OpenZFS: co je nového

Pavel Šnajdr
LinuxDays 2017

OpenZFS: co je nového

- stable release 0.7.0

- Encryption
- dRAID
- Sequential Scrubs and Resilvers
- VDEV Removal
- ZFS Channel Programs
- Allocation Classes

OpenZFS 0.7.0

ABD

Slab → memory fragmentation
→ wasted memory

OpenZFS 0.7.0

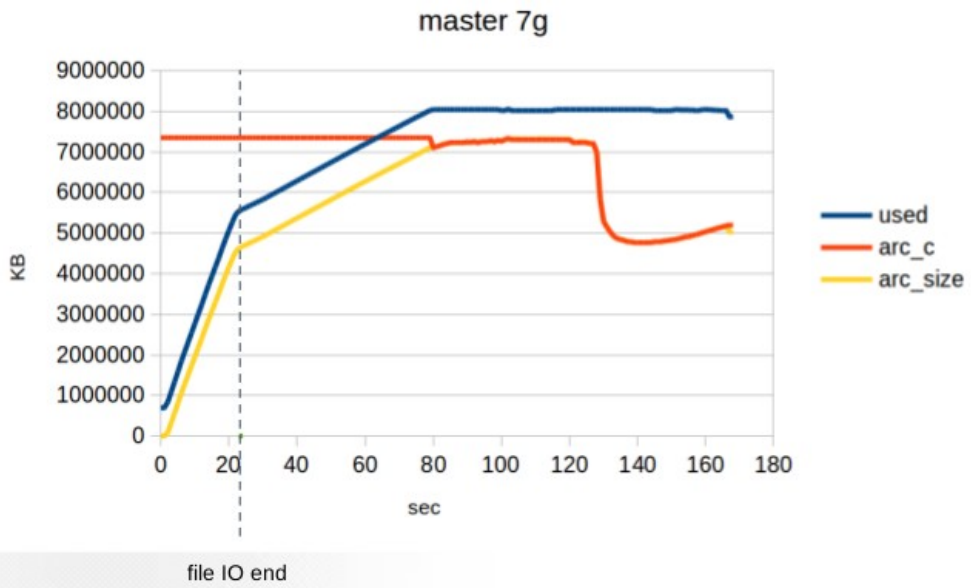
ABD

Slab → memory fragmentation
→ wasted memory

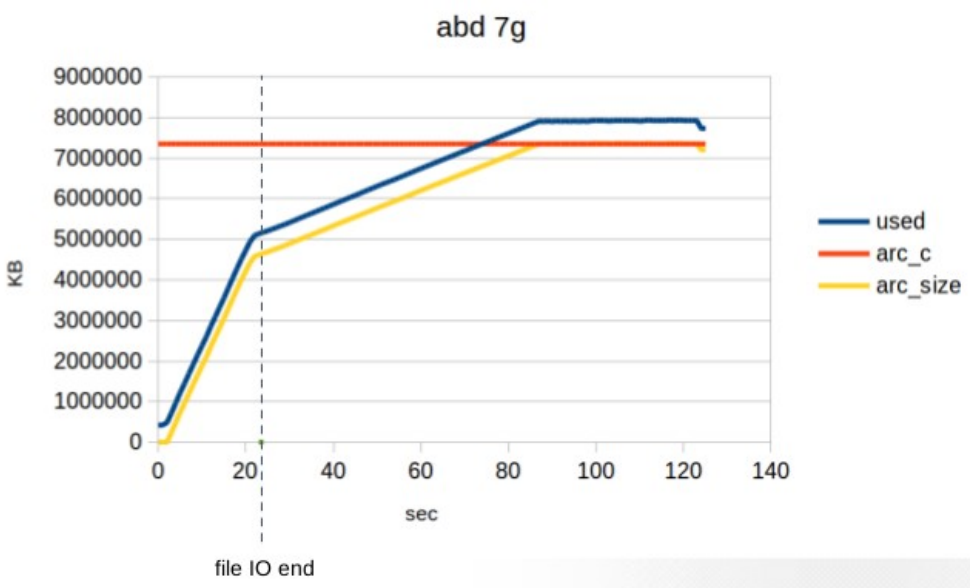
→ user data out of SLAB
scatter/gather

OpenZFS 0.7.0

ABD



file: 175906KB/s
zvol: 25509KB/s



file: 178405KB/s
zvol: 34565KB/s

OpenZFS 0.7.0

Compressed ARC

OpenZFS 0.7.0

Compressed ARC

HW optimalizace (vektorizace raidz, fletcher4)

OpenZFS 0.7.0

Compressed ARC

HW optimalizace (vektorizace raidz, fletcher4)

New checksums: SHA-512, Skein, or Edon-R

OpenZFS 0.7.0

Compressed ARC

HW optimalizace (vektorizace raidz, fletcher4)

New checksums: SHA-512, Skein, or Edon-R

dnode accounting

OpenZFS 0.7.0

Compressed ARC

HW optimalizace (vektorizace raidz, fletcher4)

New checksums: SHA-512, Skein, or Edon-R

dnode accounting

Basic JBOD management

OpenZFS 0.7.0

zfs send -c

→ compressed send

OpenZFS 0.7.0

zfs send -c

→ compressed send

zfs send -t <token> | ssh ... zfs recv -s pool/dset

→ resumable send/recv

OpenZFS 0.7.0

zfs send -c

→ compressed send

zfs send -t <token> | ssh ... zfs recv -s pool/dset

→ resumable send/recv

zpool iostat -w | -l | -r

→ request histogram | latency | request size

OpenZFS 0.7.0

zfs send -c
→ compressed send

zfs send -t <token> | ssh ... zfs recv -s pool/dset
→ resumable send/recv

zpool iostat -w | -l | -r
→ request histogram | latency | request size

zpool scrub -s
→ scrub pause

OpenZFS Encryption

Ted': Nad/pod ZFS (dmcrypt/ecryptfs)

OpenZFS Encryption

Teď: Nad/pod ZFS (dmcrypt/ecryptfs)

Native ZFS Encryption?

- výkon
- čistější implementace
- snadnější správa
- zálohy nepotřebují klíče

OpenZFS Encryption

Teď: Nad/pod ZFS (dmcrypt/ecryptfs)

Native ZFS Encryption?

- výkon
- čistější implementace
- snadnější správa
- zálohy nepotřebují klíče

→ block level @ ZFS stack

OpenZFS Encryption

zfs get | set
encryption=off

OpenZFS Encryption

```
zfs get | set
        encryption=off    \
             | aes-128-ccm     \
             | aes-192-ccm     \
             | aes-256-ccm     \
             | aes-128-gcm     \
             | aes-192-gcm     \
             | aes-256-gcm
```

OpenZFS Encryption

```
zfs get | set
          encryption=off     \
                  | aes-128-ccm         \
                  | aes-192-ccm         \
                  | aes-256-ccm | on    \
                  | aes-128-gcm         \
                  | aes-192-gcm         \
                  | aes-256-gcm
```

OpenZFS Encryption

```
zfs get | set
        encryption=off     \
                | aes-128-ccm         \
                | aes-192-ccm         \
                | aes-256-ccm | on    \
                | aes-128-gcm         \
                | aes-192-gcm         \
                | aes-256-gcm

        keyformat=raw | hex | passphrase
        pbkdf2iters=350000 (> 100k)
```

OpenZFS Encryption

```
zfs get | set
        encryption=off     \
                | aes-128-ccm         \
                | aes-192-ccm         \
                | aes-256-ccm | on    \
                | aes-128-gcm         \
                | aes-192-gcm         \
                | aes-256-gcm

        keyformat=raw | hex | passphrase
        pbkdf2iters=350000 (> 100k)
        keylocation=prompt | file://
```

OpenZFS Encryption

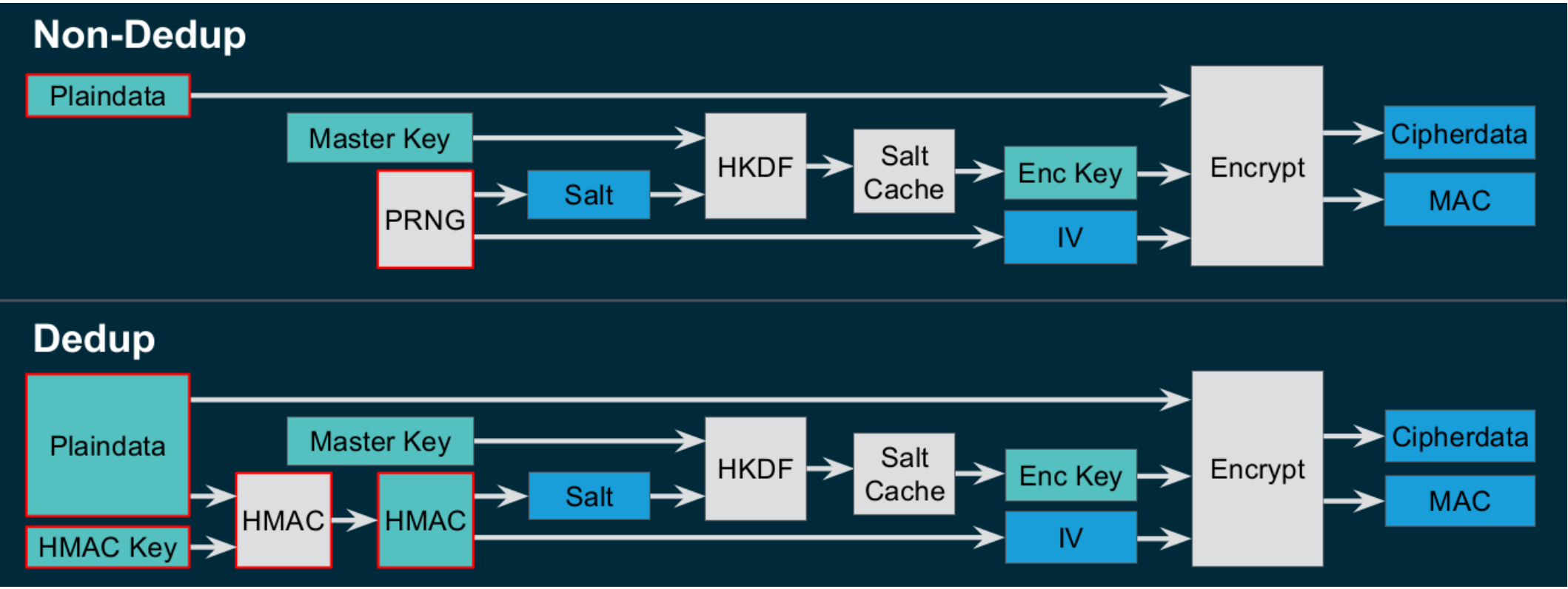
zfs get keystatus

zfs load-key [-nr] [-L keylocation] -a | filesystem

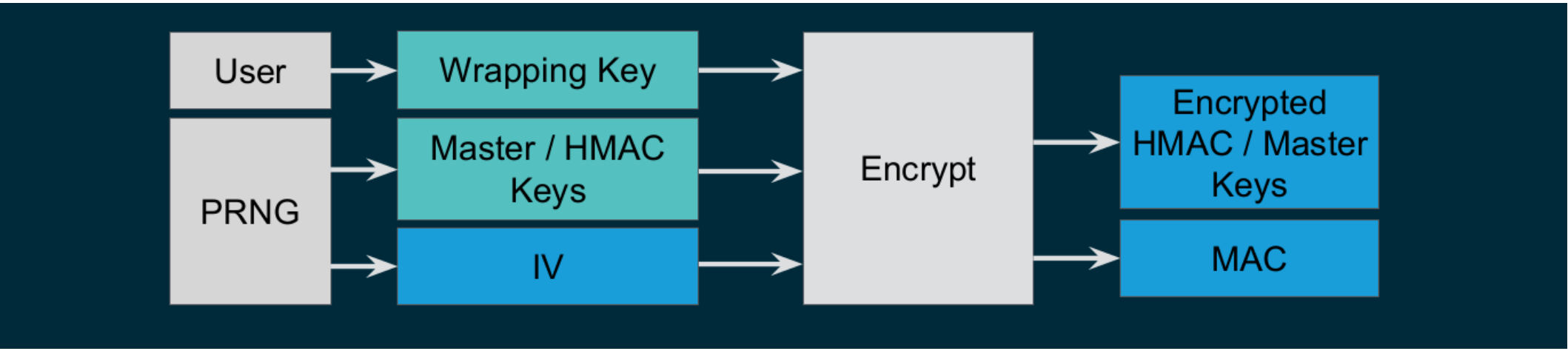
zfs unload-key [-r] -a | filesystem

zfs change-key -i [-l] [-o options] filesystem
    options: keyformat, keylocation, pbkdf2iters

OpenZFS Encryption



OpenZFS Encryption



OpenZFS Encryption

Problems

copies=2

dedup leaks, limited to same key

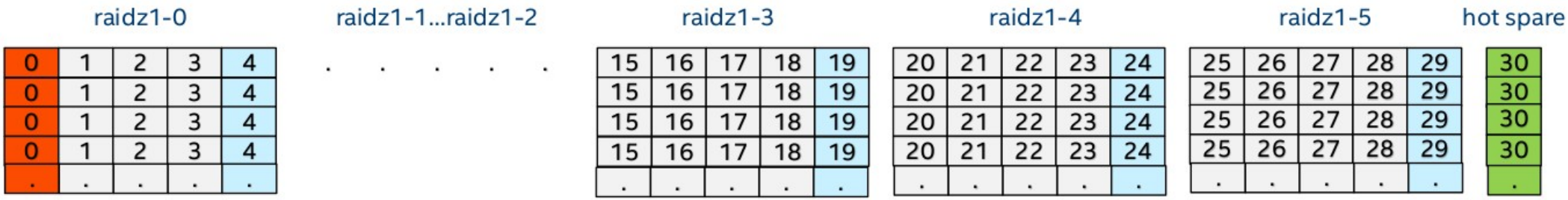
CRIME

→ compression=off

OpenZFS dRAID

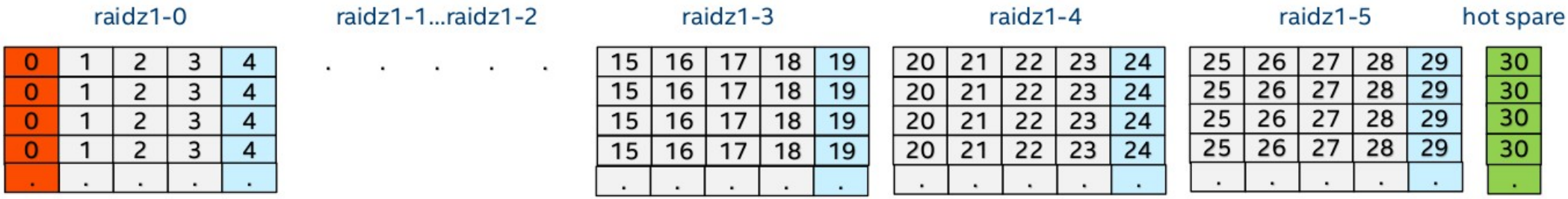
OpenZFS dRAID

Traditional RAIDZ

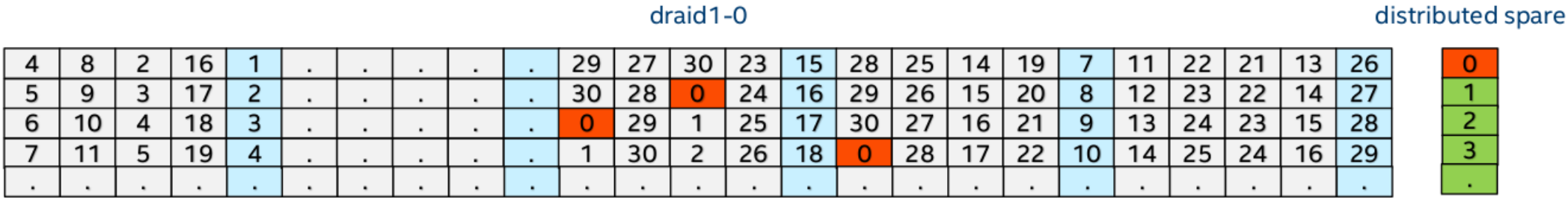


OpenZFS dRAID

Traditional RAIDZ

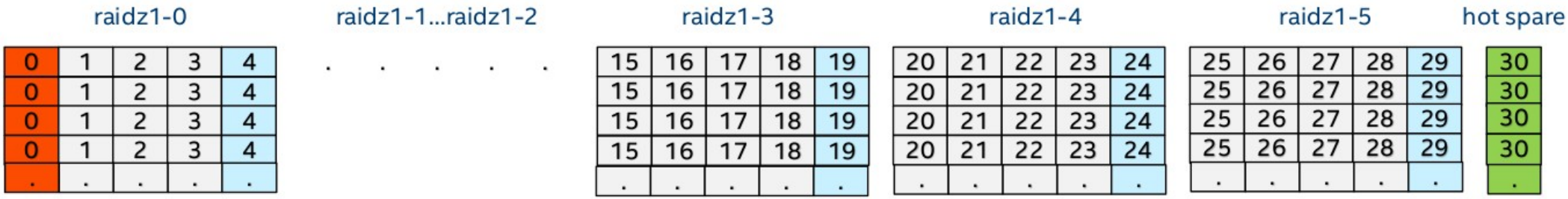


Declustered RAID

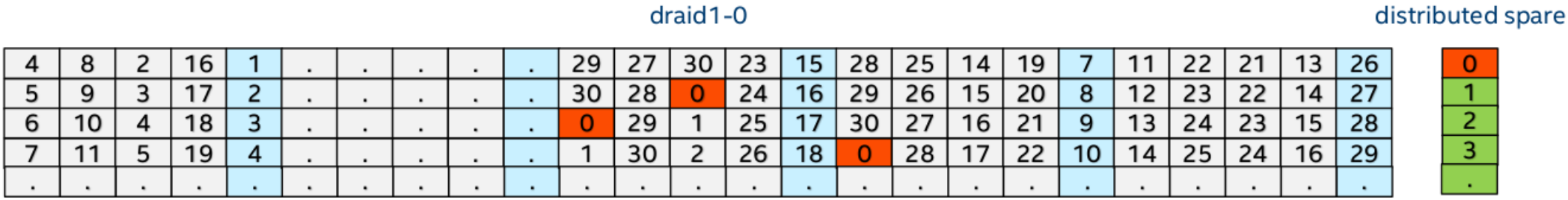


OpenZFS dRAID

Traditional RAIDZ

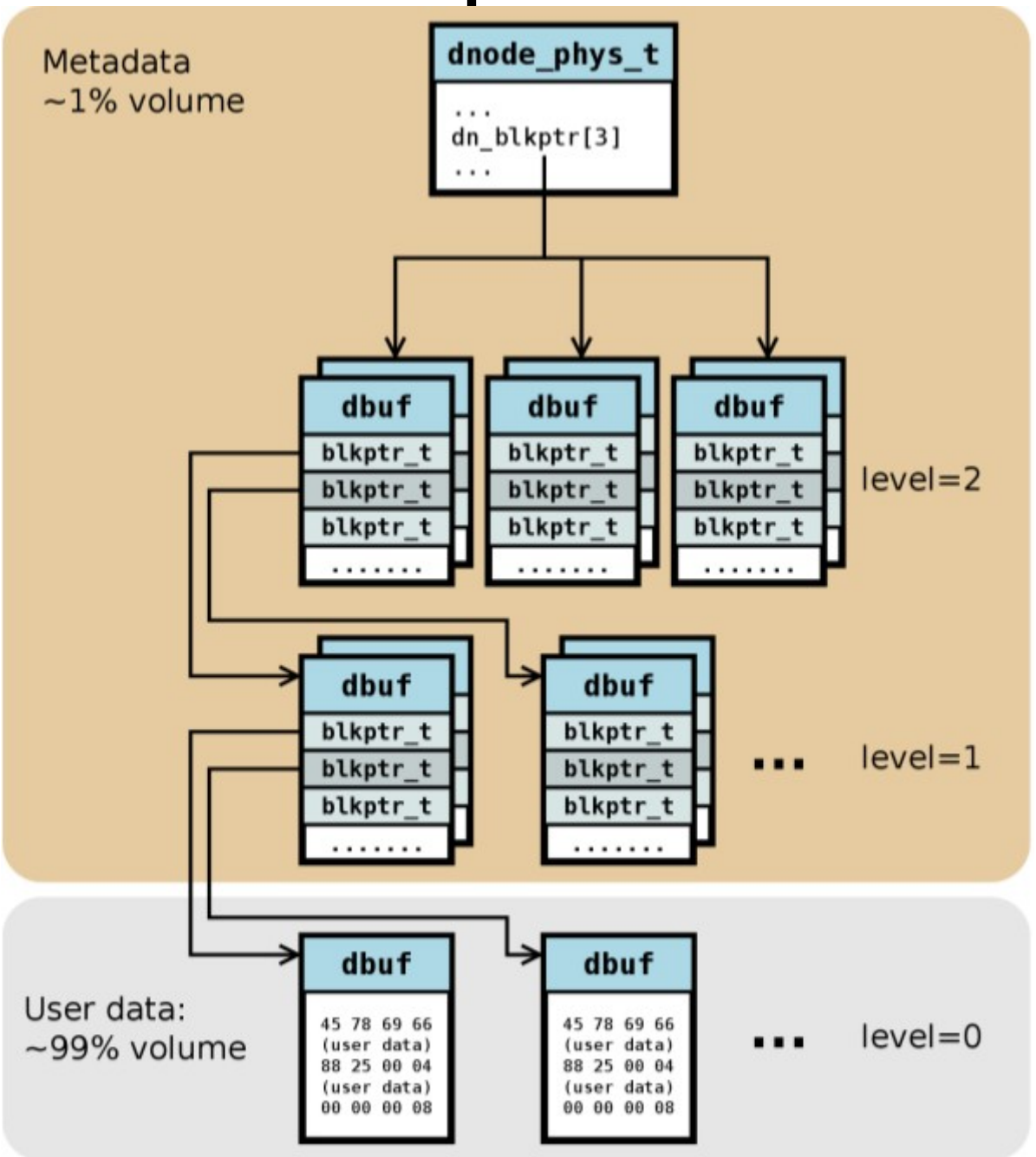


Declustered RAID

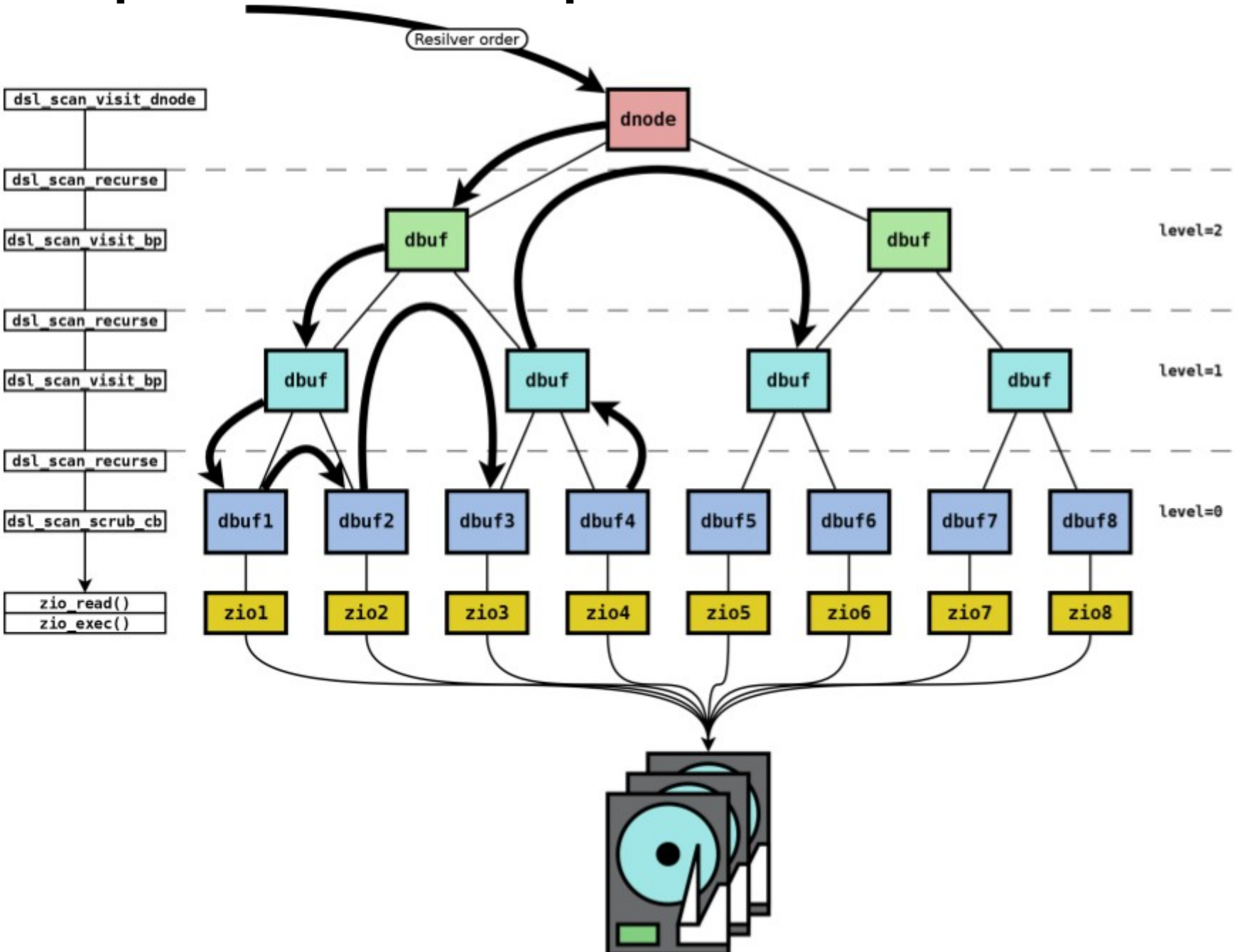


https://github.com/zfsonlinux/zfs/pull/5841

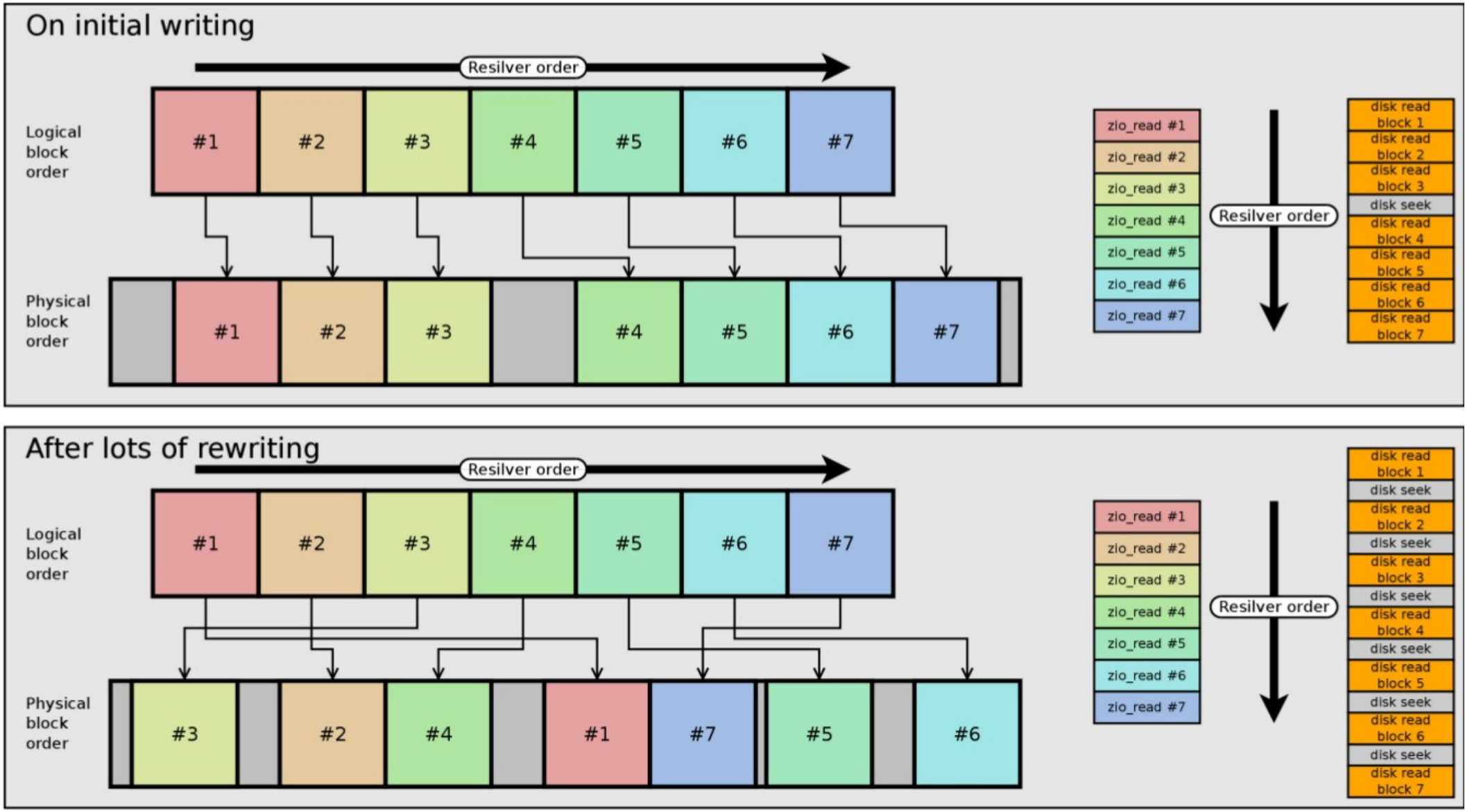
OpenZFS Sequential Resilver



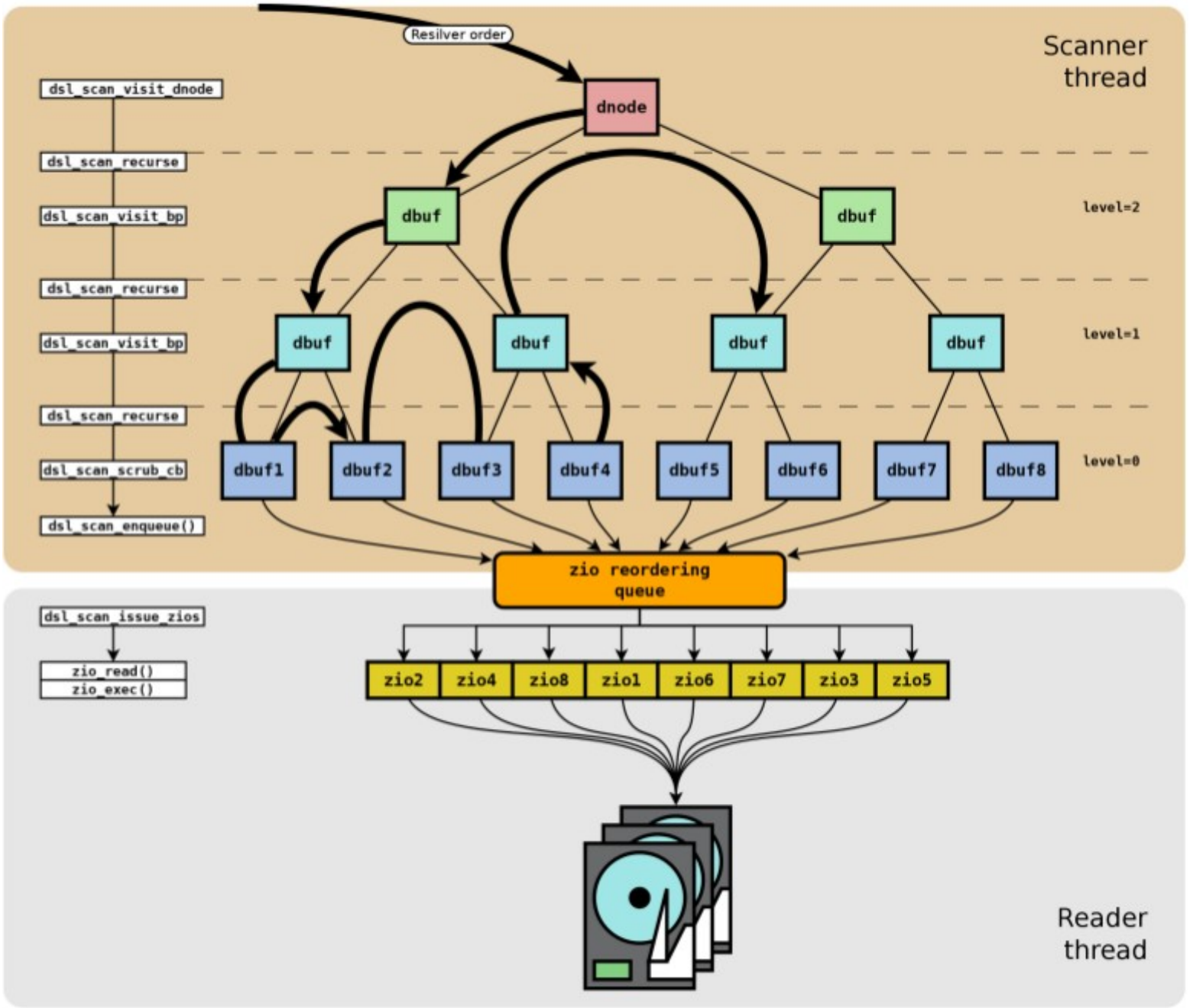
OpenZFS Sequential Resilver



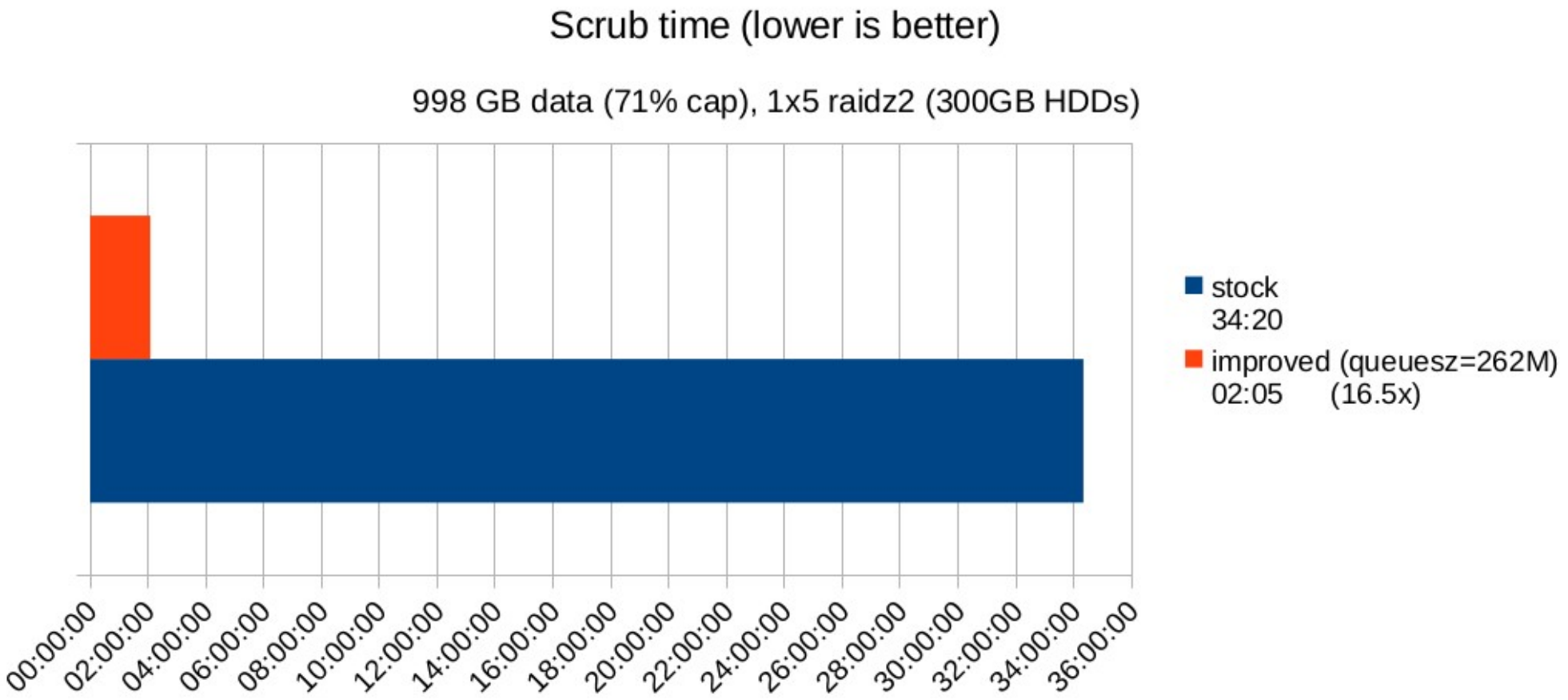
OpenZFS Sequential Resilver



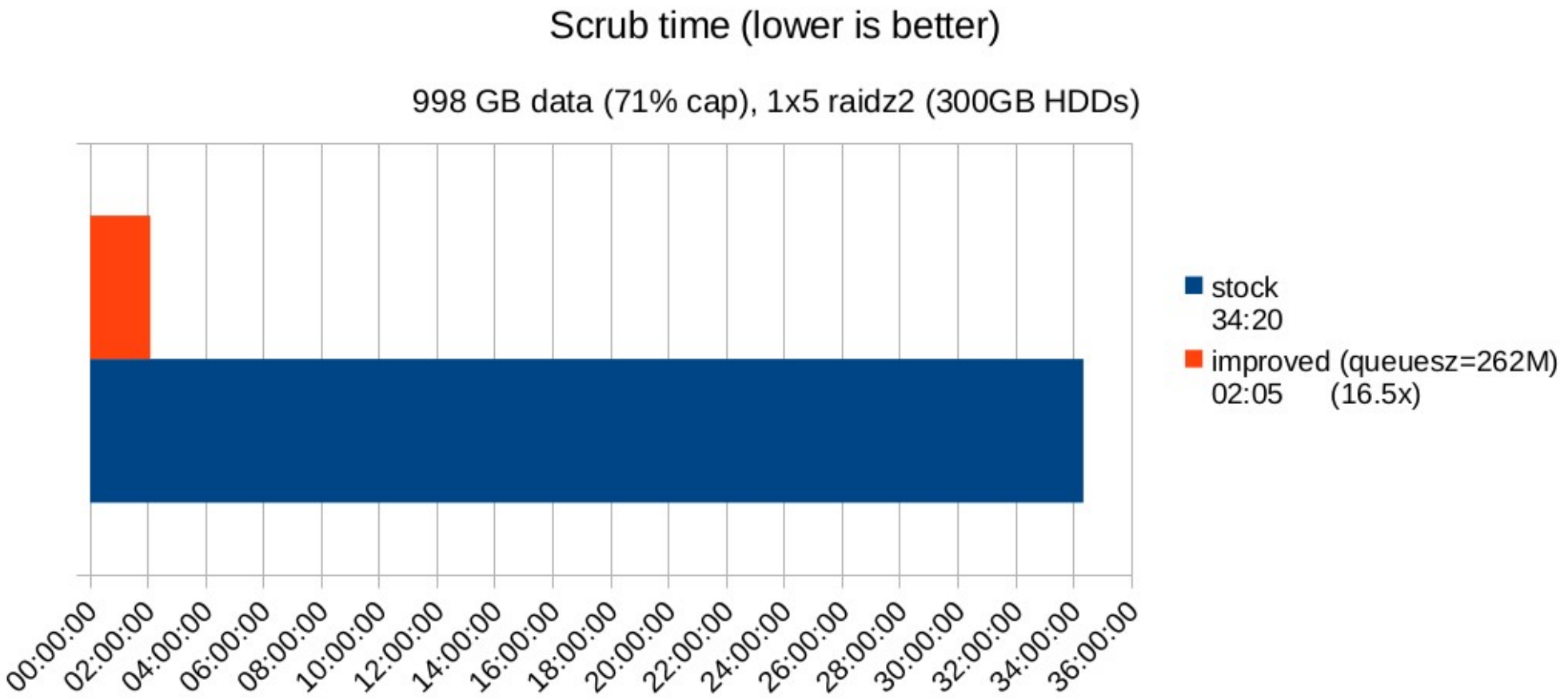
OpenZFS Sequential Resilver



OpenZFS Sequential Resilver



OpenZFS Sequential Resilver



https://github.com/zfsonlinux/zfs/pull/6256

Device evacuation/removal

zpool remove disk

Device evacuation/removal

zpool remove disk

→ virtual VDEV + indirection table

zfs remap

Device evacuation/removal

zpool remove disk

→ virtual VDEV + indirection table

zfs remap

https://github.com/openzfs/openzfs/pull/251

OpenZFS Channel Programs

Lua 5.2

zfs program [-t timeout] [-m memlimit] pool script

OpenZFS Channel Programs

Lua 5.2

zfs program [-t timeout] [-m memlimit] pool script

https://github.com/zfsonlinux/zfs/pull/6558

https://www.illumos.org/issues/7431

OpenZFS Channel Programs

Lua 5.2

zfs program [-t timeout] [-m memlimit] pool script

https://github.com/zfsonlinux/zfs/pull/6558

https://www.illumos.org/issues/7431

Open Context Channel Programs
    https://www.illumos.org/issues/8677

Metadata Allocation Classes

https://github.com/zfsonlinux/zfs/pull/5182

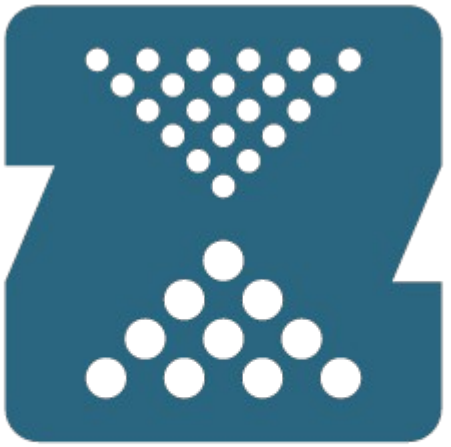
```
$ zpool list -o name,segregate_log,segregate_special
   NAME         SEGREGATE_LOG  SEGREGATE_SPECIAL
   demo                   off                 on
   spill-test             off                off

[root@ssu1_oss2]# zpool list -C ssu_1ost1
NAME             SIZE   ALLOC  FREE   CAPACITY
--------------  -----  -----  -----  --------
ssu_1ost1        72.7T  56.0T  16.8T  77.0%
  draid2-0       72.7T  56.0T  16.8T  77.0%
    normal       58.2T  55.8T  2.48T  95.8%
    special      2.25T   217G  2.04T  9.43%
    unassigned   12.2T      0  12.2T      -
```

OpenZFS Developer Summit

24. – 25. 11. 2017, San Francisco

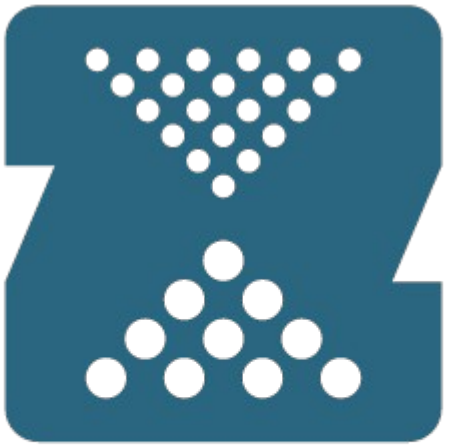
A proposal for 1,000x better dedup performance



OpenZFS

QA

?



OpenZFS