ZFS: Tips and Tricks

From Proxmox VE

Contents

Using ZFS Storage Plugin (via Proxmox VE GUI or shell)

After the ZFS pool has been created, you can add it with the Proxmox VE GUI or CLI.

Adding a ZFS storage via CLI

To create it by CLI use:

```
pvesm add zfspool <storage-ID> -pool <pool-name>
```

Adding a ZFS storage via Gui

To add it with the GUI: Go to the datacenter, add storage, select ZFS.

Misc

QEMU disk cache mode

If you get the warning:

```
qm start 4016
kvm: -drive file=/data/pve-storage/images/4016/vm-4016-disk-1.raw,if=none,id=drive-virtio1,aio=native,cache=none: could not open disk image /data/pve
```

or a warning that the filesystem do not supporting O_DIRECT, set the disk cache type of your VM from **none** to **writeback**.

Example configurations for running Proxmox VE with ZFS

Install on a high performance system

As of 2013 and later, high performance servers have 16-64 cores, 256GB-1TB RAM and potentially many 2.5" disks and/or a PCIe based SSD with half a million IOPS. High performance systems benefit from a number of custom settings, for example enabling compression typically improves performance.

- If you have a good number of disks keep organized by using aliases. Edit /etc/zfs/vdev_id.conf to prepare aliases for disk devices found in /dev/disk/by-id/ :

```
# run 'udevadm trigger' after updating this file
alias a0        scsi-36848f690e856b10018cdf39854055206
alias b0        scsi-36848f690e856b10018cdf3ce573fdeb6
alias a1        scsi-36848f690e856b10018cdf40f5b277cbc
alias b1        scsi-36848f690e856b10018cdf43a5db1b99b
alias a2        scsi-36848f690e856b10018cdf4575f652ad0
alias b2        scsi-36848f690e856b10018cdf47761587cec
```

Use flash for caching/logs. If you have only one SSD, use parted of gdisk to create a small partition for the ZIL (ZFS intent log) and a larger one for the L2ARC (ZFS read cache on disk). Make sure that the ZIL is on the first partition. In our case we have a Express Flash PCIe SSD with 175GB capacity and setup a ZIL with 25GB and a L2ARC cache partition of 150GB.

- edit /etc/modprobe.d/zfs.conf to apply several tuning options for high performance servers:

```
# ZFS tuning for a proxmox machine that reserves 64GB for ZFS
#
# Don't let ZFS use less than 4GB and more than 64GB
options zfs zfs_arc_min=4294967296
options zfs zfs_arc_max=68719476736
#
# disabling prefetch is no longer required
options zfs l2arc_noprefetch=0
```

- create a zpool of striped mirrors (equivalent to RAID10) with log device and cache and always enable compression:

```
zpool create -o compression=on -f tank mirror a0 b0 mirror a1 b1 mirror a2 b2 log /dev/rssda1 cache /dev/rssda2
```

- check the status of the newly created pool:

```
root@proxmox:/# zpool status
  pool: tank
 state: ONLINE
  scan: none requested
config:
        NAME        STATE     READ WRITE CKSUM
        tank        ONLINE       0     0     0
          mirror-0  ONLINE       0     0     0
            a0      ONLINE       0     0     0
            b0      ONLINE       0     0     0
          mirror-1  ONLINE       0     0     0
            a1      ONLINE       0     0     0
            b1      ONLINE       0     0     0
          mirror-2  ONLINE       0     0     0
            a2      ONLINE       0     0     0
            b2      ONLINE       0     0     0
        logs
          rssda1    ONLINE       0     0     0
        cache
          rssda2    ONLINE       0     0     0

errors: No known data errors
```

Using PVE 2.3 on a 2013 high performance system with ZFS you can install Windows Server 2012 Datacenter Edition with GUI in just under 4 minutes.

Troubleshooting and known issues

ZFS packages are not installed

If you upgraded to 3.4 or later, zfsutils package is not installed. You can install it with apt:

```
apt-get install zfsutils zfs-initramfs
```

Grub boot ZFS problem

- Symptoms: stuck at boot with an blinking prompt.
- Reason: If you ZFS raid it could happen that your mainboard does not initial all your disks correctly and Grub will wait for all RAID disk members - and fails. It can happen with more than 2 disks in ZFS RAID configuration - we saw this on some boards with ZFS RAID-0/RAID-10

Boot fails and goes into busybox

If booting fails with something like

```
No pool imported. Manually import the root pool
at the command prompt and then exit.
Hint: try: zpool import -R /rpool -N rpool
```

is because zfs is invoked too soon (it has happen sometime when connecting a SSD for future ZIL configuration). To prevent it there have been some suggestions in the forum. Try to boot following the suggestions of busybox or searching the forum, and try ONE of the following:

a) edit /etc/default/grub and add "rootdelay=10" at GRUB_CMDLINE_LINUX_DEFAULT (i.e. GRUB_CMDLINE_LINUX_DEFAULT="rootdelay=10 quiet") and then issue a # update-grub

b) edit /etc/default/zfs, set ZFS_INITRD_PRE_MOUNTROOT_SLEEP='4', and then issue a "update-initramfs -k 4.2.6-1-pve -u"

Snapshot of LXC on ZFS

If you can't create a snapshot of an LXC container on ZFS and you get following message:

```
INFO: rsync: set_acl: sys_acl_set_file(archiv, ACL_TYPE_DEFAULT): Operation not supported (95)
```

you can run following commands

```
zfs create -o mountpoint=/mnt/vztmp rpool/vztmp
zfs set acltype=posixacl rpool/vztmp
```

Now set /mnt/vztmp in your /etc/vzdump.conf for tmp

Replacing a failed disk in the root pool

This assumes you have a RAIDZ***x*** setup, where ***x***≥1. If you have RAIDZ0 with a failed disk, your pool is FAILED, no DEGRADED and no recovery is possible. First, check your pool status:

```
root@pve:~# zpool status -v
  pool: rpool
 state: ONLINE
  scan: resilvered 143G in 15h22m with 0 errors on Fri Oct 14 10:59:46 2016
config:

        NAME        STATE     READ WRITE CKSUM
        rpool       ONLINE       0     0     0
          raidz1-0  ONLINE       0     0     0
            sda2    ONLINE       0     0     0
            sdb2    ONLINE       0     0     0
            sdc2    ONLINE       0     0     0

errors: No known data errors
```

If your disk has already failed, one of the disks will likely show OFFLINE, and the raidz***x***, rpool and state entries will likely read DEGRADED instead of ONLINE. This example shows a disk that has started logging SMART errors but which has not completely failed yet.

Assuming the failing disk is /dev/sdb2, first take the disk offline:

```
root@pve:~# zpool offline rpool /dev/sdb2
root@pve2:~# zpool status -v
  pool: rpool
 state: DEGRADED
status: One or more devices has been taken offline by the administrator.
        Sufficient replicas exist for the pool to continue functioning in a
        degraded state.
action: Online the device using 'zpool online' or replace the device with
        'zpool replace'.
  scan: resilvered 143G in 15h22m with 0 errors on Fri Oct 14 10:59:46 2016
config:

        NAME        STATE     READ WRITE CKSUM
        rpool       DEGRADED     0     0     0
          raidz1-0  DEGRADED     0     0     0
            sda2    ONLINE       0     0     0
            sdb2    OFFLINE      0     0     0
            sdc2    ONLINE       0     0     0

errors: No known data errors
```

Now replace the physical disk (if it's hot-swappable, otherwise you'll have to reboot which could be interesting if it's /dev/sda that's failing).

Before you can rebuild the ZFS pool, you need to partition the new disk. Proxmox uses a GPT partition table for all ZFS-root installs, with a protective MBR, so we want to clone a working disk's partition tables, copy the GRUB boot partition, copy the MBR, and rerandomize the GUIDs before letting ZFS at the disk again.

Copy the partition table from /dev/sda to /dev/sdb:

```
root@pve:~# sgdisk --replicate=/dev/sdb /dev/sda
```

(Make sure you get the devices in the right order: you're invoking sgdisk on the WORKING disk, replicating TO the new disk.)

If you are upgrading to larger disks, now is the time to go in with parted(8) and manually resize/recreate the partitions. (Have fun.)

Ensure the GUIDs are randomized otherwise the kernel and ZFS will get really, really confused:

```
root@pve:~# sgdisk --randomize-guids /dev/sdb
The operation has completed successfully.
```

Install the Grub on the new disk, to ensure that it will boot.

```
grub-install /dev/sdb
```

Then replace the disk in the ZFS pool, assuming the new disk has also shown up as /dev/sdb:

```
root@pve:~# zpool replace rpool /dev/sdb2
Make sure to wait until resilver is done before rebooting.
```

Ensure the RAIDZ array is rebuilding:

```
root@pve:~# zpool status -v
  pool: rpool
 state: DEGRADED
status: One or more devices is currently being resilvered.  The pool will
        continue to function, possibly in a degraded state.
action: Wait for the resilver to complete.
  scan: resilver in progress since Mon Oct 17 13:02:00 2016
    147M scanned out of 298G at 4.46M/s, 19h1m to go
    47.3M resilvered, 0.05% done
config:

        NAME             STATE     READ WRITE CKSUM
        rpool            DEGRADED     0     0     0
          raidz1-0       DEGRADED     0     0     0
            sda2         ONLINE       0     0     0
            replacing-1  OFFLINE      0     0     0
              old        OFFLINE      0     0     0
              sdb2       ONLINE       0     0     0  (resilvering)
            sdc2         ONLINE       0     0     0

errors: No known data errors
```

Once "zpool status" finally shows nothing but ONLINE, it is safe to reboot.

Glossary

- ZPool is the logical unit of the underlying disks, what zfs use.

- ZVol is an emulated Block Device provided by ZFS
- ZIL is ZFS Intent Log, it is a small block device ZFS uses to write faster
- ARC is Adaptive Replacement Cache and located in Ram, its the Level 1 cache.
- L2ARC is Layer2 Adaptive Replacement Cache and should be on an fast device (like SSD).

Further readings about ZFS

- http://wiki.illumos.org/download/attachments/1146951/zfs_last.pdf
- http://zfsonlinux.org/faq.html
- http://wiki.complete.org/ConvertingToZFS
- https://www.freebsd.org/doc/handbook/zfs.html (even if written for freebsd, of course, I found this doc is extremely clear even for less "techie" admins [note by m.ardito])
- https://pthree.org/2012/04/17/install-zfs-on-debian-gnulinux/ (and all other pages linked there)

and this has some very important information to know before implementing zfs on a production system.

- http://www.solarisinternals.com/wiki/index.php/ZFS_Best_Practices_Guide

Very well written manual pages

```
man zfs
man zpool
```

See also

ZFS on Linux

Storage: ZFS

Retrieved from "https://pve.proxmox.com/mediawiki/index.php?title=ZFS:_Tips_and_Tricks&oldid=9795"

Category: HOWTO

- This page was last edited on 24 May 2017, at 14:09.