ZFS Optimierung und Todsünden

ZFS Funktionen für SSDs

- `zpool get autotrim <poolname>`: Auslesen ob „autotrim“ eingeschaltet ist
- `zpool trim <poolname>`: Pool manuell trimmen. Nicht alle SSDs können das, aber die meisten
- `zpool status -t <poolname>`: Auslesen TRIM status
- `zpool set autotrim=on <poolname>`: generell trimmen (autotrim einschalten, wenn manuelles trimmen erfolgreich)
- `zfs set logbias=throughput <dataset/zvol>`: Eventuell bei großen Files spannend
- `zfs set sync=disabled  <dataset/zvol>`: kann Leistung steigern, auf Kosten der Sicherheit
- Es gibt weniger Müll an SAS SSDs als SATA am Markt
- Enterprise SSDs verwenden! 480G, 960G, 1920G, 3840G ....

ZFS mit Spindeln

- Mit mehr disks skalierst die Bandbreite
- Mit mehr vdevs skalierst IOps
- SAS (vollduplex) ist bei gleicher Bandbreite schneller als SATA (Halbduplex)!
- HBA muss den Linkspeed der Disk unterstützen, neuere Disk funktioniert an altem HBA und anders herum, Links Speed kann man u.a. mit lshw auslesen
- SATA Disks können an SAS Controller betrieben werden, anders herum nicht kompatibel

U.2 sticht SAS aus

- Viele Mainboards unterstützen SATA, SAS und U.2
- Direkte Anbindung an PCIe Bus, also kein 12G Limit

ZFS mit Cache und oder LOG

- `zpool add cache <poolname> diskxpx`: Eventuell Lesecache auf Boot NVMe vom PVE oder ganze Disk
- `zpool add log <poolname> mirror diskxpx  diskxpx`: Eventuell Logdevice auf Boot NVMe´**s** vom PVE oder ganze Disks im **Spiegel plus USV!**
- `zfs set sync=always <poolname>`: ohne Sync, kein Log

Todsünden

> **HW RAID Controller nutzen und mehrere LUNs als Raid VDEVs nutzen**
> Datenverust, kein Smart, keine Metriken

> **Single VDEV Raid z. B. auf Hardware Raid oder iSCSI/FC SAN**
> Bei Beschädigung kann keine weitere Disk die Redundanz, bzw. die Konsistenz bestätigen. Wenn ein VDEV zu viele Fehler liefert, lehnt ZFS es für einen weiteren Import dauerhaft ab!!! Eventuell zpool import -fFX poolname probieren, destruktiv für Rollback auf letzten guten Zustand, sehr riskant

> **Deduplizierung aktivieren**
> kann nicht mehr ausgeschaltet werden, ggf. in neueren ZFS Versionen optimiert

> **ZFS_ARC zu klein / zu groß**
> schlechte Leistung, bei Cache und LOG mehr ARC notwendig

> **Pool über 80% Füllstand**
> Leistungseinschränkung bis Stillstand bei ca 96%

> **Snapshot Bereinigung / Management aus dem Blick verlieren**
> führt ggf. unbeobachtet zu vollem Raid

> **Snapshots auf Backups**
> völlig kontraproduktiv, führt zu sehr vollem Raid

> **Anfügesnapshot durch schlechtes Management verlieren**
> Replikation benötigt neuen Platz und Zeit

Druckbetankung
dafür ist ZFS nicht gemacht, was mehr kann, braucht auch mehr Leistung

PVE auf selbem RAID wie Systeme (OS auf Datenpool)
Last auf VMs führt im Härtefall bis zum Kernelcrash

Kein ECC RAM
Datenbeschädigung die nicht in jedem Fall durch Scrub und Resilver behoben werden kann (selten)

Proxmox Datastore auf ZFS Dataset Directory setzen
Proxmox erstellt eventuell beim Booten den Ordner selbst und ZFS kann dann nicht mounten, Rollback setzt alle virtuellen Disks im Dataset zurück

Platzangaben vertrauen
Recordsizes, Volblocksizes, Compressions, Special Devices und Co. machen eine Vorhersage fast unmöglich und rechtfertigen Thin Provisioning

zpool upgrade zu früh oder zu spät
pool upgrade erweitert features die ggf. beim aktuellen Proxmox ISO noch nicht verfügbar, jedoch im Upgrade verfügbar sind. Im Recoveryfall kann man nur readonly mounten!

Festplatte ersetzen bei Proxmox ZFS Boot Systemen
zpool replace sollte nur nach Übernahme der Partitionstabelle einer funktionierenden Disk auf die dritte Partition ausgeführt werden, zusätzlich mit proxmox-boot-tool init/format auf die zweite Partition.

iSCSI LUNs mit Proxmox
nicht empfehlenswert, bestenfalls mit dem Free-/TrueNAS Plugin. Wie lange das supportet ist weiss keiner.

Import Proxmox Pool in Recovery Umgebung

Achtung Eigentümer des Pools ändert sich, daher zpool import -fa benutzen, beim nächsten boot wieder zpool import -fa und exit, /etc/pve ist in Recoveryumgebung nicht vorhande, da aus FUSE generiert

Billige USB Adapter und Gehäuse
Manche Adapter haben alle die gleiche Kennung, daher ggf. Raid erstellen und importieren über zpool import /dev/disk/by-path oder by-partuuid

Raid ohne dryrun erweitern
'zpool add **-n** rpool **mirror** sdc3 sdd3' wäre korrekt, wobei es tödlich wäre hier das wort mirror zu vergessen

Konsumer SSDs für Cache und LOG
Waren im Dauerbetrieb schon nach weniger als zwei Jahren am Limit, währen Enterprise SSDs nach 9 Jahren erst bei 30% waren. Getestet mit Samsung Consumer und Intel Enterprise

VMs nicht trimmen
Discard in VMs gibt mit SSD Emulation freie Blöcke an ZFS zurück, ansonsten sehr viel Platz verschenkt!

Startsektor der Partition nicht oktal

Manche Festplatten starten ab Block 34, manche auf Block 2048. Bei Startblock 34 sollte die erste Partition bis Block 2047 gehen, damit die nächste Partition direkt bei 2048 anfängt

Alles auf einmal machen
Betrieb, Scrub, Trim Backup und Replikation treiben die Load in die Höhe, nach fest kommt ab:)

Nur auf ZFS vertrauen
Auch wenn wir nun weit über 15 Jahre ZFS produktiv fahren, machen wir immer zusätzlich Replikate und Backups. Nur so kann wirklich nichts mehr passieren. Replikate bei denen die VM zum Snapshotten heruntergefahren

> wurde, dürften die absolut beste Sicherung sein und immer die erste Wahl bei Problemen

Update 2024 zum Thema Volblocksize

zfs create -o volblocksize=8k rpool/8k -V 1G

#Warning: volblocksize (8192) is less than the default minimum block size (16384).To reduce wasted space a volblocksize of 16384 is recommended.

rpool/16k **1.11G** 15.2T 81.4K -
rpool/8k **1.49G** 15.2T 81.4K -

NAME STATE READ WRITE CKSUM

rpool ONLINE 0 0 0

raidz1-0 ONLINE 0 0 0
scsi-20100000000000005cd2e475ae4b5651-part3 ONLINE 0 0 0
scsi-20100000000000005cd2e4f1bd4e5651-part3 ONLINE 0 0 0
scsi-20100000000000005cd2e4ef8f4b5651-part3 ONLINE 0 0 0
scsi-20100000000000005cd2e4889d4b5651-part3 ONLINE 0 0 0

Bei RaidZ2 wäre der Overhead über 200%

Inzwischen warnt ZFS, jedoch sieht das keine in der PVE GUI!

Emfpehlung Volblocksize für ZVOLs

Raid 1 / 10 8-16k

Raid5(Z1) 16k

Raid6+(Z2-3) 32k

Empfehlung Recordsize für Datasets

LXC 128k

Mailarchiv 16k

Proxmox Backup Server 1024k

Version #13
Erstellt: 2023-01-04 21:43:13 UTC von Admin
Zuletzt aktualisiert: 2025-08-27 21:08:19 UTC von Christian Zengel (sysops GmbH)