napp-*it*

napp-it Z-RAID vCluster v.2
User's Guide

Setup on OmniOS
with napp-it Pro complete

(require napp-it 20.06 or newer)

published: 2021, Mar 12 (c) napp-it.org

Content:

Preword

Solarish and ZFS are very reliable and easy to understand ZFS storage solutions. There are systems with a unique long uptime. Usually your main concern should be:

Keep it simple !

Modern server hardware is very reliable. The part that fails mostly is a disk and for that you have raid, hotspare and hotplug. The propability of a failure of other parts is very low.

Maintenance is the problem. Today you cannot allow to keep a system online without up to date security fixes. From time to time you need to upgrade the OS. Such a security fix or an OS update can work flawless but there is no guarantee. While ZFS can go back to the former OS state with its concept of boot environments, you cannot allow intensive tests for a production system after an update or upgrade or after enabling new options.

This is a serious problem and the main reason for a ZFS Cluster!

Maintanance with zero downtime !

This is the real important aspect of a dualhead ZFS Cluster. You can upgrade the inaktive head and do intensive tests. When everything is stable, failover the pool and services to allow maintenance or modifications of the other head.

And yes. If the hardware fails, you can manually or automatically failover between heads or storage Jbods. See this as a nice add-on.

1. Levels of Availability

On a basic server system, any software or hardware failure or problem during OS updates or security patches can lead to a service outage. To improve availability, you must be prepared or you must add redundancy.

1.1 Improved Availability Level1 (single server)

First step to improve availability is to take care about power stability. If you use a redundant PSU with an UPS where you put the UPS to one of the PSU units, you can survive a power outage or a UPS failure.

Next is using a Raid that allows disks to fail, ideally any two disks like a Raid-Z2 with a hot or cold spare disk.

Third step is a backup of your datapool ideally via replication to a second storage system on a different physical location. If your OS setup is not very basic (what it should prefer for a storage system), you can replicate your current OS environment/ BE to your datapool with a napp-it replication job. For disaster recovery, you need a minimal OS environment with napp-it (either via reinstall or a basic cold spare bootdisk that you have prepared) where you can restore and activate the former boot environment. This allows a OS disaster recovery or OS distribution within minutes. Data can be restored after a disaster (fire, theft, pool lost) from backup.

1.2 Improved Availability Level2 (second standby server with async replication)

This requires two server where the second can be used as a backup and failover system. Data of both are kept in sync via zfs replication. This allows a syncronisation of both datapools even on a Petabyte system under high load with a delay down to a minute. If the first server fails, you can stop the replication and switch services manually to this backup system. Data modifications between last replication are lost then. If the first system is active again you can switch back after a backward replication based on last snap.

On most problems, your datapool remains intact and the problem is related to hardware or OS. If your second server has enough free disk bays you can also simply move the disks and import the pool to switch services based on current data. Biggest advantage of such a configuration is its simplicity compared with most cluster solutions.

1.3 Improved/High Availability Level3 (single cluster, two server, single storage)

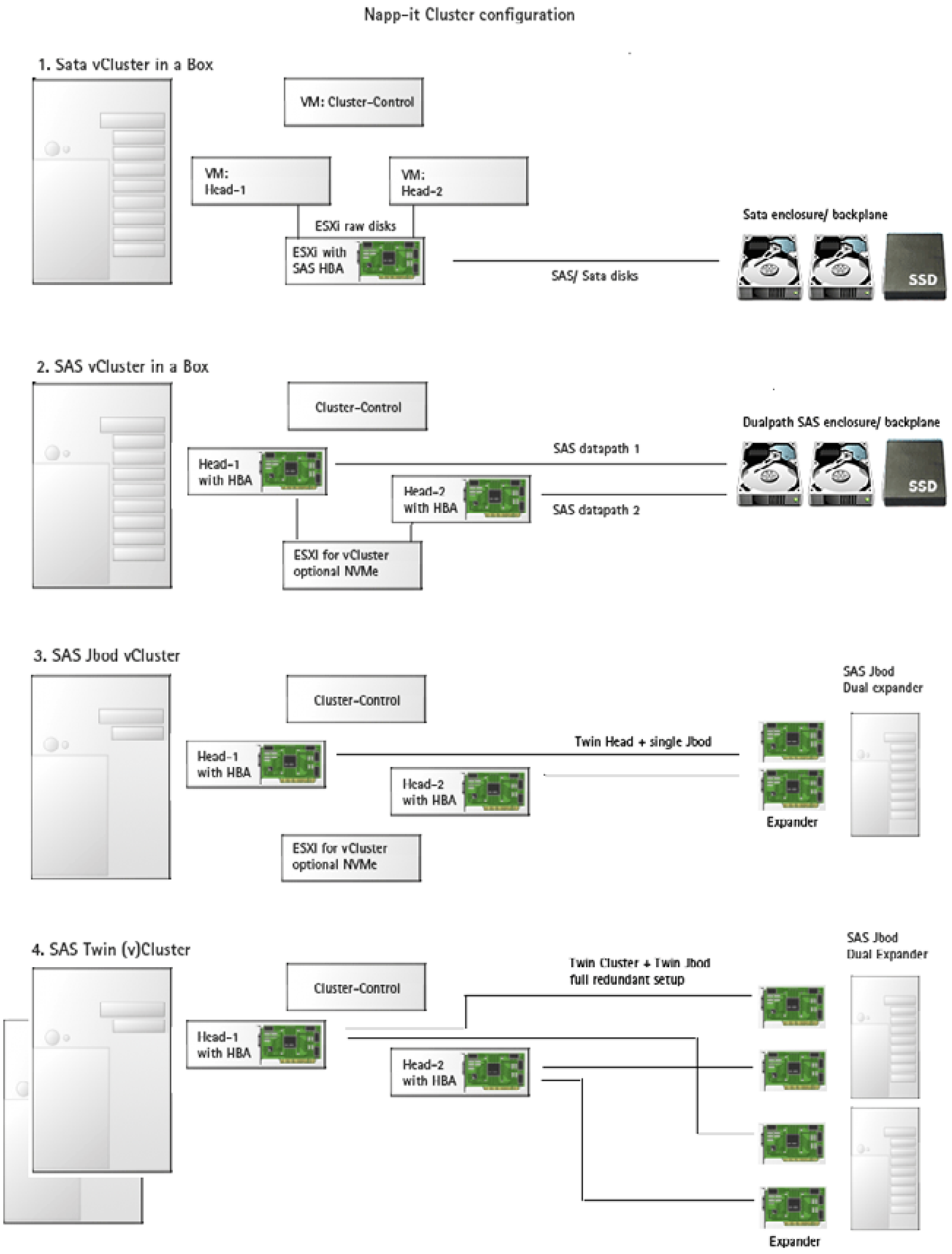
A ZFS cluster requires two heads with a common storagepool also known as „Cluster in a Box" solution. In many configurations storage was connected via dual path SAS or iSCSI either to one or the other head. A cluster management software is needed to control that only one active head has access to the storage and to manage storage and service failover either manually or managed (active/passive) or automatically when the the master server fails (active/active in auto mode). With such a solution, you can survive a failure of a head that provides services and you can do maintenance to a server ex on updates and security patches after you switched services on the fly to the second failover system. This solution does not allow a storage/Jbod failure as storage is not redundant

1.4 Improved/High Availability Level4 (twin cluster, two server, dual storage, Z-Raid)

A ZFS Twin vCluster allows the outage of a whole vCluster. It either works with two iSCSI storage boxes that offer their LUNs to the first and second head which builds a network raid-1 pool (Z-Raid) over them. You can also use two multipath SAS Jbods units each with a twin expander setup. On a failover the second head imports the network mirrored pool or the SAS boxes. A realtime network mirrorring and the dual SAS Jbod solution allows a failure of a server head and a whole storage box failure with services always based on most current data (realtime storage mirror to a second location/ Jbod) offered by the remaining head.

Napp-it supports all four availability levels in a very simple way using Solarish ZFS with its in the OS and ZFS integrated NFS/SMB handling for service switchover and napp-it for server, pool, ip, job and user failover and ip management. ESXi and vCluster reduce complexity and costs massively.

2. vCluster in a Box with Dualpath SAS Options (or ESXi shared Sata disks)



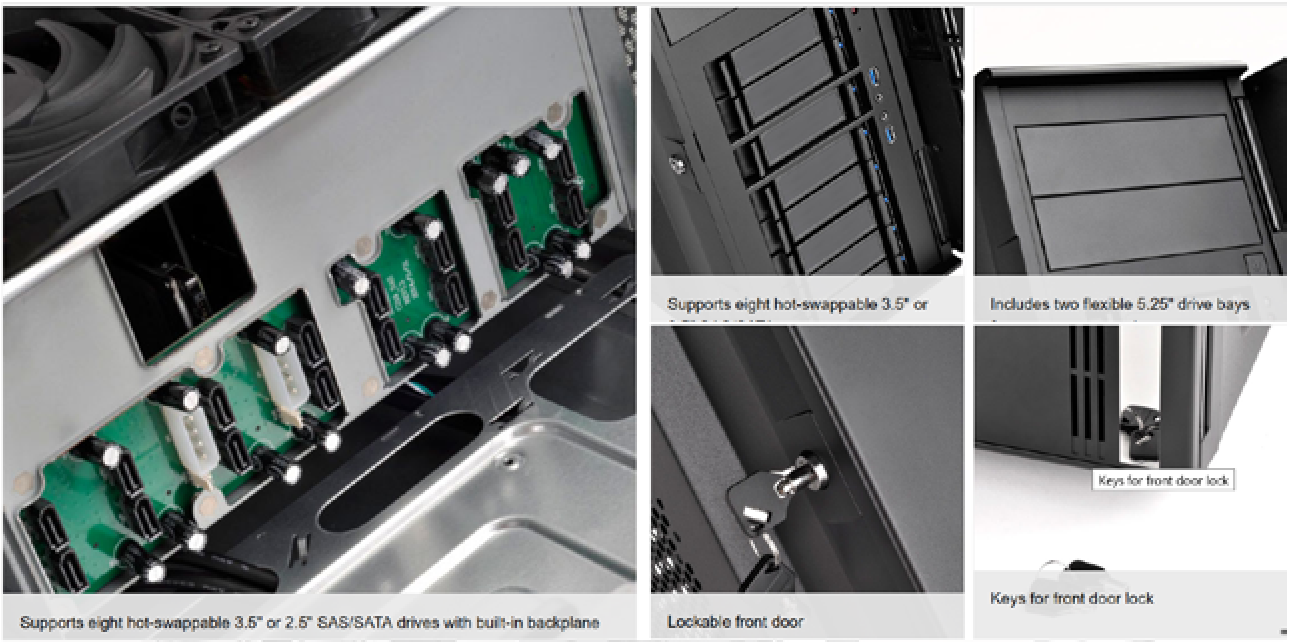
Example of a Lab, SoHo or cost sensitive „vCluster in a Box“ on ESXi

Silverstone CS 380 with 8 x Dual SAS Backplane

Silverstone CS 380

Standard ATX Motherboard
2 x 2,5“ drive Bays

8 x Dualpath SAS Backplane


Supports eight hot-swappable 3.5" or 2.5" SAS/SATA drives with built-in backplane
Supports eight hot-swappable 3.5" or
Includes two flexible 5.25" drive bays
Lockable front door
Keys for front door lock
Keys for front door lock

For a vCluster in a Box with SAS disks you need:

- An ESXi AiO system with 2 x SAS HBAs
- Two storage VMs, each with one of the SAS HBAs in pass-through mode
- up to 8 SAS disks

Connect first HBA to datapath 1 of the SAS disks, the other to datapath2

- One Cluster-Control VM

Example of a Lab, SoHo or cost sensitive „vCluster in a Box" on ESXi

Any case with empty 5,25" Slots

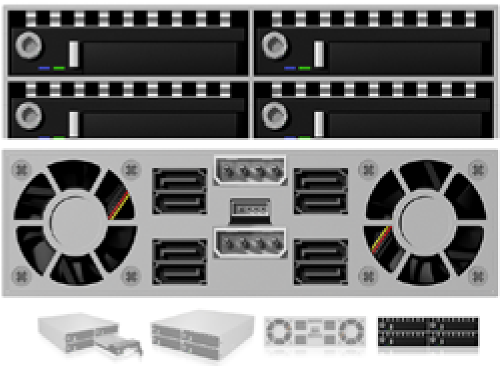
1 x Icybox IB-554   (4 x 3,5")
1 x Icybox IB-2222 (4 x 2,5")

2 x Icybox IB-2222 (8 x 2,5")

Icy Box IB-554 SSD Backplane

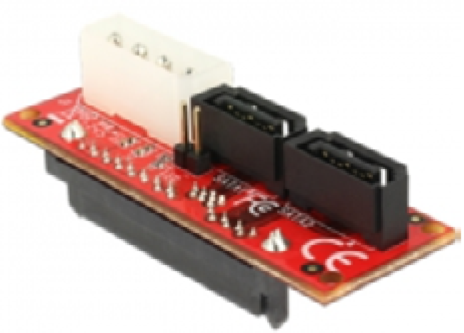
Mpio SAS Backplane for a 4 x 3,5" SAS disks
in a 3 x 5,25" Bay

Icy Box IB-2222 SSK Backplane



Mpio SAS Backplane for 4 x 2,5" SAS disks
in a single 5,25" Bay

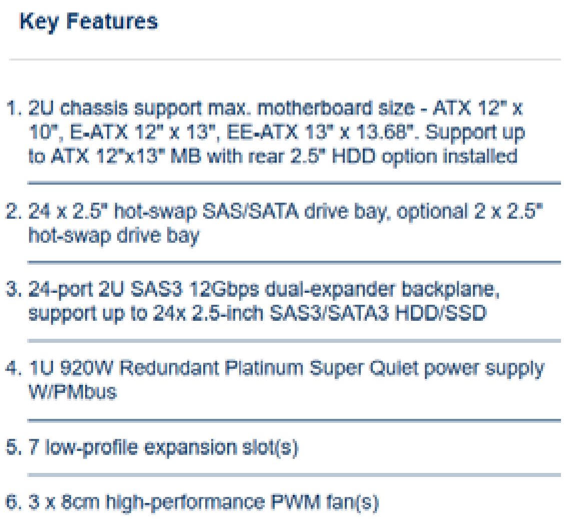
Icy Dock MB991IK-B
1 x 2,5" SAS enclosure > 2 x Sata



Delock 62469
Dualpath SAS -> 2x Sata

Example of a production „vCluster in a Box“

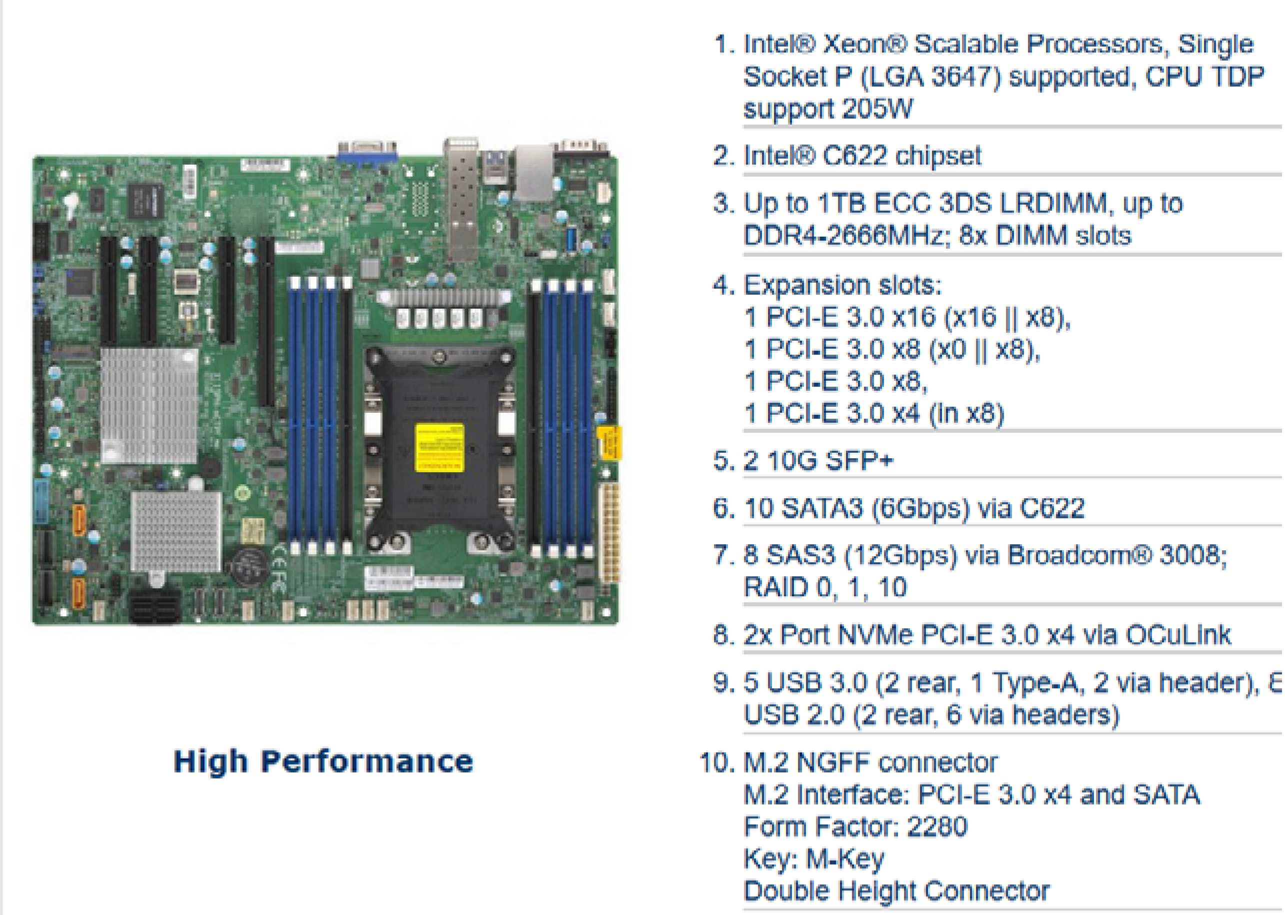
SuperMicro SuperChassis 216BE2C-R920LPB



Key Features

1. 2U chassis support max. motherboard size - ATX 12" x 10", E-ATX 12" x 13", EE-ATX 13" x 13.68". Support up to ATX 12"x13" MB with rear 2.5" HDD option installed
2. 24 x 2.5" hot-swap SAS/SATA drive bay, optional 2 x 2.5" hot-swap drive bay
3. 24-port 2U SAS3 12Gbps dual-expander backplane, support up to 24x 2.5-inch SAS3/SATA3 HDD/SSD
4. 1U 920W Redundant Platinum Super Quiet power supply W/PMbus
5. 7 low-profile expansion slot(s)
6. 3 x 8cm high-performance PWM fan(s)

Suitable Mainboard SuperMicro X11SPH-nCTPF



1. Intel® Xeon® Scalable Processors, Single Socket P (LGA 3647) supported, CPU TDP support 205W
2. Intel® C622 chipset
3. Up to 1TB ECC 3DS LRDIMM, up to DDR4-2666MHz; 8x DIMM slots
4. Expansion slots:
   1 PCI-E 3.0 x16 (x16 || x8),
   1 PCI-E 3.0 x8 (x0 || x8),
   1 PCI-E 3.0 x8,
   1 PCI-E 3.0 x4 (in x8)
5. 2 10G SFP+
6. 10 SATA3 (6Gbps) via C622
7. 8 SAS3 (12Gbps) via Broadcom® 3008; RAID 0, 1, 10
8. 2x Port NVMe PCI-E 3.0 x4 via OCuLink
9. 5 USB 3.0 (2 rear, 1 Type-A, 2 via header), 8 USB 2.0 (2 rear, 6 via headers)
10. M.2 NGFF connector
    M.2 Interface: PCI-E 3.0 x4 and SATA
    Form Factor: 2280
    Key: M-Key
    Double Height Connector

High Performance

needed Extras:
For a Cluster with two nodes you need an LSI HBA each.
As the mainboard has one LSI 9300 HBA, you need another BroadCom 9300-8i HBA

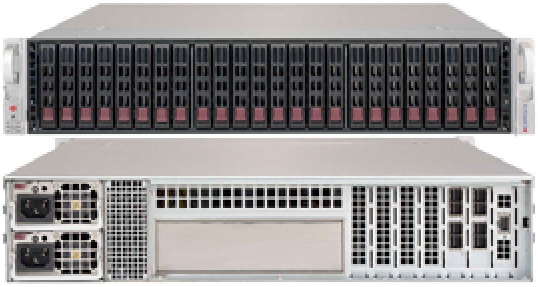
As bootdisk (ESXi and local datastore for the cluster nodes: an Intel DC SSD or Optane 900 or better (280GB min))

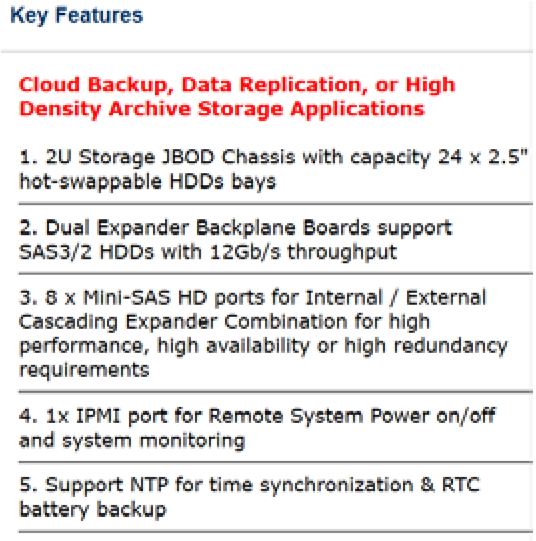
Example of a production Twin „vCluster in a Box“ (allows a whole ESXi or Head failure)

SuperMicro SuperChassis 216BE2C-R920LPB

The setup is similar to the above vCluster with the difference that you do not use the internal storage bays but use a dedicated Jbod box, select one from https://www.supermicro.com/products/chassis/JBOD/index.cfm to allow storage access simulationsly from two ESXi boxes via external 12G SAS

ex SuperMicro SC216BE2C-R741JBOD Jbod with Dual Expander





Key Features

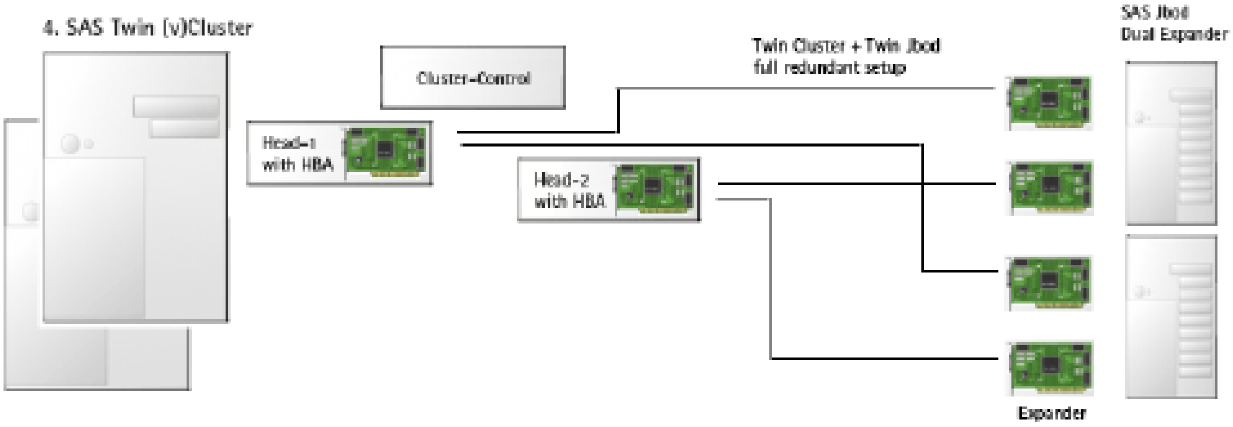
Cloud Backup, Data Replication, or High Density Archive Storage Applications

1. 2U Storage JBOD Chassis with capacity 24 x 2.5" hot-swappable HDDs bays
2. Dual Expander Backplane Boards support SAS3/2 HDDs with 12Gb/s throughput
3. 8 x Mini-SAS HD ports for Internal / External Cascading Expander Combination for high performance, high availability or high redundancy requirements
4. 1x IPMI port for Remote System Power on/off and system monitoring
5. Support NTP for time synchronization & RTC battery backup

Connect one expander to the storage VM on ESXi server 1 (external SAS) and the other expander to the storage VM on ESXi server 2 (external SAS). This gives both ESXi boxes storage access. A failover of services and storage is even possible on a whole outage of an ESXi box with all VMs.

Single point of failure remains this Jbod box. While a failure is very unlikely as there are only disks and the expander with a redundand PSU you may not allow this. In this case you need two Jbod. Connect both to one of the SAS connectors of each storage VM HBA. This will give both storage VMs access to both Jbods. To allow a whole Jbod failure you must select a pool layout that allows a failure ex a 4way mirror vdevs with 2 disks on each Jbod. You will not only get unique security and read performance but even a whole Jbod failuire will keep a pool intact with a mirror per vdev. You can also setup head-1 on hardware for ultimate performance and use a vCluster for head-1, controlserver and optionally backup.

Cabling of a Twin (v)Cluster



vCluster concepts

„vCluster in a Box"

The idea behind is that you need only one ESXi AiO server with two VMs (head1 and head2) that provides iSCSI, NFS and SMB services in a Cluster environment with a failover between them for maintenance or on a OS failure. Disks are shared between the two VMs, either as ESXi shared Sata/SAS/NVMe disks or preferable via SAS multipathing or optionally shared iSCSI targets. This concept is called „vCluster in a Box" (virtual Cluster in a Box). Failover management can be done manually or managed via a ClusterControl VM to switch pools, ip, jobs and user.

„Twin (v)Cluster"

As an extension you can double this and use two ESXi server (or use hardware for every node), each with a head VM to provide iSCSI, NFS or SMB services and a Cluster-Control VM for failover management. Each ESXi server can additionally hold a Storage server VM to provide iSCSI targets that you can access from both clusters to allow a realtime network mirror of data. A Twin Cluster requires common access to multipath SAS in a single or dual Jbod setup or optionally two independent NAS/SAN storageserver for iSCSI. Failover management is done with one of the two Cluster-Control VMs. A Dual Jbod setup requires a dual expander configuration in each Jbod.

Twin-Cluster with iSCSI solutions can be quite complicated to setup, tune and maintain. There is no support from napp-it for such setups. Use them only with a qualified staff or a service contract. Prefer multipath SAS.

Suggested hardware for the ESXi server („vCluster")

Lab use and ESXi raw disk access

You can use any ESXi capable server with16GB RAM or more and an LSI HBA. Support for pass-through (vt-d) is optional but not required. Add the disks to one VM (cluster-control) as raw disks and to the other VMs via „add existing hard disk". For a manual failover you can use ESXi free and napp-it free or napp-it Pro Complete and a Cluster-Control VM for a fully managed manual or auto failover.

Office and department filer and Dualpath SAS in a vCluster

Use any ESXi capable Xeon server with at least 32 GB ECC RAM (64GB-128GB recommended). As bootdevice for ESXi you can use a small SSD that you can use as a local datastore for the VMs. Add an LSI HBA for head-1 and head-2 (pass-through mode) and dualpath enclosures and SAS disks for storage, Slog and L2Arc.

Large capacity filer and Dualpath SAS („Petabyte storage") in a vCluster

Use any ESXi capable Xeon server with at least 64 GB ECC RAM (64GB-128GB recommended). As bootdevice for ESXi you can use a small SSD that you can use as a local datastore for the VMs. Add an LSI HBA for head-1 and head-2 and one or two Jbod dualpath enclosures in a dual expander configuration and SAS disks for storage, SAS Slog and L2Arc. For high performance filer usage, add two 10G nics in pass-trough mode.

Barebone Options

You can install Cluster-Control, head-1 and Head-2 in a barebone setup. It is possible to combine a backupserver and Cluster-Control. Such a HA config requires two heads and a backupserver and allows any of the heads to fail. In a Dual Jbod setup with dual expander, it additionally allows a Jbod failure.

Twin Storage/ Jbod Option

A Cluster consists of two heads with access to a single storage box. If you want to allow a full Jbod failure, you need two Jbods, each with a dual expander setup. You can the connect each Jbos to each head. You need a pool layout that allows a full Jbod failure ex a 4way mirror with 2 disks on each Jbod.

iSCSI Option

You can use iSCSI Luns with shared access instead of multipath SAS. Main advantage is that you are not limited by 10m cable length like with SAS. This allows a full storage mirrorring between locations. You can use raw disks ar a whole filesystem/pool (zvol) as source of a LUN. Handling is not as easy and performance is not as good as SAS dualpath solutions.

Dualpath NVMe

Currently NVMe is mainly a singlepath technology. Only one head can access NVMe directly.
As a workaround you can provide NVMe disks as shared ESXi disks in a vCluster setup or you can use iSCSI to access NVMe disks.

There are now upcoming dualpath NVMe devices and appliances to build pure NVMe HA solutions.

Required OS software

ESXi 6.7u2 or newer (free or any other ESXi license, does not matter). For the head and control VMs, you can use my free ESXi ZFS server template with napp-it 18.02free on OmniOS 151028 stable. For a manual failover no additional software is required.

For managed manual/auto failover with user and job failover you need at least the Cluster-Control VM with napp-it Pro Complete. There are also Twin (Filer and Backup), Triple (HA Cluster, and backup), Quad (HA Cluster, filer and Backup) and Location licenses (all server on a location like a campus) available.

Preword!

A failover- or software problem or a hanging server process can results in a situation where both heads want access to a pool simultaniously. This can corrupt a pool. Napp-it cares about such a critical situation with three methods. First is the ZFS property multihost that is set to on to avoid multiple mounts - manually or automatically. Second protection is that the former head is rebooted via a remote control command prior a failover (more secure and faster than a pool export/import that may be blocked in some situations. It also frees the HA ip in any case). Prior a failover, the second head checks if the first server is offline. This can be a problem on a hanging system. This is why you can add an additional protection, an independent kill method (stonith). In a production environment you should care about a stonith mechanism (Esxi can force reboot a VM via SSH or you can use IPMI for server reset) to force a reboot of the former head with two independent methods.

In a failover Cluster, backup is as important as on a basic storage system.
Use the HA ip as source for a replication to continue a replication with either head active.

ZFS setting multihost=on

It is highly recommended to set the ZFS property multihost to on with shared pools. This is a protection against a concurrent mount of a ZFS pool on two servers. It is not perfect enough to be the only protection but is a very good additional protection especially without stonith and against accidental manual mounts.
https://wiki.lustre.org/Protecting_File_System_Volumes_from_Concurrent_Access

2.1. Cluster-Setup, Lab example (vCluster with SAS)

Cluster setup consists of the steps (follow step by step):

a.) - provide the server nodes h1, h2, and control (virtualized as a vCluster or barebone)
b.) - provide shared disk access to h1 and h2 (SAS Dualpath, any ESXi shared disks or iSCSI)

c.) - configure cluster-control
d.) - build an appliance group on control for remote access of h1 and h2
e.) - configure heads

f.) - configure manual and auto failover

2.1.a Setup ESXi and nodes

- Use for ex. a Silverstone CS 380 with a SuperMicro server mainboard and two SAS HBAs (onboard or via PCIe)
- Install ESXi 6.7u1+ onto a boot SSD (240GB or more)
- import three instances of the napp-it ova template, name them control, h1 and h2
- add one of the LSI HBA to h1, the other to h1 as a pass-through device

2.1.b Setup Dualpath SAS disks

- Connect one port of the SAS disks to h1/HBA-1 and the other disk port to h2/HBA2.
  Both heads h1 and h2 can access disks simultaniously.

2.1.c Setup Cluster-Control (control)

set hostname and ip	hostname = control, set management and lan ip

- on control, open menu System > Appliance Cluster > Cluster Settings and configure:

Settings head-1	
LAN IP head1	enter the ip on h1 for lan access (use static adresses)
Man IP head 1	enter the ip on h1 for management access, can be the same like lan (use static adresses)
Net Link head 1	name of the nic link ex vmxnet3s and add :3 ex vmxnet3s0:3 (HA ip will use this link)
Stonith gead 1	VM reset command for head 1 via ipmi or ESXi cli (for tests enter echo 1)
Settings head-2	
LAN IP head2	enter the ip on h2 for lan access (use static adresses)
Man IP head 2	enter the ip on h2 for management access, can be the same like lan (use static adresses)
Net Link head 2	name of the nic link ex vmxnet3s and add :3 ex vmxnet3s0:3 (HA ip will use this link)
Stonith head 2	VM reset command for head 2 via ipmi or ESXi cli (for tests enter echo 1)
Settings control	
LAN IP control	enter the ip on control for lan access (use static adresses)
Man IP control	enter the ip on control for management access, can be the same like lan (static)
Stonith check	ESXi CLI command to check if Stonith is available (for evaluations use: echo test)
Stonith reply	answer of check ex test
Failover mode	use manual for tests
HA settings	
HA service ip	enter the ip under whom you want to access NFS or SMB (switched on failover)
HA netmask	enter the netmase for HA service ip
set mode	of control to cluster-control in menu About > Settings

2.1.d Build an Appliance Group on control to allow remote-access of h1 and h2

- On Cluster-Control open menu Extensions > Appliance Group > ++add appliance
  add the ip (management ip) of h1 and h2 and the HA ip (napp-it 19.06 pro) all together as a cluster.
  This allows a replication from Cluster-Control or a backup server that is failover safe.

Test remote control: Open menu System > Appliance Cluster. You should see state of h1 and h1

2.1.e Setup head-1 (h1) and head-2 (h2)

head-1
- Set hostname and ip   hostname = h1, set management and lan ip
- Create a datapool with name „zraid-1" from your Dualpath SAS disks, export the Pool (to avoid auto-mount)
  You always need a pool zraid-1 for moving jobs, users, or other settings ex iSCSI
- Switch h1 to mode „Cluster-Failover" in About > Settings
- optionally: create a filesystem rpool/h1 and share it via SMB as readonly (to check active head via smb)

head-2
- Set hostname and ip   hostname = h2, set management and lan ip
- Switch h2 to mode „Cluster-Failover" in About > Settings
- optionally: create a filesystem rpool/h2 and share it via SMB as readonly (to check active head via smb)

2.1.f Configure failover

Open menu System > Appliance Cluster
- Menu must show state of h1and h1 otherwise you made an error during past steps
- Node h1 has mounted HA pool zraid-1, node h2 shows pool zraid-1 as importable
  control is in mode cluster-control, h1 and h2 in mode cluster-failover

Open menu System > Appliance Cluster > Failover Settings

For basic tests, save defaults.
You can edit failover settings when requires

2.2 Test manual failover

Open menu System > Appliance Cluster > Manual Failover
Under action, set h1 to master and confirm

- This disables the zraid pool on the other head h2 (with a forced reboot to ensure a pool dismount) and
- mount pool zraid pool on h1
- enable ha ip on h1

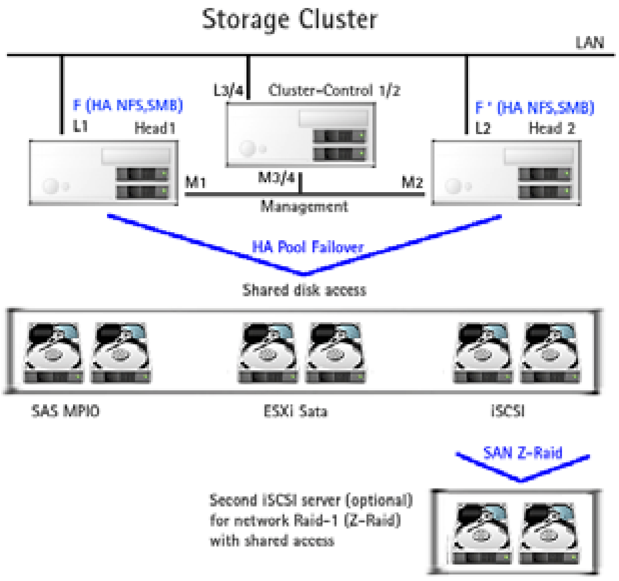
iSCSI, NFS and SMB are availabled over the HA ip (optionally: you can check active head via SMB)

Open menu System > Appliance Cluster > Manual Failover again
Under action, set h2 to master and confirm

- This disables the zraid pool on the other head h1 (with a forced reboot to ensure a pool dismount) and
- mount pool zraid pool on h2
- enable ha ip on h2

Comstar iSCSI, NFS and SMB are available over the HA ip after an outage of around 20s.
Mount a HA zraid-pool only via „manual Failover". If you create or mount a pool in standalone mode,
this pool is auto mounted after reboot! what must be avoided!

3. Cluster in a Box Cabling



Such a cluster requires two server (head1 and head2) with a common storagepool also known as „Cluster in a box" solution. Such a setup can be done with two barebone server and multipath SAS disks with a datapath to both servers. Additionally you need a cluster control software for failover. Such a software can run on a head or on a dedicated machine.

We use a virtual „vCluster in a Box" with one ESXi server with the three VMs Head1, Head2 (to provide HA for iSCSI, NFS and SMB) and Cluster-Control for failover management. For the shared disks, we can use Sata but prefer SAS disks, connected to an LSI HBA assigned to ESXi (no pass-through). This allows to add the disks as raw-disk to cluster-control and then additionally to Head1 and Head2. Best is using multipath SAS with two HBAs in pass-through mode.

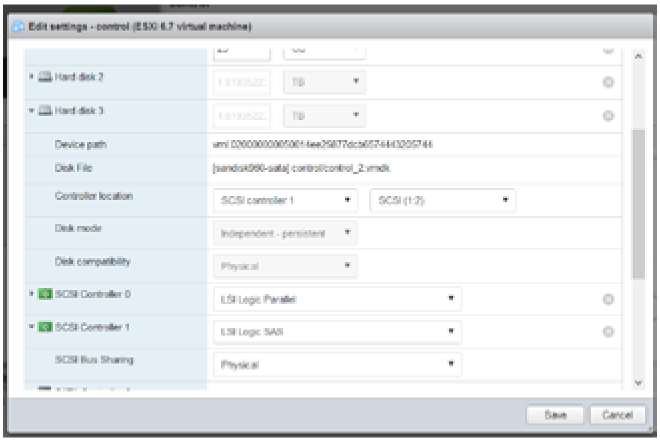
3.1 Disk-Setup of the vCluster-Control-1 VM with ESXi shared disk access

Import the napp-it template (or setup the VM manually) and name the VM cluster-control or control. This is the VM where you add the raw Sata or SAS disks that are connected to an LSI HBA (assigned to ESXi). Only on SAS disks ESXi shows size and manufacturer. Use vmxnet3 vnics as they are faster than the e1000 vnic.

As a first step, your controller VM needs a second virtual SCSI controller with the type **LSI Logic SAS and SCSI-Bus sharing=physical.** Check/ add SCSI Controller 1.

Then add the disks in the VM settings with menu Add hard disk > new raw disk.
Edit the settings for the new raw disk: assign the disk to the SCSI controller1 (SAS mode),
Disk mode = Independent - persistent and Disk compatibility = Physical.

You may first set compatibility mode to virtual to switch Disk mode, then set compatibility mode back to Physical.


Edit settings - control (ESXi 6.7 virtual machine)

Hard disk 2
Hard disk 3
Device path: vml.02000000005001 4ee25877dcb5574443205744
Disk File: [sandisk960-sata] control/control_2.vmdk
Controller location: SCSI controller 1 ▾ SCSI (1:2) ▾
Disk mode: Independent - persistent ▾
Disk compatibility: Physical ▾
SCSI Controller 0: LSI Logic Parallel ▾
SCSI Controller 1: LSI Logic SAS ▾
SCSI Bus Sharing: Physical ▾

SCSI controller = physical
This allows controller sharing between ESXi machines

SCSI controller = virtual
This allows controller sharing only on this ESXi machine

Disk mode=Independent-persistent
Disk behaves like real disks, no ESXi snaps/ logs

Disk compatibility=physical
Allows the VM to access hardware directly

If you use Sata disks instead of SAS disks, ESXi shows only an identifiert but under ZFS you will see manufacurer, serial and size as this is raw physical disk access. You should prefer SAS disks as this shows model and capacity in ESXi in menu add raw disk. For best performance prefer SAS disks in a dual HBA setup.

3.2 Setup of the vCluster-Control-2 VM (optional)

In a „vCluster in Box" you only need the VMs head1, head2 and control.
If you want a full Storage and Service Failover with disks from iSCSI targets, you need two ESXi servers, each with a storage control VM and either head1 or head2 and two iSCSI storage servers to provide iSCSI targets (virtual server on either of the ESXi machines or physical server). This setup is called „Twin vCluster" with a realtime Raid over iSCSI targets in a network. (no support from napp-it for iSCSI and Twin Cluster setups)

Settings of cluster-control2 are identical to control1. Use vmxnet3 vnics as they are faster than the e1000 vnic.
Set any Cluster-control servers to this mode in About > Settings

3.3 Setup of Head1 (services iSCSI, NFS and SMB) with ESXi shared disk access

Setup of this VM is quite identical to the control VM with the difference that you do not add raw disks (they are assigned to the control VM) but add the disks via menu
Add hard disk > Existing hard disk where you select the disks from the control-VM.

Import the napp-it template (or setup the VM manually) and name the VM cluster-control or control. This is the VM where you add the raw Sata or SAS disks that are connected to an LSI HBA (assigned to ESXi). Only on SAS disks ESXi shows size and manufacturer. Use vmxnet3 vnics as they are faster than the e1000 vnic.

As a first step, your controller VM needs a second virtual SCSI controller with the type
LSI Logic SAS and SCSI-Bus sharing=physical. Check/ add SCSI Controller 1.

Then add the disks in the VM settings with menu Add hard disk > Existing hard disk.
You find the disk in the folder of the control-VM (ex control_2.vmdk).
Edit the settings for the disk: assign the disk to the SCSI controller1,
Disk mode = Independent - persistent and Disk compatibility = Physical.

You may first set compatibility mode to virtual to switch Disk mode, then set
compatibility mode back to Physical.

Be careful: Re-check now if all shared disks are assigned to scsi-1 (sas mode) with
disk-mode = independent-persistent and disk compatibility = Physical. If not, delete the disk
(delete from datastore) and re-create !! On ESXi hickups, reboot ESXi.

Option with SAS multipath.

You need an LSI HBA in pass-through mode on head-1 and head-2 and a dualpath enclosure for 2,5" or 3,5" disks.
Connect one datapath to HBA-1, the other to HBA-2 to give both heads disk access. This is the preferred method.

additional settings
Create a filesystem root/h1, share it via SMB and set to readonly
Use it to check active head (via HA ip)

Edit /boot/defaults/loader.conf
abd set reboot delay from 10s to 2s

3.4 Setup of Head2 with ESXi shared disk access

In a „vCluster in a Box" solution, the VM head2 is on the same ESXi server as head1 and control. In a „Twin vCluster" where you want full head redundancy, you must place the head2 VM together with the control2 VM to the second ESXi server. Setup is identical to head-1.

Re-check now **if all shared ESXi disks are assigned to scsi-1 (sas mode and SCSI-Bus sharing=physical.) with disk-mode = independent-persistent and disk compatibility = Physical. If not, delete the disk (delete from datastore) and re-create !!** Set heads to Cluster Failover mode in About > Settings.

Option with SAS multipath.

You need an LSI HBA in pass-through mode on head-1 and head-2 and a dualpath enclosure for 2,5" or 3,5" disks. Connect one datapath to HBA-1, the other to HBA-2 to give both heads disk access. This is the preferred method. For redundant storage you need two Jbods (dual expander)

additional settings
Create a filesystem root/h2, share it via SMB and set to readonly
Use it to check active head (via HA ip)

Edit /boot/defaults/loader.conf
abd set reboot delay from 10s to 2s

3.5 Build an appliance group from all servers in the cluster (control, h1, h2, backup)

Build an Appliance Group over control, h1 and h2
(Control VM: Menu Extensions > Appliance Group > ++ add appliance)

3.6 Pool and Service setup for „vCluster in a box"

Current state: Boot the VMs and all of them can see the shared disks.
We must now strictly ensure that only one VM is allowed to manipulate the disk (the Master). Only one VM must be allowed at a time to write to disks or create/ import the datapool and provide NFS/SMB services.

First step: create a pool zraid-1 on head 1. in standalone mode. Only use head1 for access.

Second step: switch server mode in About > Settings
- Switch mode for head 1 and head 2 from standalone to Cluster-Failover
- Switch mode for control from standalone to cluster-control

On the clustercontrol VM (Cluster Control functionality requires a complete license)
Setup your Cluster settings in menu System > Appliance Cluster > Cluster settings.
Setup your Failover settings in menu System > Appliance Cluster > Failover settings.

Third step: Create an appliance group on control VM

- menu Extensions > Appliance Group: ++ add appliance

additional settings
Create a filesystem root/h, share it via SMB and set to readonly
Use it to check active head (via HA ip)

Edit /boot/defaults/loader.conf
abd set reboot delay from 10s to 2s

3.7 Shared disk access with dualport/multipath SAS disks

Shared raw disks access via ESXi storage features allows a single box vCluster solution. For high performance demands you can use multipath SAS with a dualpath SAS enclosure or a Jbod storage unit in a dual expander configuration where each SAS disk is connected to both expanders. You can then connect one expander to head-1 and the other to head-2 to give both shared access to all disks. With napp-it Cluster-Control only one head can access disks at a time. From a performance view, shared raw or SAS disk access is faster than iSCSI.

If you want full storage redundancy (allow a full failure of a Jbod case and a outage of a head) you need two Jbod boxes, both with a dual expander solution and two or four external SAS ports. Connect them to either head and build a pool with a layout that allows a full storage failure with a given level of redundancy even when one Jbod fails (ex 4 way mirror with two disks on each Jbod). You can use one physical head (main head) with SAS disk access while you can virtualize the second head with HBA passthrough as this one is only needed for maintenance and cluster control.

4.0 Manual Cluster failover („by hand" with napp-it free)

You can build a fully free Cluster based on ESXi free and napp-it free. In such a configuration you prepare the two heads h1 and h2 for failover without the Cluster-Control node. Both heads are in standalone mode.

Create a pool from your shared disks on head-1. Add a dedicated ip to your LAN nic.
Access NFS and SMB shares always over this ip. Give head-2 shared disk access.

For a manual failover you must now disable or switch the LAN ip for access to services and export the pool on head-1 (Never mount a pool simultaniously on both heads, requeck that the pool is exported).

On head-2 you can then import the pool with shares and ZFS settings intact. You need to add the same local users with same uid/gid and SMB groups (or use Active Directory) to allow SMB access for your users. Enable/switch the ip for LAN access. Your users can now access the shares on head-2. Only a short outage is seen from a user side. Napp-it Cluster-Control helps to automate this failover with user an jobs intact to make a switch fast and secure.

4.1 Managed manual Cluster failover

use menu System > Appliance Cluster > Manual Failover

4.2 Auto Failover

Enable the Cluster agent (VM control) in menu Services > Z-Raid Cluster.
(start the agent on Cluster-Control only)

The agent initiates a „set to master" ex after bootup or when no head is in a master role.
It selects an available head (prefers h1) for the master role. When a master fails, the auto-failover process is initiated for the second standby head. Such an auto failover can last 30-40s. While an auto failover is processed, a manual failover is blocked.

During an auto failover it is essential that the former master is down when the other head mounts the pool. A hanging system at the end of the remote controlled down process can theoretically lead to a situation where both heads access the pool.
To avoid such a situation in any case, use a Stonith mechanism. This must be a command that can kill/ reset a system independently from its OS state. A Stonith command can be an esxcli command via SSH to reset the VM or a reset via ipmitools.

4.3 Switch users between h1 and h2

A manual failover or a running cluster-service syncs users to the zraid-1 pool.
In failovers ettings you can enable to restore them after a failover from the zraid-1 pool

4.4 Switch jobs between h1 and h2

When you create a snap or scrub job on a master node, you can set the job to Cluster-mode.
This allows a continuation of the jobs after a failover.

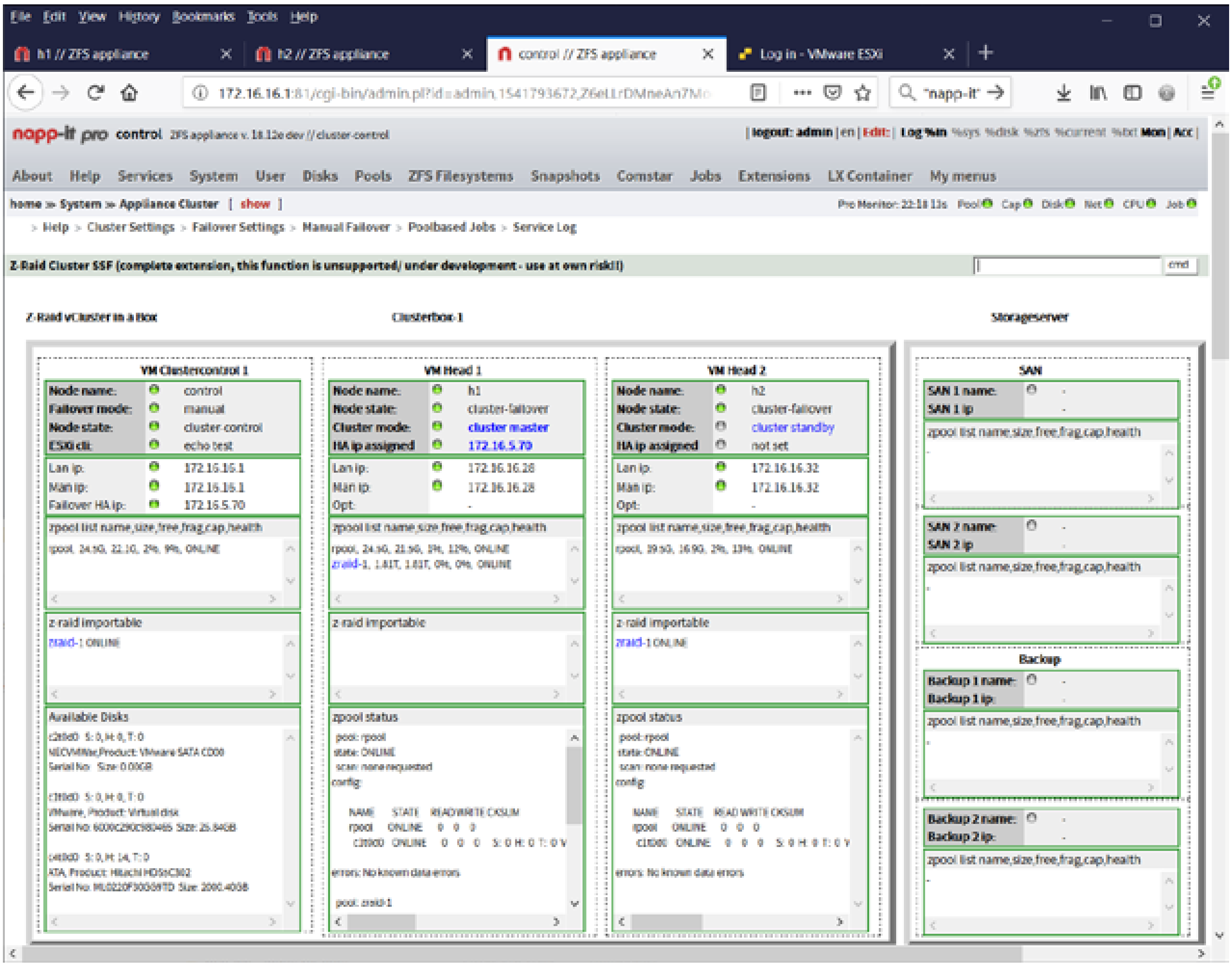
For replications, use a backup server where you build an appliance group with h1 and h2 and the HA ip as source.
- set h1 as master and add h1 with its HA ip as a group member on your backup server
- failover to h2 and add h2 with its HA ip as a group member on your backup server (h1 and h2 with same ip)

- create a replication job on a backup server from active master. The job remains valid after a failover due same HA ip when both heads have the same hostname. For different hostnames copy the group file in /var&weg-gui_logs/ group from h1 when master to h2. Prefer a job per filesystem (avoid recursive)

4.5 Switch from Cluster to Standalone mode

in About > Settings. Never to mount a poo simultaniously on both heads!!

4.6 Cluster State



File Edit View History Bookmarks Tools Help
h1 // ZFS appliance
h2 // ZFS appliance
control // ZFS appliance
Log in - VMware ESXi
172.16.16.1:81/cgi-bin/admin.pl?id=admin,1541793672,Z6eLLrDMneAn7Mo
"napp-it"
napp-it pro control ZFS appliance v. 18.12e dev // cluster-control
| logout: admin | en | Edit: | Log %in %sys %disk %zfs %current %txt Mon | Acc |
About Help Services System User Disks Pools ZFS Filesystems Snapshots Comstar Jobs Extensions LX Container My menus
home » System » Appliance Cluster [show]
Pro Monitor: 22:18 13s Pool Cap Disk Net CPU Job
> Help > Cluster Settings > Failover Settings > Manual Failover > Poolbased Jobs > Service Log
cmd

Z-Raid Cluster SSF (complete extension, this function is unsupported/ under development - use at own risk!!)

Z-Raid vCluster in a Box — Clusterbox 1 — Storageserver

VM Clustercontrol 1

Node name:	●	control
Failover mode:	●	manual
Node state:	●	cluster-control
ESXi cli:	●	echo test
Lan ip:	●	172.16.16.1
Man ip:	●	172.16.16.1
Failover HA ip:	●	172.16.5.70

zpool list name,size,free,frag,cap,health
rpool, 24.5G, 22.1G, 2%, 9%, ONLINE

z-raid importable
zraid-1 ONLINE

Available Disks
```
c2t0d0  S: 0, H: 0, T: 0
NECVMWar,Product: VMware SATA CD00
Serial No:   Size: 0.00GB

c1t0d0  S: 0, H: 0, T: 0
VMware, Product: Virtual disk
Serial No: 6000C290C980465 Size: 25.84GB

c4t0d0  S: 0, H: 14, T: 0
ATA, Product: Hitachi HDS5C302
Serial No: ML0220F30000TD Size: 2000.40GB
```

VM Head 1

Node name:	●	h1
Node state:	●	cluster-failover
Cluster mode:	●	cluster master
HA ip assigned	●	172.16.5.70
Lan ip:	●	172.16.16.28
Man ip:	●	172.16.16.28
Opt:		-

zpool list name,size,free,frag,cap,health
rpool, 24.5G, 21.5G, 3%, 12%, ONLINE
zraid-1, 1.81T, 1.81T, 0%, 0%, ONLINE

z-raid importable

zpool status
```
 pool: rpool
state: ONLINE
 scan: none requested
config:

    NAME     STATE   READ WRITE CKSUM
    rpool    ONLINE     0     0     0
     c2t0d0  ONLINE     0     0     0   S: 0 H: 0 T: 0 V

errors: No known data errors

 pool: zraid-1
```

VM Head 2

Node name:	●	h2
Node state:	●	cluster-failover
Cluster mode:	○	cluster standby
HA ip assigned	○	not set
Lan ip:	●	172.16.16.32
Man ip:	●	172.16.16.32
Opt:		-

zpool list name,size,free,frag,cap,health
rpool, 39.5G, 16.9G, 2%, 13%, ONLINE

z-raid importable
zraid-1 ONLINE

zpool status
```
 pool: rpool
state: ONLINE
 scan: none requested
config:

    NAME     STATE   READ WRITE CKSUM
    rpool    ONLINE     0     0     0
     c1t0d0  ONLINE     0     0     0   S: 0 H: 0 T: 0 Y

errors: No known data errors
```

SAN

SAN 1 name:	○	-
SAN 1 ip		-

zpool list name,size,free,frag,cap,health
-

SAN 2 name:	○	-
SAN 2 ip		-

zpool list name,size,free,frag,cap,health
-

Backup

Backup 1 name:	○	-
Backup 1 ip:		-

zpool list name,size,free,frag,cap,health
-

Backup 2 name:	○	-
Backup 2 ip:		-

zpool list name,size,free,frag,cap,health

Extending or debugging Manual Failover

If you want to debug or enhance the manual failover script:
„/var/web-gui/data/napp-it/zfsos/03_System and network=-lin/11_Appliance_Cluster/04_Manual_Failover/action.pl“

common library:
/var/web-gui/data/napp-it/zfsos/_lib/illumos/cluster-lib.pl

Remote control (sub grouplib_remote_call_zraid):
/var/web-gui/data/napp-it/zfsos/_lib/illumos/grouplib.pl

Extending or debugging Auto Failover
If you want to debug or enhance the autofailover agent:
- Stop the agent in menu services > Zraid Cluster
- Start the service manually at console as root

perl /var/web-gui/data/napp-it/zfsos/_lib/scripts/agents/agent_zraid_ssf.pl
end the script with ctrl-c

Cluster failover supports NFS and SMB services with AD users or local users connected.
Other services (ex www, iSCSI) would require addition failover scripts

5. Twin Cluster solution

Twin Cluster solutions are a method to allow a whole Cluster failure and/or a failure **of a storage box.**
You can combine physical heads with vCluster and one or two SAS Jbods or use iSCSI storage.

Twin Cluster and iSCSI solutions can be quite complicated to setup, tune and to do maintenance.
There is no support from napp-it for setup and maintenance. Prefer the multipath SAS solution.

In a Twin Cluster setup you want not only redundancy for the heads that provide NFS and SMB services but full redundancy on a whole server or storage level. In this configuration you need two head units, each with a cluster-control instance with shared disk access from the two heads with the options

5.1 Twin vCluster, single storage

In this setup you use two ESXi server systems. Each system holds a head-vm with an HBA and external SAS ports in pass-through mode and a cluster-control VM. Shared access to disks can be achieved by a single dual expander multipath SAS Jbod case that you can connect to both ESXi systems. Such a system will allow a full outage of a ESXi server system as the second server system with ist own control instance can initiate a failover and can provide storage services, ex https://www.supermicro.com/products/chassis/2U/826/SC826BE2C-R741JBOD

As an alternative to SAS you can use shared access to FC/iSCSI Luns on a SAN server. Acess the Luns over the ESXi initiator or Comstar initiator. (Check both for performance in your environment)

5.2 Twin Cluster, single storage (not virtualised)

This is the same concept like 5.1 but you use a physical server for head-1.
The failover system can be a vCluster with head-2 and cluster-control.

As an Slog use a high performance SAS Enterprise SSD, ex WD Ultrastar DC SS 530

5.3 Twin Cluster, dual/redundant SAS storage

This is an extension where you use two SAS Jbod cases, each with a dual expander. You can the connect both Jbods with both heads. This will allow a full storage (Jbod) or full ESXi or head failure.

ex https://www.supermicro.com/products/chassis/2U/826/SC826BE2C-R741JBOD

For high performance or an Slog use SAS Enterprise SSD, ex
WD Ultrastar DC SS 530

5.4 Twin Cluster, dual/redundant iSCSI storage (realtime network/location mirror)

This is an extension where you use a vCluster or Cluster on two physical locations. Storage is provided by iSCSI LUNs (can be a LUN from a whole pool). A Z-Raid pool with NFS and SMB services is created from a realtime/ network mirror over these LUNs. Performance of the network mirrorred pool depends heavily on network performance and latency. Main use case is realtime backup and failover to a second location.

5.5 Setup iSCSI targets with napp-it from a filesystem or whole pool

Step 1: enable Comstar target (and optional initiator) services in menu Services > Comstar
Step 2: create a pool and a filesystem ex netraid1, optionally remove 10% fres reservation
Step 3: create an iSCSI share in menu ZFS filesystems. This share is also called a LUN or iSCSI disk.

In the row of the filesystem ex netraid1, click on zfs unset (column iSCSI). You can now create an iSCSI Share (zvol + LU+Target+View) from this filesystem. Select a size up to 90% of the poolsize. Confirm to activate the Lun. If you need you can increase the size of the zvol (ex when pool size grows) ex via „zfs set volsize=230G pool/vol"

You can also create Zvols, Logical Units, Targets and Views manually in menu Comstar. If you use the mechanism in menu filesystem this is a one click and set mechanism and if you replicate the zvol you can simply activate it again in menu filesystems as the GUID of the iSCSI LUN is part of the zvol name then.
If you use thin provisioned Zvols on a pool you should care about the data used by other filesystems to avoid a sudden filesystem full problem.

5.6 Switch from Cluster mode to Standalone mode

Switch all nodes to standalone mode in About > settings

5.7 Setup iSCSI initiator with napp-it

You need an Initiator to connect iSCSI Luns. An Initiator Service is included with many Operating systems like ESXi, any Solarish and Windows. If you want to connect a LUN from Solaris or OmniOS you must

1. enable the initiator service in menu Services > Comstar
2. select a detection method for network LUNs. For auto-detection use „Sendtarget Discovery"
   Enable „Sendtarget Discovery" in menu Comstar > Initiator > Sendtarget Discovery
   You only need to activate and to enter the ip address of the SAN server that provides Targets and LUNs.

After you have enabled Sendtarget Discovery, LUNs are automatically detected and offered like a local disk. You can now create a pool from this basic disk. Redundancy is not needed as the source of the LUN is a pool with redundancy. If you want an Slog you can create a dedicated LUN ex from a partition on an Intel Optane and add this LUN as an Slog. As an example 1 have added am Intel DC 3610 as a LUN build from the raw disk

6. Licenses

To build a Z-Raid vCluster (network Raid), you need at least

- ESXi 6.7u1 (free is ok)
- Three OmniOS VMs ClusterControl, Head-1 and Head-2

- it is strongly suggested to add a backupserver. You canuse the backup server for Cluster-Control

OmniOS stable/long term stable is OpenSource/free (a support contract with access to OS developpers is avail-able at https://omniosce.org/invoice.html). Only Cluster Control needs a napp-it Pro complete license. Head-1 and Head-2 can be napp-it free. If you want all features of napp-it Pro on all nodes, SAN and backup systems, you can aquire a Twin, Quad Cluster or Location license for all of your machines.

A initial setup of napp-it comes with a 30day eval period. For short tests you can request 2day keys online at https://www.napp-it.org/extensions/evaluate_en.html - for a longer eval period, just ask.

For a Twin Cluster or a single Cluster based on iSCSI you need one or two iSCSI server.
You can use physical OmniOS, OpenIndiana or Solaris servers or you can virtualise them on ESXi.

7. Hardware example for vCluster on ESXi

My vCluster testenvironment (I use the SFP+ version of the mainboard):
https://www.supermicro.com/products/system/2U/5029/SSG-5029P-E1CTR12L.cfm

This is a SuperMicro barebone with 10G and onboard SAS HBA. You only need to add a CPU, RAM and disks.
I added an Intel Xeon Silver, 32GB RAM, a 120 GB boot SSD for ESXi (Intel DC3510) and an Intel Optane 900P that I use as an ESXi datastore for the VMs and a 20GB vmdk for Slog and another one for L2Arc.

You can also use such a vCluster to evaluate a hardware based SAS Cluster setup. You need two LSI HBA. Use each of them in pass-through mode to either head-1 and head-2. Then use a Disk case with dual SAS and SAS disks for example https://www.raidsonic.de/products/internal_cases/backplanes/index_en.php?we_objectID=1151

This is a SAS backplane for 4 x 3,5" SAS disks with multipath capability. You can use two 4 x Sata to mini SAS cables to connect the disks simulataniously to the HBAs on head-1 and head-2.

7.1 Hardware examples for physical Cluster

For a ZFS Cluster that is build on dedicated hardware you need

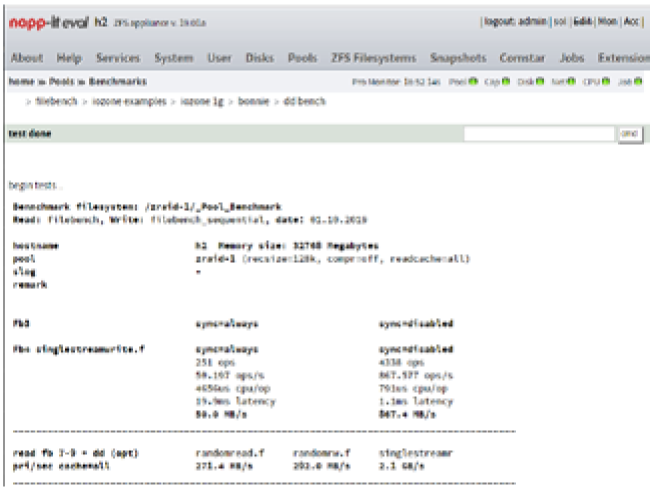
- 1 x server for Head-1 with two external SAS ports. If you want auto-mode, the server should allow remote reset (Stonith) either via ESXi (Twin vCluster) as base or via ipmitools on a barebone system
example: ipmitool -I lanplus -H 172.19.10.21 -U ADMIN -P ADMIN chassis power cycle

- 1 x server for Head-2 with two external SAS ports. If you want auto-mode, the server should allow remote reset (Stonith) either via ESXi (Twin vCluster) as base or via ipmitools on a barebone system
example: ipmitool -I lanplus -H 172.19.10.21 -U ADMIN -P ADMIN chassis power cycle

- 1 x Storage Jbod box with external SAS connectors, best in a dual expander configuration example
https://www.supermicro.com/products/chassis/2U/826/SC826BE2C-R741JBOD

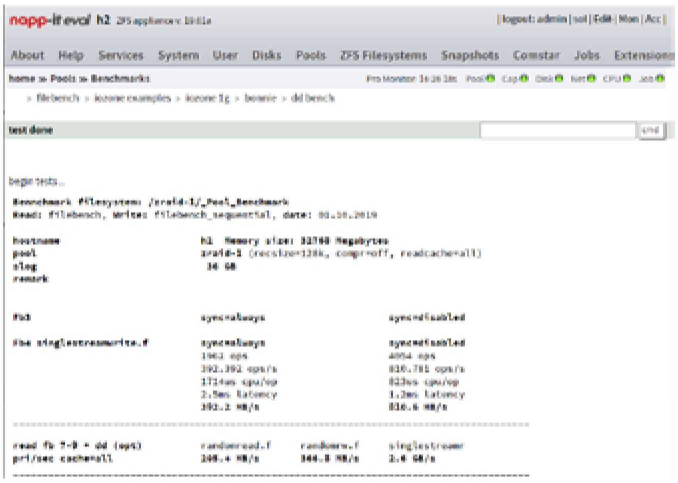
7.2 High performance HA storage/ Slog

You can build a high capacity/ high performance HA storage using dualpath SAS disks that is very fast sequentially and with enough RAM for read/write caching also fast on random loads but is worse on secure sync write.


napp-it eval h2
|logout: admin | sol | Edit | Mon | Acc |
About Help Services System User Disks Pools ZFS Filesystems Snapshots Comstar Jobs Extension
home » Pools » Benchmarks
» filebench » iozone examples » iozone 1g » bonnie » dd bench
test done
cmd

```
begin tests ..

Benchmark filesystem: /sraid-1/_Pool_Benchmark
Read: filebench, Write: filebench_sequential, date: 01.10.2019

hostname                      h2  Memory size: 32768 Megabytes
pool                          sraid-1 (recsize=128k, compr=off, readcache=all)
slog                          -
remark


Fb3                           sync=always                  sync=disabled

Fb4 singlestreamwrite.f       sync=always                  sync=disabled
                              251 ops                      4338 ops
                              50.197 ops/s                 867.577 ops/s
                              4656us cpu/op                791us cpu/op
                              19.9ms latency               1.1ms latency
                              50.0 MB/s                    867.4 MB/s
______________________________________________________________________________

read fb 7-9 + dd (opt)        randomread.f    randomrw.f   singlestreamr
pri/sec cache=all             271.4 MB/s      202.0 MB/s   2.1 GB/s
______________________________________________________________________________
```

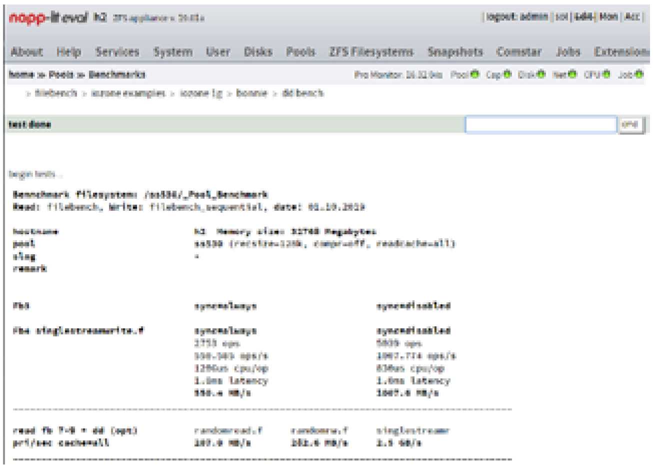
3 disks in a Raid-0 without Slog
Async write: >800 MB/s, sync write: 50 MB/s


napp-it eval h2
|logout: admin | sol | Edit | Mon | Acc |
About Help Services System User Disks Pools ZFS Filesystems Snapshots Comstar Jobs Extensions
home » Pools » Benchmarks
» filebench » iozone examples » iozone 1g » bonnie » dd bench
test done
cmd

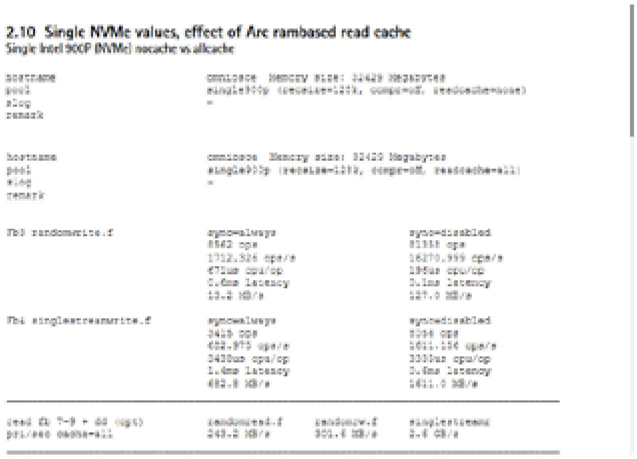
```
begin tests ..

Benchmark filesystem: /sraid-1/_Pool_Benchmark
Read: filebench, Write: filebench_sequential, date: 01.10.2019

hostname                      h1  Memory size: 32768 Megabytes
pool                          sraid-1 (recsize=128k, compr=off, readcache=all)
slog                           16 GB
remark


Fb3                           sync=always                  sync=disabled

Fb4 singlestreamwrite.f       sync=always                  sync=disabled
                              1962 ops                     4054 ops
                              392.392 ops/s                810.781 ops/s
                              1714us cpu/op                823us cpu/op
                              2.5ms latency                1.2ms latency
                              392.2 MB/s                   810.6 MB/s
______________________________________________________________________________

read fb 7-9 + dd (opt)        randomread.f    randomrw.f   singlestreamr
pri/sec cache=all             249.4 MB/s      344.8 MB/s   2.4 GB/s
______________________________________________________________________________
```

same 3 disks in a Raid-0 with an Slog (SS530)
Async Write > 800 MB/s, sync write nearly 400 MB/s

If you need HA storage with a very high random sync or async write performance, you can build an array from dualpath 12G SAS SSDs WD Ultrastar SS530. This is one of the fastest dualpath enterprise SAS SSDs, see datasheet. These disks are available with a different random write performance/ write endurance, as a 400GB model (use it as an Slog) or up to 15 TB for regular storage. With these SAS SSDs (3DW or 10 DW) you can get around 70-80% of the Intel Optane regarding sync write (currently the best Slog). As Optane is not available as dualpath SAS, the WD SS530 400 GB is a perfect Slog. Use a 12G HBA as on a 6G HBA they are around 20% slower.


napp-it eval h2
|logout: admin | sol | Edit | Mon | Acc |
About Help Services System User Disks Pools ZFS Filesystems Snapshots Comstar Jobs Extension
home » Pools » Benchmarks
» filebench » iozone examples » iozone 1g » bonnie » dd bench
test done
cmd

```
begin tests ..

Benchmark filesystem: /ss530/_Pool_Benchmark
Read: filebench, Write: filebench_sequential, date: 01.10.2019

hostname                      h2  Memory size: 32768 Megabytes
pool                          ss530 (recsize=128k, compr=off, readcache=all)
slog                          -
remark


Fb3                           sync=always                  sync=disabled

Fb4 singlestreamwrite.f       sync=always                  sync=disabled
                              2753 ops                     5039 ops
                              550.589 ops/s                1007.774 ops/s
                              1290us cpu/op                830us cpu/op
                              1.8ms latency                1.0ms latency
                              550.4 MB/s                   1007.6 MB/s
______________________________________________________________________________

read fb 7-9 + dd (opt)        randomread.f    randomrw.f   singlestreamr
pri/sec cache=all             187.9 MB/s      262.6 MB/s   2.5 GB/s
______________________________________________________________________________
```



2.10 Single NVMe values, effect of Arc rambased read cache
Single Intel 900P (NVMe) nocache vs allcache

```
hostname                      omniosce  Memory size: 32429 Megabytes
pool                          single900p (recsize=128k, compr=off, readcache=none)
slog                          -
remark


hostname                      omniosce  Memory size: 32429 Megabytes
pool                          single900p (recsize=128k, compr=off, readcache=all)
slog                          -
remark


Fb3 randomwrite.f             sync=always                  sync=disabled
                              8562 ops                     81355 ops
                              1712.326 ops/s               16270.999 ops/s
                              671us cpu/op                 190us cpu/op
                              0.6ms latency                0.1ms latency
                              13.2 MB/s                    127.0 MB/s

Fb4 singlestreamwrite.f       sync=always                  sync=disabled
                              3415 ops                     8058 ops
                              682.973 ops/s                1611.134 ops/s
                              2430us cpu/op                3331us cpu/op
                              1.4ms latency                0.6ms latency
                              682.8 MB/s                   1611.0 MB/s
______________________________________________________________________________

read fb 7-9 + dd (opt)        randomread.f    randomrw.f   singlestreamr
pri/sec cache=all             243.2 MB/s      301.4 MB/s   2.4 GB/s
______________________________________________________________________________
```

A pool from a single DC SS 530 (3DW, 400 GB)
Async write: 1007 MB/s, sync write: 550 MB/s

napp-it AiO under ESXI
with disks in pass-through mode

A pool drom a single Intel Optane 900 NVMe
Async Write 1611 MB/s, sync write 680 MB/s

Optane in a barebone setup (there are problems using Optane as pass.through device

Fazit:
WD Ultrastar DC SS 530/540 or Seagate Nytro is a perfect alternative for Optane when you need dualpath/ HA

Connect one SAS port from expander-1 to head-1 and the other from expander-2 to Head-2
As an extension you can use two Jbod server. Connect each to Head-1 and Head-2.
You can then build a pool ex from 1x1 or 2x2 mirrors that allow an outage of a whole Jbod

For high performance or an Slog use SAS Enterprise SSD, ex
WD Ultrastar DC SS 530

- 1 x Cluster Control server. There is no special hardware needed but you can combine the Cluster-Control functionality with a backups system. This is a suggested setup.

optionally: Second backup system on a different physical location/ building

btw
ipmitools are inncluded in OmniOS, install via pkg install ipmitool

8. Todo and checklist

to build an evaluation or production ready and cost sensitive ZFS vCluster for 4 or 8 SAS disks
Please follow the checklist step by step and confirm every point

8.1 Lab hardware example

Step	Needed/ Action	min costs	confirm	remark
Server hardware, ex https://napp-it.org/doc/downloads/napp-it_build_examples.pdf				
1	1 x Mainboard, vt-d, 10G, SAS, 32 GB RAM ex SuperMicro X11SSH-CTF and a G44/Celeron with Sata or M.2 bootdisk 240GB min	800 Euro/$ 100 Euro/$		
2	1 x second LSI HBA ex LSI 9207, LSI 9003	200 Euro/$		
3	1 x Computer case Silverstone CS380, PSU or ICY Dock IB 554SSK SAS enclosure	200 Euro/$		
Storage Hardware				
4	1 x ICY Dock IB 554SSK (Dualpath 4 x 3,5") optionally Icy Box IB-2222 SSK (Dualpath 4 x 2,5")	opt		for 8 disks use two enclosures
5	n x SAS disks ex HGST Ultrastar HE or SAS SSDs	--		
6	optional SAS Slog/L2Arc ex WD Ultrastar SS 530	opt		

8.2 Get needed software

Step	Needed/ Action	min costs	confirm	remark
7	ESXi 6.7u1 download iso and create an USB stick (Rufus)	free		
8	napp-it + OmniOS 151028 ZFS server template download newest from napp-it.org	free		
9	napp-it free napp-it Pro complete (cluster control)	free		manual failover depends
	min costs for a hardware without disks	**1300 Euro/$**		
	For a ready to use managed cluster add SAS disks and napp-it Pro complete			**at least required on Cluster-Control**

8.3 Basic Setup Cluster AiO (my All-In-One configuration)

Step	Needed/ Action		confirm	remark
10	2 x MPIO SAS cabling of the Icy Dock Case (green ports to HBA1 and blue ports to HBA2) insert your SAS disks into the enclosure	SAS-> 4 x Sata adapter cables		
11	Boot ESXi bootstick (from step 7) and install ESXi onto the Sata or M.2 bootdrive			
12	Connect your Browser to the ESXi management interface (ip see console at step 11)			
13	Configure ESXi to allow pass-through for your HBAs ESXi menu Host > Manage, reboot ESXi			

8.4 switch back from Cluster mode to standalone mode

Step	Needed/ Action	confirm	remark
21	head (1/2) with zraid-1 imported: Switch to standalone mode in menu About settings.		

A failover from control is only possible from/to a node in cluster-failover mode

Attention: in Standalone mode it is possible to import the zraid pools on both heads
Care about: This can corrupt a pool !!

8.5 Backup and replication

If you want to replicate a HA pool like zraid-1 (prefer a replication per filesystem) you need a backupserver from whom you must create an appliance group to the Cluster (both heads and the HA ip).

On a failover, the HP ip switches to the other head h1. To continue the replication from this head, it must behave like the former head h1 with same group key.. This requires that you use the HA ip as replication source. From the view of the backupserver you can access both heads in ha mode. For any other use case you can use the hostnames h1 or h2.

9. FAQ and remaining questions

Q: When should I use local disks (either SAS multipath or shared ESXi Sata,SAS,NVMe) and when iSCSI?
A: Local disks are always faster. Use iSCSI mainly if you want a realtime net-mirror with a remote location.
For high performance, think of a local multipath SAS Cluster with async replication to the external location.

Q: When should I use a vCluster and when a barebone Cluster?
A: When a vCuster is not fast enough, use barebone systems

Q: When should I use a single Cluster and when a Twin Cluster?
A single Cluster operates with shared access to a single storage system/ Jbod/ ESXi disks. It allows a failure or maintenance of a head. If the storage box fails (example due a PSU failure), your system is offline.

A Twin Cluster operates with two independent storage systems, either two Jbod cases, each with a twin expander in a crossover setup (both heads can access both Jbod ex with 4 way mirrors and 2 x 2 disks on each) . Use this setup if your Cluster must allow a head and/or a complete storage box failure.

Q: Do I need a Windows AD server for HA SMB filer use?
A: No, you can use a filer with local users. Napp-it can failover local users with permissions intact.

Q: What distances are allowed between head and storage?
A: With SAS your can use cables up to 10m. With iSCSI there is not limit.

Q: I need a secure write behaviour. Where do I need the Slog
A: All ZFS writes are going over the rambased write cache. On a crash the content is lost. In a Cluster situation you need sync write with an Slog that is part of the failover pool. In a vCluster you can use shared disk access to an NVMe Slog like the Intel Optane. In a Twin Cluster SAS setup, you need a Slog mirror with an Slog on each Jbod what means you need a SAS Slog ex a ZeusRAM or 12G Enterprise SAS SSD with high 4k write iops ex
WD Ultrastar DC SS530, see
https://www.anandtech.com/show/13149/hgst-launches-ultrastar-dc-ss530-ssds-3d-tlc-nand-1536-tb-sas-12-gbps
For iSCSI based storage where you share a whole pool as a LUN, you need an slog with sync write for the local pool to protect local writes and an Slog for your storage head to protect client writes.

Q: Can I share RAM between VMs in a vCluster?
A: Yes you can with a single HBA assigned to ESXi. This can reduce needed overall RAM.

If you use a multipath SAS setup with two HBA, one for each head in pass-through mode you can't.
Hardware pass-through requires that RAM is rederved to a VM. Calculate enough RAM for each head (12GB min).

Q: Suggest me a basic HA config that is fast, cheap and reliable with one common storage.
A: Use vCluster with shared SAS disks, one HBA under ESXi control or a HBA per head in pass-through mode with each head connected to one SAS datapath.

Q: Suggest me an HA config that is fast and reliable and allows a full Cluster, head and Jbod outage.
A: Use TwinCluster (virtual or physical) with a HBA in pass-through mode on each head. and two Jbod cases with dual expander. Connect the two Jbod cases with head-1 and head-2 and create a redundand pool ex from 4way mirrors where 2 disks are on Jbod1 and two disks are on Jbod2. Even on a complete Jbod failure a regular mirror remains intact.

Q: Must I name head-1 and head-2 different or can I give them same hostname
A: Both can have reasons. For replications and different names like h1 and h2, copy the appliance group key ex for a backupserver from /var/web-gui/_log/groups on h1 to the same location on h2. Add an DNS entry then for the HA ip and a desired DNS hostname like filer.

Q: How do you support failover iSCSI via Comstar (require napp-it 19.06 Pro)
In napp-it 19.06 Pro you find a new menu Full save/clear/restore under the Comstar menu. This allows to save, clear and restore all Comstar settings without a reboot. In the failover settings you find a corresponding entry to use this mechanism for a manual ot auto failover.

Q: My stonith command does not run
If you start a script or programm from napp-it, it runs under the napp-it account. If you need root permissions, start the programm via ‚sudo program' or ‚sudo bash script.sh'. As sudo use iths own path you may need to call ‚sudo /path/program'

Q: I need more than one HA ip.
Use the failover pre/post-script in failover settings

To add a new ip (post) ip:3

```
sudo ifconfig vmxnet3s0:3 plumb
sudo ifconfig vmxnet3s0:3 172.19.30.76 netmask 255.255.0.0 up
```

To unplumb this ip:3 (pre):

```
sudo ifconfig vmxnet3s0:3 unplumb
```

9.1 Appliance Z-RAID with SSF (Storage and Service Failover)

Z-RAID vs RAID-Z

ZFS and Raid-Z is a perfect solution when you want data protection against bitrot with end to end data checksums and a crash resistent CopyOnWrite filesystem with snaps and versioning. ZFS Raid-Z protects against disk failures. ZFS replication adds a ultrafast method for async backups even when files are open.

ZFS and Z-RAID is the next step as it adds realtime sync between independent Storage Appliances where ZFS protects against a whole Appliance failure. It also adds Availabilty for NFS and SMB services as it allows a full Storage Server failure with a manual or automatic Pool and NFS/SMB service failover with a shared virtual ip.

RAID-Z

Traditionally you build a ZFS pool from RAID-Z or Raid-1 vdevs from disks. To be protected against a disaster like a fire or flash, you do backups and snaps for daily access of previous versions to access deleted or modified files. In case of a disaster, you can restore data and re-establish services based the last backup state.
Main Problem: there is a delay between your last data state and the backup state. You can reduce the gap with ZFS Async Replication but the problem remains that backup is never up to date. An additional critical point are open files. As ZFS Replication is based on snaps, the last state of a replication is like a sudden poweroff what means that files (or VMs in a virtualisation environment) may be in a corrupted state on the backup.

Another problem is time to re-establish services like NFS or SMB on a server crash. If you need to be back online in a short time, you use a second backup system that is able and prepared to takeover services based on the last backup state. As on Solarish systems, NFS and SMB are integrated in the OS/Kernel/ZFS with the Windows security identifier SID as an extended ZFS attribute (Solarish SMB), this is really troublefree. Even in a Windows AD environment, you only need to import a pool, takeover the ip of the former server, and your clients can access their data with all AD permission settings intact without any additional settings to care about.

Z-RAID (network raid)

A Unix sytsem with ZFS requires less maintenance and can have a long uptime. But from time to time you must update the system software or do maintenance. If you cannot allow a service outage, you need to failover services to a second system with the current data state intact. A second standby system with replicated or based on a backup is not enough. This is the use case of a napp-it ZFS Cluster that allows a failover to a second node either in a vCluster in a Box or in a Twin Cluster. (Full redundant system incl. Cluster Control and Storage failover)

9.2 Differences to other Failover solutions

like PaceMaker + Corosync/Hearbeat or RSF-1

A typical OpenSource Cluster solution on Linux is build on PaceMaker as Cluster Resource Manager and Corosync or Heartbeat as a Cluster Communication Manager. This is a very flexible concept as PaceMaker allows to control any sort of services and resources on any node with a lot of possible modifications, scripts and settings to control the services that are involved in the failover process. While you can install PaceMaker on Solaris it is a quite complex solution and requires either a sercice contract with SLA or an IT department to support such a solution inhouse. On ZFS, RSF-1 is a typical commercial Cluster solution with support.

Z-RAID SSF vClusters are different and the approch is „as simple and cheap as possible" and „as powerfull as needed". It reduces the Cluster Resources that are controlled for a failover to „ZFS Pools with NFS and SMB services". Other services can be added via pre or post failover scripts. On Solarish you have the unique situation these are pure ZFS resources as SMB and NFS services are fully integrated into the OS/Kernel and ZFS so you do not need a dedicated Cluster Resource Manager for them. For SMB you have the additional advantage that Windows SID in an AD environment are used in Solarish ZFS. A Pool failover in a Z-Raid Cluster between two heads that are both AD members or with local users allows a full service failover where all Windows permissions stay intact without any extra efforts or problems. This allows a high performance/ high available but affordable filer that is easy to setup and maintain.

Solarish ZFS itself is the best of all Resource Manager as a Z-Pool failover (export/import) between the heads automatically controls and enables these services from pool and ZFS properties itself. You only need to add failover management and Cluster Communication that are provided by napp-it. A Cluster based on the napp-it Z-Raid SSF concept with two heads (Master and Slave), different disk options and a dedicated control instance reduce complexity massively.

The napp-it Cluster solutions offers Cluster communication and failover management.
It gives you full control about shared disk and configuration options and covers use cases from a simple „vCluster in a Box" with Sata disks to realtime location mirrorring = realtime backup.

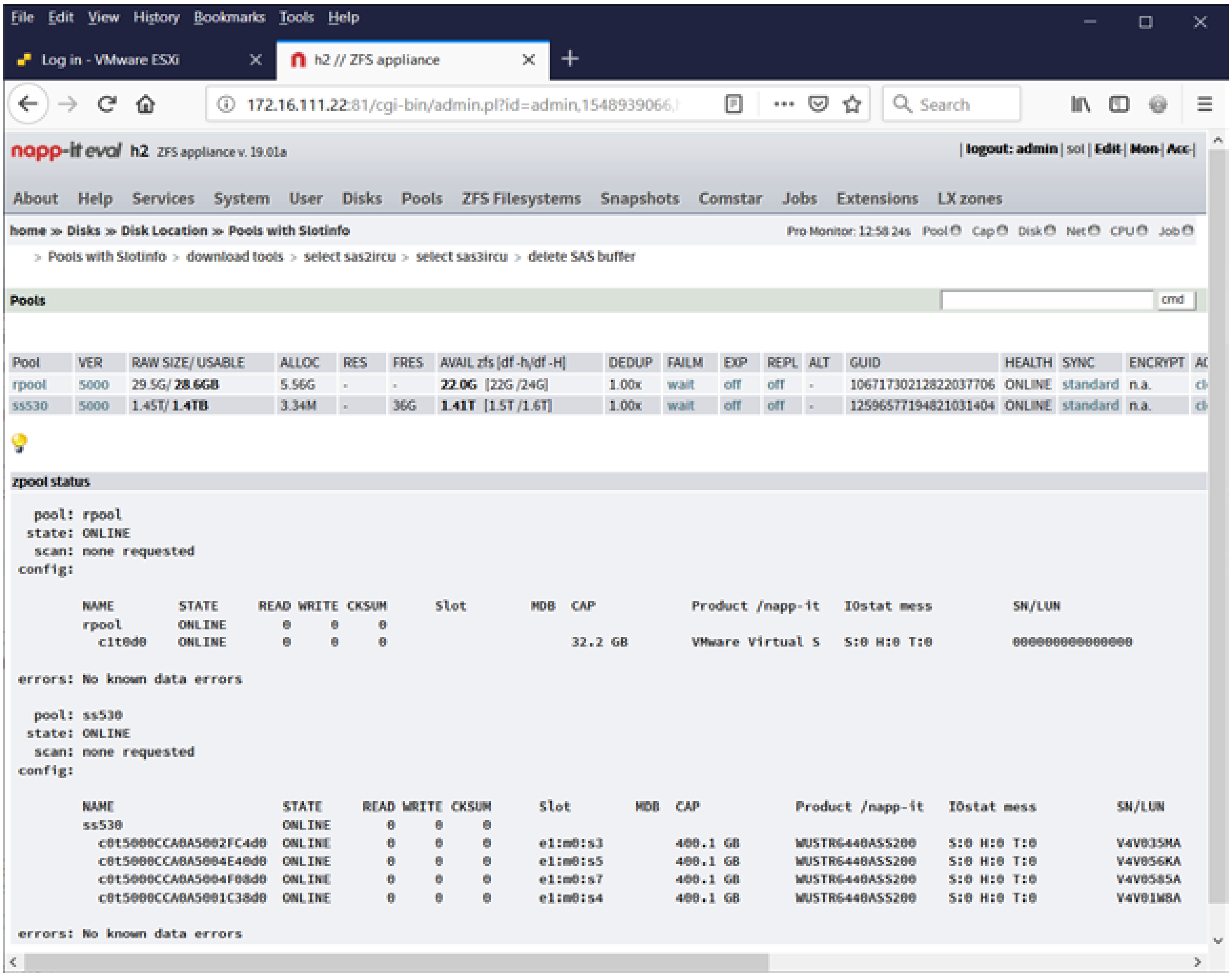
9.2 WD Ultrastart SS 530 performance (Raid-0, one to four SSD)

High performance dualpath 12G SAS SSD (up to 16 TB, over 300k iops, powerloss protection)

```
Benchmark: Write: filebench_sequential, Read: filebench, date: 01.31.2019

host                            h2
pool                            ss530 (recsize=128k, compr=off, readcache=all)
slog                            -
remark                          one SS 530-400 3dpwd basic


Fb3                             sync=always                     sync=disabled

Fb4 singlestreamwrite.f        sync=always                     sync=disabled
                                1854 ops                        4526 ops
                                370.477 ops/s                   905.133 ops/s
                                3000us cpu/op                   1152us cpu/op
                                2.7ms latency                   1.1ms latency
                                370.3 MB/s                      904.9 MB/s

read fb 7-9 + dd (opt)          randomread.f     randomrw.f     singlestreamr
pri/sec cache=all               211.8 MB/s       316.9 MB/s     2.6 GB/s


host                            h2
pool                            ss530 (recsize=128k, compr=off, readcache=all)
slog                            -
remark                          two SS 530-400 3dpwd in Raid-0


Fb3                             sync=always                     sync=disabled

Fb4 singlestreamwrite.f        sync=always                     sync=disabled
                                2088 ops                        7320 ops
                                417.092 ops/s                   1461.970 ops/s
                                3482us cpu/op                   1041us cpu/op
                                2.4ms latency                   0.7ms latency
                                416.9 MB/s                      1461.8 MB/s

read fb 7-9 + dd (opt)          randomread.f     randomrw.f     singlestreamr
pri/sec cache=all               210.4 MB/s       166.8 MB/s     2.6 GB/s


slog                            -
remark                          three SS 530-400 3dpwd in Raid-0


Fb3                             sync=always                     sync=disabled

Fb4 singlestreamwrite.f        sync=always                     sync=disabled
                                3350 ops                        8641 ops
                                669.985 ops/s                   1727.906 ops/s
                                1110us cpu/op                   1081us cpu/op
                                1.5ms latency                   0.6ms latency
                                669.8 MB/s                      1727.7 MB/s

read fb 7-9 + dd (opt)          randomread.f     randomrw.f     singlestreamr
pri/sec cache=all               258.0 MB/s       200.2 MB/s     2.1 GB/s


slog                            -
remark                         four SS 530-400 3dpwd in Raid-0
```

```
Fb3                             sync=always                     sync=disabled

Fb4 singlestreamwrite.f        sync=always                     sync=disabled
                                3467 ops                        10284 ops
                                693.313 ops/s                   2056.568 ops/s
                                1109us cpu/op                   931us cpu/op
                                1.4ms latency                   0.5ms latency
                                693.1 MB/s                      2056.4 MB/s
_________________________________________________________________________________________

read fb 7-9 + dd (opt)          randomread.f     randomrw.f     singlestreamr
pri/sec cache=all               213.8 MB/s       169.4 MB/s     2.6 GB/s
_________________________________________________________________________________________
```

Result: This is an virtualized environment. A pool from these SSDs offer excellent performance on read, write and sync write. With their dualpath SAS ports they are perfect for a high performance filer/cluster.


File Edit View History Bookmarks Tools Help
Log in - VMware ESXi
h2 // ZFS appliance
172.16.111.22:81/cgi-bin/admin.pl?id=admin,1548939066,
Search
napp-it eval h2 ZFS appliance v. 19.01a
| logout: admin | sol | Edit | Mon | Acc |
About Help Services System User Disks Pools ZFS Filesystems Snapshots Comstar Jobs Extensions LX zones
home » Disks » Disk Location » Pools with Slotinfo
Pro Monitor: 12:58 24s Pool Cap Disk Net CPU Job
> Pools with Slotinfo > download tools > select sas2ircu > select sas3ircu > delete SAS buffer
Pools
cmd

Pool	VER	RAW SIZE/ USABLE	ALLOC	RES	FRES	AVAIL zfs [df -h/df -H]	DEDUP	FAILM	EXP	REPL	ALT	GUID	HEALTH	SYNC	ENCRYPT	AC
rpool	5000	29.5G/ **28.6GB**	5.56G	-	-	**22.0G** [22G /24G]	1.00x	wait	off	off	-	10671730212822037706	ONLINE	standard	n.a.	cl
ss530	5000	1.45T/ **1.4TB**	3.34M	-	36G	**1.41T** [1.5T /1.6T]	1.00x	wait	off	off	-	12596577194821031404	ONLINE	standard	n.a.	cl

zpool status

```
  pool: rpool
 state: ONLINE
  scan: none requested
config:

        NAME        STATE     READ WRITE CKSUM      Slot        MDB  CAP            Product /napp-it   IOstat mess          SN/LUN
        rpool       ONLINE       0     0     0
          c1t0d0    ONLINE       0     0     0                       32.2 GB        VMware Virtual S   S:0 H:0 T:0          0000000000000000

errors: No known data errors

  pool: ss530
 state: ONLINE
  scan: none requested
config:

        NAME                       STATE     READ WRITE CKSUM      Slot        MDB  CAP            Product /napp-it   IOstat mess          SN/LUN
        ss530                      ONLINE       0     0     0
          c0t5000CCA0A5002FC4d0   ONLINE       0     0     0      e1:m0:s3         400.1 GB       WUSTR6440ASS200    S:0 H:0 T:0          V4V035MA
          c0t5000CCA0A5004E40d0   ONLINE       0     0     0      e1:m0:s5         400.1 GB       WUSTR6440ASS200    S:0 H:0 T:0          V4V056KA
          c0t5000CCA0A5004F08d0   ONLINE       0     0     0      e1:m0:s7         400.1 GB       WUSTR6440ASS200    S:0 H:0 T:0          V4V0585A
          c0t5000CCA0A5001C38d0   ONLINE       0     0     0      e1:m0:s4         400.1 GB       WUSTR6440ASS200    S:0 H:0 T:0          V4V01W8A

errors: No known data errors
```

Pool Layout with four SS 530

Care about

more docs

napp-it Homepage:
http://www.napp-it.org

Feature sheet
http://www.napp-it.org/doc/downloads/featuresheet.pdf

Build examples
https://www.napp-it.org/doc/downloads/napp-it_build_examples.pdf

How to setup the napp-it ZFS storage server
http://www.napp-it.org/doc/downloads/napp-it.pdf

How to setup napp-in-one (virtualized storage server on ESXi)
http://www.napp-it.org/doc/downloads/napp-in-one.pdf

Intel Optane performance
https://www.napp-it.org/doc/downloads/optane_slog_pool_performane.pdf

Performancetuning with 10G and SMB2
http://napp-it.org/doc/downloads/performance_smb2.pdf

Download napp-it ToGo (ready to use images for a barebone setup or ESXi template)
http://napp-it.org/downloads/index_en.html

Howto setup OmniOS manually
http://napp-it.org/downloads/omnios_en.html