PROXMOX VE MIT CEPH & ZFS

ZUKUNFTSSICHERE INFRASTRUKTUR IM RECHENZENTRUM

Alwin Antreich
Proxmox Server Solutions GmbH
FrOSCon 14 | 10. August 2019

Willkommen!

Alwin Antreich

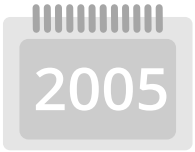
Software Entwickler @ Proxmox

15 Jahre in der IT als

System / Netzwerk Administrator

Proxmox Server Solutions GmbH

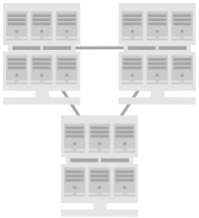
Aktive Community



Proxmox seit 2005
in Wien (AT)

Globales Partnernetz

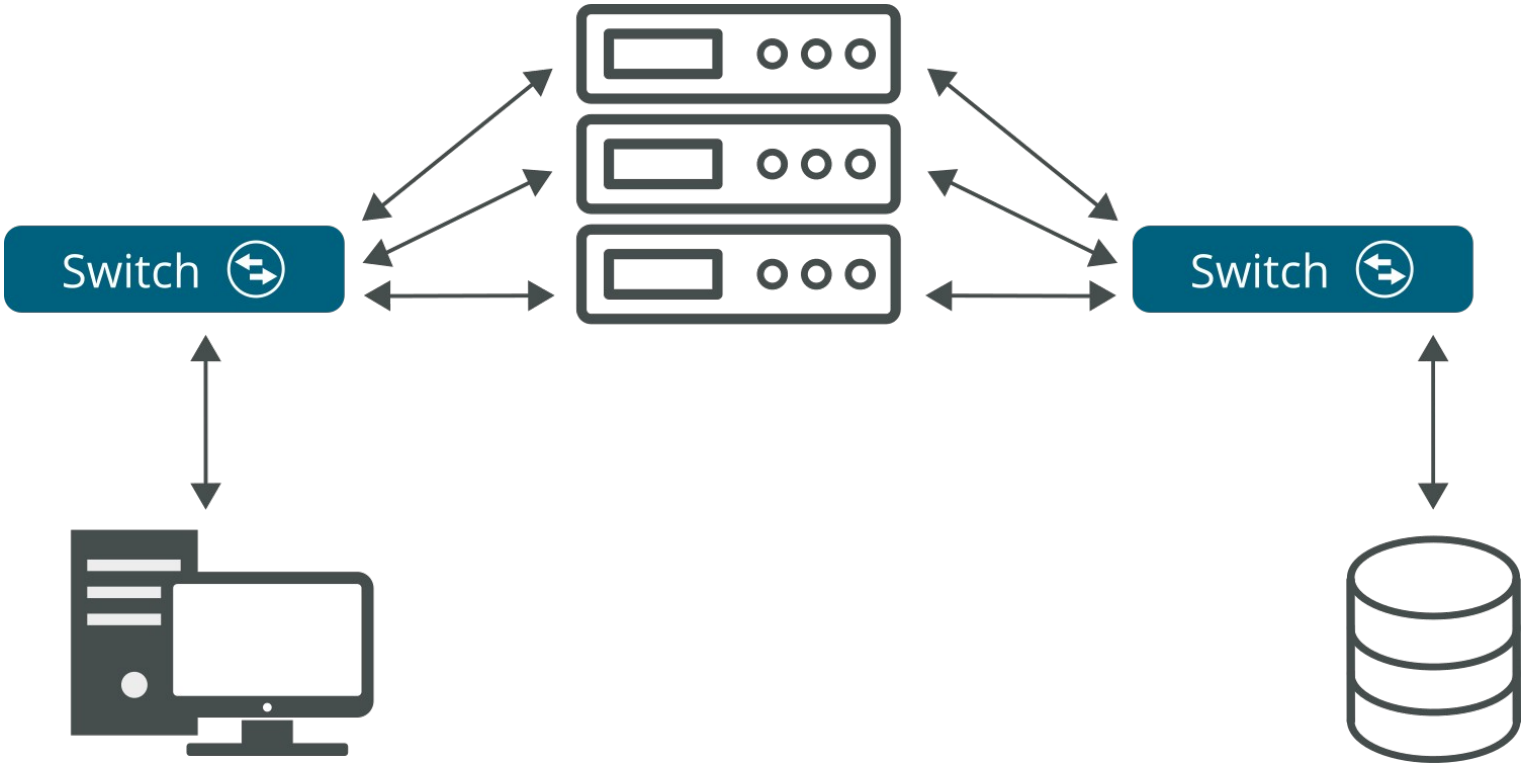
Proxmox Mail Gateway
(AGPL,v3)



Proxmox VE (AGPL,v3)

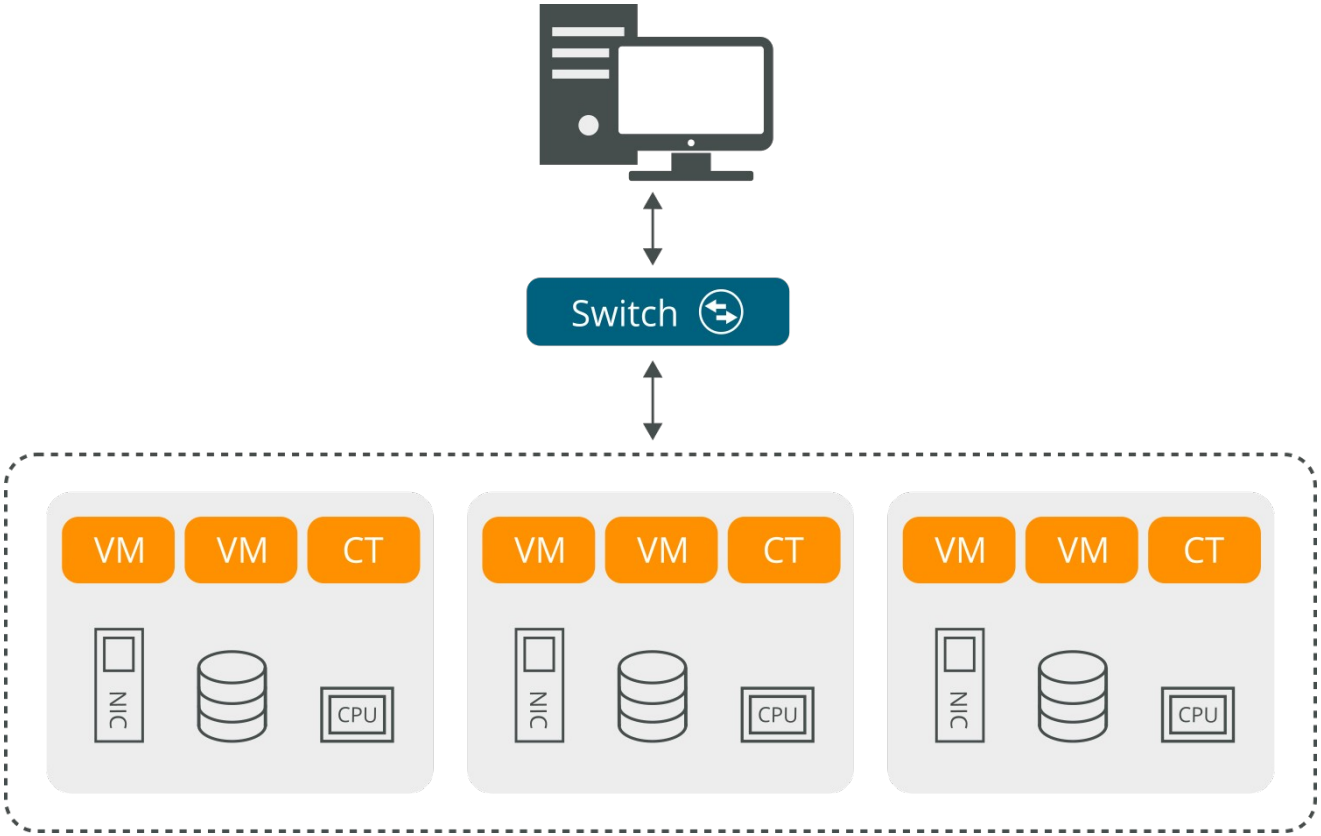
Enterprise
Support & Services

Traditionelle Infrastruktur



Hyperkonvergenz

Hyperkonvergente Infrastruktur

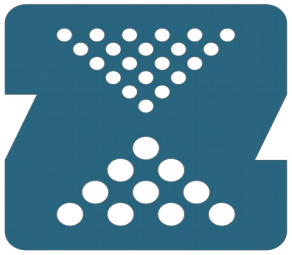


Voraussetzung für Hyperkonvergenz

CPU / RAM / Netzwerk / Storage

Verwende immer genug von allem.

PROXMOX



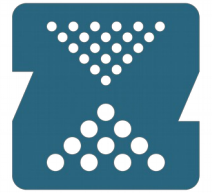
OpenZFS

ceph

Was ist ‚das'?

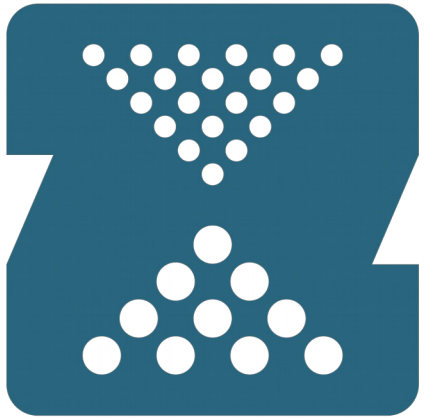
- Ceph & ZFS - software-basierte Storagelösungen
- ZFS lokal / Ceph verteilt im Netzwerk
- Hervorragende Performance, Verfügbarkeit und Skalierbarkeit
- Verwaltung und Überwachung mit Proxmox VE
- Technischer Support für Ceph & ZFS inkludiert in Proxmox Subskription

https://ceph.io/

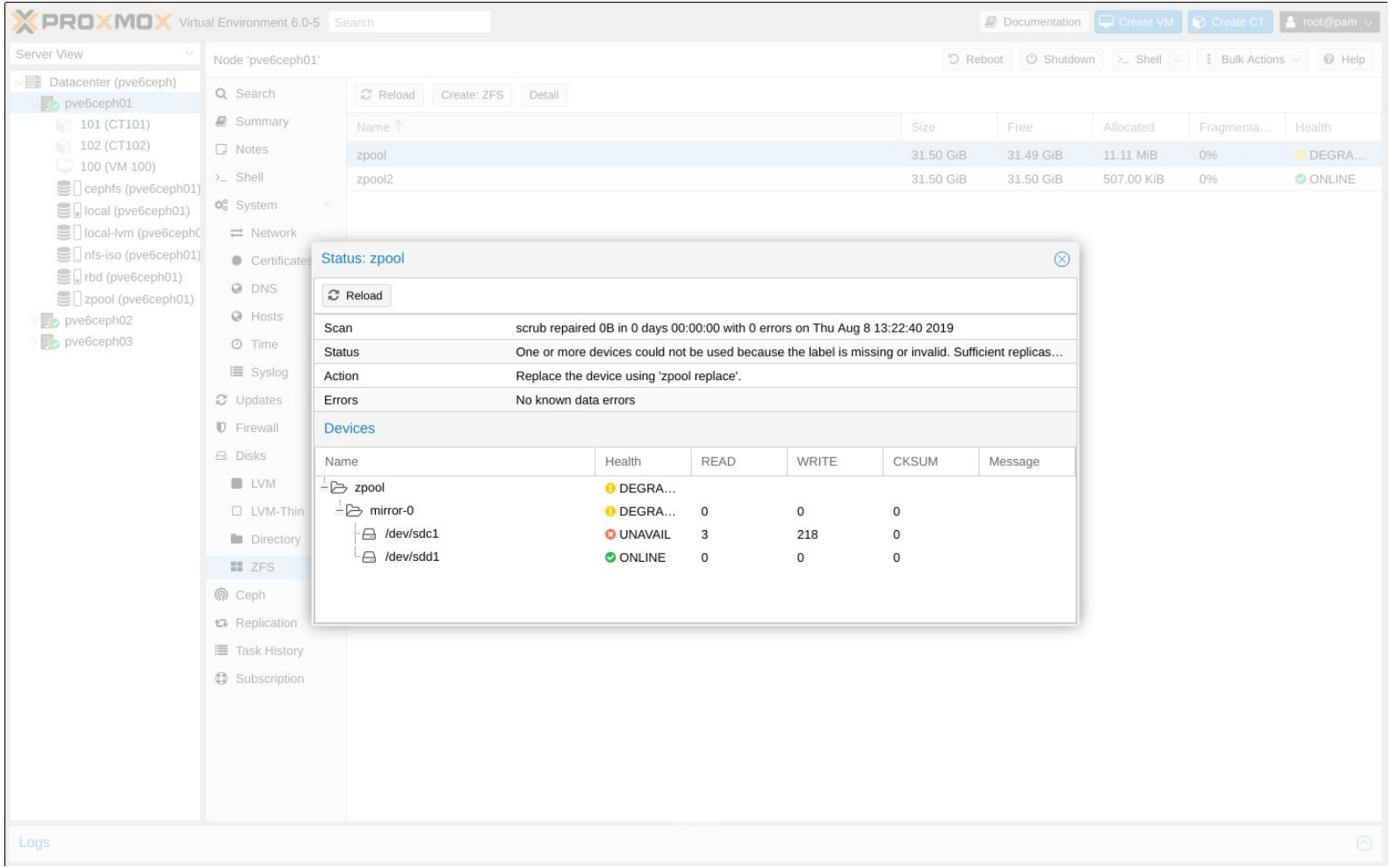


OpenZFS

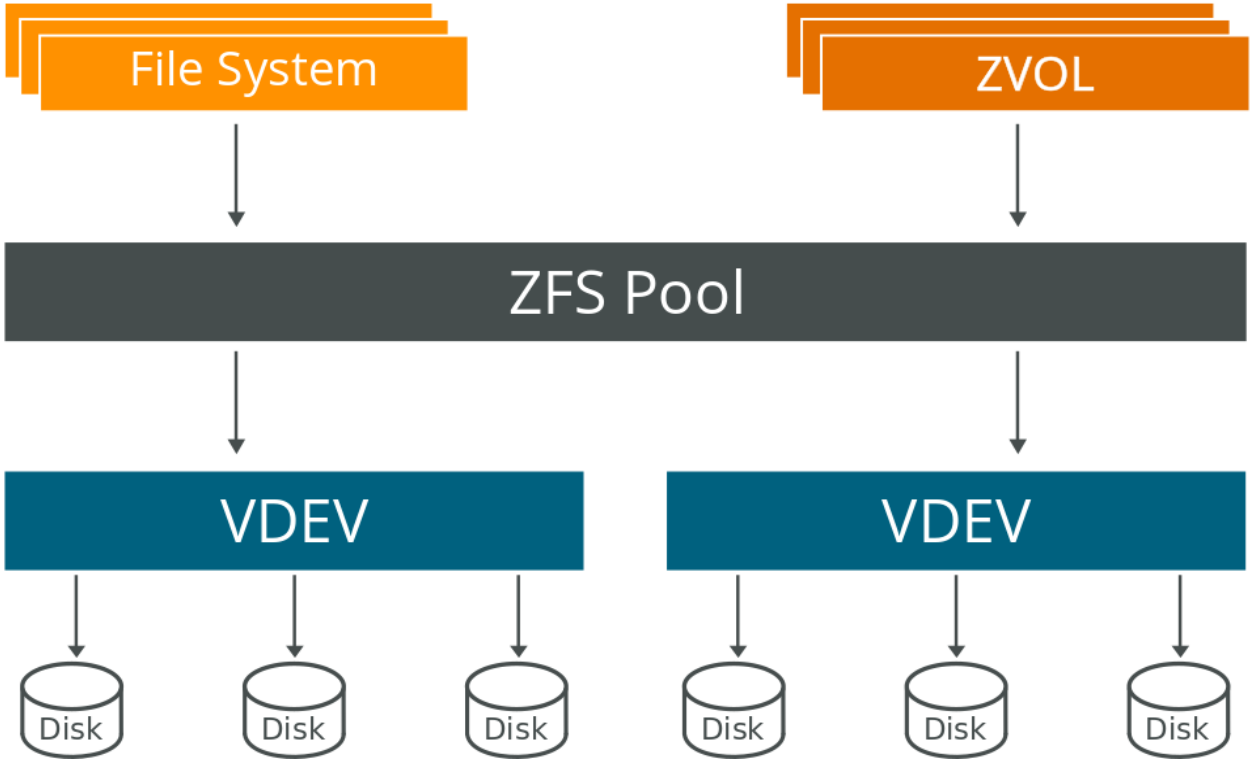
http://open-zfs.org/



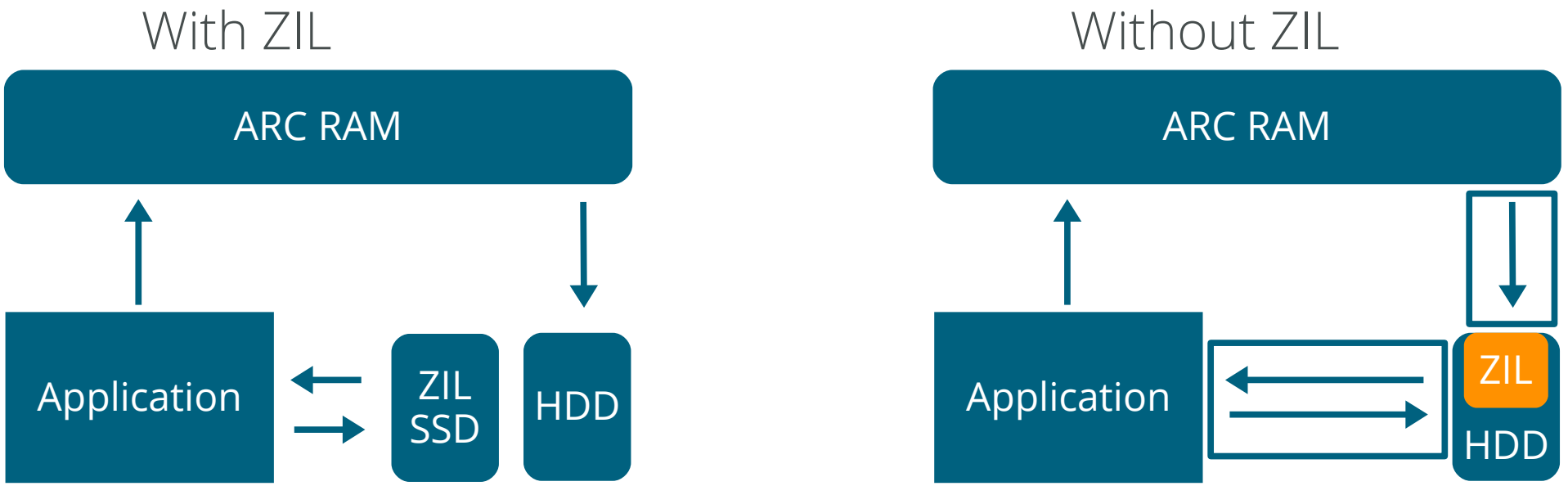
OpenZFS


PROXMOX Virtual Environment 6.0-5
Search
Documentation
Create VM
Create CT
root@pam
Server View
Node 'pve6ceph01'
Reboot
Shutdown
Shell
Bulk Actions
Help
Datacenter (pve6ceph)
pve6ceph01
101 (CT101)
102 (CT102)
100 (VM 100)
cephfs (pve6ceph01)
local (pve6ceph01)
local-lvm (pve6ceph0
nfs-iso (pve6ceph01)
rbd (pve6ceph01)
zpool (pve6ceph01)
pve6ceph02
pve6ceph03
Search
Summary
Notes
Shell
System
Network
Certificates
DNS
Hosts
Time
Syslog
Updates
Firewall
Disks
LVM
LVM-Thin
Directory
ZFS
Ceph
Replication
Task History
Subscription
Reload
Create: ZFS
Detail
Name
Size
Free
Allocated
Fragmenta…
Health
zpool
31.50 GiB
31.49 GiB
11.11 MiB
0%
DEGRA…
zpool2
31.50 GiB
31.50 GiB
507.00 KiB
0%
ONLINE
Status: zpool
Reload
Scan
scrub repaired 0B in 0 days 00:00:00 with 0 errors on Thu Aug 8 13:22:40 2019
Status
One or more devices could not be used because the label is missing or invalid. Sufficient replicas…
Action
Replace the device using 'zpool replace'.
Errors
No known data errors
Devices
Name
Health
READ
WRITE
CKSUM
Message
zpool
DEGRA…
mirror-0
DEGRA…
0
0
0
/dev/sdc1
UNAVAIL
3
218
0
/dev/sdd1
ONLINE
0
0
0
Logs

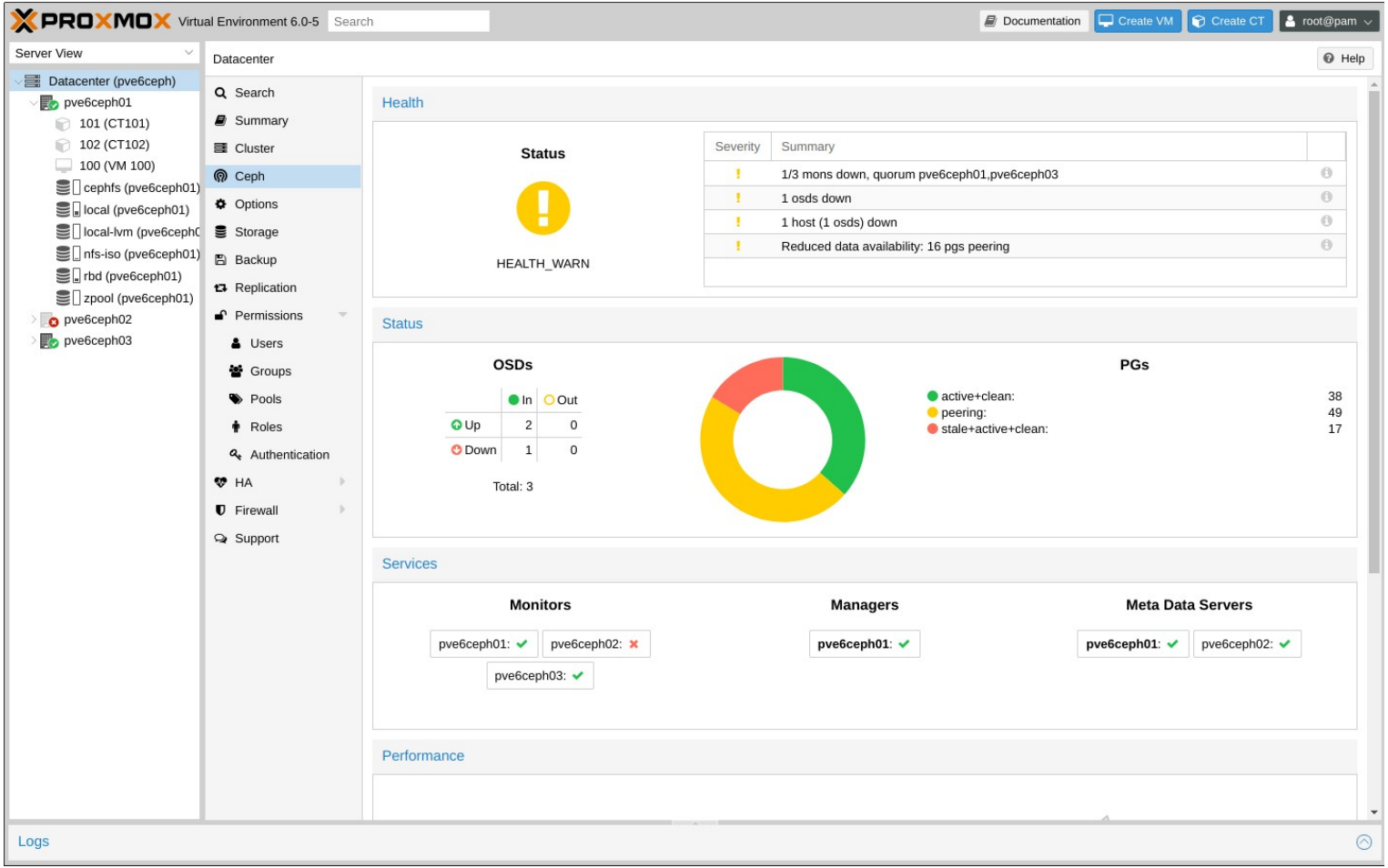
ZFS Architektur



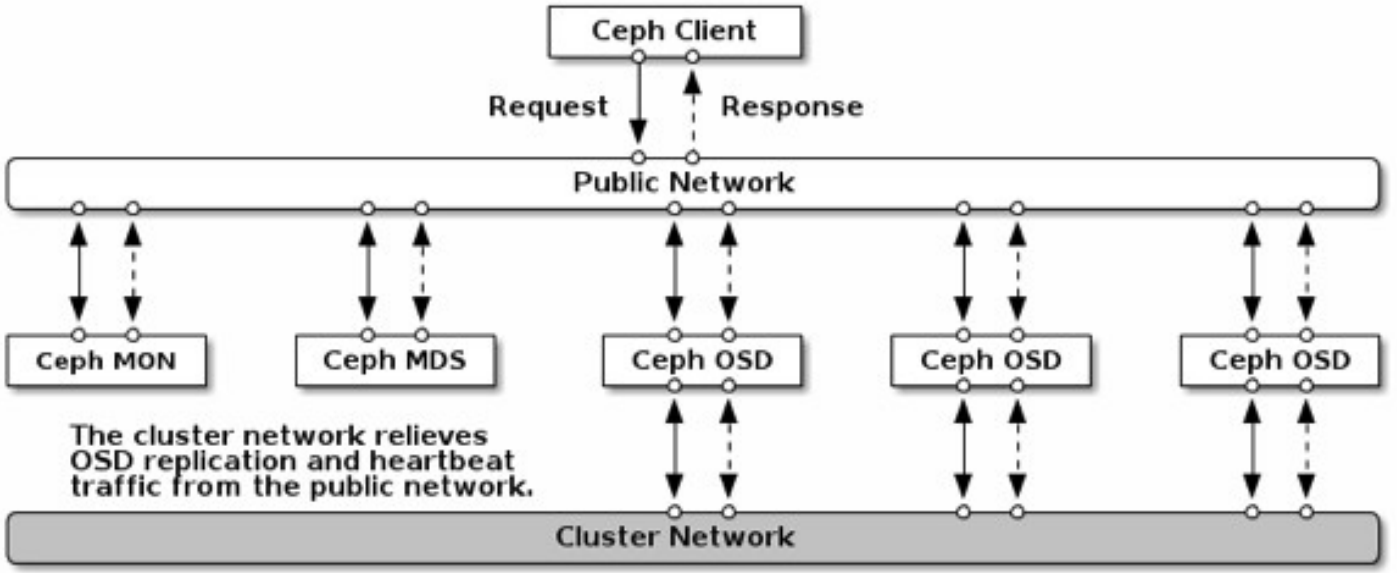
ZFS ARC, L2ARC and ZIL



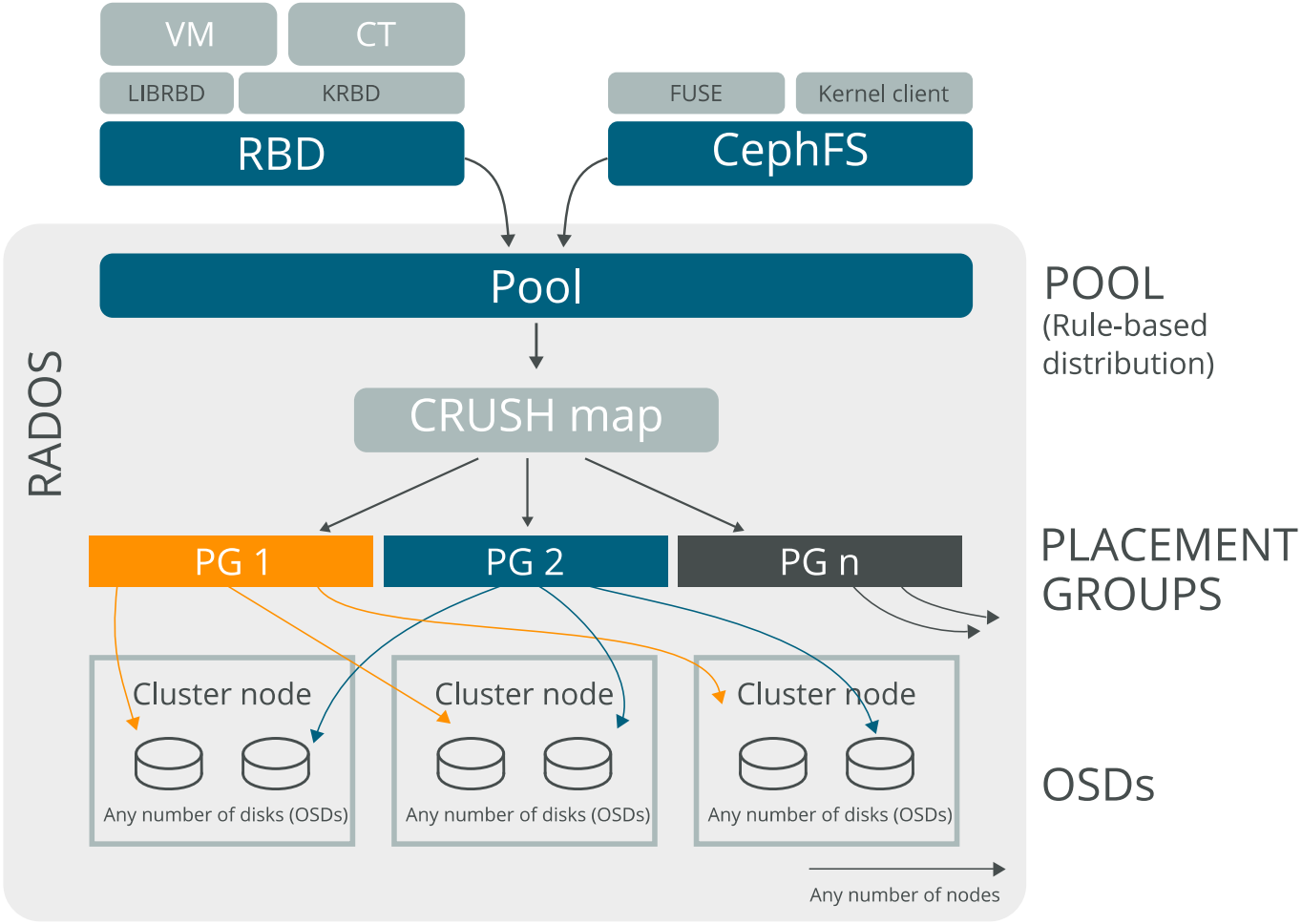
ceph


PROXMOX Virtual Environment 6.0-5
Search
Documentation
Create VM
Create CT
root@pam
Server View
Datacenter
Help
Datacenter (pve6ceph)
pve6ceph01
101 (CT101)
102 (CT102)
100 (VM 100)
cephfs (pve6ceph01)
local (pve6ceph01)
local-lvm (pve6ceph0
nfs-iso (pve6ceph01)
rbd (pve6ceph01)
zpool (pve6ceph01)
pve6ceph02
pve6ceph03
Search
Summary
Cluster
Ceph
Options
Storage
Backup
Replication
Permissions
Users
Groups
Pools
Roles
Authentication
HA
Firewall
Support
Health
Status
HEALTH_WARN
Severity
Summary
1/3 mons down, quorum pve6ceph01,pve6ceph03
1 osds down
1 host (1 osds) down
Reduced data availability: 16 pgs peering
Status
OSDs
In
Out
Up
2
0
Down
1
0
Total: 3
PGs
active+clean:
38
peering:
49
stale+active+clean:
17
Services
Monitors
pve6ceph01:
pve6ceph02:
pve6ceph03:
Managers
pve6ceph01:
Meta Data Servers
pve6ceph01:
pve6ceph02:
Performance
Logs

Ceph Network



Ceph Docs: https://docs.ceph.com/docs/master/


VM
CT
LIBRBD
KRBD
RBD
FUSE
Kernel client
CephFS
Pool
CRUSH map
RADOS
PG 1
PG 2
PG n
Cluster node
Cluster node
Cluster node
Any number of disks (OSDs)
Any number of disks (OSDs)
Any number of disks (OSDs)
Any number of nodes
POOL
(Rule-based distribution)
PLACEMENT GROUPS
OSDs

Häufige Fehler

/ Kein separiertes Netzwerk /

/ Keine Enterprise-Hardware /

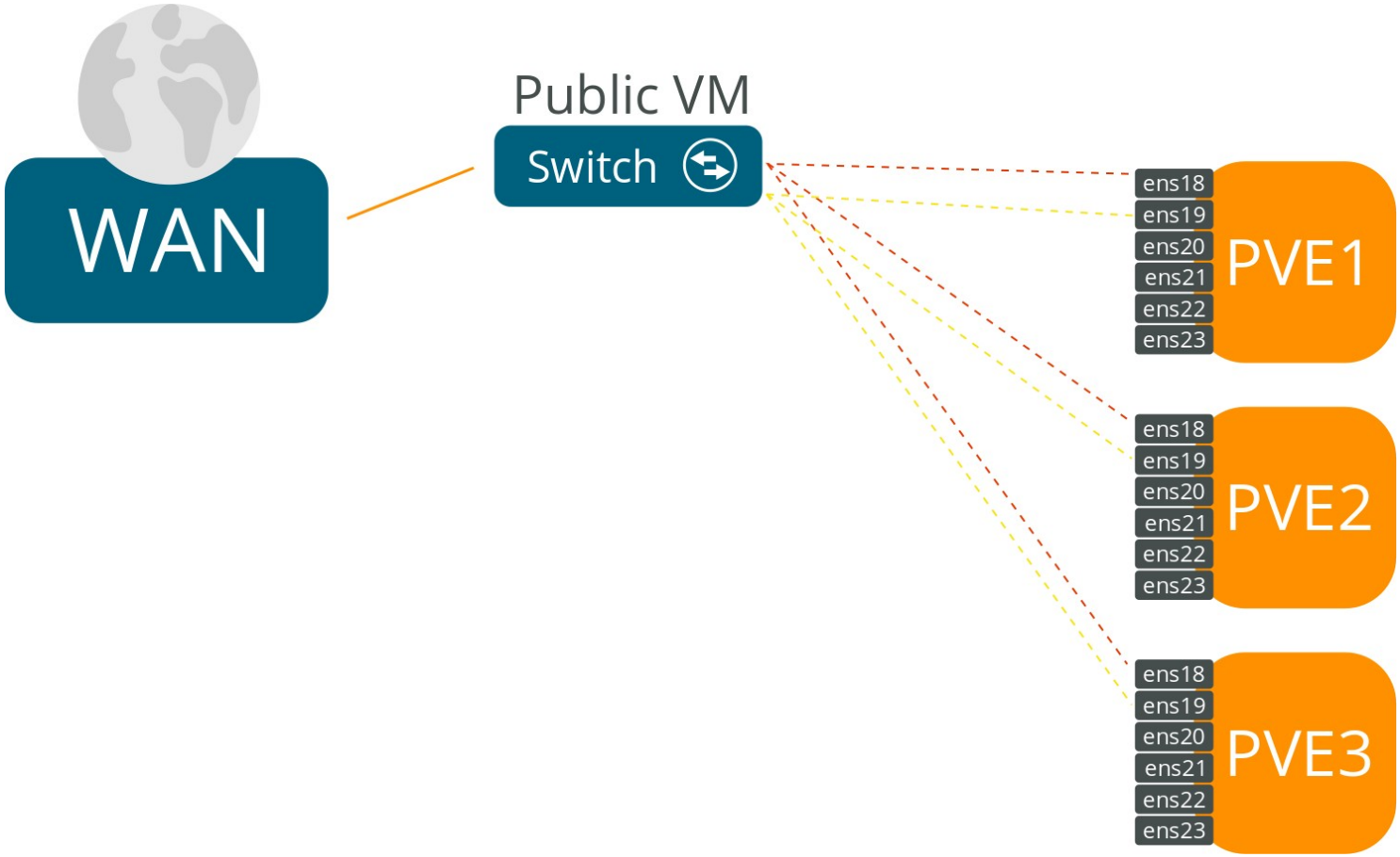
/ Over-commitment /

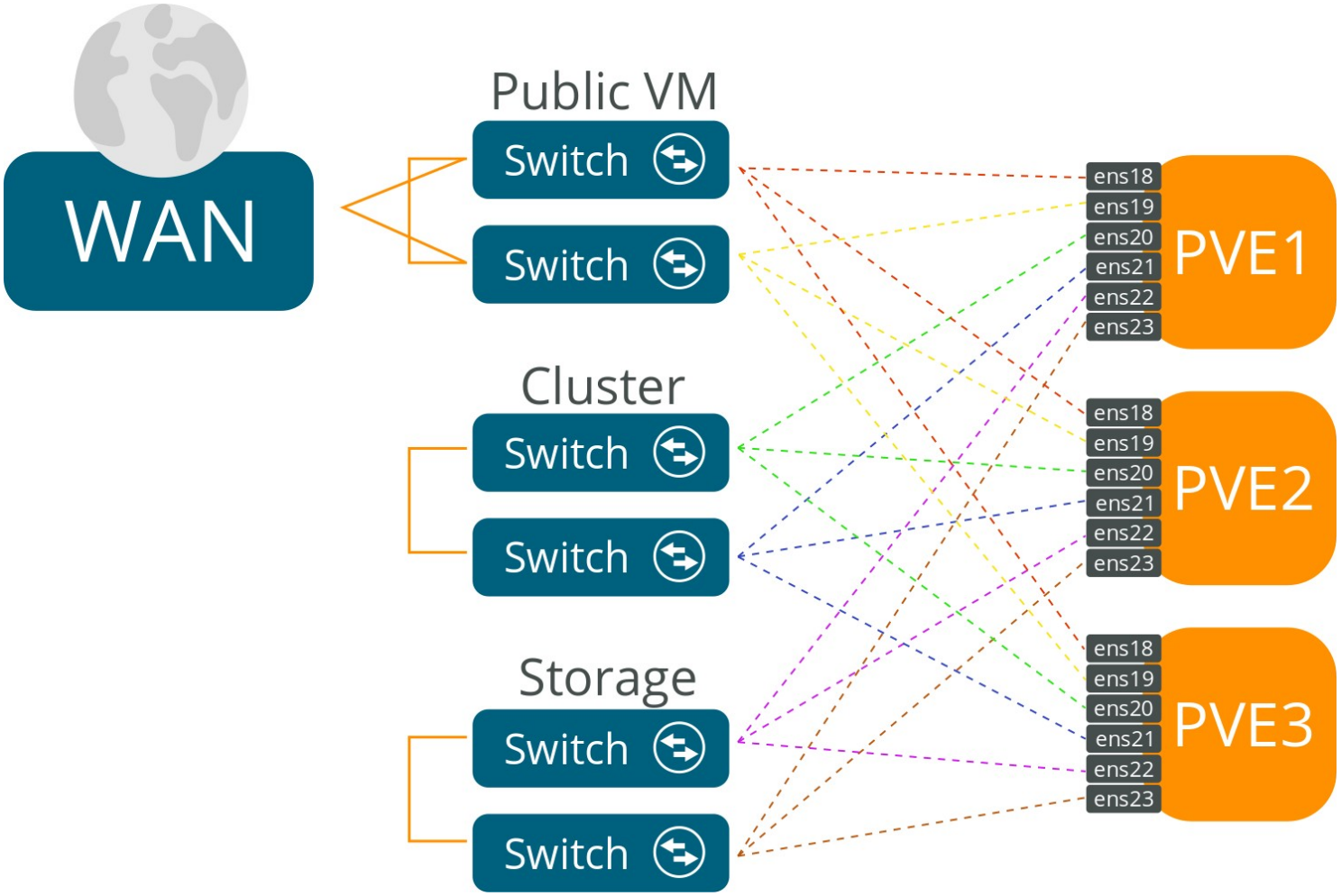
Die Basis

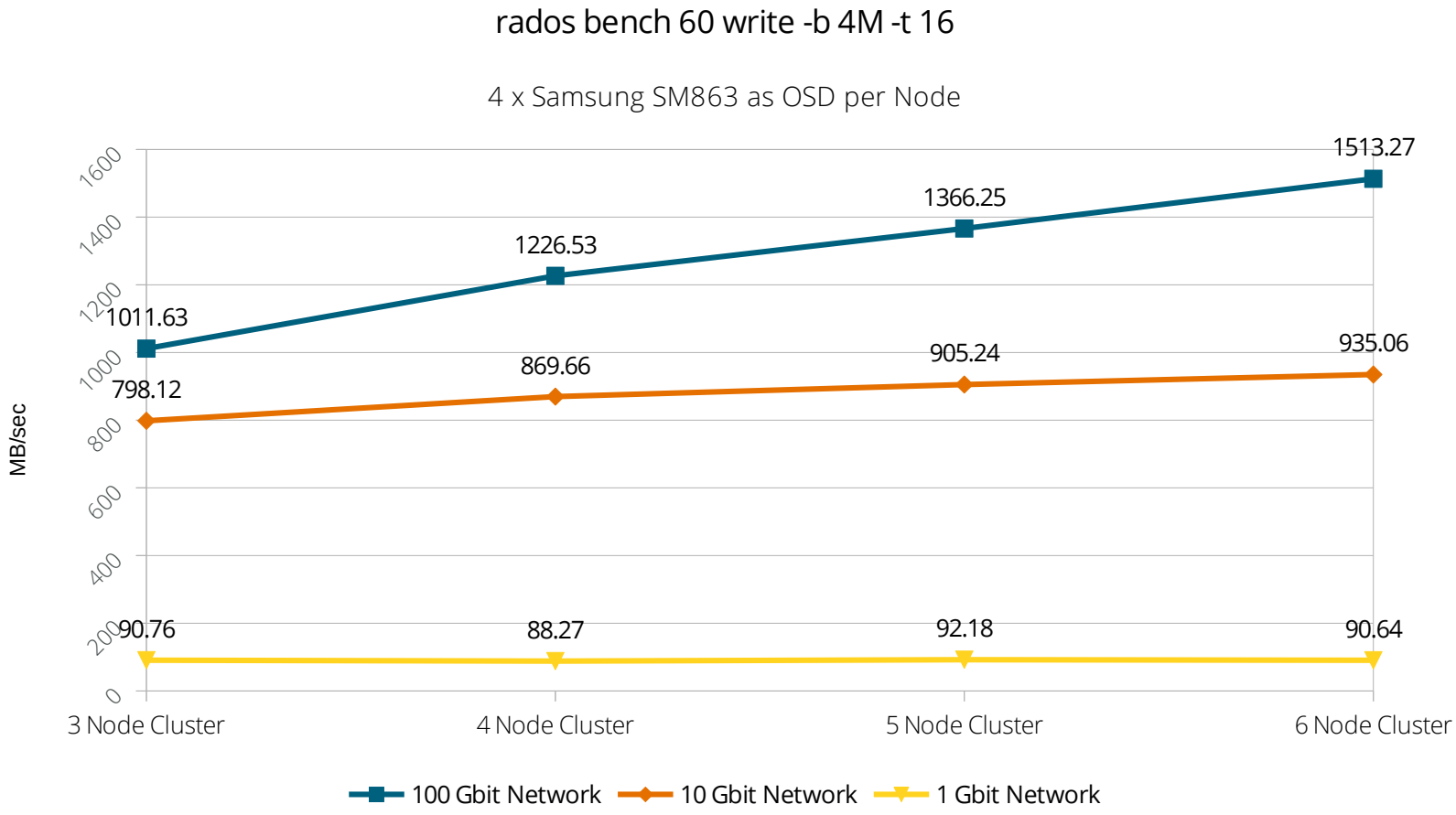
```
fio --ioengine=libaio --filename=/dev/sdX --direct=1 --sync=1 --rw=write --bs=4K --numjobs=1 --iodepth=1 --runtime=60 --time_based --group_reporting --name=baseline --output-format=normal --output=fio.log –bandwidth-log
```

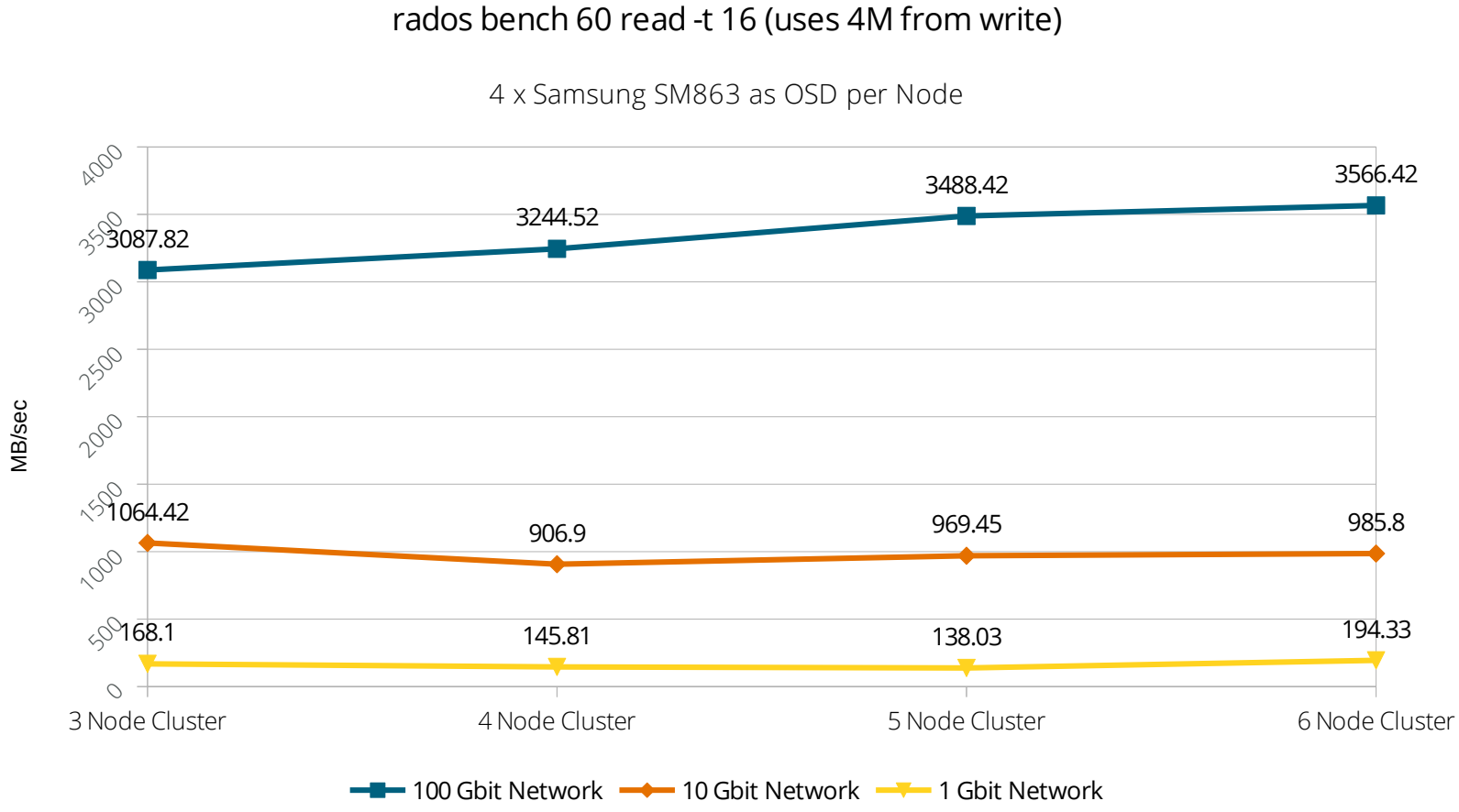
CPU:	Single Intel® Xeon® E5-2620v4 2,1 GHZ 8/16 2133
Mainboard:	Supermicro X10SRi-F S2011-3
Case:	2U Supermicro Chassis 8x Hotswap
100 Gbit NIC:	Mellanox MCX456A-ECAT ConnectX-4, x16 PCIe 3.0
Memory:	4 x 16 GB DDR4 FSB2400 288-pin REG x4 1R

Kein separiertes Netzwerk

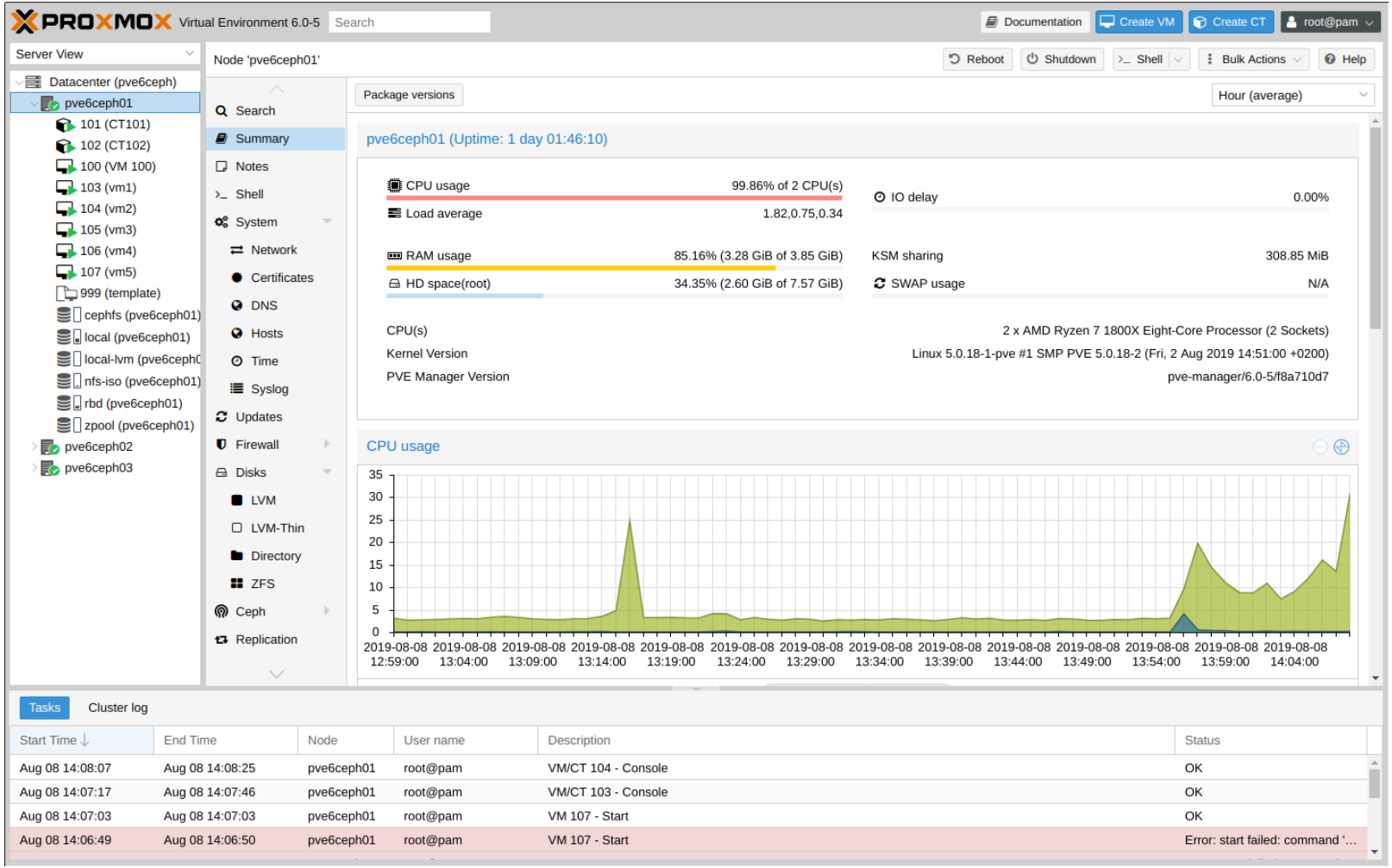

WAN
Public VM
Switch
ens18
ens19
ens20
ens21
ens22
ens23
PVE1
ens18
ens19
ens20
ens21
ens22
ens23
PVE2
ens18
ens19
ens20
ens21
ens22
ens23
PVE3


WAN
Public VM
Switch
Switch
Cluster
Switch
Switch
Storage
Switch
Switch
ens18
ens19
ens20
ens21
ens22
ens23
PVE1
ens18
ens19
ens20
ens21
ens22
ens23
PVE2
ens18
ens19
ens20
ens21
ens22
ens23
PVE3

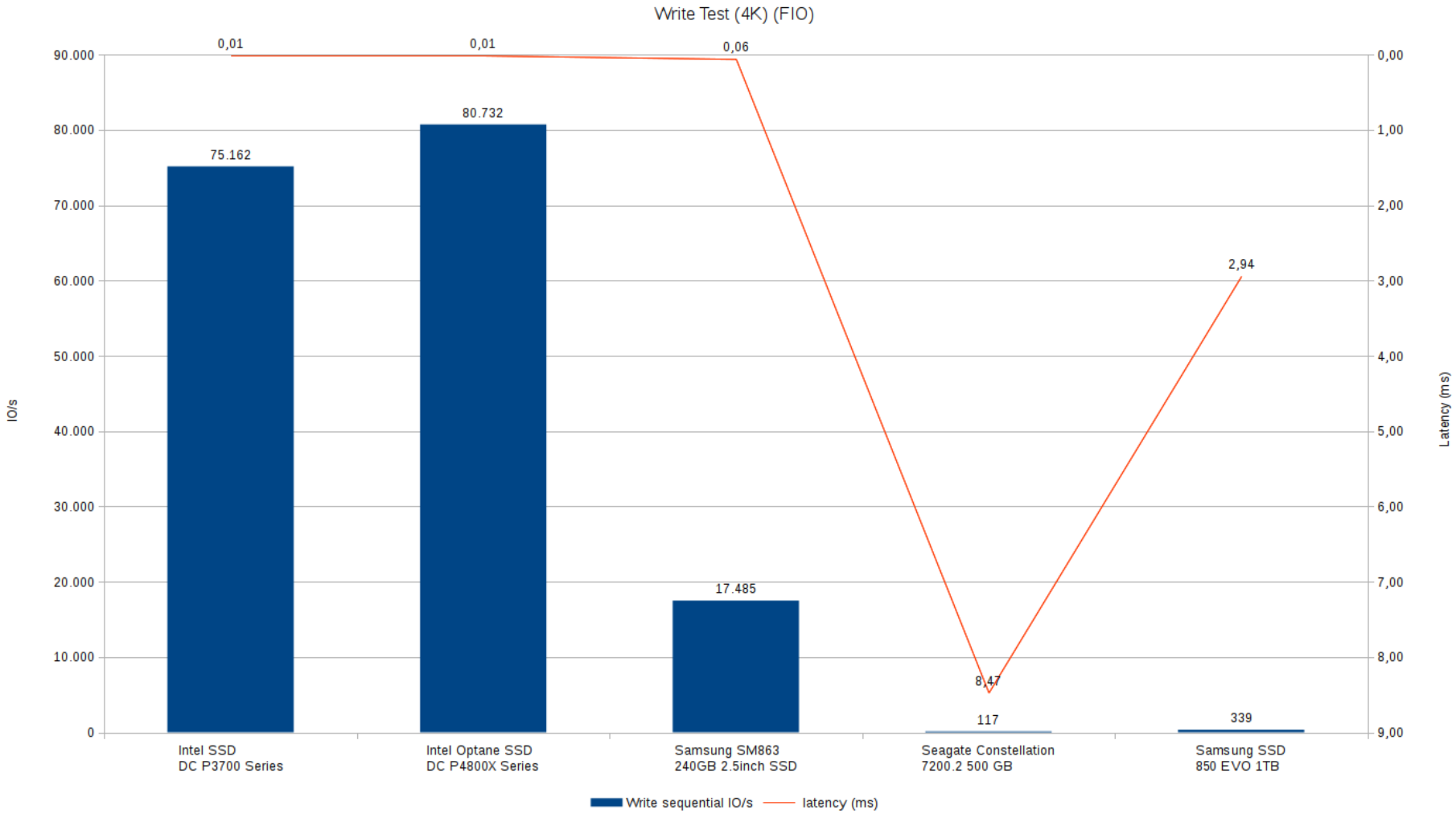

rados bench 60 write -b 4M -t 16
4 x Samsung SM863 as OSD per Node
MB/sec
1600
1400
1200
1000
800
600
400
200
0
1513.27
1366.25
1226.53
1011.63
935.06
905.24
869.66
798.12
90.76
88.27
92.18
90.64
3 Node Cluster
4 Node Cluster
5 Node Cluster
6 Node Cluster
100 Gbit Network
10 Gbit Network
1 Gbit Network


rados bench 60 read -t 16 (uses 4M from write)
4 x Samsung SM863 as OSD per Node
MB/sec
4000
3500
3000
2500
2000
1500
1000
500
0
3087.82
3244.52
3488.42
3566.42
1064.42
906.9
969.45
985.8
168.1
145.81
138.03
194.33
3 Node Cluster
4 Node Cluster
5 Node Cluster
6 Node Cluster
100 Gbit Network
10 Gbit Network
1 Gbit Network

Over-commitment

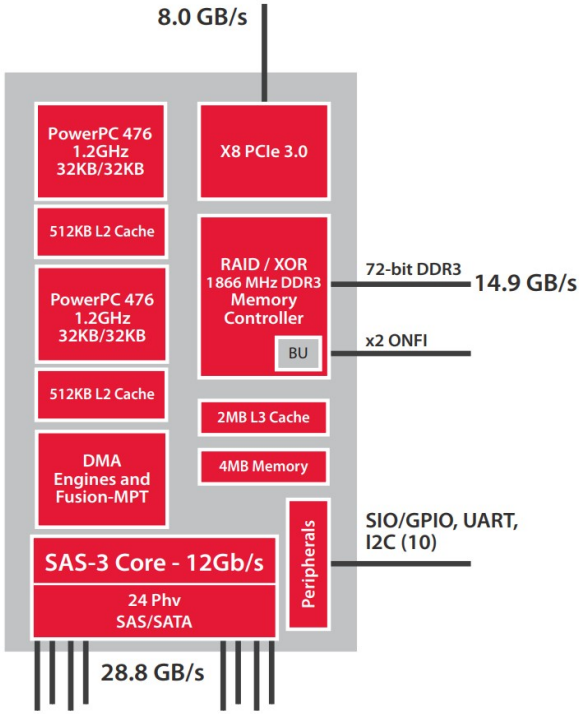

PROXMOX Virtual Environment 6.0-5
Search
Documentation
Create VM
Create CT
root@pam
Server View
Node 'pve6ceph01'
Reboot
Shutdown
Shell
Bulk Actions
Help
Datacenter (pve6ceph)
pve6ceph01
101 (CT101)
102 (CT102)
100 (VM 100)
103 (vm1)
104 (vm2)
105 (vm3)
106 (vm4)
107 (vm5)
999 (template)
cephfs (pve6ceph01)
local (pve6ceph01)
local-lvm (pve6ceph0
nfs-iso (pve6ceph01)
rbd (pve6ceph01)
zpool (pve6ceph01)
pve6ceph02
pve6ceph03
Search
Summary
Notes
Shell
System
Network
Certificates
DNS
Hosts
Time
Syslog
Updates
Firewall
Disks
LVM
LVM-Thin
Directory
ZFS
Ceph
Replication
Package versions
Hour (average)
pve6ceph01 (Uptime: 1 day 01:46:10)
CPU usage 99.86% of 2 CPU(s)
Load average 1.82,0.75,0.34
RAM usage 85.16% (3.28 GiB of 3.85 GiB)
HD space(root) 34.35% (2.60 GiB of 7.57 GiB)
IO delay 0.00%
KSM sharing 308.85 MiB
SWAP usage N/A
CPU(s) 2 x AMD Ryzen 7 1800X Eight-Core Processor (2 Sockets)
Kernel Version Linux 5.0.18-1-pve #1 SMP PVE 5.0.18-2 (Fri, 2 Aug 2019 14:51:00 +0200)
PVE Manager Version pve-manager/6.0-5/f8a710d7
CPU usage
35
30
25
20
15
10
5
0
2019-08-08 12:59:00
2019-08-08 13:04:00
2019-08-08 13:09:00
2019-08-08 13:14:00
2019-08-08 13:19:00
2019-08-08 13:24:00
2019-08-08 13:29:00
2019-08-08 13:34:00
2019-08-08 13:39:00
2019-08-08 13:44:00
2019-08-08 13:49:00
2019-08-08 13:54:00
2019-08-08 13:59:00
2019-08-08 14:04:00
Tasks
Cluster log
Start Time
End Time
Node
User name
Description
Status
Aug 08 14:08:07 Aug 08 14:08:25 pve6ceph01 root@pam VM/CT 104 - Console OK
Aug 08 14:07:17 Aug 08 14:07:46 pve6ceph01 root@pam VM/CT 103 - Console OK
Aug 08 14:07:03 Aug 08 14:07:03 pve6ceph01 root@pam VM 107 - Start OK
Aug 08 14:06:49 Aug 08 14:06:50 pve6ceph01 root@pam VM 107 - Start Error: start failed: command '…

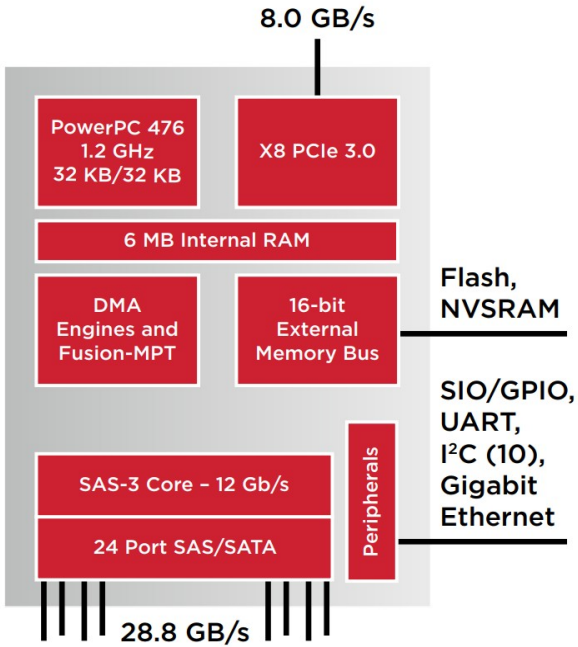
Keine Enterprise-Hardware


Write Test (4K) (FIO)
IO/s
Latency (ms)
90.000
80.000
70.000
60.000
50.000
40.000
30.000
20.000
10.000
0
0,00
1,00
2,00
3,00
4,00
5,00
6,00
7,00
8,00
9,00
0,01
0,01
0,06
8,47
2,94
75.162
80.732
17.485
117
339
Intel SSD DC P3700 Series
Intel Optane SSD DC P4800X Series
Samsung SM863 240GB 2.5inch SSD
Seagate Constellation 7200.2 500 GB
Samsung SSD 850 EVO 1TB
Write sequential IO/s
latency (ms)

RAID vs HBA

SAS3324
12Gb/s SAS
RAID-on-Chip
(ROC)





SAS3224
I/O Controller

Product Brief: https://www.broadcom.com/products/storage/

Mitmachen - Community

- Support Forum
  - https://forum.proxmox.com
- Proxmox VE Wiki
  - https://pve.proxmox.com
- Referenz-Dokumentation
  - https://pve.proxmox.com/pve-docs/
- Übersetzung
  - https://pve.proxmox.com/wiki/Translations

Mitmachen - Entwicklung

- Roadmap
  - https://pve.proxmox.com/wiki/Roadmap
- Mailing-Liste
  - https://pve.proxmox.com/cgi-bin/mailman/listinfo
- Bug Tracker
  - https://bugzilla.proxmox.com
- Quellcode
  - https://git.proxmox.com

Kontakt

Kundenportal: https://my.proxmox.com

Support Forum: https://forum.proxmox.com

Online Shop: https://shop.maurer-it.com/contact.php

Allgemeine Anfragen: office@proxmox.com

DANKE FÜR IHRE AUFMERKSAMKEIT!

Proxmox Server Solutions GmbH
Bräuhausgasse 37
1050 Vienna | Austria

office@proxmox.com
www.proxmox.com

© 2019 Proxmox. All Rights Reserved. All brand names, product names, or trademarks belong to their respective holders.