ZFS Performance mit SSD erhöhen

Ich habe hier in meinem Debian-Server in den letzten Monaten immer wieder Performance-Probleme mit dem ZFS.

Auf meinem System laufen derzeit 25 virtuelle Maschinen und deren Virtual-Disks liegen alls auf dem ZFS-Dateisystem.

Ich habe eine Weg gefunden nachträglich einen Schreib-Lese Cache in den Zpool zu integrieren.

Zuerst habe ich mir eine Enterprise SSD besorgt die für den Dauerbetrieb gemacht ist (24/7) z.B. Western Digital WD Black (NVME)

Normale SSD schaffen die sog. IOPs nicht (IO-per-second)
Die WD Black ist die derzeit schnellste Platte im Konsumer Segment mit ca. 3600 MB/Sec.

lesen
und ca. 2500 MB/sec schreiben.

Die WD NVME schafft bis zu 400.000 IOpS die Samsung PRO dagegen "nur" ca. 120.000 IOpS
Da das ZFS aufgrund der vielen parallelen Schreib- Lese- Zugriffe viele Datenblöcke in kurzer Zeit

bearbeiten möchte ist die WD eine gute Wahl.

Dazu muss man die NVME Platte in einen PCI-Express 4x Adapter stecken und einen
freien 16x oder 8x PCIe Slot im Rechner haben.

ACHTUNG ! die meisten älteren Mainboards können nicht direkt von der NVME Platte booten.
Neue Mainboards haben bereits 2 oder 3 NVME-Slots on Board.
Diese Mainboards können dann auch von der NVME-Platte booten.

Unter Linux ist es allerdings kein Problem von einer normalen Platte oder SATA-SSD zu starten und
dann im laufenden Betrieb die NVME-Disk einhängen.

Man kann die SSD einfach in den bestehenden ZFS Pool integrieren.
Zuvor muss man allerdings einige Schritte ausführen.

Zuerst auf der SSD mit parted eine neu GPT-Partition erstellen.

Dann zwei Partitionen anlegen (eine für Cache und eine fürs Log)

Bei einer 256 GB Partition habe ich 2x 120 GB genommen.

parted /dev/sde
create gpt Partition

create 120G
Toggle "FreeBSD zfs" (Nr 36)

Dies macht man für beide Partitionen.

Wenn die Partitionen angelegt sind ist es wichtig auch ein Dateisystem darauf zu erzeugen, sonst schreibt das ZFS auf die "RAW-Partition"

Die macht man mit dem Befehl

zpool create -f log1 /dev/sde1
zpool create -f cache1 /dev/sde2

Hiermit legt man zwei neue ZFSPools an mit dem namen log1 und cache1.
Dies ist nötig damit das System die Festplatte mit dem ZFS beschreibt (wie formatieren)

Danach muss man die beiden Pools wieder auflösen.

zpool destroy log1
zpool destroy cache1

Nun kann man die beiden Partitionen dem bestehenden ZFS-Pool hinzufügen.

zpool add zfspool log /dev/sde1
zpool add zfspool cache /dev/sde2

Diese Schritte wiederholt man mit der zweiten SSD (sofern vorhanden)

zpool create -f cache mirror /dev/sdd1 /dev/sde1
zpool create -f log mirror /dev/sdd2 /dev/sde2

Damit hat man zwei gespiegelte Partitionen angelegt die man dann einbinden kann.

Mit dem Befehl zpool status kann man sehen das das Cache und Log Dateisystem eingehängt wurde.

```
NAME                    STATE     READ WRITE CKSUM
 zfspool                ONLINE       0     0     0
 raidz1-0               ONLINE       0     0     0
 zfs-80e6dfxxxxxxxxxxx  ONLINE       0     0     0
 zfs-12f97xxxxxxxxxxx   ONLINE       0     0     0
 zfs-4567fxxxxxxxxxxx   ONLINE       0     0     0
 zfs-34abfxxxxxxxxxxx   ONLINE       0     0     0
 zfs-a0be4xxxxxxxxxxx   ONLINE       0     0     0
 zfs-4522xxxxxxxxxxx    ONLINE       0     0     0
 logs
 sde2                   ONLINE       0     0     0
 cache
 sde3                   ONLINE       0     0     0
```

ACHTUNG ! Wegen der Ausfallsicherheit sollte man eigentlich 2 SSDs einbauen die gespiegelt werden.
ACHTUNG ! Wenn der Cache ausfällt (die SSD versagt) wird das Raid degraded und ist nicht mehr sicher.
ACHTUNG ! Wenn die Log-Platte ausfallen sollte gehen eventuell Daten verloren, die noch nicht auf das Raidz geschrieben wurden !

Ich habe als zusätzliche Sicherheit noch stündliche Snapshots auf dem ZFS laufen, damit man im Fehlerfall das Dateisystem wieder reparieren kann.
Zusätzlich habe ich ein nächtliches Backup laufen das ALLE Daten aus dem System auf eine 10TB Backup Platte sichert.

Als zweite Backupstufe wird dann noch die Sicherung auf ein externes NAS Laufwerk geschrieben.

Im Server habe ich nur noch ein SATA-Anschluss frei daher muss eine SSD reichen.

ACHTUNG dies ist aufgrund der Konfiguration mit nur einer SSD Festplatte eine gefährliche Konfiguration und es droht Datenverlust wenn man kein Backup hat.

Ich übernehme Keinerlei Haftung für diese Anleitung.

Viel Spass noch

Euer Admin