZN0_WD-WCC4N5[redacted]33          wwn-0x50014ee[redacted]09        wwn-0x50014e[redacted]55-part9
ata-Crucial_CT250MX200SSD1_162112[redacted]CD-part1    ata-WDC_WD10EFRX-68JCSN0_WD-WCC1U[redacted]26-part1    ata-WDC_WD30EFRX-68EUZN0_WD-WCC4N5[redacted]33-part1    wwn-0x50014ee[redacted]09-part1  wwn-0x50014e[redacted]82
ata-Crucial_CT250MX200SSD1_162112[redacted]CD-part2    ata-WDC_WD10EFRX-68JCSN0_WD-WCC1U[redacted]26-part9    ata-WDC_WD30EFRX-68EUZN0_WD-WCC4N5[redacted]33-part9    wwn-0x50014ee[redacted]09-part9  wwn-0x50014e[redacted]82-part1
ata-INTEL_SSDSC2KW128G8_PHLA8230035[redacted]          ata-WDC_WD10EFRX-68PJCN0_WD-WCC4J[redacted]95          usb-General_USB_Flash_Disk_0326000000014C30-0:0          wwn-0x50014ee[redacted]7f        wwn-0x500a07[redacted]30
ata-INTEL_SSDSC2KW128G8_PHLA8230035[redacted]-part1    ata-WDC_WD10EFRX-68PJCN0_WD-WCC4J[redacted]95-part1    usb-General_USB_Flash_Disk_0326000000014C30-0:0-part1    wwn-0x50014ee[redacted]7f-part1  wwn-0x500a07[redacted]30-part1
ata-INTEL_SSDSC2KW128G8_PHLA8230035[redacted]-part2    ata-WDC_WD10EFRX-68PJCN0_WD-WCC4J[redacted]95-part9    usb-General_USB_Flash_Disk_0326000000014C30-0:0-part2    wwn-0x50014ee[redacted]7f-part9  wwn-0x500a07[redacted]30-part2
ata-INTEL_SSDSC2KW128G8_PHLA823003U[redacted]          ata-WDC_WD30EFRX-68EUZN0_WD-WCC4N[redacted]MN          usb-General_USB_Flash_Disk_0326000000014C30-0:0-part3    wwn-0x50014ee[redacted]76        wwn-0x500a07[redacted]cd
ata-INTEL_SSDSC2KW128G8_PHLA823003U[redacted]-part1    ata-WDC_WD30EFRX-68EUZN0_WD-WCC4N[redacted]MN-part1    usb__Patriot_Memory_0707C11143E39D07-0:0                 wwn-0x50014ee[redacted]76-part1  wwn-0x500a07[redacted]cd-part1
ata-INTEL_SSDSC2KW128G8_PHLA823003U[redacted]-part2    ata-WDC_WD30EFRX-68EUZN0_WD-WCC4N[redacted]MN-part9    usb__Patriot_Memory_0707C11143E39D07-0:0-part1           wwn-0x50014ee[redacted]76-part9  wwn-0x500a07[redacted]cd-part2
ata-WDC_WD10EFRX-68FYTN0_WD-WCC4J[redacted]83          ata-WDC_WD30EFRX-68EUZN0_WD-WCC4N[redacted]TZ          wwn-0x50014ee[redacted]39                                wwn-0x50014ee[redacted]76        wwn-0x55cd2e[redacted]68
ata-WDC_WD10EFRX-68FYTN0_WD-WCC4J[redacted]83-part1    ata-WDC_WD30EFRX-68EUZN0_WD-WCC4N[redacted]TZ-part1    wwn-0x50014ee[redacted]39-part1                          wwn-0x50014ee[redacted]76-part1  wwn-0x55cd2e[redacted]68-part1
ata-WDC_WD10EFRX-68JCSN0_WD-WCC1U[redacted]78          ata-WDC_WD30EFRX-68EUZN0_WD-WCC4N[redacted]EU          wwn-0x50014ee[redacted]39-part9                          wwn-0x50014ee[redacted]76-part9  wwn-0x55cd2e[redacted]68-part2
ata-WDC_WD10EFRX-68JCSN0_WD-WCC1U[redacted]78-part1    ata-WDC_WD30EFRX-68EUZN0_WD-WCC4N[redacted]EU-part1    wwn-0x50014ee[redacted]db                                wwn-0x50014ee[redacted]9c        wwn-0x55cd2e[redacted]8b
ata-WDC_WD10EFRX-68JCSN0_WD-WCC1U[redacted]78-part9    ata-WDC_WD30EFRX-68EUZN0_WD-WCC4N[redacted]EU-part9    wwn-0x50014ee[redacted]db-part1                          wwn-0x50014ee[redacted]9c-part1  wwn-0x55cd2e[redacted]8b-part1
ata-WDC_WD10EFRX-68JCSN0_WD-WCC1U[redacted]29          ata-WDC_WD30EFRX-68EUZN0_WD-WCC4N[redacted]82          wwn-0x50014ee[redacted]db-part9                          wwn-0x50014ee[redacted]9c-part9  wwn-0x55cd2e[redacted]8b-part2
ata-WDC_WD10EFRX-68JCSN0_WD-WCC1U[redacted]29-part1    ata-WDC_WD30EFRX-68EUZN0_WD-WCC4N[redacted]82-part1    wwn-0x50014ee[redacted]4c                                wwn-0x50014ee[redacted]06
ata-WDC_WD10EFRX-68JCSN0_WD-WCC1U[redacted]29-part9    ata-WDC_WD30EFRX-68EUZN0_WD-WCC4N[redacted]82-part9    wwn-0x50014ee[redacted]4c-part1                          wwn-0x50014ee[redacted]06-part1

./by-label:
data  ElTorito  rpool  sysrcd-4.9.0

./by-partlabel:
L2ARC          zfs-2afa57ef0794cdef  zfs-531592e9df91ee36  zfs-a532604abbc52342  zfs-c539d54676fce582  zfs-e15e52480ce70997
zfs-12a5f552acc24182  zfs-37cc02cc8e67973b  zfs-71793302db6c39cf  zfs-b7372f54682df765  zfs-e00af3d75b00b4ff  ZIL

./by-partuuid:
115bff4a-[redacted]a5  308ad3a2-[redacted]40  52c5fb6d-[redacted]6a  ab48d40c-[redacted]0a  c2944246-[redacted]38  efd18f4a-[redacted]88
21ca97f0-[redacted]14  401bea0a-[redacted]27  84c16995-[redacted]9e  af3b765b-[redacted]d9  c565a110-[redacted]0d  fb6ea82e-[redacted]f6
24675c4f-[redacted]d7  478808bf-[redacted]4a  85a8ee64-[redacted]91  af5a79c5-[redacted]01  c795a75f-[redacted]36
27cc952e-[redacted]0e  49c881f5-[redacted]90  8e8c556d-[redacted]80  b71b5087-[redacted]8a  e8c74ccd-[redacted]63
289636e7-[redacted]00  4aff62c2-[redacted]67  95d58a31-[redacted]b1  b74d2cde-[redacted]d2  eb844c93-[redacted]a2
29e26797-[redacted]9d  4c9c1fbe-[redacted]5b  a4bf4652-[redacted]d1  b9abe384-[redacted]2f  ede06152-[redacted]64

./by-path:
pci-0000:00:12.2-usb-0:2:1.0-scsi-0:0:0:0       pci-0000:01:00.0-sas-phy2-lun-0       pci-0000:01:00.0-sas-phy5-lun-0-part9  pci-0000:02:00.0-sas-phy1-lun-0-part1  pci-0000:02:00.0-sas-phy5-lun-0-part1
pci-0000:00:12.2-usb-0:2:1.0-scsi-0:0:0:0-part1 pci-0000:01:00.0-sas-phy2-lun-0-part1 pci-0000:01:00.0-sas-phy6-lun-0       pci-0000:02:00.0-sas-phy1-lun-0-part2  pci-0000:02:00.0-sas-phy5-lun-0-part2
pci-0000:00:16.2-usb-0:4:1.0-scsi-0:0:0:0       pci-0000:01:00.0-sas-phy2-lun-0-part9 pci-0000:01:00.0-sas-phy6-lun-0-part1  pci-0000:02:00.0-sas-phy2-lun-0         pci-0000:02:00.0-sas-phy6-lun-0
pci-0000:00:16.2-usb-0:4:1.0-scsi-0:0:0:0-part1 pci-0000:01:00.0-sas-phy3-lun-0       pci-0000:01:00.0-sas-phy6-lun-0-part9  pci-0000:02:00.0-sas-phy2-lun-0-part1  pci-0000:02:00.0-sas-phy6-lun-0-part1
pci-0000:00:16.2-usb-0:4:1.0-scsi-0:0:0:0-part2 pci-0000:01:00.0-sas-phy3-lun-0-part1 pci-0000:01:00.0-sas-phy7-lun-0       pci-0000:02:00.0-sas-phy2-lun-0-part2  pci-0000:02:00.0-sas-phy6-lun-0-part9
pci-0000:00:16.2-usb-0:4:1.0-scsi-0:0:0:0-part3 pci-0000:01:00.0-sas-phy3-lun-0-part9 pci-0000:01:00.0-sas-phy7-lun-0-part1  pci-0000:02:00.0-sas-phy3-lun-0         pci-0000:02:00.0-sas-phy7-lun-0
pci-0000:01:00.0-sas-phy0-lun-0                 pci-0000:01:00.0-sas-phy4-lun-0       pci-0000:01:00.0-sas-phy7-lun-0-part9  pci-0000:02:00.0-sas-phy3-lun-0-part1  pci-0000:02:00.0-sas-phy7-lun-0-part1
pci-0000:01:00.0-sas-phy0-lun-0-part1           pci-0000:01:00.0-sas-phy4-lun-0-part1 pci-0000:02:00.0-sas-phy0-lun-0         pci-0000:02:00.0-sas-phy3-lun-0-part9  pci-0000:02:00.0-sas-phy7-lun-0-part9
pci-0000:01:00.0-sas-phy1-lun-0                 pci-0000:01:00.0-sas-phy4-lun-0-part9 pci-0000:02:00.0-sas-phy0-lun-0-part1  pci-0000:02:00.0-sas-phy4-lun-0
pci-0000:01:00.0-sas-phy1-lun-0-part1           pci-0000:01:00.0-sas-phy5-lun-0       pci-0000:02:00.0-sas-phy0-lun-0-part9  pci-0000:02:00.0-sas-phy4-lun-0-part1
pci-0000:01:00.0-sas-phy1-lun-0-part2           pci-0000:01:00.0-sas-phy5-lun-0-part1 pci-0000:02:00.0-sas-phy1-lun-0         pci-0000:02:00.0-sas-phy5-lun-0

./by-uuid:
0590-8E08  12756214194519879778  16420466833479614243  1c0c0c60-86a7-4ba9-88d5-ebcd684db98d  2016-10-29-18-31-25-00  251A-400B  5855e3b1-7995-40af-8820-527f7dccc58e  FCF8-ECB8
root@[redacted]:/dev/disk#
```

Snapshots

- Provide a glimpse of the dataset at the time taken
- Can be used to roll back a dataset to the point in time the snapshot was created
- Mountable
- They do take up space
- The space used is only a delta from the most recent snapshot
- Not automatically deleted, so they need to be managed
- There are existing tools to automatically manage snapshots
  - zsnapd
  - zfs-auto-snapshot
- Can be enabled/disabled per dataset

Snapshots for Offsite Backup

- The “zfs” tool provides a send function and a receive function allowing snapshots to be sent between pools
- The pools do not need to be on the same machine, the receiver can even be a dataset under another pool

  (e.g. send from <pool>/dataset@<snapshot> to <pool_2>/dataset/dataset_2)
- Most common transport is via ssh, but any tool that lets you send and receive data can be used (mbuffer is another common tool)

Tuneables

- There are almost 230 tuneable parameters for the kernel module alone
- There are over 75 tuneable parameters for each dataset, more when you are dealing with enabling non-standard or new features

Compression and Deduplication

- Native filesystem level compression
  - lz4
  - lzjb
  - gzip
  - zle
- Depupication is RAM intensive (1GB of RAM for every 1TB of deduplicated data space)
- Both can help you squeeze more storage out of your disk

Optimisation for all ZFS (Kernel)

- Tune the ARC size to fit your needs
- Tune metaslab performance for spreading writes across vdevs
- Tune ARC/L2ARC performance
- Tune TRIM limits for SSD storage

Easy Tuning for Most Purposes

- Create pools using ashift=12
- Enable lz4 compression
- Set recordsize to 128k
- Disable atime,dev,exec,suid as needed (atime is a big saver)
- Set logbias to latency
- Set sync to “standard” or “disabled”

Optimisation for MySQL/MariaDB

- This is for innodb only MyISAM is left to people who know this RDBMS better
- recordsize=16k
- primarycache=metadata
- logbias=throughput

Optimisation for PostgreSQL

- Use separate datasets for data and WAL
- recordsize=8k
- primarycache=metadata
- logbias=throughput

Optimisation for running VM's

- Controversy over using vdev's versus qcow2 files
- Different approaches require different optimisation
- VDEV's should be created with a recordsize that refelcts the FS that will run on the VM, have logbias=throughput, and primarycache=metadata and volmode=full
- Using qcow2 files on dedicated datasets is the recommended way. The datasets should have a recordsize that matches the FS that will be used in the VM

Running ZFS in a [hosted] VM

- Use a single disk vdev
- Still use an SLOG device
- Worry more about file compression and RAM usage than underlying storage.

Resources

- Manpages
  - zpool
  - zfs
  - zdb
  - zpool-features
  - zfs-module-parameters
  - zfs-events
- Online
  - OpenZFS wiki http://open-zfs.org/wiki/Main_Page
  - Arch Linux wiki https://wiki.archlinux.org/index.php/ZFS
  - ZFS on Linux FAQ https://github.com/zfsonlinux/zfs/wiki/FAQ


The End ...