Discover ZFS

A reliable, powerful and accessible storage

OpenZFS

OVHcloud

VERY TECH TRIP

Who am I 🗣

🐔 **ZFS pool lifeguard** @ OVHcloud (2020)

- 👪 father
- 🛠 build and use tools
- 🐍 Python Francophone community (AFPy)

📁 file storage

- **backup process**
- **virtualised system** (*Virtual machine images*)
- **database** (*specific requirements*)
- **data processing** (*cache, buffer, etc...*)

ZFS ❓

- *Zettabyte File System*

ZFS ?

- *Zettabyte File System*...or not

> “ I picked ZFS for the simplest of reasons: it sounds cool ”
>
> *Jeff Bonwick*

🗺️ Plan

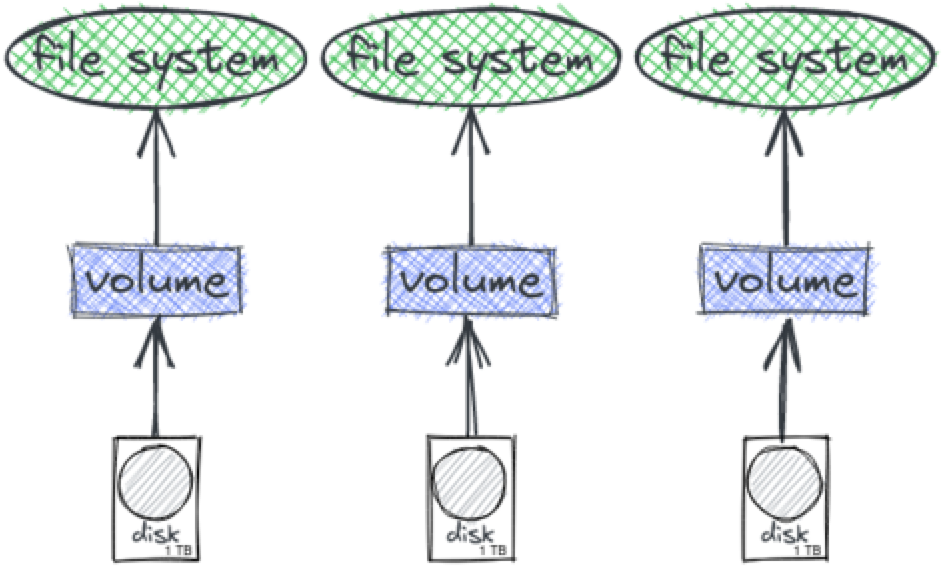
- 🕰️ History
- 💡 ZFS concepts
- 🛠️ Uses and choices at OVH
- 💩 Beware anyway… 😱

🕰️ History

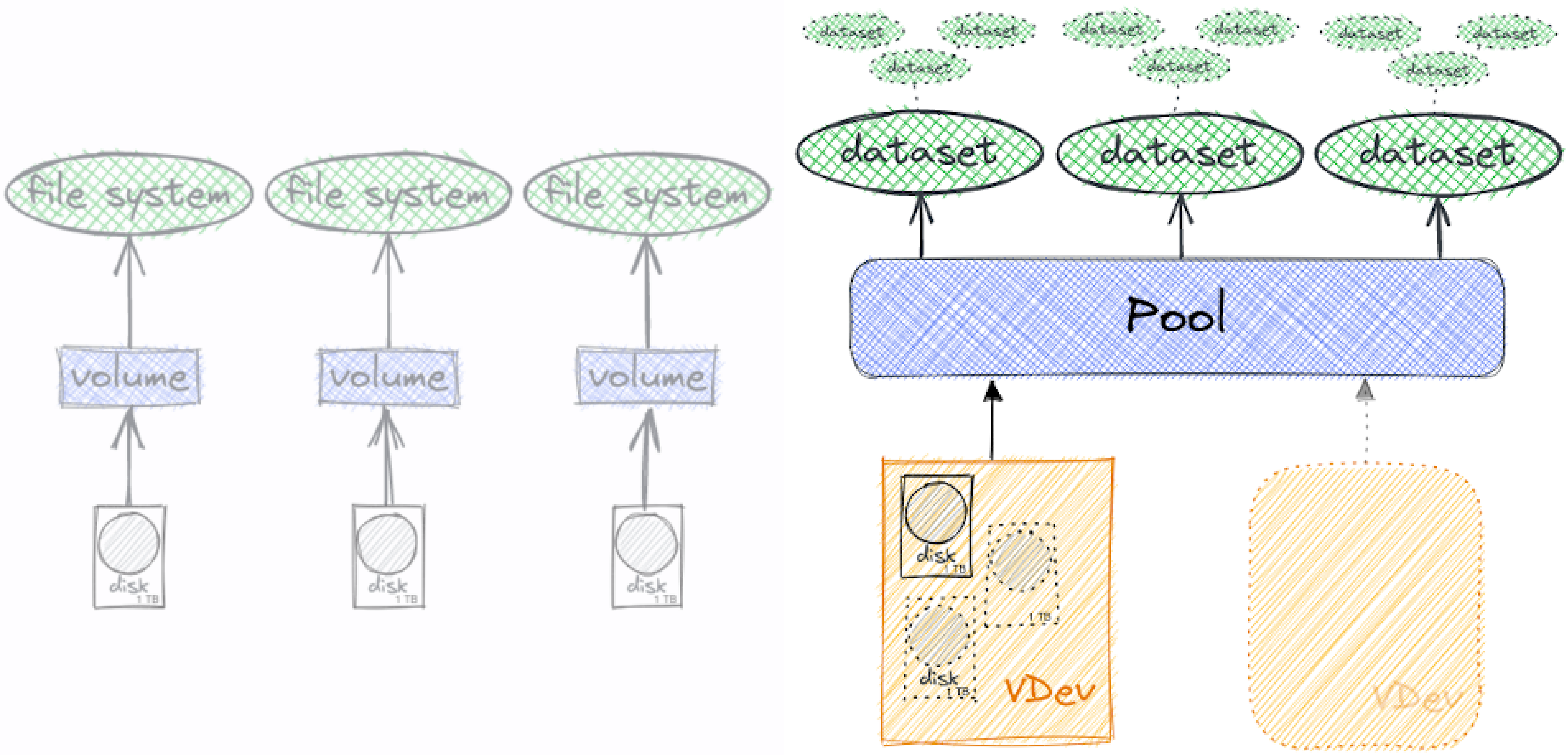
- **2001**: 🍼 **Birth** at Sun Microsystems
- **2005**: ZFS **source code** is **published**
- **2008**: ZFS is published in **FreeBSD 7.0**
- **2010**: 💰 Sun buyout by **Oracle**
- **2010**: Illumos/ OpenSolaris
- **2013**: 🍼 Birth **OpenZFS**
- **2020**: 🌋 ZFS 2.0 Code Merge **FreeBSD/Linux**

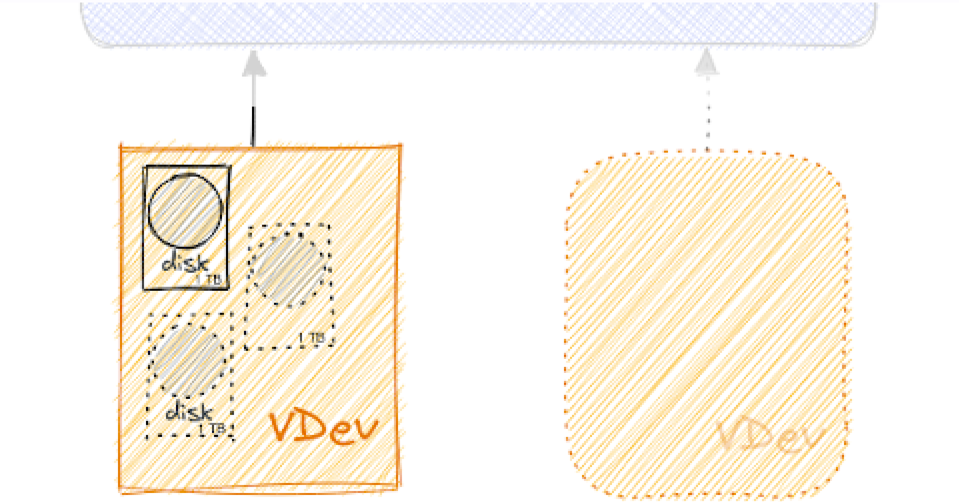
ZFS key concepts 💡

Volume Manager & File System

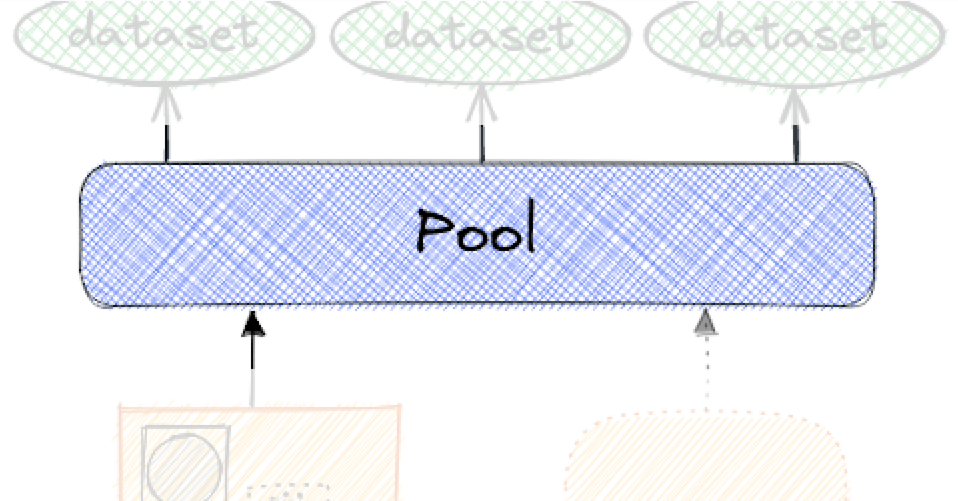


Volume Manager & File System

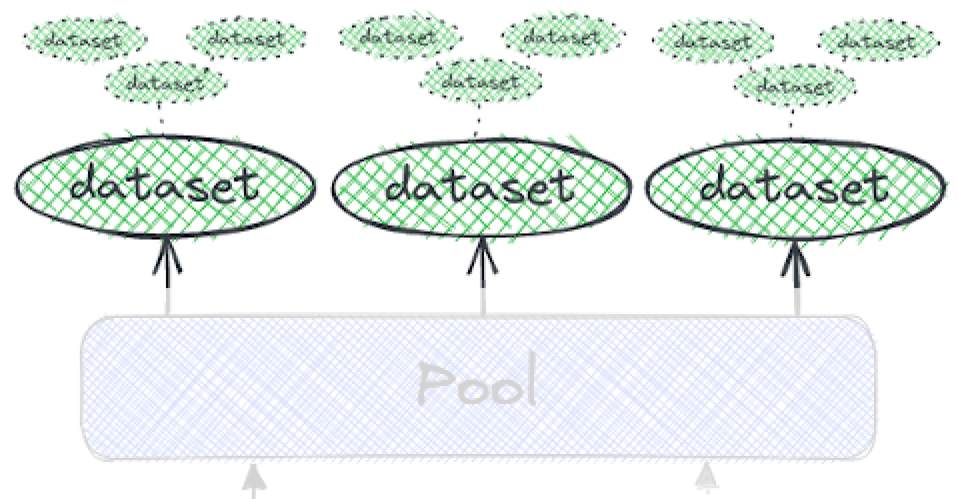




- `VDEV == Virtual DEVice`
- mirror (+2 disks)
- RAID-Z (1-3)
  - Variable block size
  - Distributed parity (~RAID5)
- Log / Cache / spare



- Consisting of `VDEV`
- Can expand / collapse (*under conditions*)
- Preventive maintenance
  - reconstruction, scrub, **data and metadata**
- Contains datasets



- **Type:** file-system, snapshot, clone, volume
- **Legacy:** nested / arborescent
- **Properties:** reservation, quota, compress°, dedup°, authorised access (ACLs), personalised, etc.

⚡ Cache

- *Adaptative Replacement Cache*
- MFU & MRU (Most Frequently/Recently Used)
  - L1 (Level 1) -> RAM
  - L2 -> disk
- ZIL (ZFS Intent Log) -> disk
  - ⚠️ persistence & redundancy
  - ➡️ PM Gandi

🎇 Copy-On-Write

- *“delete later, never modify”* 🗑️ ⏳
- ✅ consistently transactional model
  - no `fsck`, never (write hole)
- 📸 Snapshot
- 🔄 Send / receive
  - 🚀 faster than `rsync`
- ⚠️ Space management and usage

🤓 Easy administration

- Hot/online operations
  - disk manipulation
  - resilvering and scrub (*data and metadata*)
- 2 commands: `zpool` / `zfs`
- Delegation rights: `zfs allow <user> <perm> <dataset>`

At OVHcloud ?

OVHcloud

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

Baremetal

- image mirrors
  - netboot
  - installation
    - Debian
    - 180T / HDD 6TB / RAID-Z
    - 1 monthly scrub (24h)

Digital Core Databases

- MySQL & Postgres backups
  - ZFS on the ~300T replica infrastructure
  - asset: snapshoting and send/receive

Storage (*products*)

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
Backup	128	RAID-Z

Storage (Management)

- ~128 VM
- Remote backup tool (BorgBackup)
  - small volume / (3 remote sites)
- Monitoring DB (Zabbix)
  - compression / mirroring / bare metal

Storage (incidents)

- It also happens to us… 😱
- But in small proportion
- **2022:** *2 customer corruptions*
  - ➡️ backup restoration
  - ℹ️ simultaneous disk failure

🧙 Secret?

- 👪 a team that rocks
- 🛠️ good tools...

`zfswatchd`

😇 👩‍🔧 👩‍🚒 🤖

- 🕰️ 2016, in-house developed
- multi-OS daemon (python)
  - independent and autonomous
- Triggers and monitors disk management
- 👂 SMART, ZFS, OS
- 🗣️ Datacentre, operations, OS

`zfswatchd`

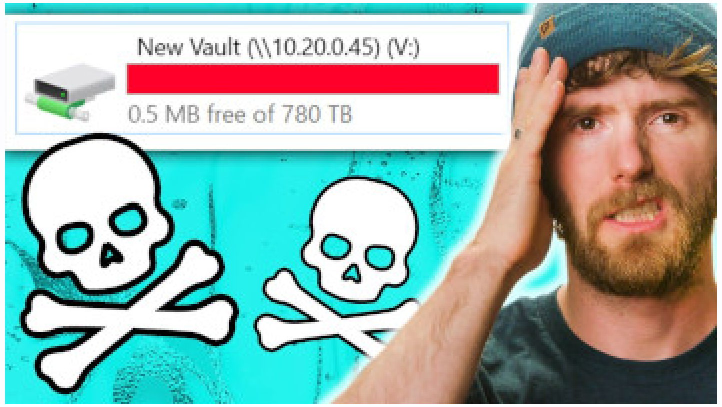
Disk intervention	Quantity
average monthly	81
average weekly	22
Total (since 2016)	15038
monthly scrub	7423

💩 Be careful... 😱

Gandi - Postmortem: 2020 September 30 storage incident

➡️ human error: HDD -> ZIL (SSD)

LTT - Our data is GONE... Again



➡️ Errors: lack of care

🤝 Thank you!

- *Matt Ahrens* & *George Wilson* for: 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, remarks…

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