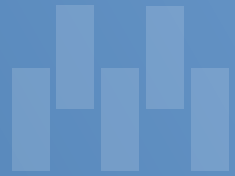
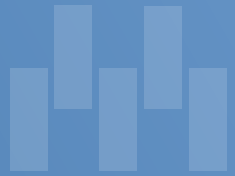
Vortrag CLT2019



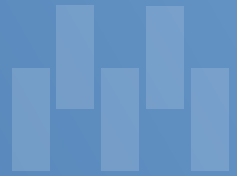
#itsatrap



- 40 Computation Nodes mit je
  - 2x Intel Xeon Gold 6150 (18C) → **1440 Kerne in Summe**
  - 384 GB RAM
  - FDR-Infiniband (56 Gbit/s)
- 1 Storage-Server („für NFS-Server“)
  - Intel Xeon Silver 4110 (Single)
  - 32 GB RAM
  - 4x 4TB HDD 3,5“ SATA im RAID10

```
root@server:~$ iostat
avg-cpu:  %user   %nice %system %iowait  %steal   %idle
          3,52    0,01    1,64    94,20    0,00   0,64

[...]
```

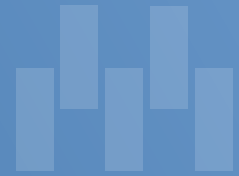


Industrie/Enterprise:

„Gute Performance, sehr gerne – aber nicht auf Kosten der Zuverlässigkeit!“

Forschung:

„Wir nehmen die 100-GbE-Dual-Port-Karte, schließen beide Ports an einen Switch an und bekommen 200 GbE quasi frei Haus. Dann sparen wir sogar das Geld für den 2. Switch und können größere CPUs einbauen!“




Hardware
Netzwerk
Software


Hardware
Netzwerk
Software

Samsung 970 Pro

Warranty	MZ-V7P512BW (512 GB) 5 Years or 600 TBW

337 GB/Tag ≙ **0,65 DWPD**

Model Code (Capacity)[1]	MZ-V7P512BW (512 GB)
General Feature	APPLICATION Client PCs
	DIMENSION (WxHxD) 80.15 X 22.15 X 2.38 (mm)
	CONTROLLER Samsung Phoenix Controller

DWPD[5]			1.3 (3 Years)	

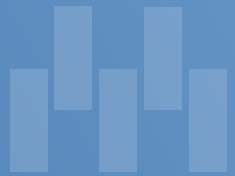
SM863a		MZ7KM240HMHQ	MZ7KM480HMHQ	MZ
Capacity[1]		240 GB	480 GB	
Performance[2]	Seq. read (128 KB)	410 MB/s	510 MB/s	
	Seq. write (128 KB)	450 MB/s	485 MB/s	
	Rand. read (4KB, QD32)	90K IOPS	95K IOPS	
	Rand. write (4KB, QD32)	10K IOPS	19K IOPS	
Average power consumption[3] (1,920 GB)			Active Re (Typ.) W, Active Write	
TBW (Terabytes written)[4]		1,540 TB	3,080 TB	
DWPD[5]			3.6 (5 Years)	

Drive
Writes
Per
Day

Enterprise SSD: ~500 MByte/s
24x Enterprise SSDs: ~12 GByte/s

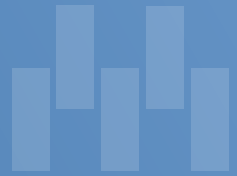
Lösung I: SAS-Expander

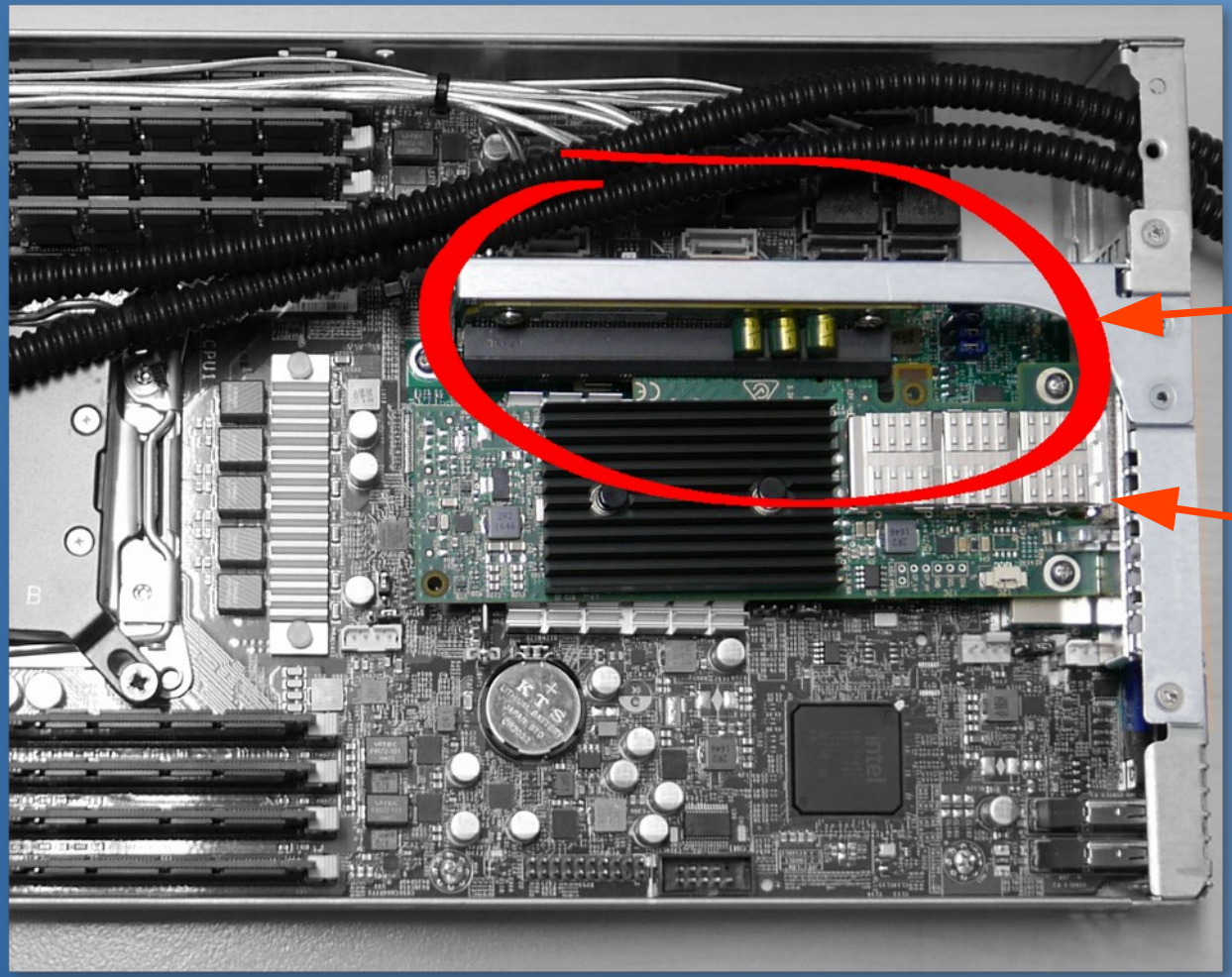
- preisgünstig
- i.d.R. 4 Ports (2 in / 2 out)
  - 1. Port (Master): 4,8 GByte/s
  - 2. Port: +10%



Lösung II: DAC-Backplane

- an PCIe angebunden
- z.B. 6 x 4 Ports
  - 4,8 Gbyte/s pro 4 Devices.
- Nachteile:
  - teurer
  - mehr PCI-Lanes benötigt





1 HE: 45mm

Riser-Card

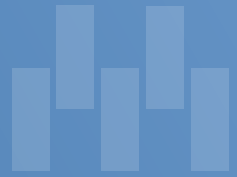
Netzwerkkarte
(Low-Profile: ~70mm)

100-GbE-Netzwerkkarte an PCIe 3

PCIe 3: $985 \frac{MByte}{s * Lane}$

x8 $985 \frac{MByte}{s * Lane} \times 8\,Lanes \times \frac{8 \frac{bit}{Byte}}{1024 \frac{M}{G}} \approx \mathbf{61{,}5\,Gbit/s}$

x16 $985 \frac{MByte}{s * Lane} \times 16\,Lanes \times \frac{8 \frac{bit}{Byte}}{1024 \frac{M}{G}} \approx \mathbf{123{,}1\,Gbit/s}$



NVMe(-SSD)
U.2-Format

(SATA-)**SSD**

NVMe(-SSD)
M.2-Format

+ HHHL-PCIe-Karte

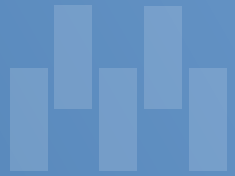
Testsystem

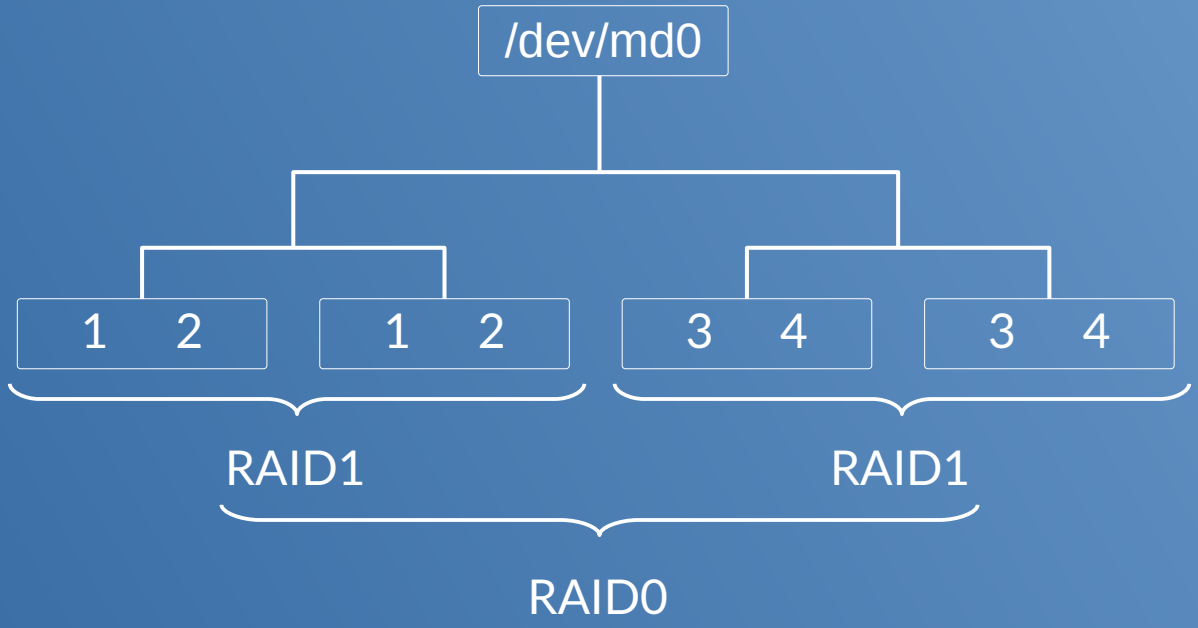
- 2x Intel Xeon Silver 4116 (12C@2,1GHz)
- 64 GB RAM
- Mellanox Connect-X5 Dual 100 GbE-Karte (am PCIe x16)

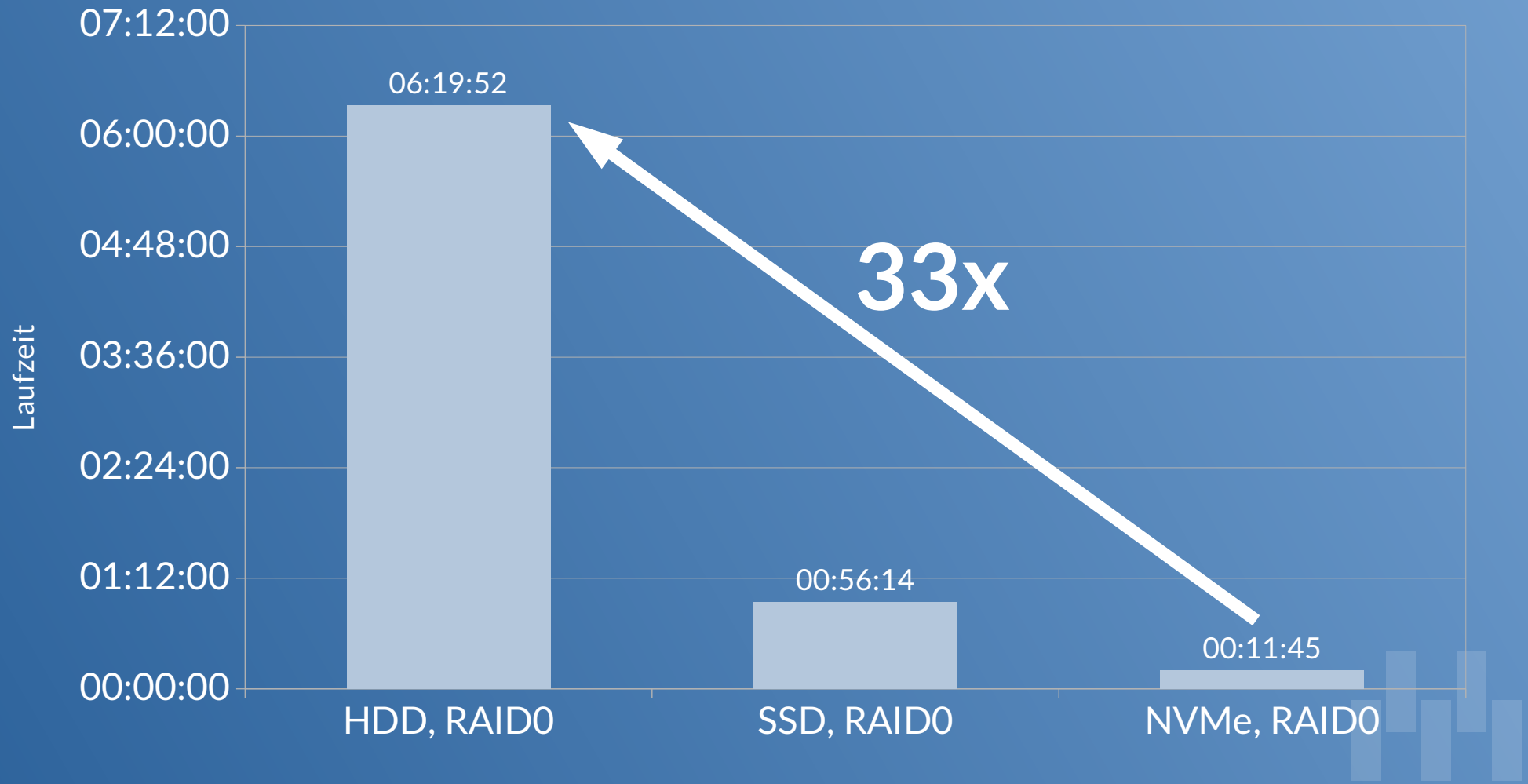
4x

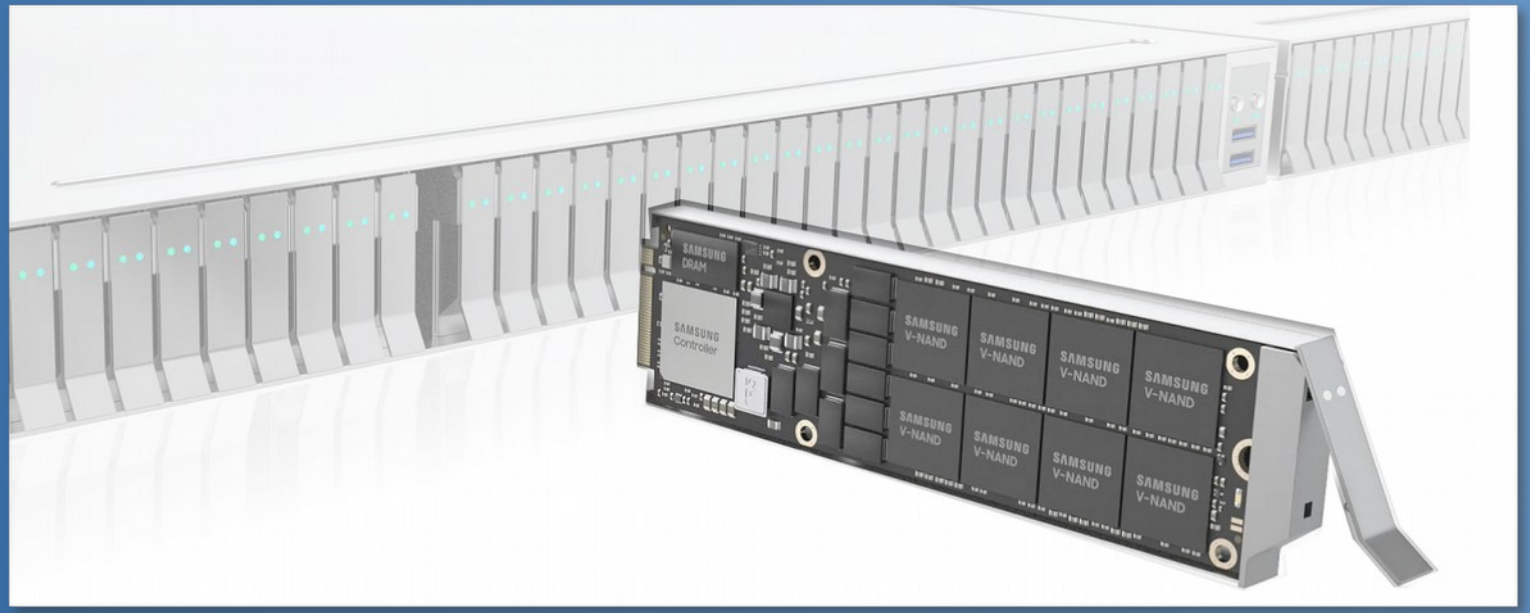
```
fio --rw={write,randwrite,read,randread} --size={1,10,100,1000,10000}M --numbjobs={1,3,5,10} --bs=1M \
    --fallocate=none --refill_buffers --direct=1
```

Mittelwert aus 3 Durchläufen
Je *BW* und *IOPS* gespeichert → 480 Messergebnisse
Single-Disk + RAID0




/dev/md0
1 2
1 2
3 4
3 4
RAID1
RAID1
RAID0


07:12:00
06:00:00
04:48:00
03:36:00
02:24:00
01:12:00
00:00:00
Laufzeit
06:19:52
00:56:14
00:11:45
33x
HDD, RAID0
SSD, RAID0
NVMe, RAID0



NVMe(-SSD)
NF1-Format

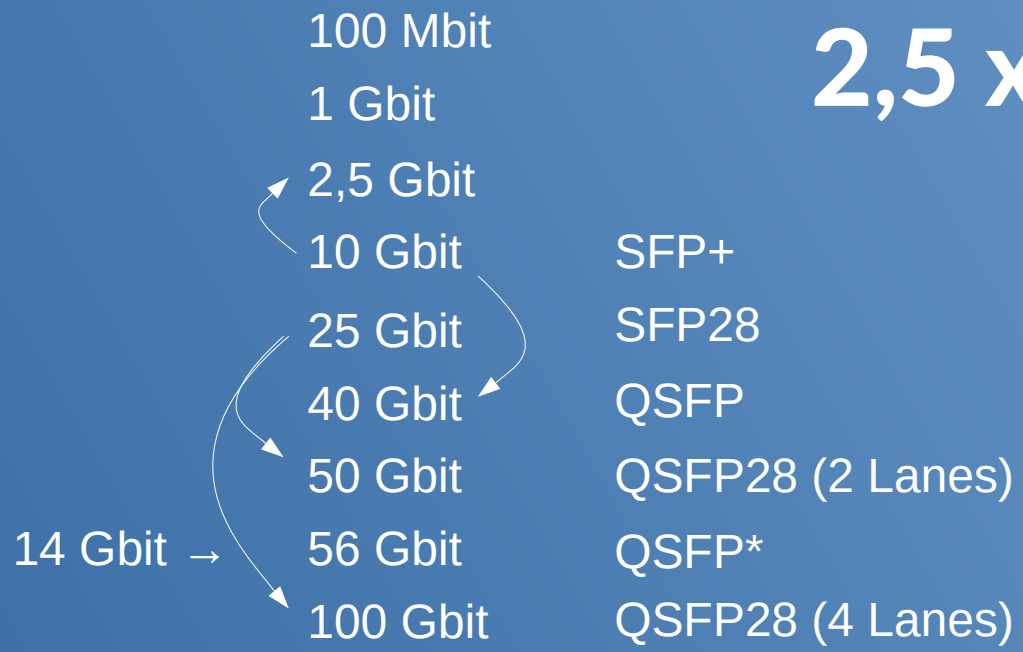
https://www.samsung.com/semiconductor/ssd/nf1-ssd/


Hardware
Netzwerk
Software


Hardware
Netzwerk
Software

Netzwerkgeschwindigkeiten bei Ethernet

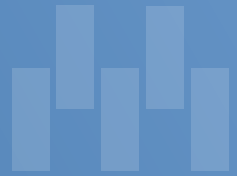
Oder warum 25 GbE cooler ist als 40 GbE



2,5 x 4 = 10

Infiniband

- Bussystem → HPC-Bereich
- geringe Latenz (max 2µs) → Protokollstack wird in Netzwerkhardware ausgelagert
- kurze Strecken
- zwei Firmen: Mellanox (Nvidia) und Intel
- nutzt QSFP-Ports :-)
- QDR (32 Gbit/s), FDR (56 Gbit/s), EDR (100 Gbit/s)
- RDMA

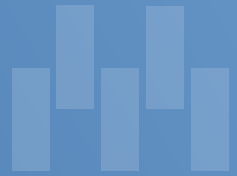


RDMA

Prolog: DMA

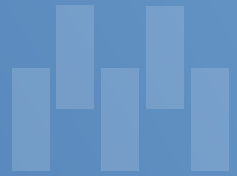
RDMA

- Erweitert DMA um Netzwerk
- Kommt ursprünglich aus der Infiniband-Welt
- auch u.a. für Ethernet verfügbar (RoCE)
- Braucht eine „verlustfreie“ Übertragung (10^{-15})



Fibre Channel

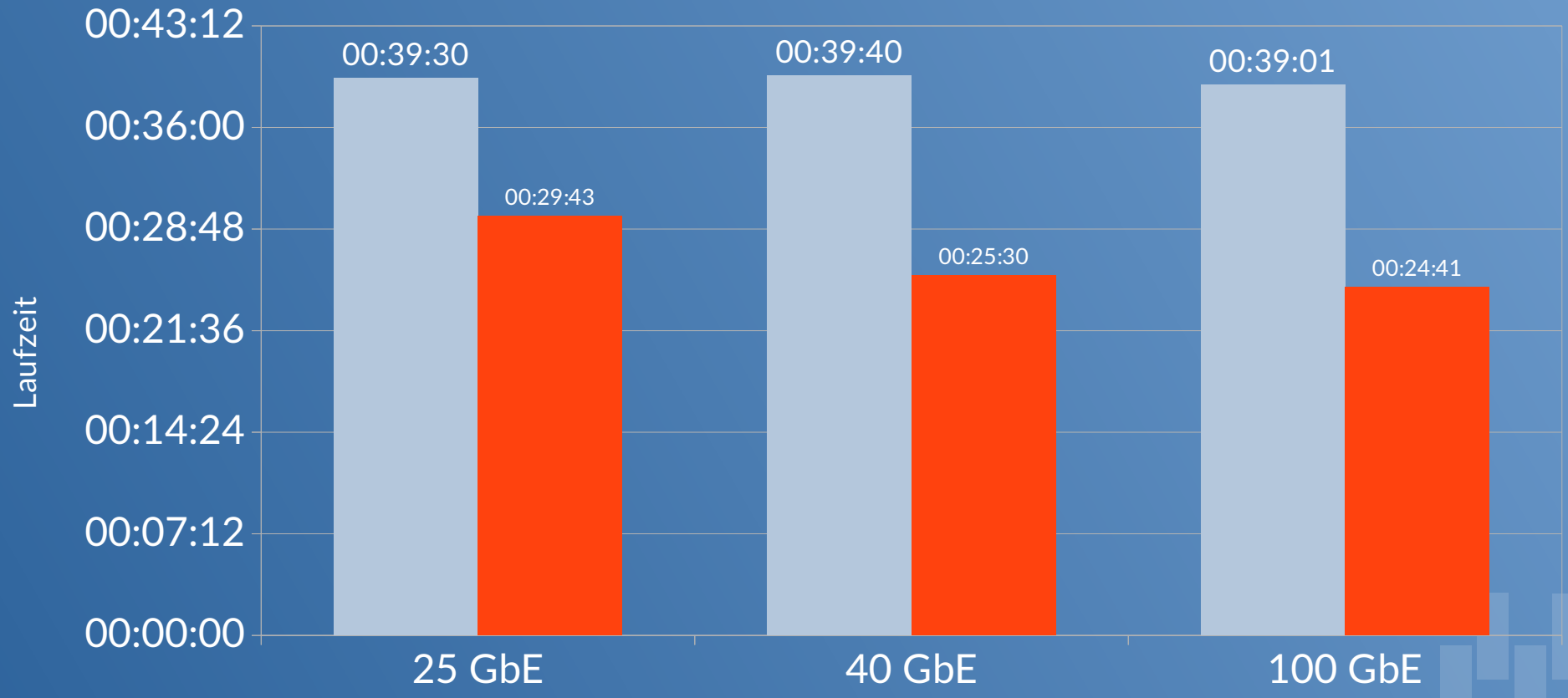
- Schnittstelle für Speichernetzwerke → SAN
- Heute üblich: 4-32 Gbit/s
- Garantierte Latenz, garantierte in-order-delivery




Hardware
Netzwerk
Software

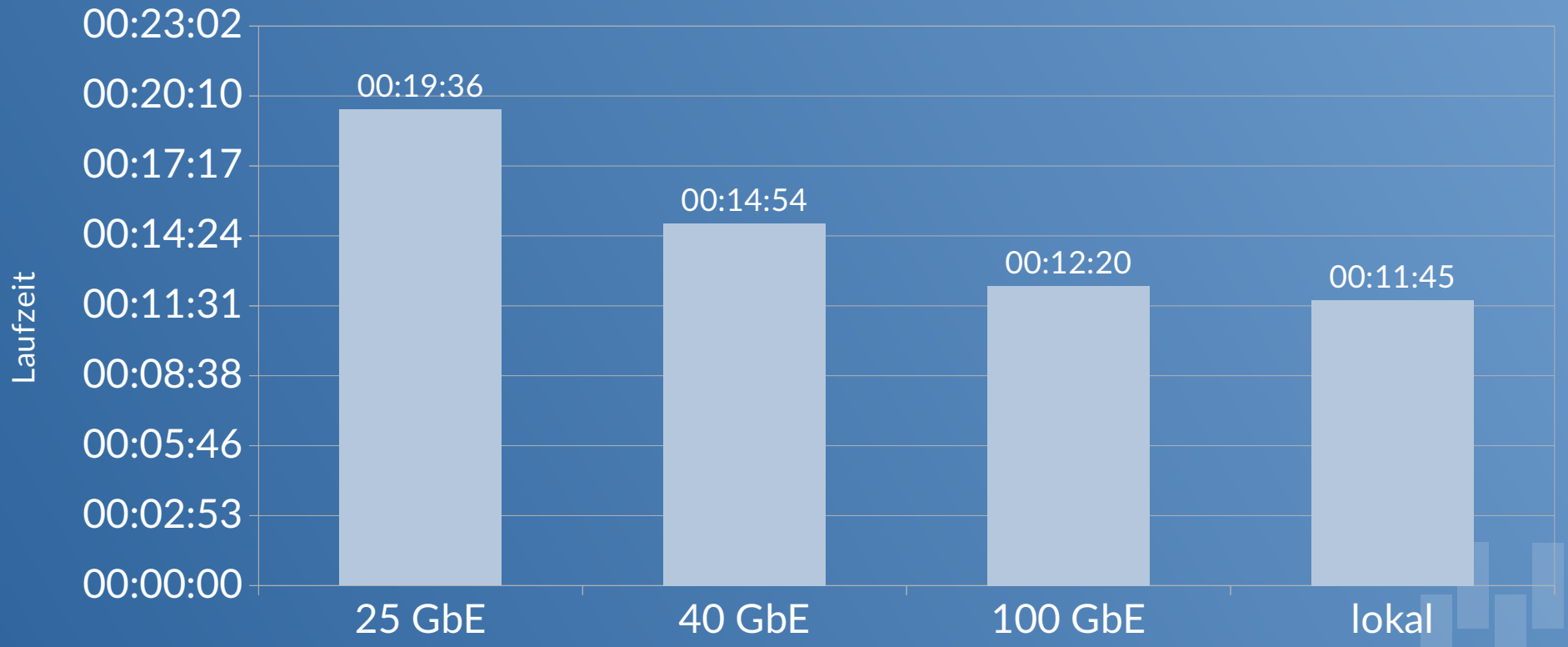

Hardware
Netzwerk
Software

NFS over RDMA



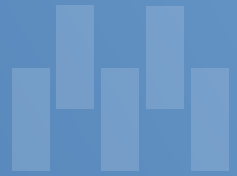
NVMe over Fabrics

(3x NVMe-SSD im RAID0)

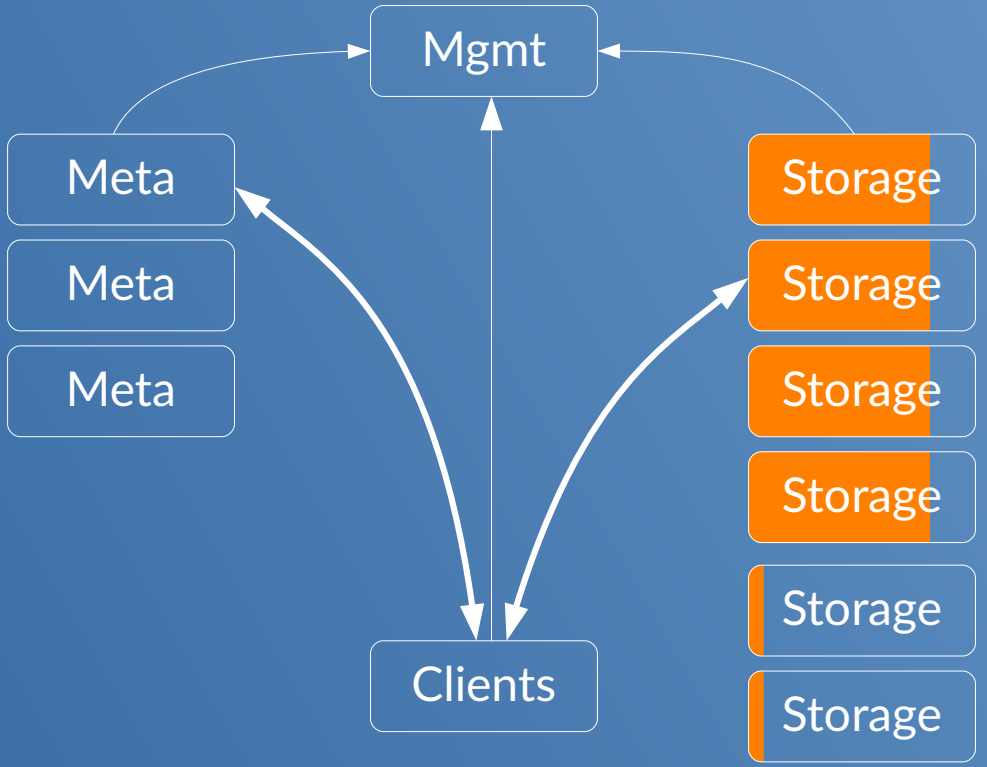


Ceph und High Performance

- Verteilt + kein SpoF (Crush-Algorithmus)
- läuft wunderbar mit Linux
- „Zuverlässigkeit auf unzuverlässiger Hardware“
  - nicht auf High-End-Hardware optimiert
- kein Infiniband-Support, RDMA-Support nicht optimal
- Herausforderung bei vielen Zugriffen auf die gleichen Daten



HPC-Dateisystem: **BeeGFS**



Menzel IT GmbH
Charlottenburger Str. 33A
13086 Berlin

030 5130 444 00
post@menzel-it.net