napp-it

napp-*it*

ZFS Storage Server
Build examples

Content:

1. Checklist

Basically you can use any desktop or serversystem as a webmanaged ZFS NAS appliance.
The minimal hardware demand for a 64 bit Solarish like OmniOS, OpenIndiana or Oracle Solaris is

- 2 GB RAM (ECC or non-ECC) for any Poolsize - but you should not enable dedup with low RAM
  + RAM for wanted read/write cache (2-128GB. prefer ECC)
- a 64 bit CPU (AMD or Intel). If you can decide, prefer frequency over number of cores.
- Sata (AHCI mode)
- a nic (Realtek but prefer Intel or Chelsio)
- a 100GB bootdisk
- disks for a datapool, should be at least two disks for a mirror

These are the minimal demands. If you care about your data and want a better performance than pure disk performance, then you should add

- ECC RAM. This is important as a RAM failure is the only type of failures ZFS cannot protect against.
But even without ECC, ZFS is more secure than older filesystems without ECC as it protects against all problems in the chain controller <-> disk, powerloss problems and undo of unwanted modifications due snaps.
If you have the option, use always ECC for any storage system. From hardware this requires ECC support from CPU and mainboard. Use an Intel mainboard with a serverchipset like C232 or C236 and a ECC capable CPU up from an Intel G4400 (socket 1151) or a Xeon (any socket).

- Add more RAM. While Solarish works with 2 GB, more RAM is used as a readcache that can dramatically improve readperformance. With more RAM you can achieve that > 80% of all reads are delivered from RAM.

- Prefer solutions with 10G Intel nics as they have the best driver support and performance.
You can enable bridging on your server. Your NAS acts then like a 10G switch over your 1G/10G nics.

- Add an HBA with an LSI chipset, preferred with a raidless IT firmware (ex LSI 9207) if you need either more ports or if you want to use it in an All-In-One setup where you can virtualize storage beside other VMs.

Such systems can be ordered quite cheap if you check offers from Dell example the PowerEdge T20 or T130 or HP systems like the HP G8 Microserver or the ML10 Gen9. Main problem of these systems are a limited expandability regarding PCI-e slots, disk bays or available CPU options.

If you want to overcome these limitations, you can built your own setup or buy built to order systems that are assembled by a distributor or vendor as a ready to use system.

I will concentrate on mainboards from SuperMicro as this is the number one option for ZFS storage with a complete storage product line with the following use cases

1.
Silent Solutions with up to 8 x 3,5" diskbays, 8 x 2,5" backplane and 2 x 2,5" bootdisks and optionally some PCI-e NVMe disks. Such a system can be placed near the desktop in an office or team environment

2.
19" cases for expanderless SAS/Sata storage with up to 24 disks
With 10TB Sata disks you can go up to over 200 TB raw storage or with a 24 x 2,5" case you can build high performance SSD only storage with up to 90 TB when using the 3,8 TB Samsung SSD.

3.
19" cases for SAS solutions. With a 90 x 3,5" bay SuperMicro case and 8 TB SAS disks, you can achieve around 500 GB per case. For a Petabyte you need two cases. There are also options with 72 x 2,5" SAS disks.

With an expander you can use Sata disks as SAS is tunnelling the Sata protocol but you should prefer SAS3 in a production environment. Professional storage vendors deny support for such a config.

2. Silent/ low budget cases

2.1 Silverstone CS 380

http://www.silverstonetek.com/product.php?area=en&pid=709

Case suitable for ITX, uATX and ATX mainboards
with 3 x 120mm fan and 8 x Sata/ SAS dualpath hotplug backplane

possible Addons:
2,5" Backplane (8 x Sata SSD in 2 x 5,25" Slots)
ex SuperMicro Sata Mobile Rack M14TQC (or other Sata/SAS mobile racks)
http://www.supermicro.com/products/accessories/mobilerack/CSE-M14TQC.cfm

Use case:
SoHo low power/ low noise system with up to 16 hotplug disks ex for a fast ZFS pool
from up to 8 SSDs (VM datastore) and a second 8 disks based pool for filer and backup
up to low budget HA systems with two virtualized storage heads under ESXi
each with an SAS HBA in pass-through mode to dualpath SAS disks)

Mainboard:
For a low budget NAS (sub 500 $/Euro) with socket 1151, G44xx or G45xx CPU with a small M.2
for OS + 8 onboard Sata https://www.supermicro.nl/products/motherboard/Xeon/C236_C232/X11SSH-F.cfm
also available with additional LSI SAS HBA for AiO systems with ESXi + local datastore + Slog with an Intel M.2 Optane 800P and optionally10G onboard.

A perfect mainboard would be one of the SuperMicro X10 SDV line
with 10G networking and LSI HBA with16 x SAS/Sata onboard (sub 1000 $/Euro system)
ex https://www.supermicro.nl/products/motherboard/Xeon/D/X10SDV-2C-7TP4F.cfm
or with more powerful CPUs, https://www.supermicro.nl/products/motherboard/Xeon3000/#1667

2.1b A silent case: Fractal Design Define R5 (black, white, titanium)



From vendor specifications

„The Define R5 case reaches the highest level of silent computing through strategically placed dense sound-absorbing material, ModuVent™ fan vent covers and finely tuned Dynamic Series fans.

- ATX, Micro ATX, Mini ITX motherboard compatibility
- 7 expansion slots
- 2 - 5.25" bays (removable)
- 8 - 3.5" HDD positions (can also accommodate 2.5" units); 2 - 2.5" dedicated SSD unit positions
- 4 - ModuVent™ plates - three in the top and one in the side
- 9 - fan positions (2 Fractal Design Dynamic GP14 140mm fans included)
- Filtered fan slots in the front and bottom
- CPU coolers up to 180mm in height
- ATX PSUs up to 190/170 mm with a bottom 120/140mm fan installed; when not using any bottom
- fan location longer PSUs up to 300mm can be used
- Graphics cards up to 310 mm in length with the top HDD cage installed; with the top cage removed, graphics cards up to 440 mm in length may be installed
- 20 - 35 mm of space for cable routing behind the motherboard plate
- Velcro straps included for easy cable management
- Front door can switch opening direction via dual mounting system
- Left side panel features Quick Release System for easy access and provides a secure closure of side panel
- Right side panel features smart captive thumbscrews so no thumbscrews are lost
- Colours available: Black, Titanium (black case, titanium front panel), White
- Case dimensions (WxHxD): 232 x 451 x 521mm
- Case dimensions - with feet/screws/protrusions: 232 x 462 x 531mm
- Net weight: 10.7 kg

possible Addons: 2,5" Backplane
SuperMicro Mobile Rack M28SACB-OEM

1. 8x 2.5" Hot-swap SAS3/SATA3 HDDs
2. Overheat LED and Alarm
3. Drive Activity / Failure LED
4. 2x 5.25" Drive Bay Enclosure
5. 2x Mini SAS HDD Connectors
6. Fan-less Subsystem

2.2 other small cases

U-NAS ex NSC 400/600/800 (mini-ITX)
www.u-nas.com

Fractal Design Node 804 (up to uATX)

Micro ATX and Mini ITX motherboard compatibility
8 - 3.5″ HDD positions
2 - 2.5″ dedicated SSD unit positions
2 - Extra positions for either 3,5″ or 2,5″ drives
5 expansion slots
1 additional space in the front for a 12.7mm slim/slimline ODD
10 - Fan positions (3 x 120mm Silent Series R2 fans included)
Filtered fan slots in front, top and bottom
CPU coolers up to 160 mm in height
PSU compatibility: ATX PSUs up to 260 mm deep
Graphics card compatibility: Graphics cards up to 320mm in length.
Graphics cards up to 290 mm in length may be installed if a fan is installed in the lower position in the front.
Velcro strap for easy cable management
Clear Window side panel included
Colors available: Black
Case dimensions (WxHxD): 344 x 307 x 389 mm
Net weight: 6 kg
Package dimensions (WxHxD): 370 x 468 x 412 mm
Package weight: 7.7kg

Silverstone CS381 (expected Q1 2019)
uATX, FlexATX, 8 x 3,5 hotswap + 2 x 3,5""
https://www.servethehome.com/silverstone-cs381-8-bay-matx-case-shown/

Silverstone CS 380 (available)
ATX, uATX, FlexATX, 8 hot swapable Dualpath SAS
(perfect solution for a HA Cluster in a Box)
https://www.silverstonetek.com/product.php?pid=709&area=en

2.3 Expanderless 19" cases with 12 to 36 bays for 3,5" Sata and SAS disks
Build to Order system

Alternate View

Rear View

Available Color

Black

Expanderless Options:

8 disks Supermicro, available but I would prefer 12 bay versions due same size.

12 disks SuperMicro
ex SuperChassis 826BA-R920LPB, 920 Watt PSU, 3 x miniSAS connector

16 disks SuperMicro
SuperChassis 836BA-R920B, 920 Watt PSU, 4 x miniSAS connector
SuperChassis 836A-R1200B, 1200 Watt PSU, 4 x miniSAS connector

24 disks SuperMicro
SuperChassis 846A-R900B, 900 Watt PSU, 6 x miniSAS connector
SuperChassis 846A-R920B, 920W Gold PSU, 6 x miniSAS connector
SuperChassis 846A-R1200B, 1200 Watt PSU, 6 x miniSAS connector

36 disks SuperMicro
Superchassis SC847BA-R1K28LPB, 1280W with a backplane without expander (iPASS connectors)

Attention: These 19" cases are intended to use in a serverrom.
For office/ near desktop use, these systems are way too loud

2.4 Expanderless 19" cases with 24 bays for 2,5" Sata and SAS disks
Build to Order system

Key Features

1. 920W Redundant High-efficiency **Platinum Level Power Supply**
2. 2U High Density 2.5" HDD Chassis 24x 2.5" Hot-swap SAS/SATA HD bays
3. Mini-I-pass (SFF 8087) Connectivity for clean cable routing
4. 7x Low-profile Expansion Slot
5. Optional rear hot-plug 2x 2.5" HDD drive bays

24 x 2,5" expanderless:
SuperMicro SuperChassis 216BA-R920LPB, 1200W PSU, 6 x miniSAS, optional 2 x Sata bootdisk

The above Case can be combined with any mainboard and HBA.
It is intended for a serverroom as it is quite loud

2.4 SuperChassis 216BA-R920LPB barebone
Prebuild Barebone system, you only need to add CPUs and RAM

3x LSI 3008 IT Mode controller
Low Latency

Available Colors: Black

Integrated Board

Super X10DRH-iT

Key Features

1. Dual socket R3 (LGA 2011) supports Intel® Xeon® processor E5-2600 **v4**†/ v3 family; QPI up to 9.6GT/s
2. Up to **2TB**† ECC 3DS LRDIMM , up to DDR4- **2400**†MHz ; 16x DIMM slots
3. 1 PCI-E 3.0 x16, 6 PCI-E 3.0 x8 (slot 1-3 occupied by controllers)
4. Dual 10GBase-T LAN w/ Intel® X540
5. 24x 2.5" Hot-swap SAS3/SATA3 direct attached drive bays; 2x 2.5" optional hot-swap drive bays (rear)
6. SAS3 via 3x LSI 3008 controller; IT mode
7. Server remote management: IPMI 2.0 / KVM over LAN / Media over LAN
8. 3x 8cm hot-swap redundant PWM fans
9. 920W Redundant Power Supplies **Platinum Level (94%)**

2.5 Zstor Cube

Ready to use NAS/SAN system with napp-it preinstalled

http://zstor.de/en/zstor-gs-cube8-mini-cube-storage-server-en.html
http://zstor.de/de/zstor-gs-cube8-mini-cube-storage-server.html

ZSTOR® Open Storage	
Product	**Mini Cube 8x 3.5"/2.5" Bay**
Processor	One INTEL® Xeon® Processor E3-1230v6 3.5GHz, 4 cores and 8 threads
Memory	4x 16GB DDR4 UDIMM ECC 2400 MT/s total 64GB
Chipset	Intel® C236 chipet, Integrated Video Controller VGA, Integrated BMC with KVM
Host Bus Adapter	Onboard Broadcom / LSI 3008 SAS/SATA HBA
Connection	1x VGA, 2x USB 2.0 ports side, 1x USB 3.0 front, 2x USB 2.0 ports back 2x USB 3.0 ports back, 1x IPMI RJ-45 Ethernet Remote Management, 2x 1GbE RJ-45 Ethernet
Network	2x onboard 1 GbE INTEL I210 RJ-45 1x onboard IPMI RJ-45 incl. Remote Management with KVM optional 1x PCI-E 10GbE INTEL 82599ES Dual SFP+ optional 1x PCI-E 10 GbE INTEL X540 Dual RJ-45
PCI-E Slots	1x HH-HL x16 mech., x8 elec. 1x HH-HL x8 mech. x8 elec.
Disk Bay	8x 3,5"/2.5" HDD/SDD Hot-swap SATAIII 1-12TB per Disk , 1x 2,5" SSD internal for OS Boot up to 96TB on a small foot print, Mixture of HDDs and SSDs possible
Operatings System	OmniOS ZFS File System with napp-it GUI, other operating systems on request

2.6 Enclosures with Expander for up to 90 3,5" disks
Build to Order system

2.6.1 SuperChassis 847BE2C-R1K28LPB (12G)

Key Features

1. 36x (24 front + 12 rear) 3.5" hot-swap SAS/SATA drive bays supporting SAS3/2 or SATA3 HDDs with 12Gbps throughput
2. Optional rear 2.5" removable HDD
3. Redundant 1280W Platinum Level (1+1) power supplies with PMBus
4. 7x low-profile expansion slots; 7x 8cm (middle) hot-swap cooling fans and adjustable air shroud
5. E1C: Single SAS3 (12Gbps) expander backplane; E2C: Dual SAS3 (12Gbps) expanders backplane
6. Mini SAS HD (SFF 8643) connectivity on backplane

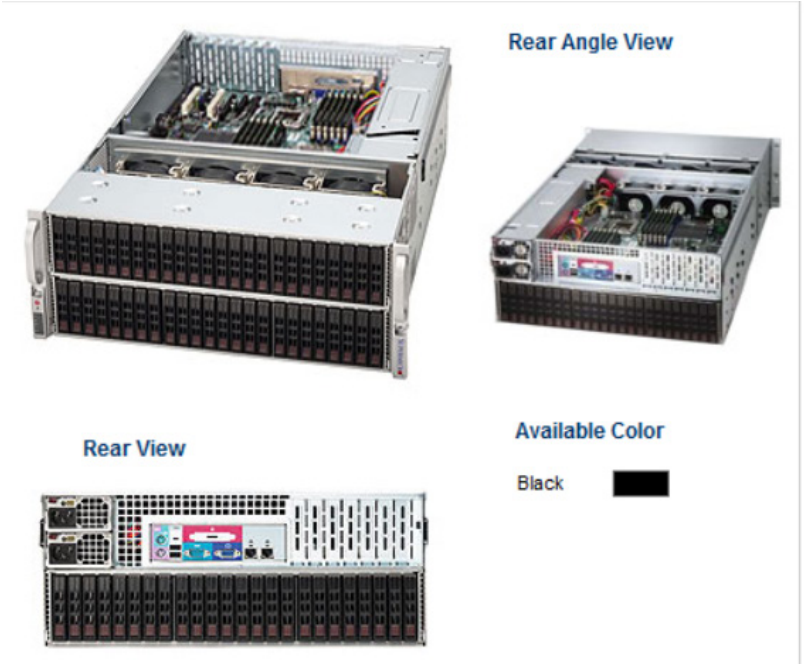
2.6.2 SuperChassis 946ED-R2KJBOD (90 disks Jbod, up to 500 TB per enclosure)

Key Features

1. Extreme High Density and High Capacity Dual-path Storage Enclosure
2. Support 90 x 3.5" or 2.5" Top Loading SAS3 12Gb/s Hot-swappable HDDs
3. Hot-swappable Expanders with redundant BMC for Remote System Power on/off and monitoring
4. Hot-swappable Tool-less Modular Design for Easy Service and Easy Maintenance
5. High Performance up to 20+ GB/s data transfer rate
6. Tool-less HDD tray with HDD LED indicator
7. SCSI Enclosure Services (SES 3.0) compliant
8. Flexible to configure up to 4 Hosts HDD Zoning and individual HDD power cycling
9. Slide Rails and Cable Management Arm included

2.7 Enclosures with Expander for up to 72 2,5" disks
Build to Order system

Supermicro SuperChassis 417E26-R1400LPB



Key Features

- High-Availability
- Extremely High Storage Capacity

1. Extra High-Density 4U Storage Chassis
2. High-Availability Features: Redundant, Hot-pluggable cooling system, Power Supplies, Hot-swap drives
3. Redundant (1+1) 1400W Gold Level power supply with PMBus function
4. 72x (48 front + 24 rear) 2.5" HDD bays E26: Dual Expander chips support SAS2 (6Gb/s)
5. 7x Low-profile expansion slots
6. 7x 8cm (middle) Hot-swap cooling fans, redundant cooling

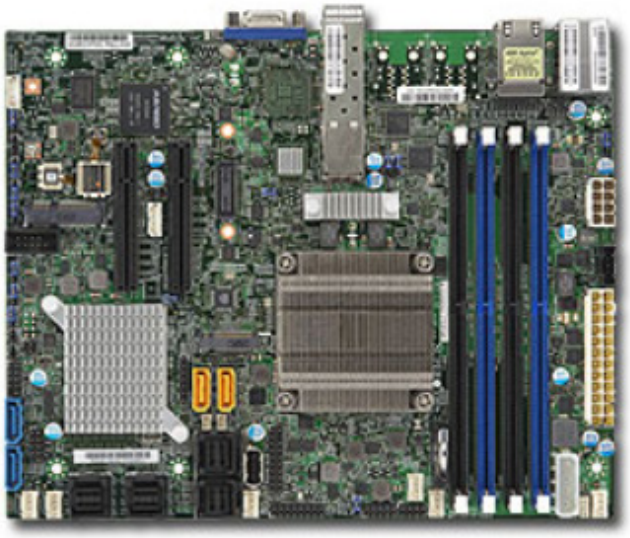
72 x 2,5" disks with expander

2.8 Variable enclosure Chenbro RM417
an ultra flexible case concept with different and selectable 3,5" or 2,5" backplanes with or without expander

Backplanes: 6Gb/s mini-SAS expanderless or mini-SAS or with expander on-board
Disks: up to 36 x 3,5" or 72 x 3,5" or like in above example 24 x 3,5" + 24 x 2,5"

3. SuperMicro Mainboard Build to Order systems

3.1 Low power, low cost System X10SDV-2C-7TP4F



Broadwell-DE, Dual 10GbE, Embedded 7-Year Product Life (Coming Soon)

Key Features

1. Intel® Pentium® processor D1508, Single socket FCBGA 1667; 2-Core, 4 Threads, 25W
2. Up to 128GB ECC RDIMM DDR4 1866MHz or 64GB ECC/non-ECC UDIMM in 4 sockets
3. Expansion slot: 2x PCIe 3.0 x8, M.2 PCIe 3.0 x4, M Key 2242/2280/ 22110, Mini-PCIe w/ mSATA support
4. Dual 10G SFP+ and dual 1GbE LAN
5. 4x SATA3 (6Gbps) ports via SoC 16 SATA3/SAS2 via LSI 2116
6. 2x USB 3.0 ports (rear), 5x USB 2.0 ports (4 via headers, 1 Type A)
7. 2x SuperDOM, 1x COM, TPM 2.0 header, GPIO and SMbus headers
8. 12V DC input and ATX Power Source

Attention:

The board works with ESXi and OmniOS, I had problems with Solaris 11.3
Drivers for the X550 are on last tests (available from OmniOS 151019)

The board supports vt-d, so you can use it in an AiO setup

The boards is also available as
X10SDV-4C-7TP4F (4C/8HT)
X10SDV-7TP4F (8C/16HT) and
X10SDV-7TP8F (16C/32HT)

3.2 more flexible low power, low cost System X11 SSH-CTF

Datacenter Optimized

Key Features

1. Single socket H4 (LGA 1151) supports Intel® Xeon® processor E3-1200 v5, Intel® 6th Gen. Core™ i3 series, Intel® Celeron® and Intel® Pentium®
2. Intel® C236 chipset
3. Up to 64GB Unbuffered ECC UDIMM DDR4 2133MHz; 4x DIMM slots
4. Expansion slots: 1 PCI-E 3.0 x8, 1 PCI-E 3.0 x2 (in x4)
5. Dual 10GBase-T LAN with Intel® X550
6. 8x SATA3 (6Gbps) via C236; RAID 0, 1, 5, 10
7. 8x SAS3 (12Gbps) via LSI 3008; RAID 0, 1, 10
8. I/O: 1x VGA, 2x COM, TPM header
9. 2x SuperDOM with built-in power
10. 5x USB 3.0 (2 rear + 2 via header + 1 Type A), 6x USB 2.0 (2 rear + 4 via headers)
11. M.2 NGFF connector

The board is based on socket 1151 what limits RAM to 64GB but it offers more flexibility regarding CPU that can range from a cheap G4400 up to a Xeon, all with ECC and vt-d support. This board offers 10G and an onboard 12G SAS controller.

The board is also available with 1G and without the SAS controller

Drivers for the X550 are on last tests (available from OmniOS 151019)

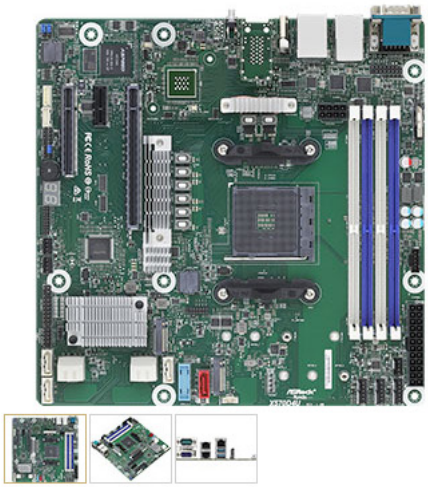
You can install OmniOS currently not from an USB installer stick.

Setup options:
- use the preconfigured napp-it systemimage
- use an Sata DVD drive to setup OmniOS manually

Needed bios settings
- Advanced > Boot: Install Windows 7 USB support: enabled
- Boot > Boot mode: Legacy

3.3 newer AMD system X570DU4 with Ryzen Pro 5000 (needed foe ECC support)

https://www.asrockrack.com/general/productdetail.asp?Model=X570D4U#Specifications



X570D4U

- Micro-ATX (9.6" x 9.6")
- Supports AMD Ryzen™ 5000 Series Desktop Processors
- 4 DIMM slots (2DPC), supports DDR4 ECC/non-ECC UDIMM
- 1 PCIe4.0 x16, 1 PCIe4.0 x8, 1 PCIe4.0 x1
- Supports 2 M.2 (PCIe4.0 x4 or SATA 6Gb/s)
- 8 SATA 6Gb/s
- 2 RJ45 (1GbE) by Intel® i210
- HDMI
- Remote management (IPMI)

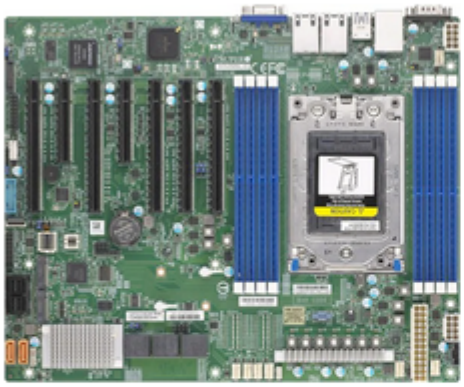
HDMI™ (High-Definition Multimedia Interface)

This motherboard supports HDMI™ (High-Definition Multimedia Interface) which is an interface standard for transferring uncompressed video data and delivering multi-channel audio through a single cable. Both video and audio data signals transferred through the HDMI™ interface are digital without being converted into analog, therefore it delivers the richest pictures and the most realistic sounds.

Up to 128 GB RAM
2 x M2
VGA and IPMI
Intel nic

3.4 AMD H12 SSL Epyc

H12SSL-C



Key Features

1. Single AMD EPYC™ 7003/7002 Series Processor
   (The latest AMD EPYC™ 7003 Series Processor with AMD 3D V-Cache™ Technology requires BIOS version 2.3 or newer)
2. 2TB Registered ECC DDR4 3200MHz SDRAM in 8 DIMMs
3. Expansion slots:
   5 PCI-E 4.0 x16
   2 PCI-E 4.0 x8
   M.2 Interface: 2 PCI-E 4.0 x4
   M.2 Form Factor: 2280, 22110
   M.2 Key: M-key
4. 8 SATA3, Broadcom 3008 SAS3 (12 Gbps) controller for 8 SAS3 ports, 2 M.2
5. 2 Gigabit Ethernet LAN Ports
6. ASPEED AST2500 BMC graphics
7. Up to 6 USB 3.0 ports
   (4 rear + 2 via header)
8. 7 PWM 4-pin Fans with tachometer status monitoring

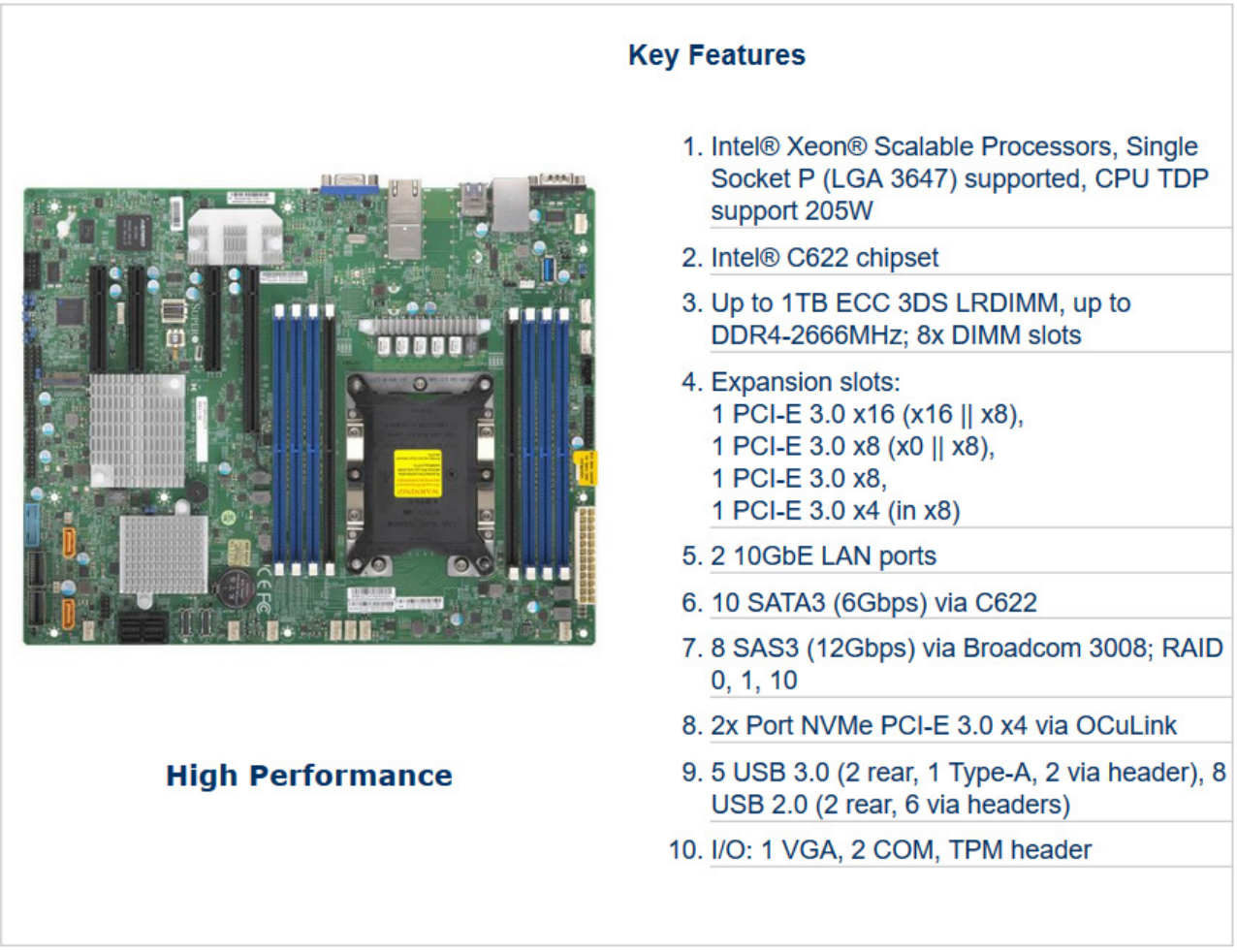
This board offers the most options as you can add up to 2 TB RAM with 7 PCI-e slots + 2*M2 for lot of NVMe/ U2 with full 4x performance, 1G/ optional 10 GB and SAS onboard.

see https://www.napp-it.org/doc/downloads/epyc_performance.pdf

NVMe

3.5 SuperMicro X11 SPH-nCTPF (SFP+) and X11 SPH-nCTF (10G-Base T)

This single Xeon socket 3647 systems are my intended main future storage platform as they offer lanes for up to 10 NVMe (2 OcuLink and M.2 onboard, up to seven more via NVMe HBA), onboard 10G and LSI 3008 and max 1TB ECC RAM at a price of around 500 Euro without CPU and RAM.

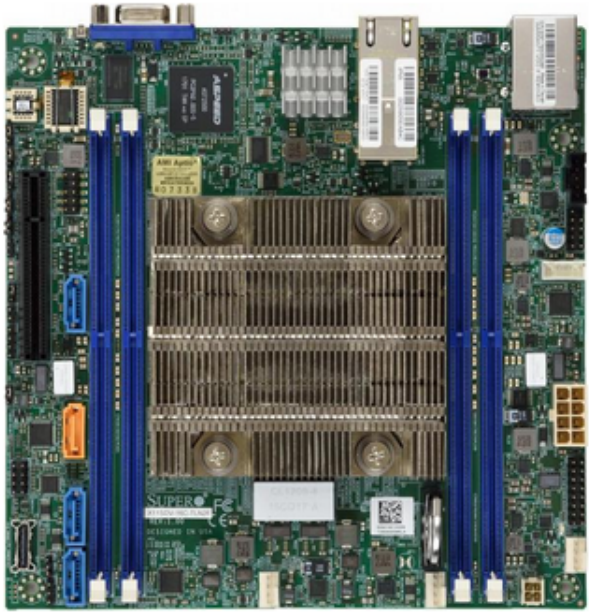


Check also newer X12/13 options from Supermicro
https://www.supermicro.com/en/products/motherboards/server-boards

3.6 SuperMicro X11SDV-4C-TLN2F miTX

This single Xeon mITX 60W system is my upcoming suggestion for an upperclass mITX storage system. It comes with a 4core/ 8 threads Intel Xeon, max 512GB ECC RAM and 2 x 10G Base-T onboard. You can connect 4 x Sata and one U.2 NVMe at the Oculink port (or alternatively 4 x Sata via Oculink). An additional 4x PCI-e port can hold an additional SAS adpater for more disks.

The same board is available in different editions with more nics or cores.
see https://www.servethehome.com/supermicro-x11sdv-4c-tln2f-review-with-intel-xeon-d-2123it/



Skylake-D
Quad Core
NVMe
Dual 10GbE

Key Features

1. Intel® Xeon® Processor D-2123IT, 4-Core, 8 Threads, 60W
2. System on Chip
3. Up to 256GB Registered ECC RDIMM or 512GB ECC LRDIMM, DDR4-2133MHz in 4 DIMM slots
4. 1 PCI-E 3.0 x8
   1 PCI-E 3.0 x4 NVMe Internal Port via OCuLink
5. Dual LAN with 10GBase-T with Intel® X557
6. Up to 8 SATA3 (6 Gbps) ports; RAID 0,1,5,10
   4 SATA ports via OCuLink (or PCIE3.0 x4 for NVMe)
7. 2 USB 3.0 ports (rear)
   2 USB 2.0 ports (via header)
8. 12V DC or ATX Power Input

4. Add-Ons

NVMe
To connect NVMe disks you have two options:

1. direct connect ex via Oculink or PCIe adapters
For full performance you need 4x lanes per NVMe. If you use passive PCIe adapters ex 16x for 4 NVMe you need a mainboard that supports 16x -> 4x4X bifurcation

2. connect via NVMe switch (Broadcom P411W-32P)
https://docs.broadcom.com/doc/BC00-0443EN

This is a 16 PCIe Gen 4 HBA alike adapter for to 8 x4, 16 x2, or 32 x1 hot plug capable NVMe connects.

L2Arc
L2ARC readcache device. This is an SSD to extend the rambased readcache. As RAM is much faster and the L2Arc requires a few percent RAM for organisation, first try to add RAM if you need more Cache. If you add an L2Arc it should be much faster than your pool, even sequentially as you need to write to the pool and the cache

L2Arc
If you need powerloss protection on writes with enabling sync, you should add an Slog.
An Slog must offer very low latency, very high iops and internal powerloss protection,
Good examples are Intel S3700 or Intel P3600/3700 or the new P3608.

LogiLink mounting brackets for internal 2,5" SSDs
http://logilink.eu/Suche/AD0014?seticlanguage=en

Mobile Racks/ backplanes for 5,25" bays
https://www.supermicro.nl/products/chassis/mobileRack/
http://www.icydock.com/goods.php?id=175

Adapter for 2,5" SSDs in a 3,5" backplane
http://www.icydock.com/goods.php?id=195

Sata OS independent Raid-1 enclosure (2 x 2,5" in a 3,5" enclosure)
http://www.icydock.com/goods.php?id=211

HBA controller, use LSI/Avaga HBA, best with raidless IT firmware, see
http://www.avagotech.com/products/server-storage/host-bus-adapters/

12G HBA

LSI 9300, 9305 and 9400 (9304 not on all supported OSes check your OS release)
http://www.avagotech.com/products/server-storage/host-bus-adapters/sas-9300-8e

ATTO ExpressSAS H12xxx (new: ATTO, a media specialist supports Illumos up from 2018)
https://www.atto.com/products/adapters

network adapters (prefer Intel or Chelsio), see
Intel X540, X550

40G network adapter
Intel X710 single or dual QSFP+ (driver i40e)

40G+, prefer Chelsio

see https://illumos.org/hcl/

5. Examples

5.1 Small workgroup filer, backup or Lab use

like HP Microserver G8, ML10 G9 orDell T20, T130

example: HP Proliant ML10-G9 with an Intel G4400 (vt-d capable) and 16 GB ECC

5.1.1 Barebone NAS setup

Your mainboard offers 4-6 Sata ports in AHCI mode and at at least 2 PCI-e slots

- Use a small Sata SSD as system disks, prefer
  a small enterprise class SSD like Intel DC S351x

- Use a pool from a mirror of two 3,5" disk (use 24/7 NAS disks, up to 10TB, I prefer
  HGST Ultrastar) or a Raid-Zn from up to 5 disks. **Do NOT use** MSR/ archive disks.

or use or add a pool from a mirror of 2 SSDs up to 3,8TB or a Raid-Zn. Prefer enterprise class SSDs like a Samsung SM/PM 863 or use upper class desktop SSDs with overprovisioning like the Sandisk Extreme Pro or add at least a manual overprovisioning to a new or secure erased SSD of about 10% to keep write performance high under load.

Options:
- Add a backplane with 4 x SSDs to one 5,25" bay or if you have two 5,25" bays,
  add a SSD backplane for 8 SSSs (see 2.1)

- Add a 3,5" hotplug capable drive bay where you can hot insert/remove a 3,5"
  disk for backups. Create a single disk pool and sync important data from your data-
  pools to this disk (autojob). Replace this backup pool regularly with other disks.

- 10G adapter (use an Intel X520 or X540)
- NVMe as an L2ARC (prefer more RAM)

5.1.2 Napp-in-one (ESXi + virtualized NAS + other operating systems)

Lab use

- Use an Sata SSD for ESXi and a local datastore where you place OmniOS on it
- add an LSI HBA in raidless IT mode in pass-through mode for OmniOS with disks
- use an SSD only mirror for VMs and a 3,5" disks for filer and backup use
- you can skip the Slog device and add an L2ARC if you cannot add/afford more RAM

6. Silent workgroup filer for video editing or lab/ office use with high capacity

6.1 Fractal Design R5 case, silent, low power demands (or other small cases)

- use a mainboard from the X10 SDV series with a 2 core, 4core, a core or 16 core CPI and 16 x SAS
- use an Sata bootdisk like the Intel enterprise class S3510-80

High capacity pool
- add up to 8 internal 3,5" disks like HGST Ultrastar 24/7 NAS (never use SMR archive disks)
Example: 6 disks in Raid-Z2. This gives up to 40 TB when using 10 TB disks

High performance/iops pool
- add a 4 or 8 bay 2,5" backplane into the 5,25" slots
- create an SSD pool from mirrors of 2 SSDs up to 3,8TB or Raid-Z2 of up to 8 SSD.
With high iops SSD you can use Raid-Z for a higher capacity, no need for Raid-10 alike setups

Prefer enterprise class SSDs like Intel S3500/3610/3700 or Samsung SM/PM 86. New champion is the WD Ultrastar SSD SS 530, a dualpath SAS SSD with up to 15TB, 320k write iops (4k) and 2 GB sequentially

Ultra high performance/ iops pool
-Create a pool from a mirror of two NVMe disks or WD Uptrastar DC SS 530
Optionally: any combination of the above pools

6.2 Napp-in-one (ESXi + virtualized NAS + other operating systems)

- Use an Sata SSD for ESXi and a local datastore where you place the napp-it ZFS appliance template onto
- Pass-through the SAS controller to OmniOS to have real disk access for ZFS with native drivers
- Use an SSD only pool without Slog for VMs and regular disks for filer/ backup use
- Use enough RAM (count 4 GB for ESXi and OmniOS, add the needs for your VMs and then add the amount of RAM that you want to use as ZFS readcache. RAM can go from 8 GB to 128 GB.

care about: Pass-through of NVMe does not work at the moment but you can use them as ESXi vdisks

6.3 Fractal Design R5 case, silent, ultra high iops demands

- use a mainboard from the X10 SR (single Xeon) or DR series (dual Xeon)
- with onboard SAS HBA or HBAs as PCI-e devices (ex LSI 9207, LSI 9003)
- with onboard 10G or with an additional Intel X520 or X540

- use an Sata bootdisk like the Intel enterprise class S3510-80

- create a pool from a mirror or Raid-Zn with up to 6 NVMe disks.
Use Intel P750. P3600, P3700 or P3608

and/or
- create an SSD pool with up to 16 SSDs in a Raid-10 or Raid-Zn config.
With high iops SSD you can use Raid-Z for a higher capacity, no need for Raid-10 alike setups
- add as much RAM as affordable/possible as readcache to improve read performance

An SSD only pool from SSDs does not require an additional L2ARC device unless you do not use a device that performs much better than your pool SSDs like an P3608 with regular SSDs in a pool.

An SSD only pool does not require an additional Slog device for sync writes. This may be only useful if your pool SSDs lack powerloss protection with an Slog that performs much better on sync writes than your pool SSDs with ultra low latency, high write iops and powerloss protection.

7. Very high capacity, price sensitive (Sata disks)

Use an expanderless 19" case with a miniSAS backplane (up to 36 x disks)

- add HBA controller according to the needed port numbers
ex: a 24 bay case require one 16 port HBA and one 8 port HBA or 3 x 8 port HBA.

Example with a 24 bay 19" case:

- mainboard from the X10SDV line with 16 port HBA onboard
- 10G nics onboard (very new, drivers for OmniOS are not yet available but on the way)
- an additional LSI 9207 that comes with IT mode firmware or another HBA

7.1 Use case: high capacity with iops as a concern

- use a multi Raid-10 setup (Create a pool from a mirror vdev, add mirror vdevs) as
iops performance scale with number of vdevs and sequential performance with number of disks

Create a pool with a raid-1 vdev of 4TB HGST Ultrastar, add more mirror vdevs up to 11 vdevs.
This means 22 disks with a capacity of 44 TB and the writre iops of 11 disks and the read iops of 22 disks.
You can calculate around 150 iops per disk (limited by disk rpm and latency) what means 1650 write iops from disks and 3300 read iops from disks. Storage values are higher with read and write caching and compress.

The sustained sequential write performance can go up to around 11 x 200 MB/s (2200 MB/s) and the sequential sustained read performance (as ZFS can read from both disks of a mirror simultaniously) up to 4400 MB/s.

Readcaching is the key for performance, so add RAM. Only if you cannot add or afford more RAM and require a larger readcache, add an L2ARC to extend the RAM. Be aware that an L2ARC SSD readcache is much slower than RAM readcache. It also requires some percent of RAM to organize the L2ARC. As every new write must go to the pool and to the L2ARC you should care of write performance of an L2ARC as well or it may affect write throughput negative.

Good L2ARC: Intel S3610-400 or an NVMe like a Intel P750-400

7.2 Use case: highest capacity (filer or backup)

- use a two Raid-Z2 or Z3 setup (Create a pool from a Z2 vdev, add Z2 vdevs)
  create a pool with a 10 disk z2 or an eleven disk z3 vdev, add a second identical vdev now or later.
  I would avoid to use much more disks per vdev as this affects resilver or scrub time negatively
- add at least one hotspare disk.

Such a setup can go up to 200 TB with 10TB Sata disks

- optionally add an L2ARC or ZIL device (see 6.1)
- optionally create one pool from SSDs (high iops) and one from 3,5" disks (high capacity)
- optionally use a SuperMicro X10 SR/DL board with up to 6 slots for NVMe disks for a second high iops pool

7.3 Napp-in-one (ESXi + virtualized NAS + other operating systems)

Lab use or as a backup/ failover system

-similar 6.2

8. Ultra high capacity or HA capable with SAS disks

Use 19" cases with expander
Use for HA configurations or capacity ranges from 200 TB to multiple Petabyte

8.1 High availability storage

based on two storageheads and a storagenode with dualpath SAS

- use two 19" systems (storageheads) with an external SAS connector
- use one Jbod storagenode with two external SAS connectors (each for one of the SAS ports)

Connect both storageheads via SAS to your storagenode. As SAS disks offer two ports, you can connect a single disk to both heads simultaniously. You need a HA software like RSF-1 from high-availability.com to manage the failover from one head to the other in case the primary head fails.

An option are SuperMicro Twin servers with two mainboards in one case.

This allows an immediate service failover of services even on a complete storagehead failure.
If you also want to allow a complete storagenode failure, you nead two heads and two nodes.

8.2 High capacity storage

Use a 19" case like a SuperChassis 847E26-R1400LPB with expander for 36 disks and a mainboard like a SM X11 SSH-CTF with a G4400 or Xeon, 64GB RAM that comes with 19G and onboard 12G SAS.

9 Petabyte storage

based on a storageheads and one or more 90 bay Jbod cases

- use a case like a SuperChassis 946ED-R2KJBOD (90 disks JBOD, up to 900 TB raw per enclosure)
- add up to 90 12G SAS disks like the HGST Ultrastar HE8 or He10, 512e, SAS12G SAS disks

From your 90 bays, use at least 2 hotspares what leaves a maximum of 88 disks
If you organize the pool in 8 vdevs with 11 disks, this gives a max usable capacity of 720 TB per case.
If you use new 10TB HE, you can get 720TB usable per enclosure (up to 900 TB raw).

If you want a more performance orientated setup you can use 14 vdevs from 6 disks each.
With 84 disks with room for hotspares, you have 560TB usable per enclosure.

- add a 19" Storagehead with a board like the X10 DRH-iT, 128-512 GB RAM, Dual Xeon and one ore more 12G SAS HBA SAS controller with external SAS connectors according to the number of enclosures**.**

You need a storagehead with a lot of RAM, at least one 12G external 2 SAS port (LSI/Avage 9300 8e).

For performance, plan two SAS ports per enclosure ex 2 x LSI 9300-8e
If you need a large L2ARC cache, add an Intel P3608

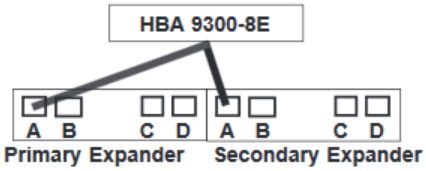
- Use one or more 10G links for uplink to a switch or use an Intel X710 (QSFP+/ 40G) as a 40G uplink
Drivers for the new X710/ 40G QSFP+ are on the way.

Such a setup has an enormous throughput. If you want to add an L2ARC for a very large readcache, you must use one that does not limit throughput on new writes as they must go to the pool and to the L2ARC. An 1.6 TB Intel P3608 with up to 5GB/s read throughput and 850k read iops may give you what you want. Enable sequential L2ARC caching in napp-it tunings. With 1.6 TB cache, use at least 256GB RAM.

Such an enclosure gives up to 900TB raw.
- If you need HA, use two storageheads with RSF-1 in an MPIO setup

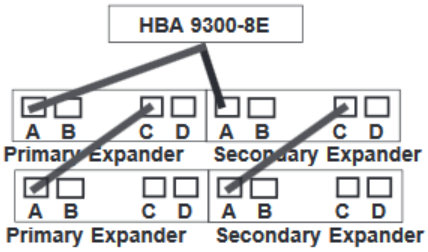
Some SAS abling options:

Single Host (LSI 9300-8E) with One JBOD



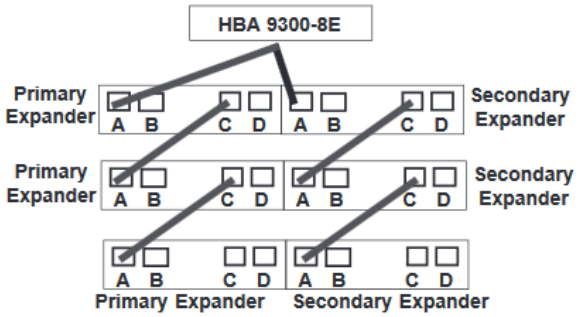
Single Host Cascading with One JBOD

Single Host (LSI 9300-8E) with Two JBODs



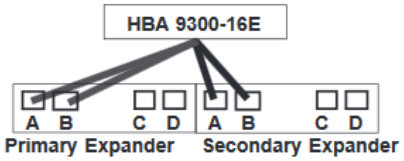
Single Host Cascading with Two JBODs

Single Host (LSI 9300-8E) with Three JBODs



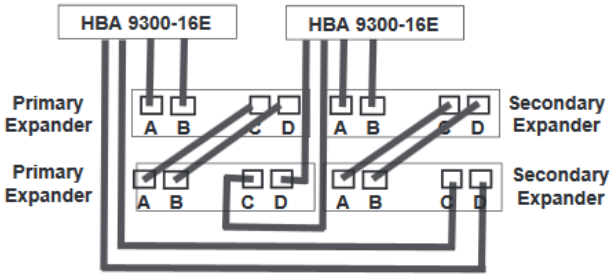
Single Host (LSI 9300-8E) with Three JBODs

Single Host (LSI 9300-16E) with One JBOD

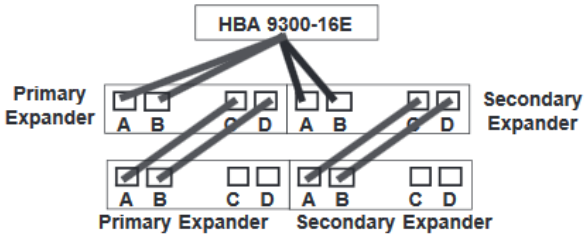


Single Host Cascading with One JBOD

Dual Host (LSI 9300-16E) with Two JBODs



Single Host (LSI 9300-16E) with Two JBODs



Special settings with many disks:
If you use the napp-it tuning options with a short timeout like 8s, you may find connectivity problems like high iostat failure rates. Increase timeout settings then to a higher value. Solaris default timeout is 60s. This may be too high as this is the time ZFS is waiting for a disk to answer.

If you do not use MPIO, optionally disable MPIO in your /kernel/drev/"controller".conf

10. Redundant Storage with data and service failover (Z-RAID SSF v.2)

This solution is beta in the current dev edition, release expected napp-it 2019.01

High availability solutions are usually based on a Dualpath Mpio SAS storage with redundant SAS controller and two storage heads. They allow a complete head or controller failure without a service outage as they offer services over a swichable failover ip.

Advantage of a typical HA solution (Cluster in a Box)
- best performance (SAS datapath)
- allow a head failure without a service outage
On a Master head failure, the standby Slave imports the pool and continue to offer services over same ip.

Problem of a typical HA solution
- costs and complexity as you need shared SAS storage and at least two physical servers

Z-RAID SSF vCluster vs traditional Cluster in a box solutions

The main idea behind a napp-it ZFS vCluster is to reduce complexity and costs of a ZFS Cluster but to maintain most of its performance and availability. This is achieved by the following

- use ESXi and storage virtualisation: One All-in- One instead of two servers/ mainboards
- use the features of ESXi to share virtual disk controllers and raw disks or vdisks between VMs
- use ZFS itself for NFS and SMB service management. This is possible due the kernelbased services on Solarish
- use napp-it for failover management and Stonith

Use cases for Z-RAID SSF
- affordable high available/ high performance NFS/SMB filers (other services optional)

Details, see
http://www.napp-it.org/doc/downloads/z-raid.pdf

11. High performance 40G Connection between Appliances

For Z-Raid SSF performance, network performance is essential. Ethernet1G is similar to old ATA disks with around 100MB. This is not fast enough for appliance Z-RAID.

10G Ethernet can give up to 1000MB/s. This is enough for a Z-RAID and offers enough performance for a traditional disk based pool or many use cases.

The upcoming QSFP+ standard offers 40G ethernet or 4 x 10G SFP+ over a breakout cable. This is in the performance area of a local high performance pool or a very fast NVMe solution. As the price of 40G adapters are quite similar with 10G server adapters, this is the best solution beside FC/IB solutions. While FC/IB may have advantages regarding latency, iSCSI is mainstream and QSFP+ is available in a lot of ethernet switches.

For 40G connectivity you can use either a new MPO (multi-fibre push-on) connector with a copper MPO DAC cable up to 5m for local connectivity, a 12 fiber MPO optical tranceiver and MPO fiber patchcables for connectivity inside a serverroom or between different brands. For long distance connectivity you can use CWDM tranceivers with an LC fiber connector. They use 4 different colors simultaniously over a single fiber to achieve 40G over a traditional Multimode (up to 300m) or Singlemode (long distance) fiber.

40G QSFP+ for a same room Appliance <-> Appliance failover setup
- 2 x Intel XL 710 DA1 (around 450 Euro each)
- QSFP + DAC copper cable (1 to 5 m), either genuine Intel or compatible
  Intel XLDACBL1 (..BL5), Intel compatible cables up from 90 Euro

40G QSFP+ for a local Appliance <-> Appliance failover setup
- 2 x Intel XL 710 DA1 (around 450 Euro each)
- 2 x Intel Tranceiver Intel E40GQSFPSR or compatible (Intel compatible up from 200 Euro)
- MPO Patchkabel or 2 x breakout cable MPO -> 12 x LC (multimode cabling) -> MPO

40G QSFP+ for a remote Appliance <-> Appliance failover setup
- 2 x Intel XL 710 DA1 (around 450 Euro each)
- 2 x CWDM Tranceiver (several 10G links over different colors) for 10G over
  Multimode od Singlemode LWL (ask for compatibility with Intel)

40G QSFP+ with a 40G Site to Site or building to building connectivity
- 2 Switches with QSFP+ ports ex H3C
- 2 x CWDM Tranceiver (several 10G links over different colors) for 10G over
  Multimode od Singlemode LWL (ask for compatibility with your switch type)
  ex (H3C) HPEX 140 for 2km (JL286A) or 10km (JG661A) or compatible up from 800 Euro

Connect your Appliance to the QSFP+ port of a switch
- 2 x QSFP+ MPO Tranceiver for your Switch ex HPE X140 MPO (JG709A)
  or compatible up from 200 Euro
- 2 x Intel XL 710 DA1 (around 450 Euro each) for the appliances
- 2 x MPO Tranceiver Intel E40GQSFPSR or compatible (Intel compatible up from 200 Euro)
- 2 MPO Patchcable to connect the XL710 to the switch, price depends on length up from 100 Euro

If you simply want 40G between a QSFP+ switch and your storage appliance
- 1 x Intel X710 (around 450 Euro)
- 1 x MPO Tranceiver Intel E40GQSFPSR or Intel compatible up from 200 Euro
- 1 x MPO Tranceiver for your Switch ex HPE X140 MPO (JG709A) or Intel compatible up from 200 Euro
- 1 MPIO patchcable, price depends on length up from 100 Euro

prices are estimated prices in Germany without tax, example from www.flexoptix.net

12. Performance considerations
for a OmniOS/OpenIndiana/Solaris based ZFS server

If you simply want the best performance, durability and security, order a server with a very new CPU with a frequency > 3GHz and 6 cores or more, 256 GB RAM and a huge Flash only storage with 2 x 12G multipath SAS (10dwpd) or NVMe in a multi mirror setup - with a datacenter quality powerloss protection to ensure data on a powerloss during writes or background garbage collection. Do not forget to order twice as you need a backup on a second location at least for a disaster like fire, theft or Ransomware.

Maybe you can follow this simple suggestion, mostly you search a compromise between price, performance and capacity under a given use scenario. Be aware that when you define two of the three parameters, the third is a result of your choice ex low price + high capacity = low performance.

As your main concern should be a workable solution, you should not start with a price restriction but with your use case and the needed performance for that (low, medium, high, extreme). With a few users and mainly office documents, your performance need is low, even a small server with a 1.5 GHz dualcore CPU, 4-8 GB RAM and a mirror from two SSD or HD can be good enough. Add some external USB disks for a rolling daily backup and you are ready.

If you are a media firm with many users that want to edit multitrack 4k video from ZFS storage, you need an extreme solution regarding pool performance (> 2GB/s sequential read,write), network (min multiple 10G) and capacity according your needs. Maybe you come to the conclusion to prefer a local NVMe for hot data and a medium class disk based storage for shared file access and versioning only. Do not forget a disaster backup solution.

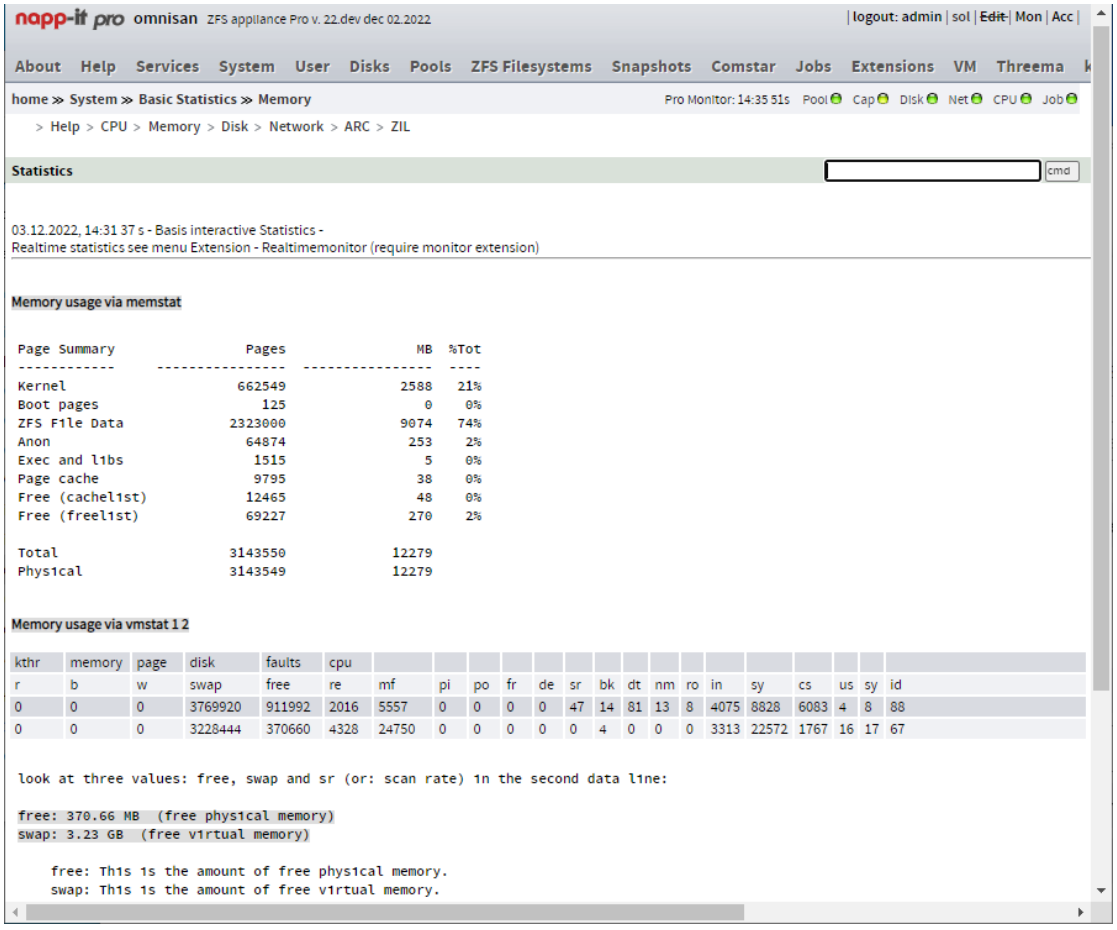
After you have defined the performance class/use case (low, medium, high, extreme),
select needed components.

CPU

For lower performance needs and 1G networks, you can skip this. Even a cheap dual/quadcore CPU is e good enough. If your performance need is high or extreme with a high throughput in a 10G network or when you need encryption, ZFS is quite CPU hungry as you see in https://www.napp-it.org/doc/downloads/epyc_performance.pdf. If you have the choice prefer higher frequency over more cores. If you need sync write (VM storage or databases) avoid encryption as encrypted small sync writes are always very slow.

RAM

Solaris based ZFS systems are very resource efficient due the deep integration of iSCSI, NFS and SMB into the Solaris kernel that was developped around ZFS from the beginning. You need less than 3 GB for a 64bit Solaris based OS itself to be stable with any pool size. Use at least 4-8 GB RAM to allow some caching for low to medium needs with only a few users.



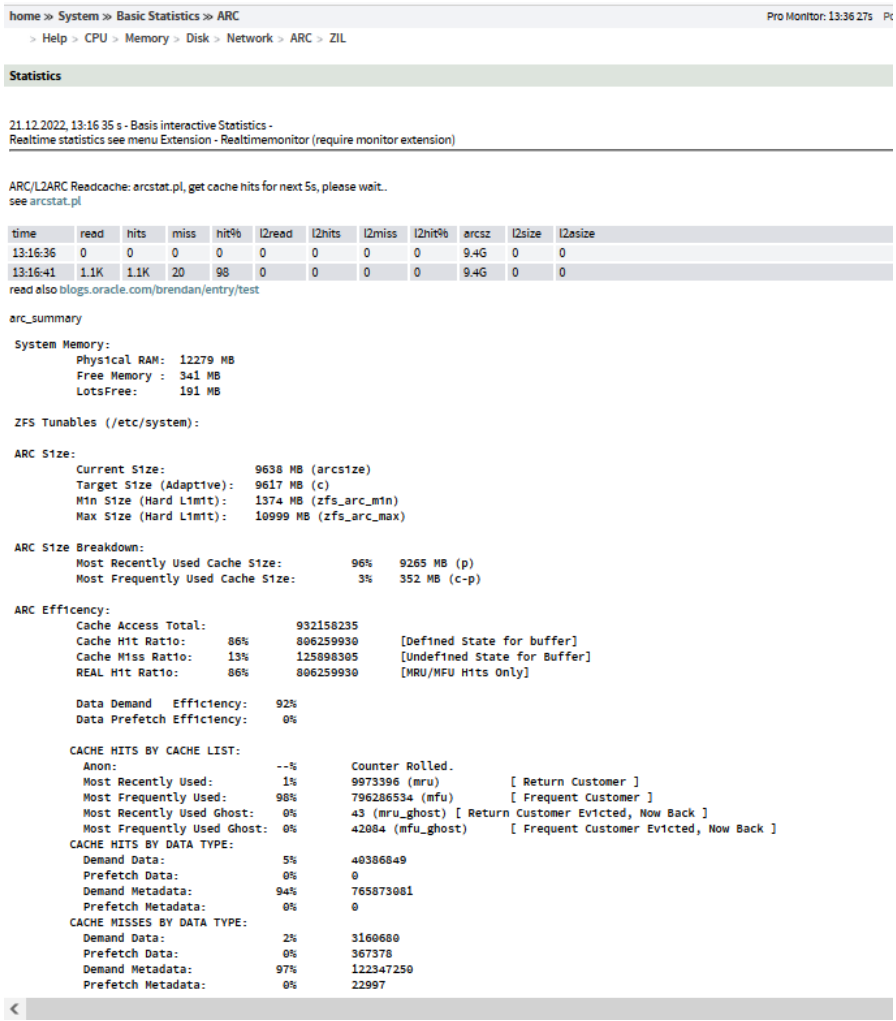
As ZFS uses most of the RAM (unless not dynamically demanded by other processes) for ultrafast read/write caching to improve performance you may want to add more RAM. Per default Open-ZFS uses 10% of RAM for write caching. As a rule of thumb you should collect all small writes < 128k in the rambased write cache as smaller writes are slower or very slow. As you can only use half of the write cache unless the content must be written to disk, you want at least 256k write cache that you can have with 4 GB RAM in a single user scenario. This RAM need for write caching scale with number of users that write concurrently so add around 0.5 GB RAM per active concurrent user.

Oracle Solaris with native ZFS works different. The rambased writecache caches last 5s of writes that can consume up to 1/8 of total RAM. In general this often leads to similar RAM needs than OI/OmniOS with Open-ZFS. On a faster 10G network with a max write of 1 GB/s this means 8GB RAM min + RAM wanted for readcaching.

Most of the remaining RAM is used for ultrafast rambased readcaching (Arc). The readcache works only for small io on a read last/ read most optimazation. Large files are not cached at all. Cache hits are therefore for matadate and small random io. Check napp-it menu System > Basic Statistic > Arc after some time of storage usage. Unless you does not have a use scenario with many users, many small files and a high volatility (ex a larger mailserver), cache hit rate should be > 80% and metadata hit rate > 90%. If results are lower you should add more RAM or use high performance storage like NVMe where caching is not so important.

If you read about 1GB RAM per TB storage, forget this. It is a myth unless you do not activate rambased realtime dedup (not recommendet or when dedup is wanted use fast NVMe as a special vdev mirror for dedup). Up from a certain amount ex 16GB, more RAM up to 256GB is reasonable depending on use case, number of users, small files or wanted cache hit rate in relation to raw pool performance but is in no way poolsize related or helpful with large files.

Arc cache hit rate



home » System » Basic Statistics » ARC — Pro Monitor: 13:36 27s

» Help » CPU » Memory » Disk » Network » ARC » ZIL

Statistics

21.12.2022, 13:16 35 s - Basis interactive Statistics -
Realtime statistics see menu Extension - Realtimemonitor (require monitor extension)

ARC/L2ARC Readcache: arcstat.pl, get cache hits for next 5s, please wait..
see arcstat.pl

time	read	hits	miss	hit%	l2read	l2hits	l2miss	l2hit%	arcsz	l2size	l2asize
13:16:36	0	0	0	0	0	0	0	0	9.4G	0	0
13:16:41	1.1K	1.1K	20	98	0	0	0	0	9.4G	0	0

read also blogs.oracle.com/brendan/entry/test

arc_summary

```
System Memory:
         Physical RAM:  12279 MB
         Free Memory :  341 MB
         LotsFree:      191 MB

ZFS Tunables (/etc/system):

ARC Size:
         Current Size:             9638 MB (arcsize)
         Target Size (Adaptive):   9617 MB (c)
         Min Size (Hard Limit):    1374 MB (zfs_arc_min)
         Max Size (Hard Limit):    10999 MB (zfs_arc_max)

ARC Size Breakdown:
         Most Recently Used Cache Size:          96%    9265 MB (p)
         Most Frequently Used Cache Size:         3%    352 MB (c-p)

ARC Efficency:
         Cache Access Total:             932158235
         Cache Hit Ratio:      86%       806259930      [Defined State for buffer]
         Cache Miss Ratio:     13%       125898305      [Undefined State for Buffer]
         REAL Hit Ratio:       86%       806259930      [MRU/MFU Hits Only]

         Data Demand   Efficiency:    92%
         Data Prefetch Efficiency:    0%

        CACHE HITS BY CACHE LIST:
          Anon:                       --%        Counter Rolled.
          Most Recently Used:          1%        9973396 (mru)          [ Return Customer ]
          Most Frequently Used:       98%        79628534 (mfu)         [ Frequent Customer ]
          Most Recently Used Ghost:    0%        43 (mru_ghost) [ Return Customer Evicted, Now Back ]
          Most Frequently Used Ghost:  0%        42084 (mfu_ghost)      [ Frequent Customer Evicted, Now Back ]
        CACHE HITS BY DATA TYPE:
          Demand Data:                 5%        40386849
          Prefetch Data:               0%        0
          Demand Metadata:            94%        765873081
          Prefetch Metadata:           0%        0
        CACHE MISSES BY DATA TYPE:
          Demand Data:                 2%        3160680
          Prefetch Data:               0%        367378
          Demand Metadata:            97%        122347250
          Prefetch Metadata:           0%        22997
```

L2Arc
L2Arc is an SSD or at best NVMe that can be used to extend the rambased Arc. L2Arc is not as fast as RAM but can increase cache size when more RAM is not an option or when the server is rebooted more often as L2Arc is persistent. As L2Arc needs RAM to organize, do not use more than say 5x RAM as L2Arc. Additionally you can enable read ahead on L2Arc that may improve sequential reads a little. (add „set zfs:l2arc_noprefetch=0" to /etc/system or use napp-it System > Tuning).

Disk types
RAM can help a lot to improve ZFS performance with the help of read/write caching. For larger sequential writes and reads or many small io it is only raw storage performance that counts. If you look at the specs of disks the two most important values are seqential transfer rate for large transfers and iops that counts when you read or write small datablocks.

Mechanical disks
On mechanical disks you find values of around 200-300 MB/s max transfer rate and around 100 iops. As a Copy on Write filesystem like ZFS is more affected by fragmentation and smaller datablocks spread over the whole pool, performance is more limited by iops than sequential values. On average use you will often not see more than 100-150 MB/s per disk. When you enable sync write on a single mechanical disk, write performance is not better than say 10 MB/s due the low iops rating.

Desktop Sata SSD
can achieve around 500 MB/s (6G Sata) and a few thousand iops. Often iops values from specs are only valid for a short time until performance drops down to a fraction on steady writes.

Enterprise SSDs
can hold their performance and offer powerloss protection PLP. Without PLP last writes are not save on a power outage during write as well as data on disk with background operations like firmware based garbage collection to keep SSD performance high.

Enterprise SSDs are often available as 6G Sata or 2 x 12G multipath SAS. When you have an SAS HBA prefer 12G SAS models due the higher performance (up to 4x faster than 6G Sata) and as SAS is full duplex while Sata is only half duplex with a more robust signalling with up to 10m cable length (Sata 1m). The best of all SAS SSDs can achieve up to 2 GB/s transfer rate and over 300k iops on steady 4k writes. SAS is also a way to use storage with more than 100 disks easily with the help of SAS expanders.

NVMe
is the fastest option for storage. The best like Intel Optane 5800x rate at 1.6M iops and 6.4 GB/s transfer rate. In general Desktop NVMe lack powerloss protection and write iops slow down on steady write so prefer datacenter models with PLP. While NVMe are ultrafast it is not as easy to use many of them as each wants a 4x pci lane connection (pci-e card, M.2 or oculink/U.2 connector). For a larger capacity SAS storage is often nearly as fast and easier to implement especially when hotplug is wanted. NVMe is perfect for a second smaller high performance pool for databases/VMs or to tune a ZFS pool with an Slog for a faster sync write on disk based pools, a persistent L2Arc or a special vdev mirror.

ZFS Pool Layout

ZFS groups one or more disks to a vdev and stripes all vdevs in a raid-0 manner to a pool to improve performance or reliability. While a ZFS pool from a single disk without redundancy rate as described above, a vdev from several disks can behave better.

Raid-0 pool (ZFS always stripes data over vdevs in a raid-0)
You create a pool from a single disk (this is a basic vdev) or a mirror/raid-Z vdev and can add more vdevs in a raid-0 configuration. Overall read/write performance from math is number of vdevs x performance of a single vdev as each must only process 1/n of data. Real world performnce is not a factor n but more 1.5 to 1.8 n depending on disks or disc caches and decreases with more vdevs. Keep this in mind when you want to decide if ZFS performance is „as expected"

A pool from a single n-way mirror vdev
You can mirror two or more disks to create a mirror vdev. Mostly you mirror to improve datasecurity as write performance of an n-way mirror is equal to a single disk (a write is done when on all disks). As ZFS can read from all disks simultaniously read performance and read iops scale with n. When a single disk rate with 100 MB/s and 100 iops a 3way mirror can give up to 300 MB/s and 300 iops. If you run a napp-it Pool > Benchmark with a singlestream read benchmark vs a fivestream one, you can see the effect. In a 3way mirror any two disks can fail without a dataloss.

A pool from multiple n-way mirror vdevs
Some years ago a ZFS pool from many striped mirror vdevs was the preferred method for faster pools. Nowaday I would use mirrors only when one mirror is enough or when an easy extension to a later Raid-10 setup ex from 4 disks is planned. If you really need performance, use SSD/Nvme as they are by far superiour.

A pool from a single Z1 vdev
A Z1 vdev is good to combine up to say 4 disks. Such a 4 disk Z1 vdev gives the capacity of 3 disks. One disk of the vdev is allowed to fail without a dataloss. Unlike other raid types like raid-5 a readerror in a degraded Z1 does not mean a pool lost but only a damaged reported file that is affected by the read error. This is why Z1 is much better and named different than raid-5. Sequential read/write performance of such a vdev is similar to a 3 disk raid-0 but iops is only like a single disk (all heads must be in position prior an io)

A pool from a single Z2 vdev
A Z2 vdev is good to combine say 5-10 disks. A 7 disk Z2 vdev gives the capacity of 5 disks. Any two disks of the vdev are allowed to fail without a dataloss. Unlike other raid types like raid-6 a readerror in a totally degraded Z2 does not mean a pool lost but only a damaged reported file that is affected by the read error. This is why Z2 is much better and named different than raid-6. Sequential read/write performance of such a vdev is similar to a 5 disk raid-0 but iops is only like a single disk (all heads must be in position prior an io)

A pool from a single Z3 vdev
A Z1 vdev is good to combine say 11-20 disks. A 13 disk Z2 vdev gives the capacity of 10 disks. Any three disks of the vdev are allowed to fail without a dataloss. There is no equivalent to Z3 in traditional raid. Sequential read/write performance of such a vdev is similar to a 10 disk raid-0 but iops is only like a single disk (all heads must be in position prior an io)

A pool from multiple raid Z[1-3] vdevs
Such a pool stripes the vdevs what means sequential performance and iops scale with number of vdevs (not linear similar to the raid-0 degression with more disks)

Many small disks vs less larger disks
Many small disks can be faster but are more power hungry and as performance improvement is not linear and failure rate scale with number of parts I would always prefer less but larger disks. The same is with number of vdevs. Prefer a pool from less vdevs. If you have a pool of say 100 disks and an annual failure rate of 5%, you have 5 bad disks per year. I you asume a resilver time of 5 days per disk you can expect 3-4 weeks where a resilver is running with a noticeable performance degration.

Special vdev
Some high end storages offer tiering where active or performance sensitive files can be placed on a faster part of an array. ZFS does not offer traditional tiering but you can place critical data based on their physical size (small io), type (dedup or metadata) or based on the recsize setting of a filesystem on a faster vdev of a ZFS pool. Main advantage is that you do not need to copy files around so this is often a superiour approach as mostly the really slow data is data with a small physical file or blocksize. As a vdev lost means a pool lost, use special vdevs always in a n-way mirror. Use the same ashift as all other vdevs (mostly use ashift=12 for 4k physical disks) to allow a special vdev remove.

To use a special vdev, use menu Pools > Extend, select a mirror (best a fast SSD/NVMe mirror with PLP) with type=special. Allocations in the special class are dedicated to specific block types. By default this includes all metadata, the indirect blocks of user data, and any deduplication tables.

The class can also be provisioned to accept small file blocks. This means you can force all data of a certain filesystem to the special vdev when you set the ZFS property „special_small_blocks" ex special_small_blocks=128K for a filesystem with a recsize setting smaller or equal. In such a case all small io and some critical filesystems are on the faster vdev others on the regular pool. If you add another vdev mirror load is distributed over both vdevs. If a special vdev is too full, data is stored on the other slower vdevs. A special vdev is mainly helpful in special use cases like a single smaller power hungry filesystem or many users, many small files and a high volatility (ex mailserver)

Slog
With ZFS all writes always go to the rambased writecache (there may be a direct io option in a future ZFS) and are written as a fast large transfer with a delay. On a crash during write the content or the writcache is lost (up to several MB). Filesystems on VM storage or databased may get corrupted while ZFS remains intact due Copy on Write. If you cannot allow such a dataloss you can enable sync write for a filesystem. This will force any write commit immediately to a faster Zil area of the pool or to a fast dedicated Slog device that can be much faster than the pool ZIL area . Despite and additionally in a second step data is written as a regular cache write. Every bit that you write is written twice, once directly and once collected in writecache. This can never be as fast as a regular write without sync. So Slog is not a performance option but a security option when you want acceptable sync write performance. The Slog is never read beside after a power outage to redo missing writes on next reboot, similar to the BBU protection of hardware raid.

Add an Slog only when you need sync write (a fileserver usually does not require sync) and buy the best that you can afford regarding low latency, high endurance and 4k write iops. Powerloss protection is a must. The Slog can be quite small (min 10GB). Widely used are the Intel datacenter Optane.

Tuning
Beside the above „physical" options you have only a few tuning options. For faster 10G+ networks you can increase tcp buffers or NFS settings in menu System > Tuning. Another option is Jumboframes that you can set in menu System > Network Eth ex to a „payload" of 9000. Do not forget to set all switches to highest possible mtu value or at least to 9126 (to include ip headers)

Another setting is ZFS recsize. For VM storage with filesystems on it I would set to 32K or 64K (not lower as ZFS becomes inefficient then). For mediadata a higer value of 512K or 1M may be faster.

13. more docs

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

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

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

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

Intel Optane, a game-changing technology
http://napp-it.org/doc/downloads/optane_slog_pool_performane.pdf