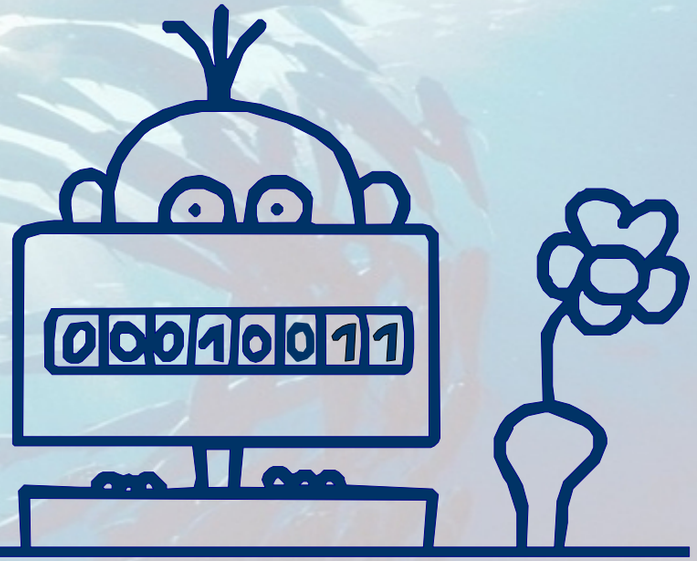
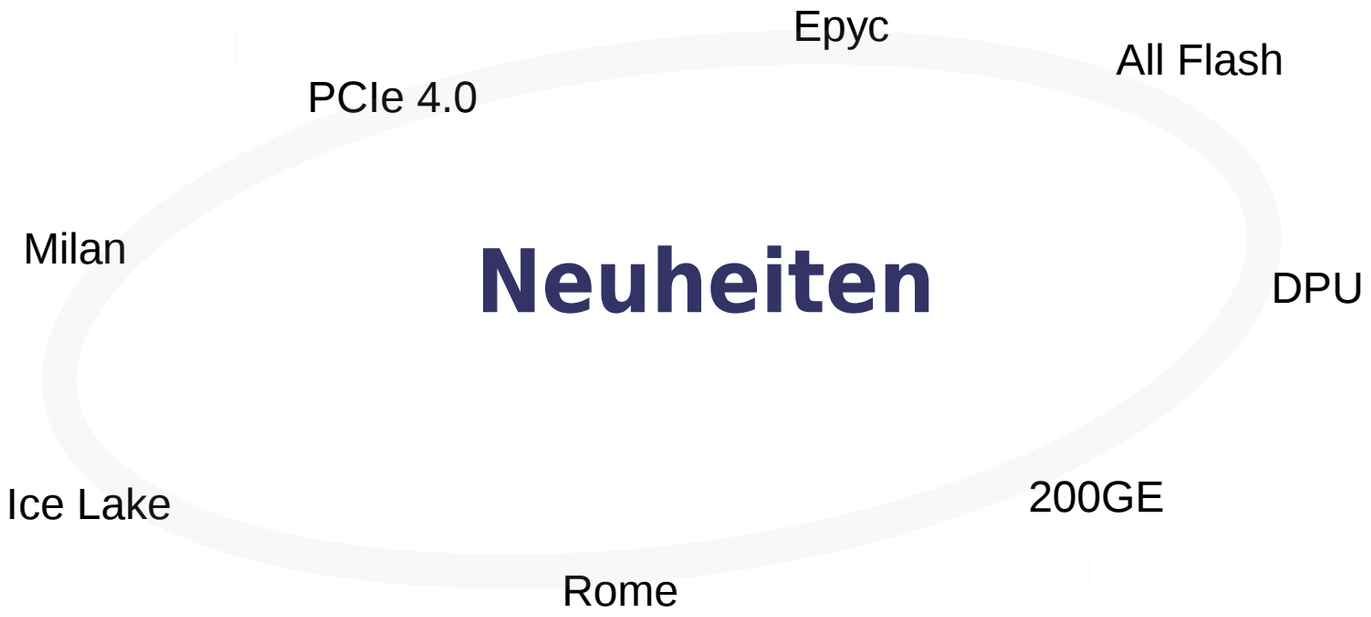
# **Hardware**
# für moderne I/O-Lösungen




OSL Aktuell

Schöneiche / Berlin • 26. Mai 2021

stefan.rulewitz@osl.eu



# Prozessoren der dritten Generation

Wie tickt die Zentrale?

## AMD Epyc 7003 Milan

- 8C/16T bis 64C/128T
- 128/160PCIE 4.0 Lanes
- 120W bis 280W TDP
- 8 Speicherkanäle @ 3200MHz
- bis 4TB RAM

## Intel Xeon Sc Ice / Cooper Lake

- 8C/16T bis 40C/80T
- 64/128 PCIE 4.0 Lanes
- bis 270W TDP
- 8 Speicherkanäle @ 3200MHz
- bis 4TB RAM

# PCI Express

Wie schnell fährt der Bus und wenn ja wieviele?

- Übergang von PCIE 3.0 zu 4.0
  - AMD Epyc seit drei Jahren
  - Xeon Ice Lake seit 2021
- Zahl der Lanes enorm gestiegen
  - AMD Epyc 128 Lanes (bis 160)
  - Xeon Sc 2.Gen 48 → 64 Lanes bei Ice Lake
- PCIE entwickelt sich zur Engstelle
- PCIE 5.0 mit CXL (Compute Express Link) angekündigt für 2022

| Version | Durchsatz via n-Lanes in GB/s | | | | |
|---|---|---|---|---|---|
| | x1 | x2 | x4 | x8 | x16 |
| 3.0 | 0,99 | 1,97 | 3,94 | 7,88 | 15,75 |
| 4.0 | 1,97 | 3,94 | 7,88 | 15,75 | 31,51 |
| 5.0 | 3,94 | 7,88 | 15,75 | 31,51 | 63,02 |

# Solid State Drives

Alles stabil im State?

- NVMe über PCIE 4.0 erreicht den Markt
  - Sättigung von PCIE 3.0 x4 erreicht
  - Raten von 7/5GByte/s und bis 1MIOPS
- Neue Bauformen
  - All Flash Gedanke → Abkehr von HDD-Bauformen
  - 2.5” → M.2 → E.1
- Verschiebung der Technologie von SLC → MLC → TLC → QLC
  - weniger Robust
  - Kompensation durch DRAM/SLC-Caches, Overprovisioning
- Aufspaltung für spezielle Anforderungen
  - Caching → Intel Optane und WD DC-ME200 Fokus auf IOPS
  - Optimierung für Lesen, Schreiben oder Mischbetrieb → Bsp. Kioxia

https://business.kioxia.com/en-emea/news/2020/20201106-1.html

# Anforderungen an vorhandene Systeme

# PCIE Lanes

Viele Spuren führen zum Erfolg

- Bisherige Systeme „kranken“ an geringer Anzahl von Lanes
  - Storage 24* x4 = 96 Lanes für NVMe SSDs
  - Netzwerk 2* x16 = 32 Lanes
  => Bedarf an 128 Lanes

- Skalierung von Xeon Scalable nutzen
  - 2-Sockel-Systeme bis 96 Lanes

- Unnötige Komponenten entfernen

- Abwärtskompatibilität ermöglicht volle Busnutzung
  - Bsp. 4.0 SSD an 3.0 Anschluss

# Kühlung

## Abwärme und Nebenwirkungen

- Abwärme moderner Prozessoren
  - Epyc Milan bis zu 280W TDP
  - Xeon Ice Lake bis 270W TDP
- SSDs mögen keine Hitze
  - Leistungsreduzierung
  - „CPU TDP is limited to 155W if using m.2 device”
- NICs heizen mit
  - 16 – 31W Abwärme per NIC
  - 2 – 2,5m/s bei  55°C
- Trend zur GPU-Beschleunigung

https://www.servethehome.com/asus-esc4000a-e10-2u-amd-epyc-nvidia-gpu-server-review/ - https://www.nvidia.com/en-us/networking/infiniband-adapters/

# SSDs

## Der Stall bestimmt die Pferde

- Auswahl an Anschlüssen
  - SATA/SAS in 2.5“ Bauform
  - U.2 in 2.5“ Bauform mit NVMe
  - PCIE-HHHL Einsteckkarten
  - M.2 mit Adapter (PCIE oder U.2)

| Standard | Modell | Max. Kapazität | lesen | | schreiben | |
|---|---|---|---|---|---|---|
| | | | Rate MB/s | IOPS 4k | Rate MB/s | IOPS 4k |
| SAS 12G | Samsung PM1643 | 30TB | 2100 | 400k | 1700 | 50k |
| U.2 | Kioxia CM6 | 12,8TB | 6900 | 1400k | 4200 | 350k |
| U.2 | Intel Optane P5800X | 3,2TB | 7200 | 1500k | 6200 | 1500k |
| HHHL | Samsung PM1735 | 12,8TB | 8000 | 1500k | 3800 | 250k |
| M.2 | Samsung 980Pro | 2TB | 7000 | 1000k | 5000 | 1000k |
| M.2 | Samsung PM9A1 | 2TB | 7000 | 1000k | 5200 | 850k |

https://business.kioxia.com/en-emea/ssd.html - https://www.samsung.com/semiconductor/ssd/

# Richtig schnelles Netzwerk

Äther oder Ether?

- Ethernet mit bis 200GBit/s
  - mit PCIE 4.0
  - alternativ 2* PCIE 3.0x16
- Angekündigt bis 400GBE
- Nur eine Technologie
  → Echte Alternative zu FC, IB, etc.
- Anbindung Network-Storage kann schneller als lokal sein

https://www.nvidia.com/en-us/networking/

# Drei Beispiele – Zwei OS

# Supermicro TwinPro HPC-Knoten #2

Dicht und Dichter?

- Vorhandene Hardware
  - 2-Sockel Xeon 4108 8C/16TP
  - 96GB RAM
  - 2* PCIE 3.0 x16
  - SATA-Controller für 6

- OSL Upgrade
  - Solaris 11.4
  - 1* Mellanox X-6 VPI Dual-100GBE
  - 5* Samsung 870 Evo SATA-SSD

- Herausforderungen
  - Beschränkung auf 2 PCIE-Steckplätze
  - Hohe Dichte vs. Kühlung
  - Solaris unterstützt verwendeten 100GE-Adapter nicht
  - Solaris unterstützt verwendete NVMe-Karte nicht

https://www.supermicro.com/en/products/system/2U/2029/SYS-2029TP-HTR.cfm

# Supermicro TwinPro HPC-Knoten #1

Dicht und Dichter?

- Vorhandene Hardware
  - 2-Sockel Xeon 4108 8C/16TP
  - 96GB RAM
  - 2* PCIE 3.0 x16
  - SATA-Controller für 6

- OSL Upgrade
  - SUSE Tumbleweed
  - 1* Mellanox X-6 VPI Dual-100GBE
  - 2* Samsung 970 Pro M.2 SSD in PCIE-Adapter mit Bifurcation (alt. PCIE-Switch Karte)

- Herausforderungen
  - Beschränkung auf 2 PCIE-Steckplätze
  - Hohe Dichte vs. Kühlung

https://www.supermicro.com/en/products/system/2U/2029/SYS-2029TP-HTR.cfm

# Supermicro TwinPro HPC-Knoten #3

Dicht und Dichter?

- Vorhandene Hardware
  - 2-Sockel Xeon 4108 8C/16TP
  - 96GB RAM
  - 2* PCIE 3.0 x16
  - SATA-Controller für 6

- OSL Upgrade
  - SUSE Tumbleweed
  - 1* Mellanox X-6 VPI Dual-100GBE
  - 5* Samsung 870 Evo SATA-SSD

- Herausforderungen
  - Beschränkung auf 2 PCIE-Steckplätze
  - Hohe Dichte vs. Kühlung

https://www.supermicro.com/en/products/system/2U/2029/SYS-2029TP-HTR.cfm

# Fujitsu PRIMERGY RX2540-M4 #9

Brot, Butter und etwas Salz dazu?

- Vorhandene Hardware
  - 2-Sockel Xeon 4110 8C/16TP
  - 128GB RAM
  - 3* PCIE 3.0 x8
  - 3* PCIE 3.0 x16
  - SAS-Controller

- OSL Upgrade
  - SUSE Tumbleweed
  - 2* Mellanox X-5 100GBE
  - 2* Samsung 980 Pro M.2 SSD in PCIE-Adapter

- Herausforderungen
  - SAS-Controller zwingend
  - Steckplatz-CPU-Zuordnung beachten

https://www.fujitsu.com/global/products/computing/servers/primergy/rack/rx2540m4/

# Was wir schon sehen können

# SSD-Performance – Solaris

## Wie hell strahlt der Stern?

| | | 1x | | 2x | | 3x | | 4x | | 5x | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | Block | MB/s | CPU % | MB/s | CPU % | MB/s | CPU % | MB/s | CPU % | MB/s | CPU % |
| Supermicro TwinPro-Node – Samsung 870 Evo SATA | | | | | | | | | | | |
| SATA RAW | 4k | 84 | 3 | 150 | 5 | 225 | 7 | 300 | 10 | 335 | 12 |
| SATA RAW | 1024k | 526 | 0 | 1048 | 0 | 1572 | 0 | 1860 | 0 | 1860 | 0 |
| SATA RAW | 8192k | 527 | 0 | 1054 | 0 | 1536 | 0 | 1916 | 0 | 1870 | 0 |

# SSD-Performance – Tumbleweed

Wie schnell taumelt das Gras?

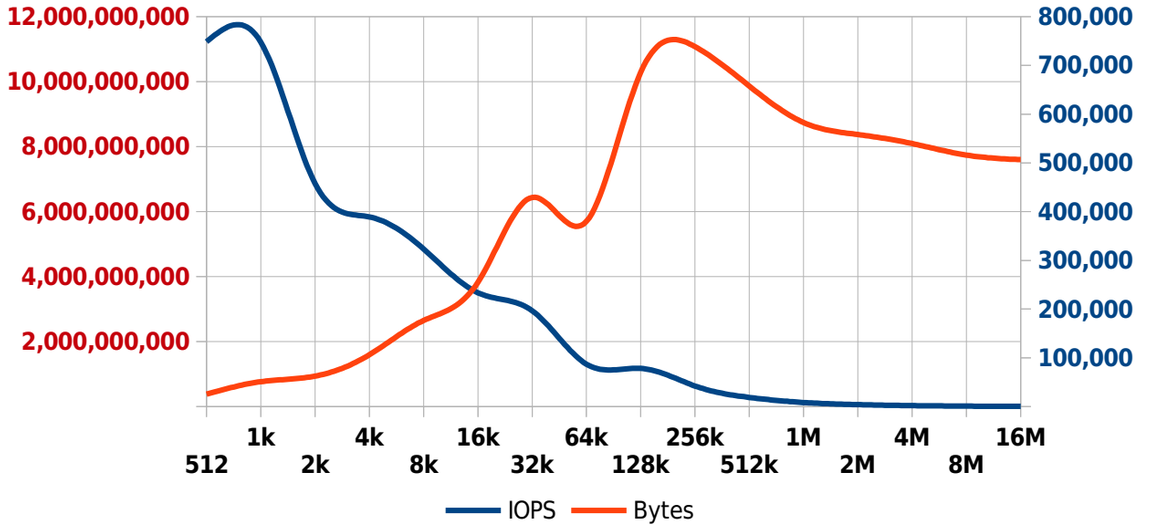
| | | 1x | | 2x | | 3x | | 4x | | 5x | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | Block | MB/s | CPU % | MB/s | CPU % | MB/s | CPU % | MB/s | CPU % | MB/s | CPU % |
| Supermicro TwinPro-Node – Samsung 970 Pro in PCIE 3.0 x8 Bifurcation Adapter | | | | | | | | | | | |
| SATA RAW | 4k | 129 | 3 | 258 | 7 | 369 | 9 | 469 | 12 | 565 | 16 |
| SATA RAW | 1024k | 510 | 4 | 1022 | 6 | 1527 | 10 | 1365 | 10 | 1855 | 13 |
| SATA RAW | 8192k | 535 | 4 | 1072 | 4 | 1605 | 7 | 1880 | 4 | 1885 | 7 |
| NMVe RAW | 4k | 382 | 3 | 840 | 5 | | | | | | |
| NMVe RAW | 1024k | 2551 | 4 | 5106 | 6 | | | | | | |
| NMVe RAW | 8192k | 2971 | 3 | 5940 | 6 | | | | | | |
| Fujitsu RX2540 M4 – Samsung 980 Pro in PCIE 3.0 x4 Adapter | | | | | | | | | | | |
| NMVe RAW | 4k | 390 | 6 | 804 | 11 | | | | | | |
| NMVe RAW | 1024k | 2618 | 7 | 5286 | 13 | | | | | | |
| NMVe RAW | 8192k | 3087 | 6 | 6370 | 12 | | | | | | |

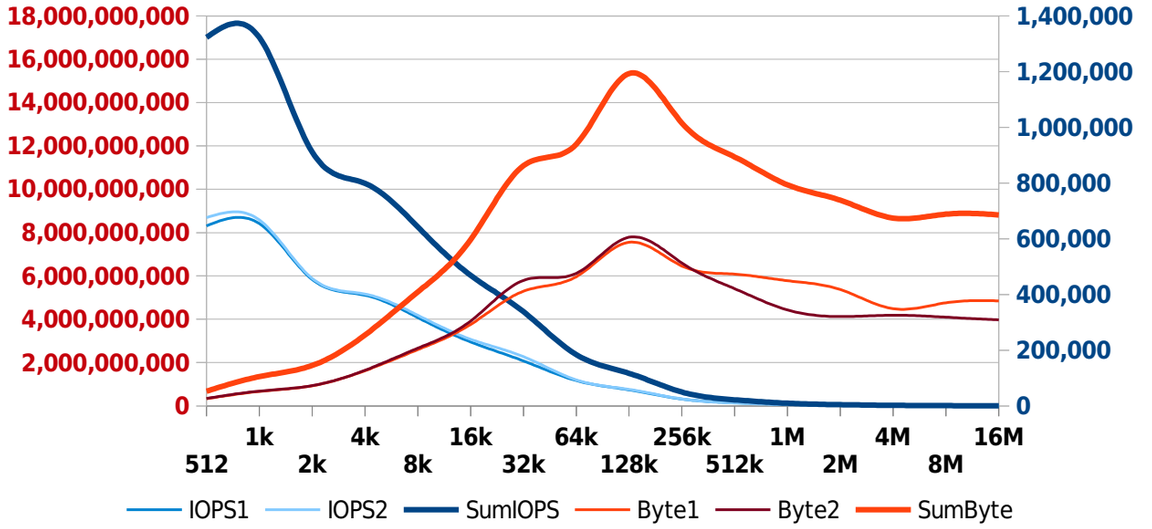
# Netzwerkperformance

Äther oder Ether?

- Tests mit 100GBit/s Ethernet
  - Single / Multipath
  - NIC ←→ Switch ←→ NIC
- Fujitsu RX2540 ←→ Supermicro HPC
  - Dual 8C/16T CPUs
  - 16 Threads
  - Blockgrößen von 512Byte bis 16MB
- Nicht linear und über Dual steigerbar
- Maximum

  Single 10,5GByte/s und 748kIOPS
  Dual 14,6GByte/s und 1,323MIOPS



# Zusammenfassung

Schnell ist möglich!

- Auch mit „älterer“ Generation von Servern sind deutliche Leistungssteigerungen möglich!
- Bis zu 1,86GByte/s per SATA-SSD unter Solaris
- Bis zu 6,3GByte/s per NVMe-SSD unter Linux
- Auslagerung von Storage über Ethernet einfach und schnell
- Netzverkehr und lokale Speicherzugriffe behindern sich nicht
- Bis zu 1,323MIOPS und 14,6GByte/s über Netzwerk

https://www.supermicro.com/en/products/system/2U/2029/SYS-2029TP-HTR.cfm | https://www.nvidia.com/en-us/networking/ | https://www.fujitsu.com/global/products/computing/servers/primergy/rack/rx2540m4/

virtualization and clustering – made simple