

# ZFS met RAID-Z als alternatief voor klassieke RAID-oplossingen

Jonas De Moor

Toegepaste Informatica - Systeem- en Netwerkbeheer  
Hogeschool Gent

*[jonas.demoor.v3741@student.hogent.be](mailto:jonas.demoor.v3741@student.hogent.be)*

15 juni 2017

## 1 Achtergrond

- Motivatie
- Onderzoeksvragen
- Opbouw van het onderzoek
- Gehanteerde methodiek

## 2 Onderzoek

- Achtergrondinformatie m.b.t. ZFS
- Architectuur van ZFS
- VDEV's & Storage Pools
- Benchmarks
- Betrouwbaarheidstesten

## 3 Conclusie

# Motivatie voor het voeren van dit onderzoek

- RAID5 'write hole'
- Relatie tussen BTRFS en ZFS
- ZFS On Linux (cf. Ubuntu 16.04 LTS)
- Interesses: Linux en Unix

- Wat zijn de grootste verschillen tussen een klassieke RAID-oplossing en ZFS RAID-Z?
- Hoe is de architectuur van ZFS opgebouwd en op welke manieren tracht het oplossingen te vinden voor de problemen die zich voordoen bij andere bestandssystemen en RAID-opstellingen?
- Hoe staat het met data-integriteit en performantie<sup>1</sup> bij ZFS onder verschillende workloads en toepassingen?

---

<sup>1</sup>Met 'performantie' wordt het aantal I/O's per seconde en de globale CPU-belasting bedoeld.

Twee grote onderdelen:

## ① Theoretisch gedeelte

- Inleiding tot RAID-niveaus
- Architectuur en ontwerpprincipes van ZFS
- Interne datastructuren en transactiemodel

## ② Praktisch gedeelte

- Storage Pools & VDEV's
- Datasets
- Performantie & Betrouwbaarheid

- Phoronix Benchmark: performantietesten op fysieke machine
  - FIO (Flexible I/O Tester): IOPS
  - FS-Mark: bestandssysteemoperaties
  - PostMark: simulatie van webserver/mailserver
  - SQLite: databankoperaties
- Virtuele Machine: betrouwbaarheidstesten
  - Wegvallen van een schijf (array van drie schijven)
  - Dataverlies door gebruikersfout
  - Bescherming tegen datacorruptie

Specificaties	
<b>Fabrikant</b>	HP
<b>Model</b>	HP Pavilion Elite HPE-310be
<b>CPU</b>	Intel Core i5 650 @ 3.2 GHz (2 Cores; 4 Threads)
<b>Geheugen</b>	10GB DDR3 @ 1333MHz
<b>GPU</b>	AMD Radeon HD 5570
<b>Interne schijven</b>	SAMSUNG HD103SJ (1TB)
	WDC WD1002FAEX-0 (1TB)
	WDC WD5000AZRX-0 (500GB)
<b>Externe schijf</b>	WD Elements 1078 (1TB)
<b>RAID Controller</b>	Intel Corporation SATA RAID Controller

**Tabel:** Specificaties van het fysieke systeem dat gebruikt werd doorheen de bachelorproef (data verkregen via lshw)

Specificaties Virtuele Machine	
<b>OS</b>	Fedora Server 25
<b>CPU</b>	4x Host CPU (Intel Core i7-4712HQ CPU @ 2.30GHz)
<b>Geheugen</b>	8GB
<b>OS-schijf</b>	20GB (/dev/sda; SATA non-hot-pluggable)
<b>Zpool schijven</b>	40GB (/dev/sdb; SATA hot-pluggable)
	40GB (/dev/sdc; SATA hot-pluggable)
	40GB (/dev/sdd; SATA hot-pluggable)
<b>NIC's</b>	VirtualBox NAT-adapter (10.0.2.15/24)
	VirtualBox Host-only Adapter (192.168.56.10/24)

**Tabel:** Specificaties van de virtuele machine die gebruikt werd voor de betrouwbaarheidstesten



## 1 Achtergrond

- Motivatie
- Onderzoeksvragen
- Opbouw van het onderzoek
- Gehanteerde methodiek

## 2 Onderzoek

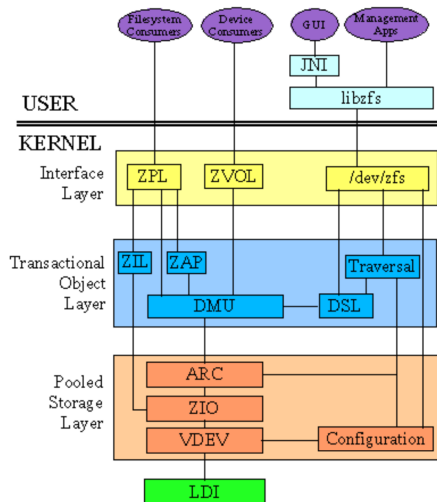
- Achtergrondinformatie m.b.t. ZFS
- Architectuur van ZFS
- VDEV's & Storage Pools
- Benchmarks
- Betrouwbaarheidstesten

## 3 Conclusie

# ZFS: een kort overzicht

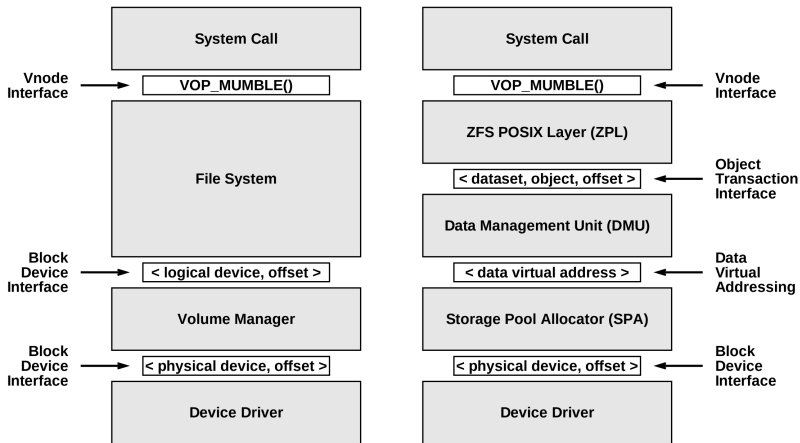
- Copy-On-Write bestandssysteem
- Ontwikkeld door Sun Microsystems (begin jaren 2000)
- Oorspronkelijk onderdeel van Solaris
- Nu: verdere ontwikkeling via OpenZFS (en Oracle)
- Ondertussen ook beschikbaar op BSD en Linux (ZFS on Linux)
- Beschikt over RAID-Z (softwarematige RAID)

# Architectuur van ZFS



Figuur: Een overzicht van de verschillende componenten van ZFS (Kendi, Onbekend)

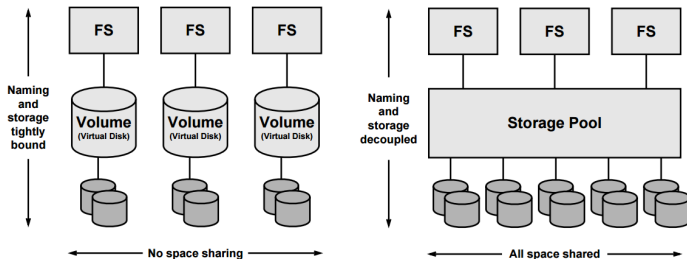
# Architectuur van ZFS



**Figuur:** Vergelijking tussen een 'traditionele' storage stack (links) en de ZFS storage stack (rechts) (Bonwick e.a., 2002)

# Storage Pools

- Abstractie voor fysieke apparaten → gegroepeerd in VDEV's
- Dynamische allocatie van opslagruimte
- Schijven kunnen worden toegevoegd zonder downtime<sup>2</sup>

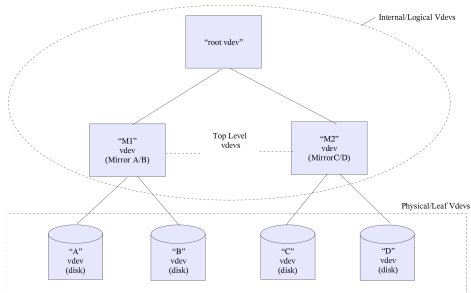


**Figuur:** Illustratie van ZFS pooled storage (rechts) t.o.v. volume-based storage (links) (Bonwick e.a., 2002)

<sup>2</sup>Afhankelijk van de situatie

# VDEV's: Virtual Devices

- Bouwstenen van storage pools
- RAID-niveaus binnen ZFS:
  - Stripes, Mirrors, RAID-Z, etc.
- Speciale VDEV's:
  - SLOG, L2ARC



**Figuur:** Conceptuele voorstelling van VDEV's in een boomstructuur (Sun Microsystems, 2006)

# Voorbeeld: zpool met een RAID-Z VDEV

```
$ zpool create storage raidz1 /dev/sda /dev/sdb /dev/sdc
$ zpool status
pool: storage
state: ONLINE
scan: none requested
config:
```

NAME	STATE	READ	WRITE	CKSUM
storage	ONLINE	0	0	0
raidz1-0	ONLINE	0	0	0
sda	ONLINE	0	0	0
sdb	ONLINE	0	0	0
sdc	ONLINE	0	0	0

```
errors: No known data errors
```

## 1 Achtergrond

- Motivatie
- Onderzoeksvragen
- Opbouw van het onderzoek
- Gehanteerde methodiek

## 2 Onderzoek

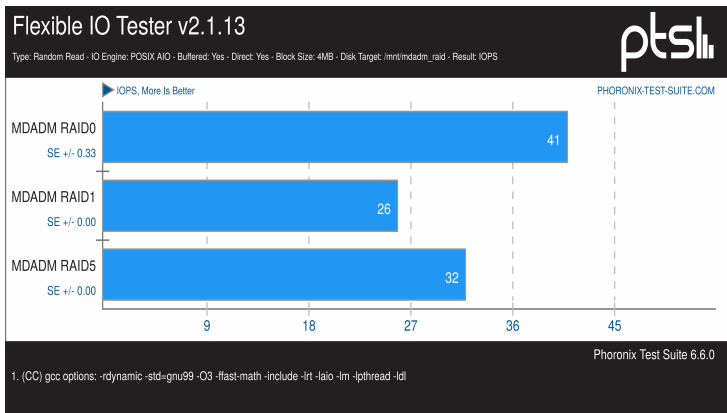
- Achtergrondinformatie m.b.t. ZFS
- Architectuur van ZFS
- VDEV's & Storage Pools
- Benchmarks
- Betrouwbaarheidstesten

## 3 Conclusie



# Benchmarks

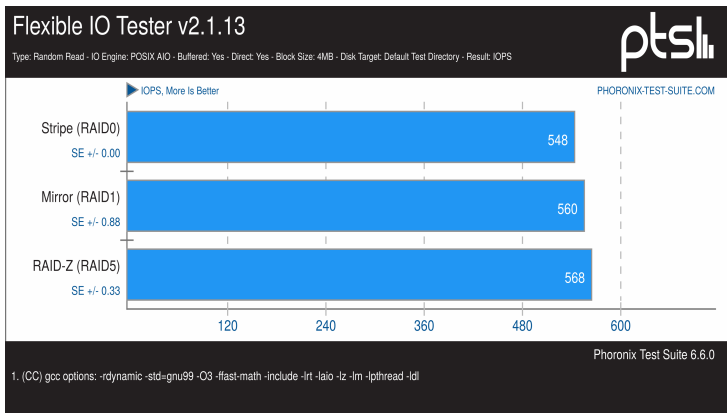
FIO-benchmark: aantal IOPS (Invoer/Uitvoer-bewerkingen per seconde)



**Figuur:** Aantal IOPS bij random read operaties (blokgrootte: 4MB), uitgevoerd op een Linux MD-opstelling

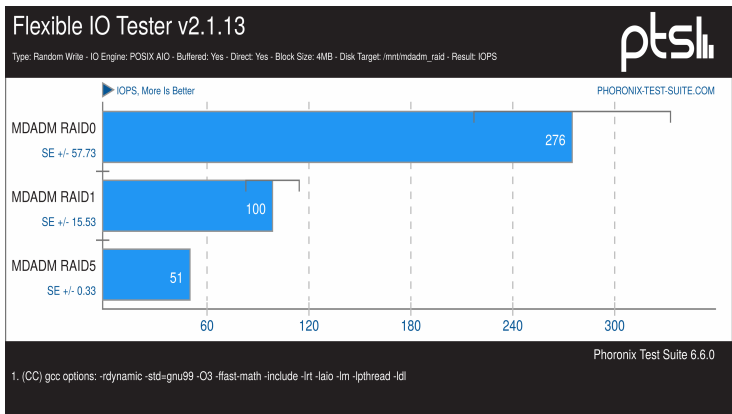
# Benchmarks

FIO-benchmark: aantal IOPS (Invoer/Uitvoer-bewerkingen per seconde)



**Figuur:** Aantal IOPS bij random read operaties (blokgrootte: 4MB), uitgevoerd op een ZFS-opstelling

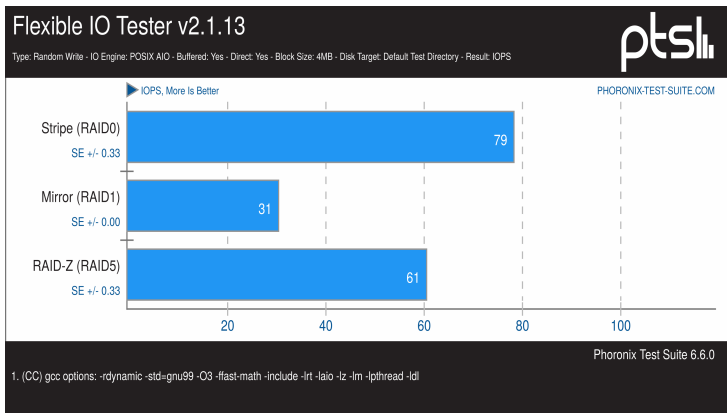
FIO-benchmark: aantal IOPS (Invoer/Uitvoer-bewerkingen per seconde)



**Figuur:** Aantal IOPS bij random write operaties (blokgrootte: 4MB), uitgevoerd op een Linux MD-opstelling

# Benchmarks

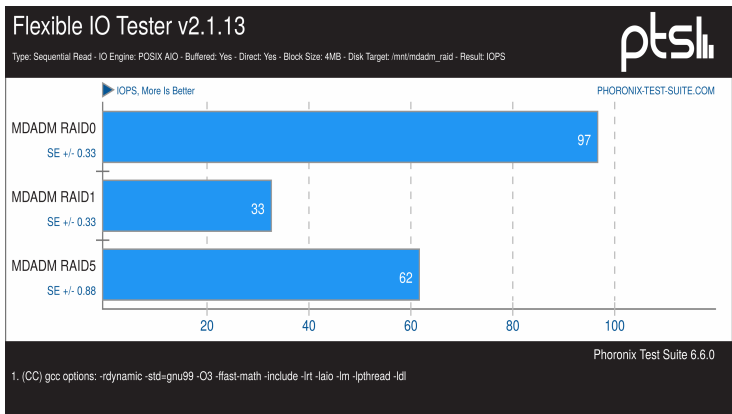
FIO-benchmark: aantal IOPS (Invoer/Uitvoer-bewerkingen per seconde)



**Figuur:** Aantal IOPS bij random write operaties (blok grootte: 4MB), uitgevoerd op ZFS-opstelling

# Benchmarks

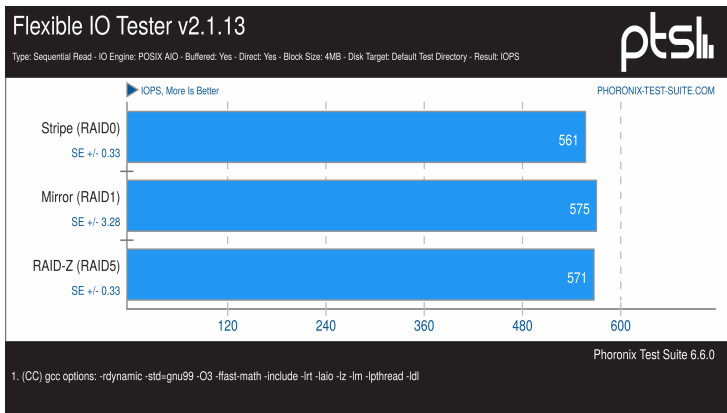
FIO-benchmark: aantal IOPS (Invoer/Uitvoer-bewerkingen per seconde)



**Figuur:** Aantal IOPS bij sequential read operaties (blokgrootte: 4MB), uitgevoerd op een Linux MD-opstelling

# Benchmarks

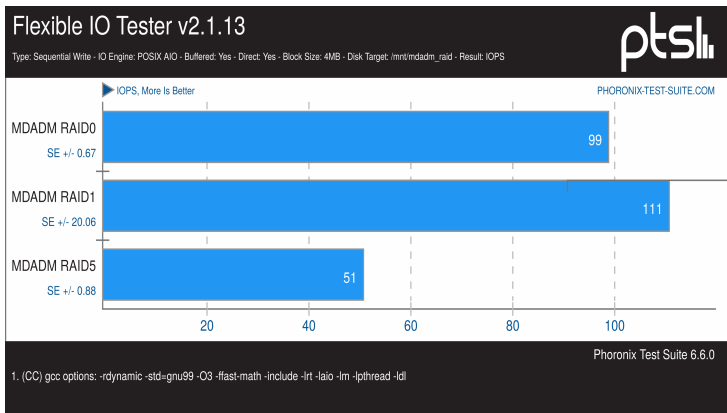
FIO-benchmark: aantal IOPS (Invoer/Uitvoer-bewerkingen per seconde)



**Figuur:** Aantal IOPS bij sequential read operaties (blok grootte: 4MB), uitgevoerd op een ZFS-opstelling

# Benchmarks

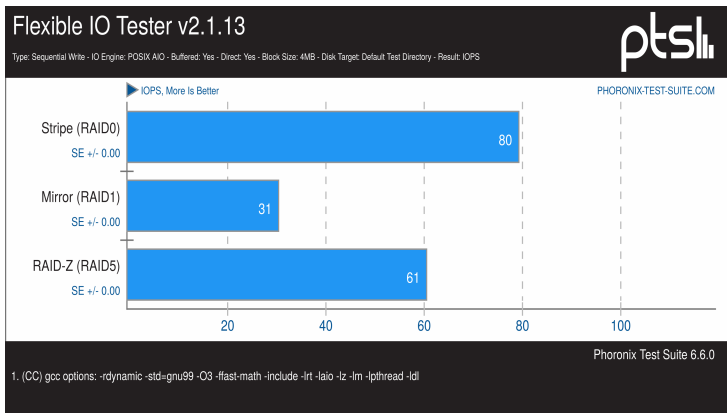
FIO-benchmark: aantal IOPS (Invoer/Uitvoer-bewerkingen per seconde)



**Figuur:** Aantal IOPS bij sequential write operaties (blokgrootte: 4MB), uitgevoerd op een Linux MD-opstelling

# Benchmarks

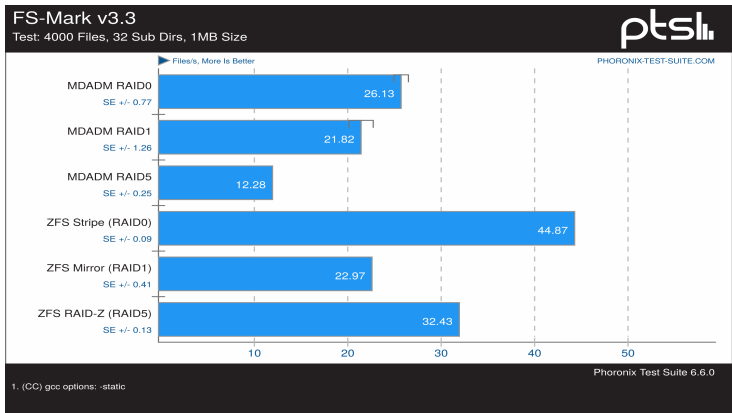
FIO-benchmark: aantal IOPS (Invoer/Uitvoer-bewerkingen per seconde)



**Figuur:** Aantal IOPS bij sequential write operaties (blokgrootte: 4MB), uitgevoerd op een ZFS-opstelling

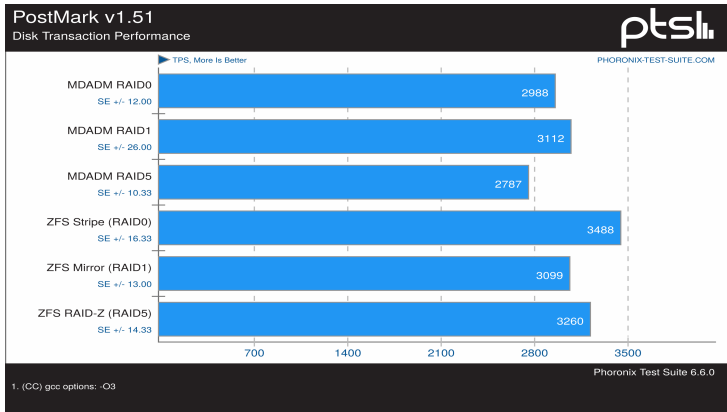


FS-Mark: algemene bestandssysteemperformantie (bestanden per seconde)



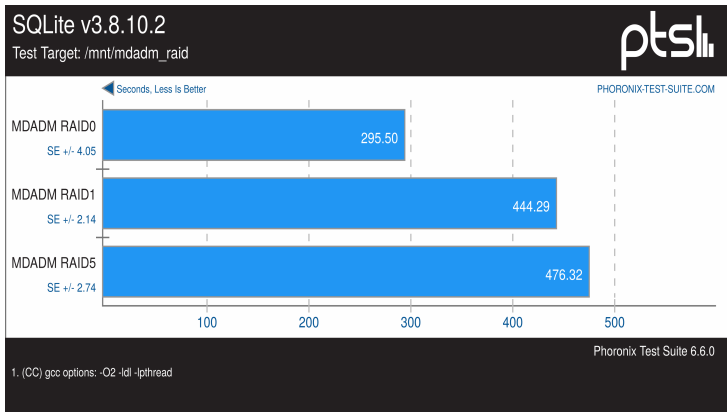
**Figuur:** Vergelijking tussen Linux MD i.c.m. XFS en ZFS inzake algemene bestandssysteemperformantie

## PostMark: Simulatie van de workload van een mail- of webserver



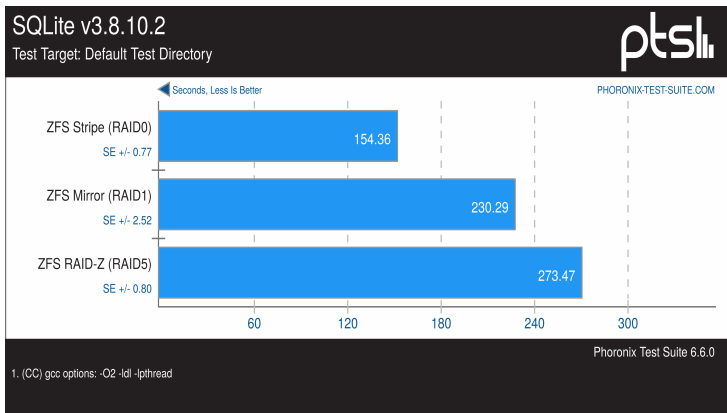
**Figuur:** Simulatie van een web- of mailserver waarbij de prestatie van respectievelijk Linux MD i.c.m. XFS en ZFS met elkaar wordt vergeleken

## SQLite: Simulatie van de workload van een databanksysteem



**Figuur:** Performantie van Linux MD i.c.m. XFS bij een groot aantal INSERT-bewerkingen op een SQLite-databank

## SQLite: Simulatie van de workload van een databanksysteem



**Figuur:** Performantie van ZFS bij een groot aantal INSERT-bewerkingen op een SQLite-databank

## 1 Achtergrond

- Motivatie
- Onderzoeksvragen
- Opbouw van het onderzoek
- Gehanteerde methodiek

## 2 Onderzoek

- Achtergrondinformatie m.b.t. ZFS
- Architectuur van ZFS
- VDEV's & Storage Pools
- Benchmarks
- Betrouwbaarheidstesten

## 3 Conclusie

- Dataverlies door een gebruikersfout<sup>3</sup>

```
$ zfs snapshot storage@$(date "+%d-%m-%Y")
$ zfs list -t snapshot
NAME                                USED  AVAIL  REFER  MOUNTPOINT
storage@31-05-2017                 0      -   69.8G  -
$ rm -f /storage/dummy_2
$ ls -lh /storage/
total 56G
-rw-r--r--. 1 root root 14G May 31 17:18 dummy_1
-rw-r--r--. 1 root root 14G May 31 17:21 dummy_3
-rw-r--r--. 1 root root 14G May 31 17:23 dummy_4
-rw-r--r--. 1 root root 14G May 31 17:24 dummy_5
```

---

<sup>3</sup>Testdata gegenereerd met: `for i in 1..5; do head -c 15GB </dev/urandom > /storage/dummy_$i; done`

- Dataverlies door een gebruikersfout

```
$ zfs rollback storage@31-05-2017
$ ls -lh /storage/
total 70G
-rw-r--r--. 1 root root 14G May 31 17:18 dummy_1
-rw-r--r--. 1 root root 14G May 31 17:19 dummy_2
-rw-r--r--. 1 root root 14G May 31 17:21 dummy_3
-rw-r--r--. 1 root root 14G May 31 17:23 dummy_4
-rw-r--r--. 1 root root 14G May 31 17:24 dummy_5
```

- Gedrag van de array bij het wegvallen van een schijf

```
# Op het hostsysteem
```

```
$ VBoxManage storageattach "Fedora Server x64" --storagectl "SATA" --port 1 --d
```

```
# Op de virtuele machine
```

```
$ dmesg
```

```
(deel van de uitvoer is weggelaten)
```

```
[ 6772.524376] ata2: exception Emask 0x10 SAct 0x0 SErr 0x4010000 action 0xe fr
```

```
[ 6772.525402] ata2: irq_stat 0x80400040, connection status changed
```

```
[ 6772.525670] ata2: SError: { PHYRdyChg DevExch }
```

```
[ 6772.525866] ata2: hard resetting link
```

```
[ 6773.198452] ata2: SATA link down (SStatus 0 SControl 300)
```



- Gedrag van de array bij het wegvallen van een schijf

```
$ zpool export storage
$ zpool import storage
$ zpool status
  pool: storage
state: DEGRADED
status: One or more devices could not be used because the label is missing or
invalid.  Sufficient replicas exist for the pool to continue
functioning in a degraded state.
action: Replace the device using 'zpool replace'.
  see: http://zfsonlinux.org/msg/ZFS-8000-4J
  scan: none requested
config:
```

NAME	STATE	READ	WRITE	CKSUM	
storage	DEGRADED	0	0	0	
raidz1-0	DEGRADED	0	0	0	
18175546172533204033	UNAVAIL	0	0	0	was /dev/sdb1
sdc	ONLINE	0	0	0	
sdd	ONLINE	0	0	0	

```
errors: No known data errors
```

- Gedrag van de array bij het wegvallen van een schijf

```
$ zpool replace storage 18175546172533204033 /dev/sdb -f
$ zpool status
  pool: storage
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 Wed May 31 19:35:54 2017
         6.88G scanned out of 105G at 227M/s, 0h7m to go
         2.29G resilvered, 6.56% done
config:
```

NAME	STATE	READ	WRITE	CKSUM	
storage	DEGRADED	0	0	0	
raidz1-0	DEGRADED	0	0	0	
replacing-0	UNAVAIL	0	0	0	
18175546172533204033	UNAVAIL	0	0	0	was /dev/sdb1/old
sdb	ONLINE	0	0	0	(resilvering)
sdc	ONLINE	0	0	0	
sdd	ONLINE	0	0	0	

errors: No known data errors

- Gedrag van de array bij het optreden van silent data corruption

```
$ sha256sum /storage/dummy_1  
fc4c5c62db504cec7b5cafa264c329416d0207da9e4a61066bb07563caf9ec2e  /storage/dumm  
  
$ zpool export storage
```

- Gedrag van de array bij het optreden van silent data corruption

```
$ zpool import storage
$ zpool status
  pool: storage
state: ONLINE
status: One or more devices has experienced an unrecoverable error. An
attempt was made to correct the error. Applications are unaffected.
action: Determine if the device needs to be replaced, and clear the errors
using 'zpool clear' or replace the device with 'zpool replace'.
  see: http://zfsonlinux.org/msg/ZFS-8000-9P
  scan: none requested
config:
```

NAME	STATE	READ	WRITE	CKSUM
storage	ONLINE	0	0	0
raidz1-0	ONLINE	0	0	0
sdb	ONLINE	0	0	5
sdc	ONLINE	0	0	0
sdd	ONLINE	0	0	0

```
errors: No known data errors
```

- Gedrag van de array bij het optreden van silent data corruption

```
$ zpool scrub storage
$ zpool status
  pool: storage
state: ONLINE
status: One or more devices has experienced an unrecoverable error. An
attempt was made to correct the error. Applications are unaffected.
action: Determine if the device needs to be replaced, and clear the errors
using 'zpool clear' or replace the device with 'zpool replace'.
  see: http://zfsonlinux.org/msg/ZFS-8000-9P
  scan: scrub repaired 39.0M in 0h0m with 0 errors on Wed May 31 21:58:32 2017
config:
```

NAME	STATE	READ	WRITE	CKSUM
storage	ONLINE	0	0	0
raidz1-0	ONLINE	0	0	0
sdb	ONLINE	0	0	640
sdc	ONLINE	0	0	0
sdd	ONLINE	0	0	0

```
errors: No known data errors
```

- Bonwick, J. e.a. (2002). *The Zettabyte Filesystem*. Verkregen van <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.184.3704&rep=rep1&type=pdf>
- Kendi, C. (Onbekend). ZFS: Enhancing the Open Source Storage System (and the Kernel). Verkregen van [https://www.blackhat.com/presentations/bh-dc-10/Kendi\\_Christian/Blackhat-DC-2010-Kendi-Enhancing-ZFS-slides.pdf](https://www.blackhat.com/presentations/bh-dc-10/Kendi_Christian/Blackhat-DC-2010-Kendi-Enhancing-ZFS-slides.pdf)
- Sun Microsystems. (2006). *ZFS on-disk specification*. Verkregen van [http://www.giis.co.in/Zfs\\_ondiskformat.pdf](http://www.giis.co.in/Zfs_ondiskformat.pdf)

Zijn er nog vragen?