




Découvrez ZFS

Un stockage fiable, puissant et accessible



Qui suis-je

  **soigneur de pool ZFS @ OVHcloud (2020)**

-  père de famille
-  construction et usage des outils
-  communauté francophone Python ([AFPy](#))

Stockage fichier

- **processus de sauvegarde**
- **système virtualisés** (*Images de machine virtuelle*)
- **base de données** (*besoin spécifiques*)
- **traitement de données** (*cache, tampon, etc...*)

ZFS ?

- *Zettabyte File System*






ZFS ?

- *Zettabyte File System* ...ou pas

“ I picked ZFS for the simplest of reasons: it sounds cool ”




Jeff Bonwick

Le plan

-  Historique
-  Principaux concepts ZFS
-  Usages et choix chez OVH
-  Faites gaffe quand même... 



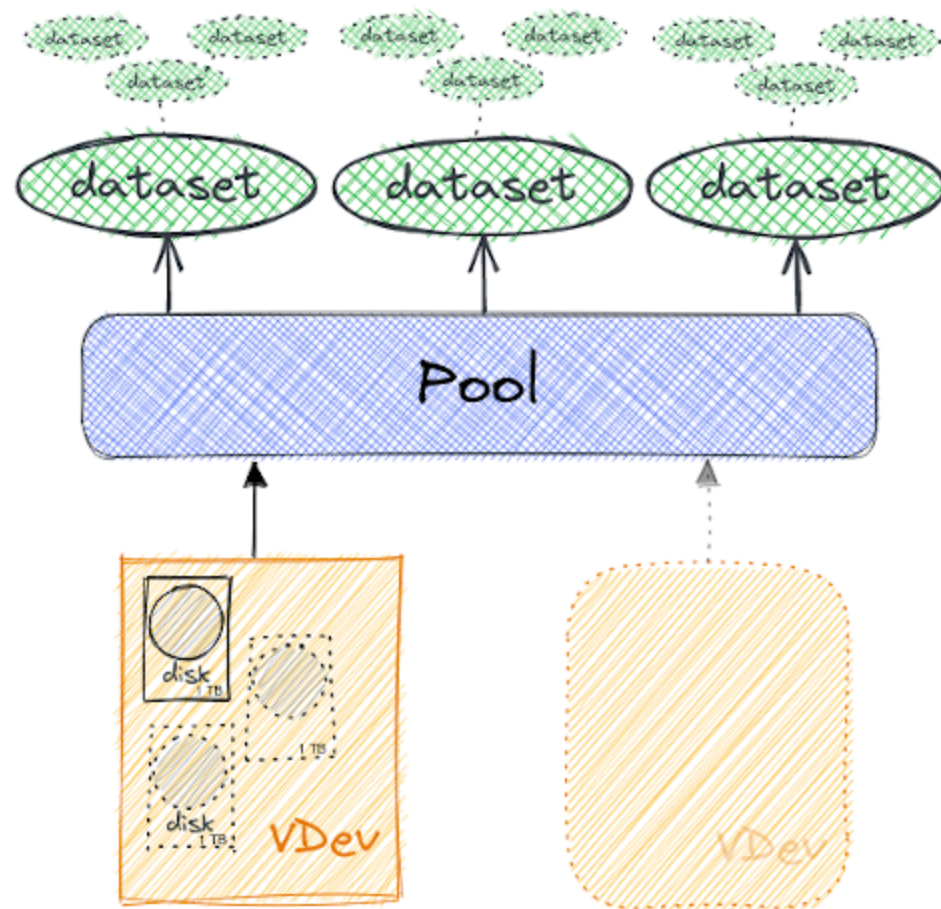
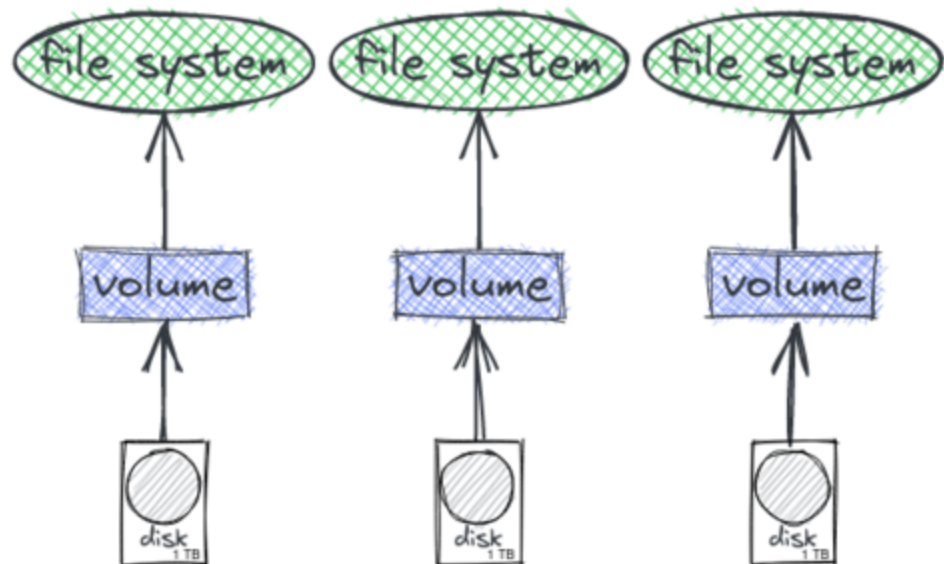
Historique

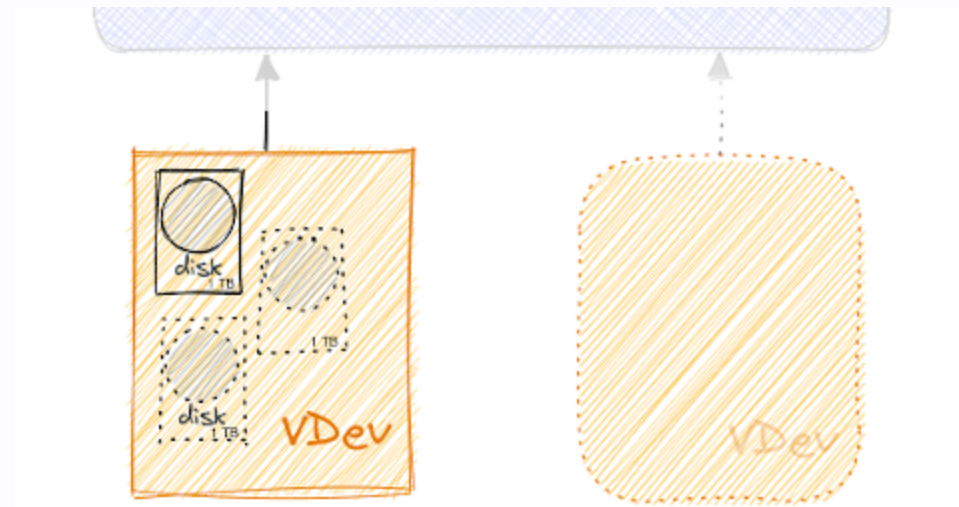
- 2001:  Naissance chez **Sun Microsystems**
- 2005: Le **code source** de ZFS est **publié**
- 2008: ZFS est publié dans **FreeBSD 7.0**
- 2010:  Rachat **Oracle**
- 2010: **illumos/ OpenSolaris**
- 2013: Naissance **OpenZFS**
- 2020:  **ZFS 2.0** Fusion du code **FreeBSD/Linux**

Principaux concepts ZFS

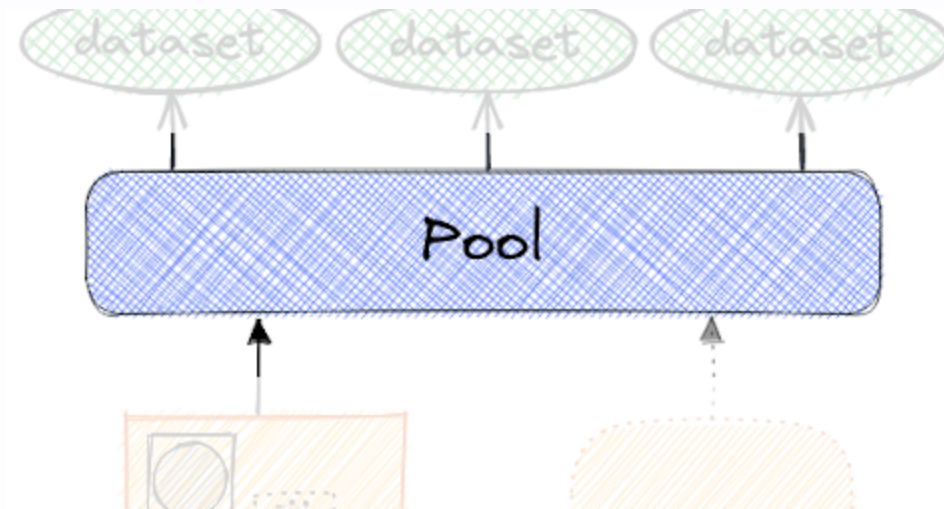


Gestionnaire de volume & système de fichiers

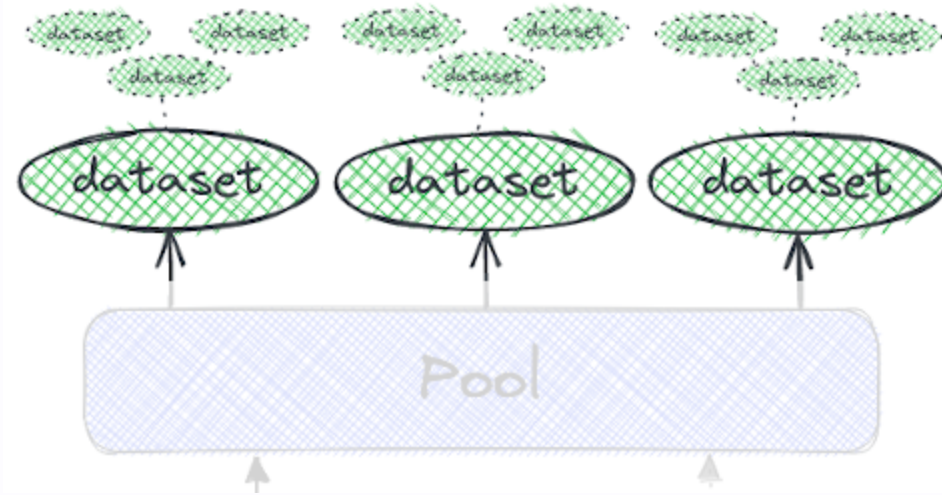




- VDEV == Virtual DEvice
- miroir (+2 disques)
- RAID-Z (1 à 3)
 - Blocs de taille variable
 - Parité distribuée (~RAID5)
- Log / Cache / *spare*





- Constitué de VDEV
- Peut s'agrandir / réduire (*sous conditions*)
- Maintenance préventive
 - reconstruction, *scrub*, **data et metadata**
- Contient des *datasets*










- **Type:** *file-system, snapshot, clone, volume*
- **Héritage:** Gigogne / arborescent
- **Propriétés:** réservation, quota, compress°, dedup°, accès autorisé (ACLs), personnalisée, etc.

Cache

- *Adaptative Replacement Cache*
- MFU & MRU (Most Frequently/Recently Used)
 - L1 (Level 1) -> RAM
 - L2 -> disque
- ZIL (ZFS Intent Log) -> disque
 -  persistence & redondance
 -  [PM Gandi](#)

Copy-On-Write

- «*efface plus tard, ne modifie jamais*»  
-  Modèle transactionnel toujours cohérent
 - pas de `fsck`, jamais (*write hole*)
-  Instantané (Snapshot)
-  Send / receive
 -  plus rapide que `rsync`
-  Gestion de l'espace et taux de remplissage

Administration simple

- Interventions à chaud / en ligne
 - manipulation de disque
 - reconstruction et *scrub* (*donnée et metadonnée*)
- 2 commandes: `zpool` / `zfs`
- Délégation de droit: `zfs allow <user> <perm>`
`<dataset>`

Chez OVHcloud ?



- *Baremetal*
- *Digital core* (Databases)
- *et Storage*

Baremetal

- mirroirs d'image
 - netboot
 - d'installation
 - Debian
 - 180T / HDD 6TB / RAID-Z
 - 1 scrub mensuel (24h)

Digital Core Databases

- sauvegardes MySQL & Postgres
 - ZFS sur l'infra replica ~300T
 - atout: snapshoting et *send/receive*

Storage (*produits*)

Product	PB used	VDev type
Datastore PCC	42	mirror
Backup storage	24	RAID-Z
Web & Mail	21	mirror
NASHA	8	mirror
Internal	0,5	mirror
<i>Backup</i>	128	RAID-Z

Storage (gestion)

- ~128 VM
- Outil de sauvegarde distant ([BorgBackup](#))
 - petit volume / (3 sites distants)
- DB de monitoring ([Zabbix](#))
 - compression / miroir / baremetal

Storage (incidents)

- Ça nous arrive aussi... 🤖
- Mais en proportion minime
- **2022: 2 corruptions clients**
 - ➡ restaurations de sauvegardes
 - ⓘ défaillance disques en simultané



Le secret ?

-  Une équipe qui assure
-  De bons outils...


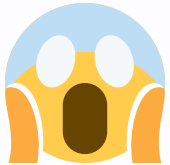
zfswatchd



- 🕒 2016, développé en interne
- démon multi-OS (python)
 - indépendant et autonome
- Déclenche et monitore la gestion des disques
- 💡 SMART, ZFS, OS
- 🗣️ Datacentre, opérations, OS

zfswatchd

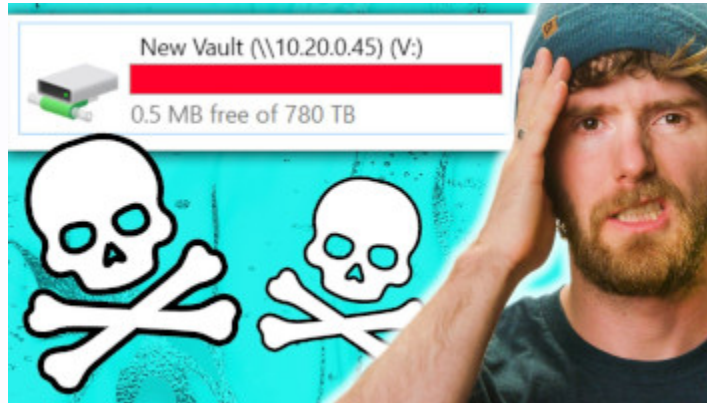
Disk intervention	Quantity
average monthly	81
average weekly	22
Total (since 2016)	15038
monthly scrub	7423

 **Faites gaffe quand
même... **

Gandi - Postmortem: 2020 September 30 storage incident

➡ Erreur humaine: HDD -> ZIL (SSD)

LTT - Our data is GONE... Again



➡ Erreurs humaines: manque de soins

Merci !

- *Matt Ahrens* & *George Wilson* pour: [OpenZFS Basics at SCALE16x](#) (March 2018)
- Ubuntu — [An overview of ZFS concepts](#)
- FreeBSD Handbook — [The Z File System \(ZFS\)](#)
- [Things Nobody Told You About ZFS](#)
- `PU.Baremetal` (*Louis*,...), `PU.Digital Core DB` (*Julien*), `PU.Webhosting` (*Maxime*, ...)
- **PU.storage team** ❤️

!? Questions , remarques,

...

Sources : `github.com/fzindovh/talk-zfs`