The Fastest Database System: How to Outpace a Shadow

- **mag. Sergej Rožman**; Abakus plus d.o.o.

- The latest version of this document is available at:

  http://www.abakus.si/

The Acme of Rocket Science

Abakus

As na disku.

The Fastest Database System
How to Outpace a Shadow

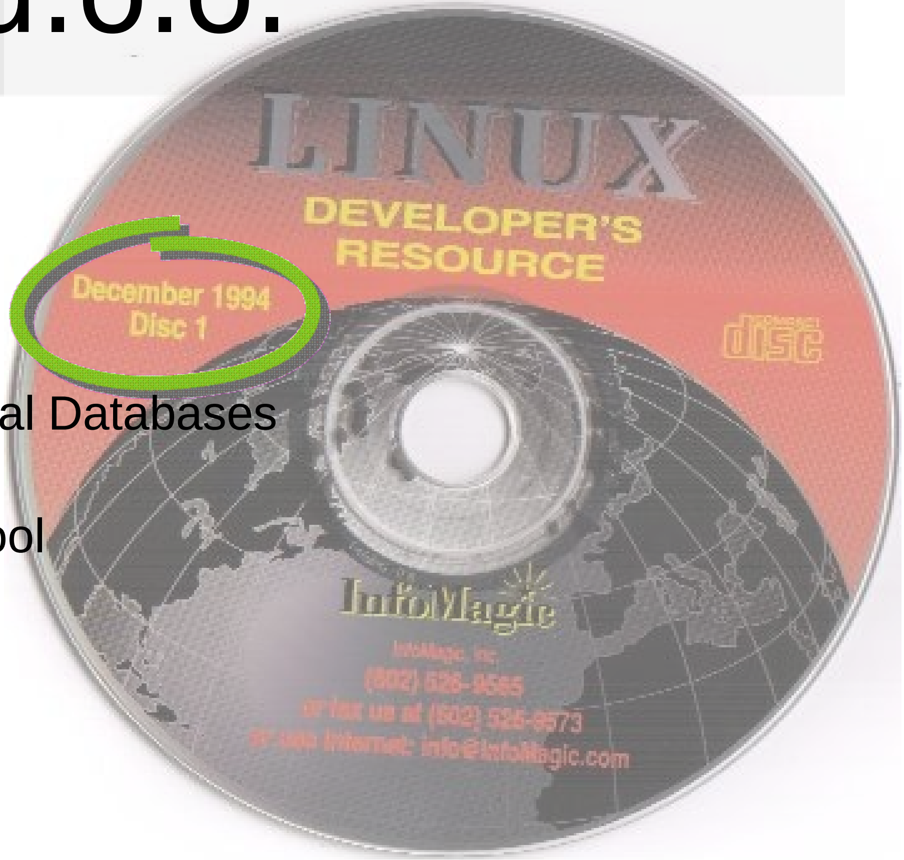
mag. Sergej Rožman

sergej.rozman@abakus.si

Make IT

2024

Abakus plus d.o.o.



History

- from 1992, ~20 employees

Applications:

- DejaVu - High Performance Architecture for Virtual Databases
- ARBITER – the ultimate tool in audit trailing
- APPM – Abakus Plus Performance Monitoring Tool

Services:

- DBA, OS administration , programming (Oracle)

Infrastructure:

- servers, SAN storage, UPS, firewalls, backup servers, virtualization

Skills & Experience:

- from 1995 GNU/Linux ***(~30 years of experience !)***
- Oracle on GNU/Linux: since RDBMS 7.1.5 & Forms 3.0 ***(before Oracle !)***
- **~35 years of experience with High-Availability !**

Customers

Gorenjska Banka

GENERALI GROUP GENERALI Zavarovalnica

Ljubljana Airport

EKDIS Ekspresno. Ekonomično.

REPUBLIKA SLOVENIJA MINISTRSTVO ZA OBRAMBO

NOVA BANKA

MILENIJUM® OSIGURANJE

KONTROLA ZRAČNEGA PROMETA SLOVENIJE

Iskra®

Hidria

Mestna občina Ljubljana

skbbanka otp group

triglav

ANDRITZ

jata emona LJUBLJANA

UNIVERZITETNA PSIHIATRIČNA KLINIKA LJUBLJANA University Psychiatric Clinic Ljubljana

BANKA SLOVENIJE

SAVA INFOND

MERKUR

TRELLEBORG

Trelleborg Slovenija, d.o.o.

SODO SISTEMSKI OPERATER DISTRIBUCIJSKEGA OMREŽJA Z ELEKTRIČNO ENERGIJO

NLB Vita Življenjska zavarovalnica

PRVA

MAGNETIK d.o.o.

ISS PEST MANAGEMENT SOLUTIONS

Mercator

MM KARTON

studio ritem

Blubit

Tiko TOVARNA KOVINSKE OPREME

ZAVOD ZA ŠPORT RS PLANICA

PH Primorska hranilnica

GOODYEAR DUNLOP SAVA TIRES

SAVA HOTELS & RESORTS

CENTROSINERGIJA

PANTEON® GROUP

Lonia

PRONET CHOOSE THE FUTURE

hit alpinea Kranjska Gora

LASERLINE

DRINA OSIGURANJE

ROS d.o.o.

INFOTRANS®

PARK POSTOJNSKA JAMA

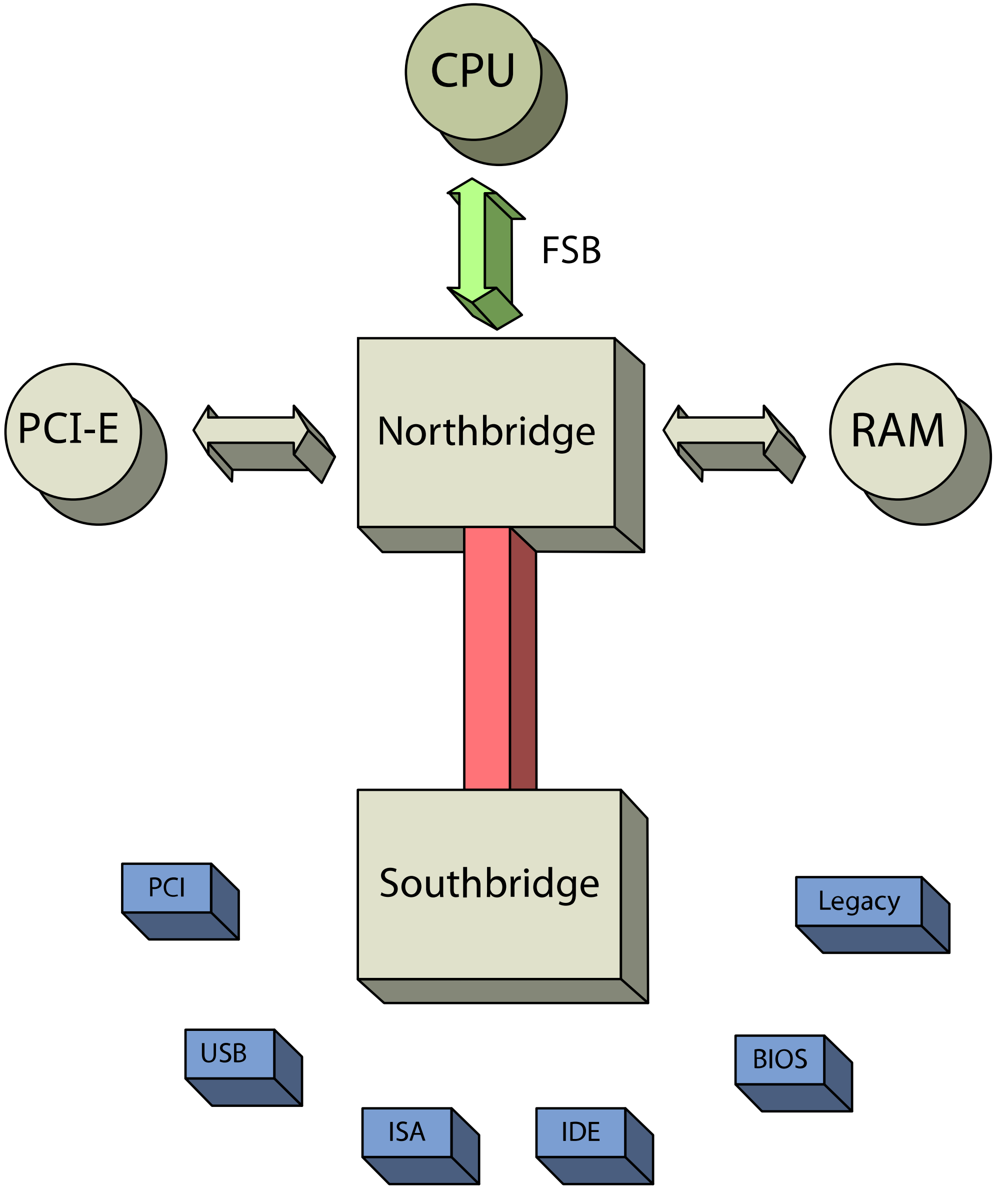
ADRIA ANKARAN HOTEL & RESORT

How to Build the Fastest DB Server?

Recipe:

- get a decent CPU (https://www.cpu-world.com/)
- use fast RAM
- take top-notch disk storage

Traditional x86 Computer Topology



CPU

high-speed devices

low-speed devices

source: https://en.wikipedia.org/wiki/Northbridge_(computing)

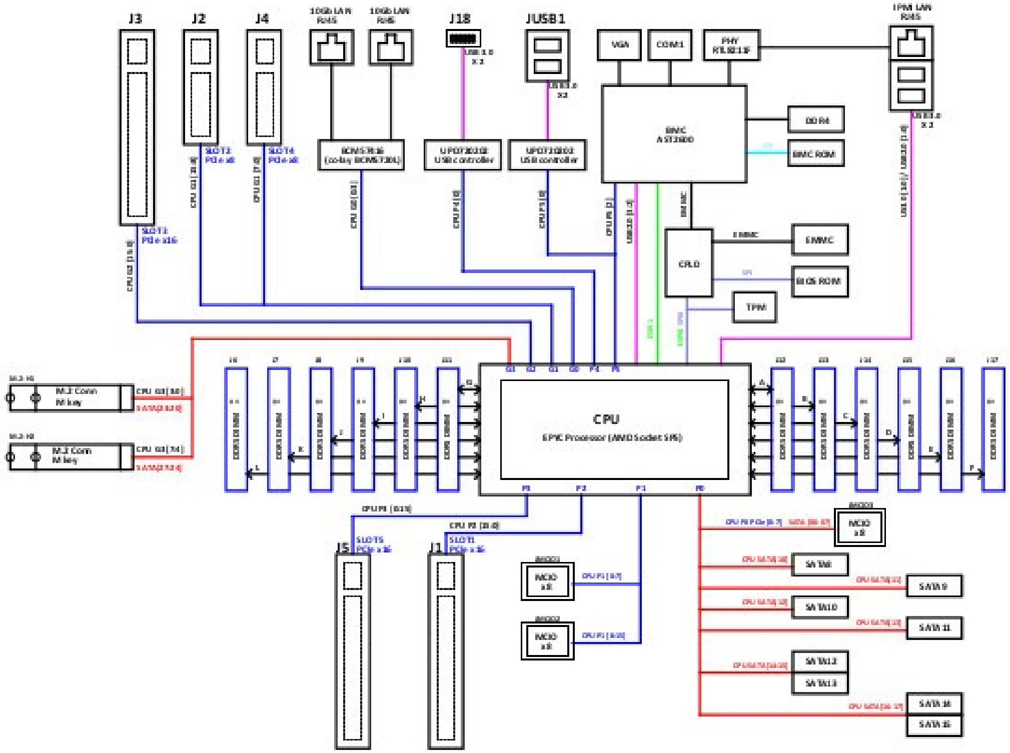
Modern x86-64 Computer Topology (NUMA)

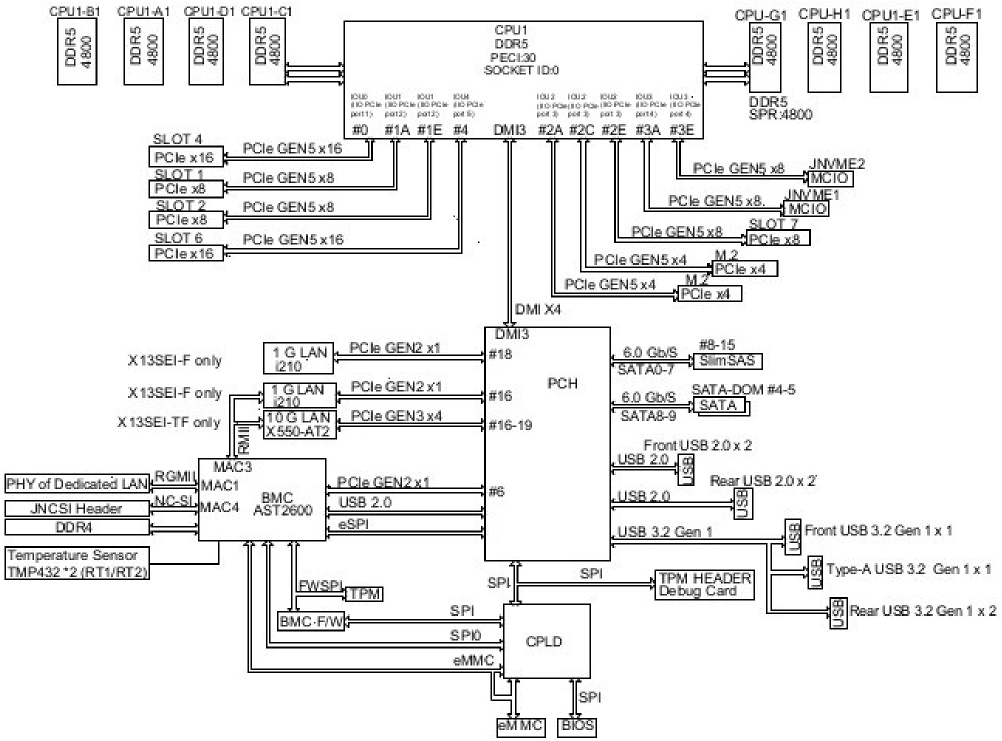
AMD EPYC

- SoC – System on a chip
- Supermicro H13SSL-NT

Intel XEON

- Supermicro X13SEI-TF





As na disku.

source: https://www.supermicro.com/support/resources/

RAM

Type	Throughput (MB/s)	Introduced
DDR-400	3200	1998
DDR2-800	6400	2003
DDR3-1600	12800	2007
DDR4-3200	25600	2014
DDR5-4800	38400	2020
DDR6-?	?	2026(?)

source: https://en.wikipedia.org/wiki/DDR_SDRAM

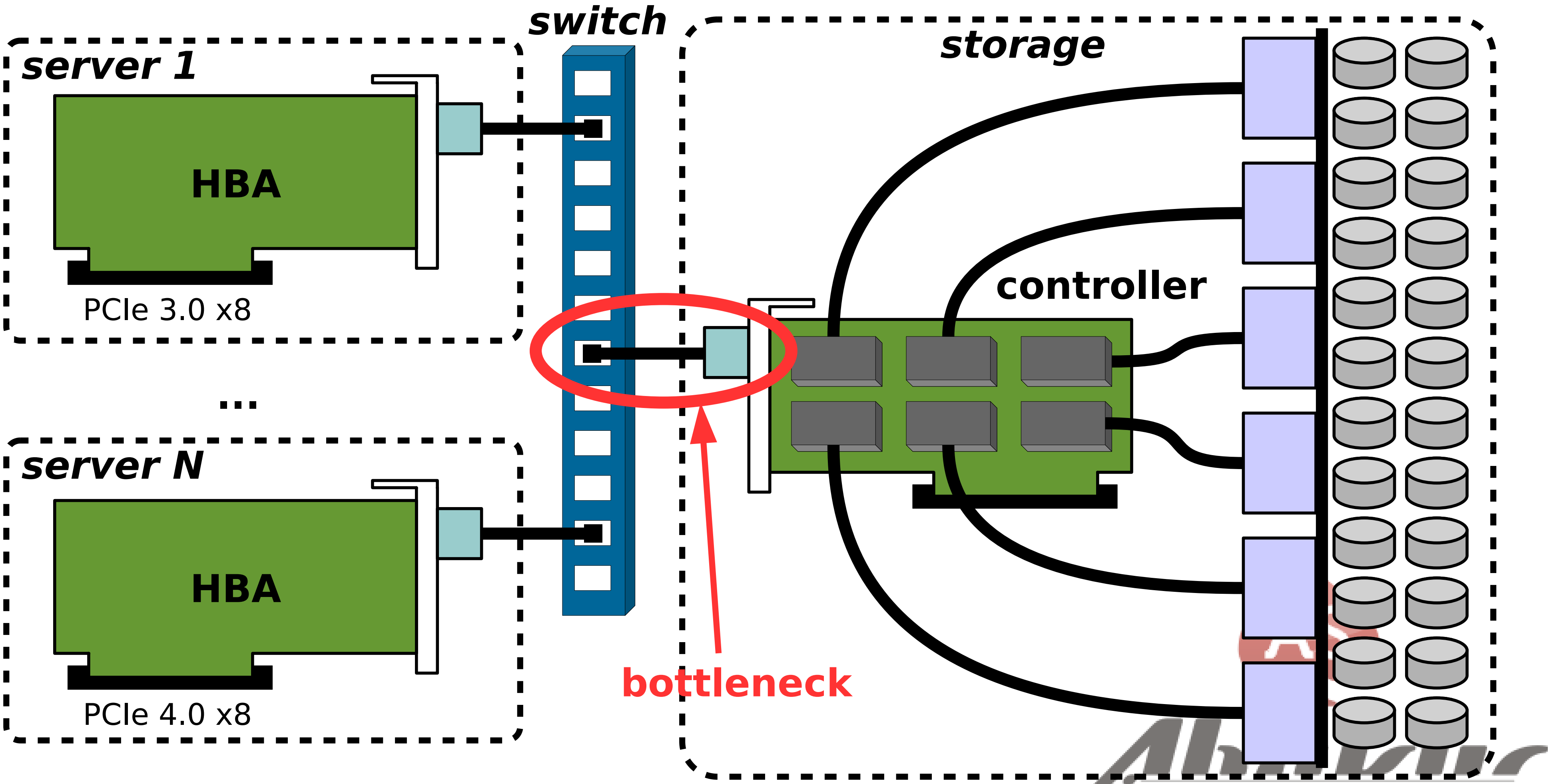
Buzzword

We have bought »all-flash (SAN) storage«.

- Which type? QLC performs badly.
- Are you using RAID5|6 again?
- How is »all-flash storage« connected to the host?

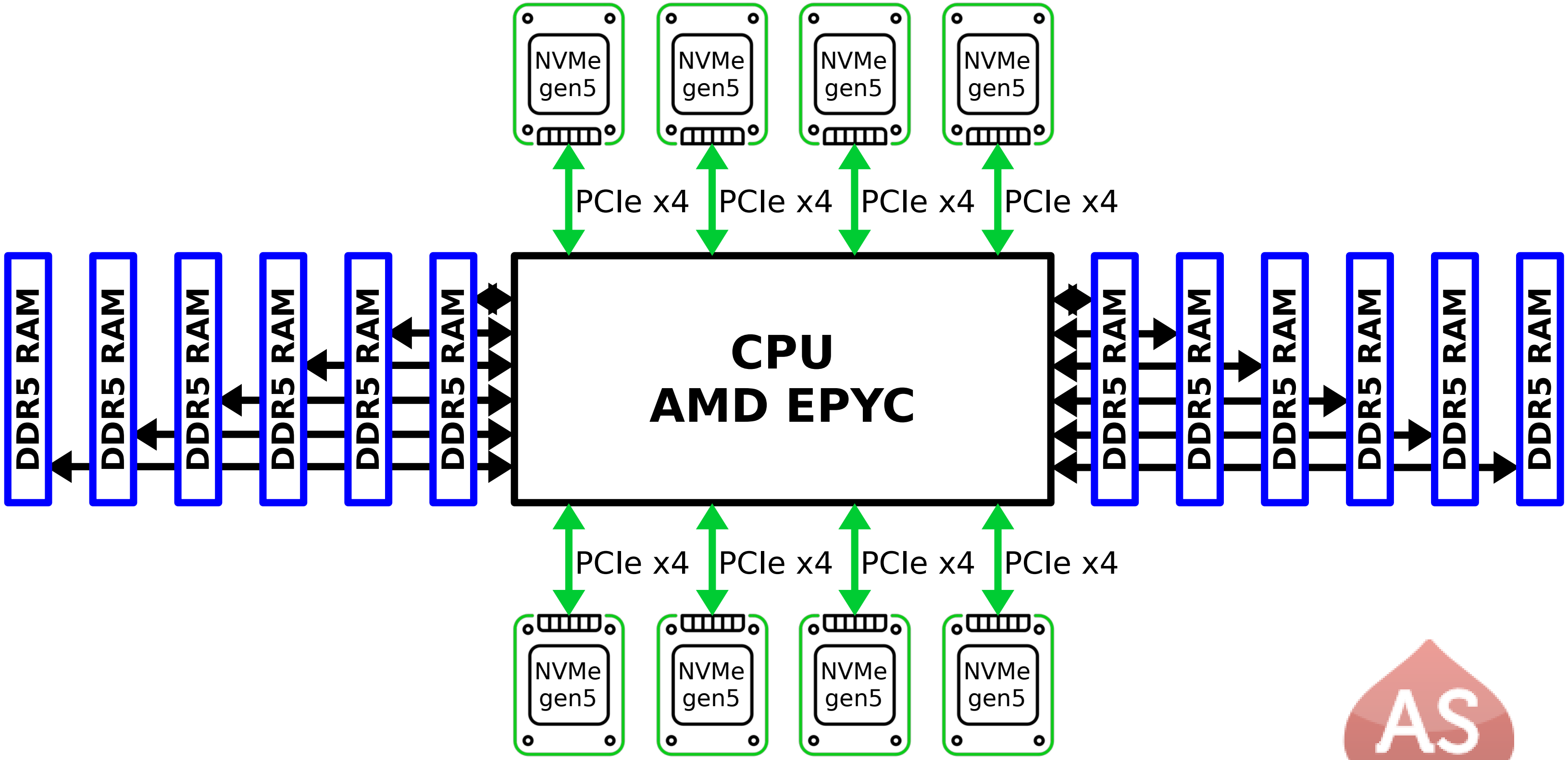
Inevitable Fact

Shared storage always leads to contention.



PCIe	switch	disks
8 + … + 16 = 24 + ...	16	24 x 1.2 = 28.8

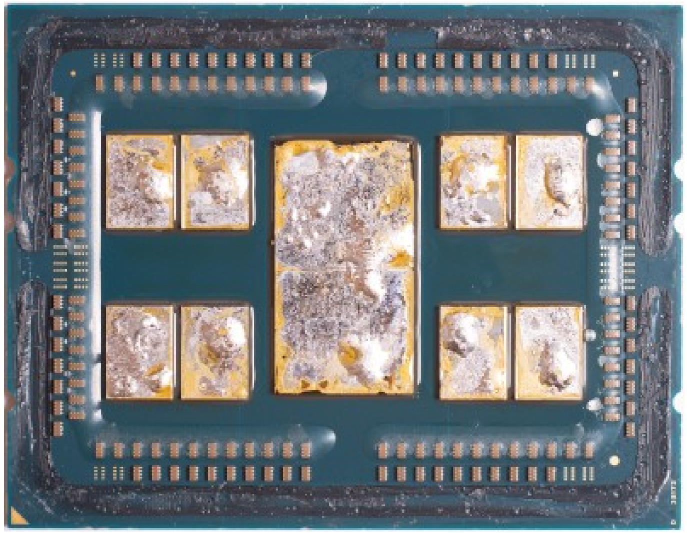
NVMe – Non-Volatile Memory Express

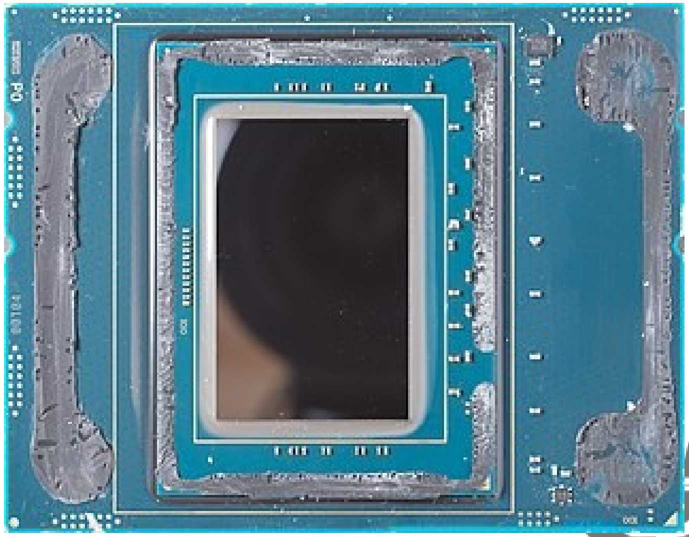


CPU Features

AMD EPYC	
Cores	<128
Memory controllers	<12
PCIe lanes	128

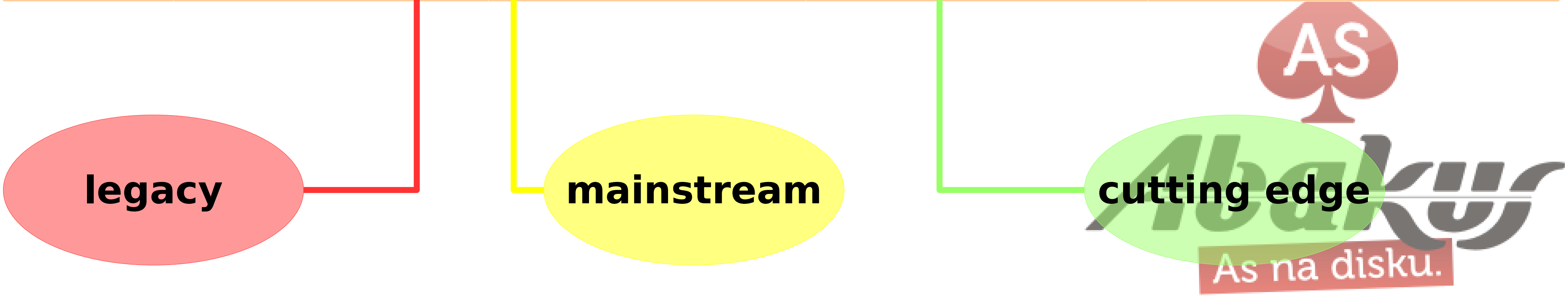
Intel XEON	
Cores	<64
Memory controllers	<8
PCIe lanes	80





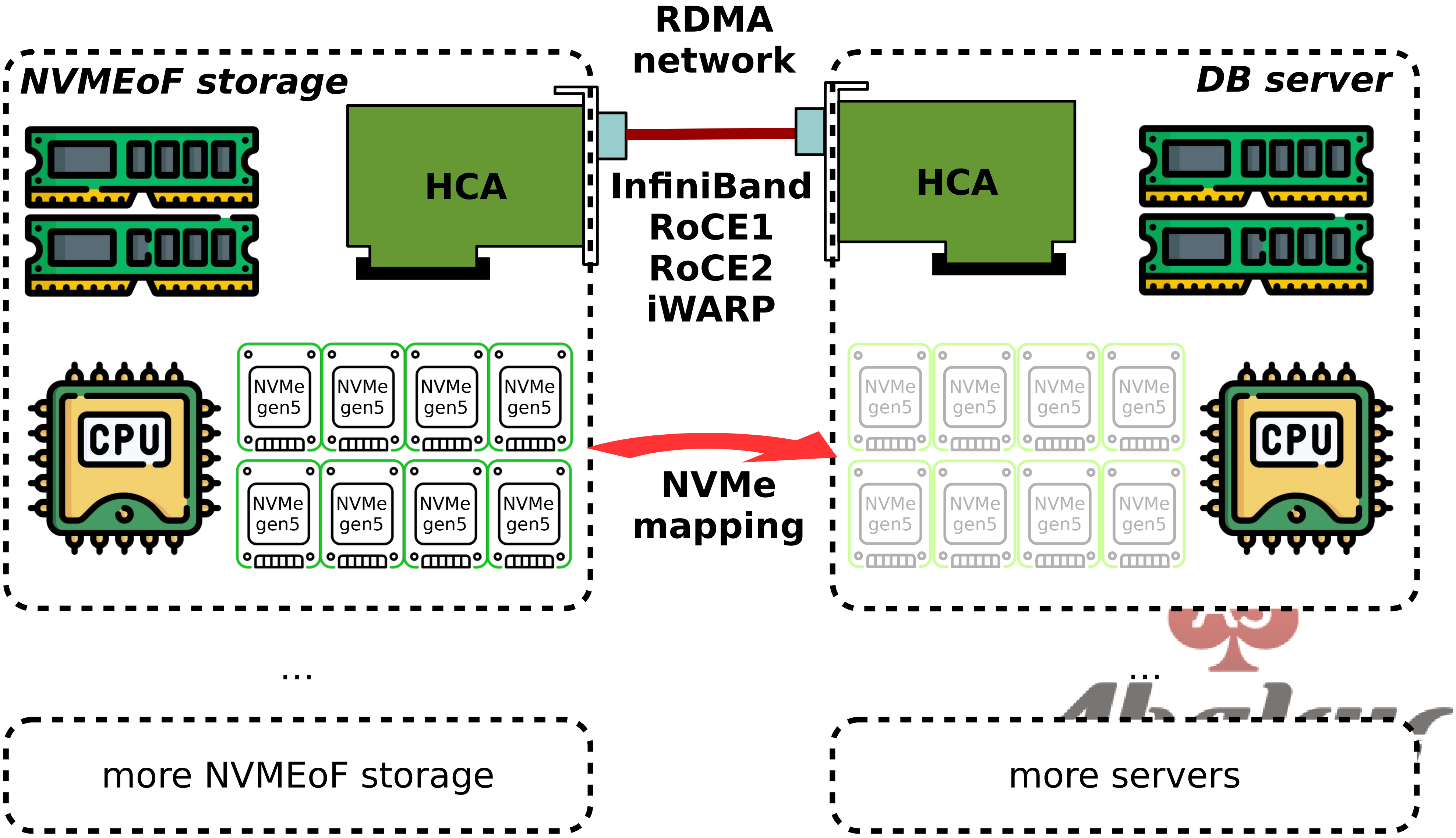
PCIe & NVMe

Version	Throughput x1 (GB/s)	Throughput x4 (GB/s)	Throughput x16 (GB/s)	Introduced
1.0	0,25	1	4	2003
2.0	0,5	2	8	2007
3.0	1	4	16	2010 (NVMe ~2015)
4.0	2	8	32	2017 (NVMe ~2019)
5.0	4	16	64	2019 (NVMe ~2022)
6.0	7,5	30	120	2022
7.0	15	60	240	2025 (planned)



source: https://en.wikipedia.org/wiki/PCI_Express & chatGPT

NVMEoF – NVMe Over Fabric



Remote Storage (SAN)

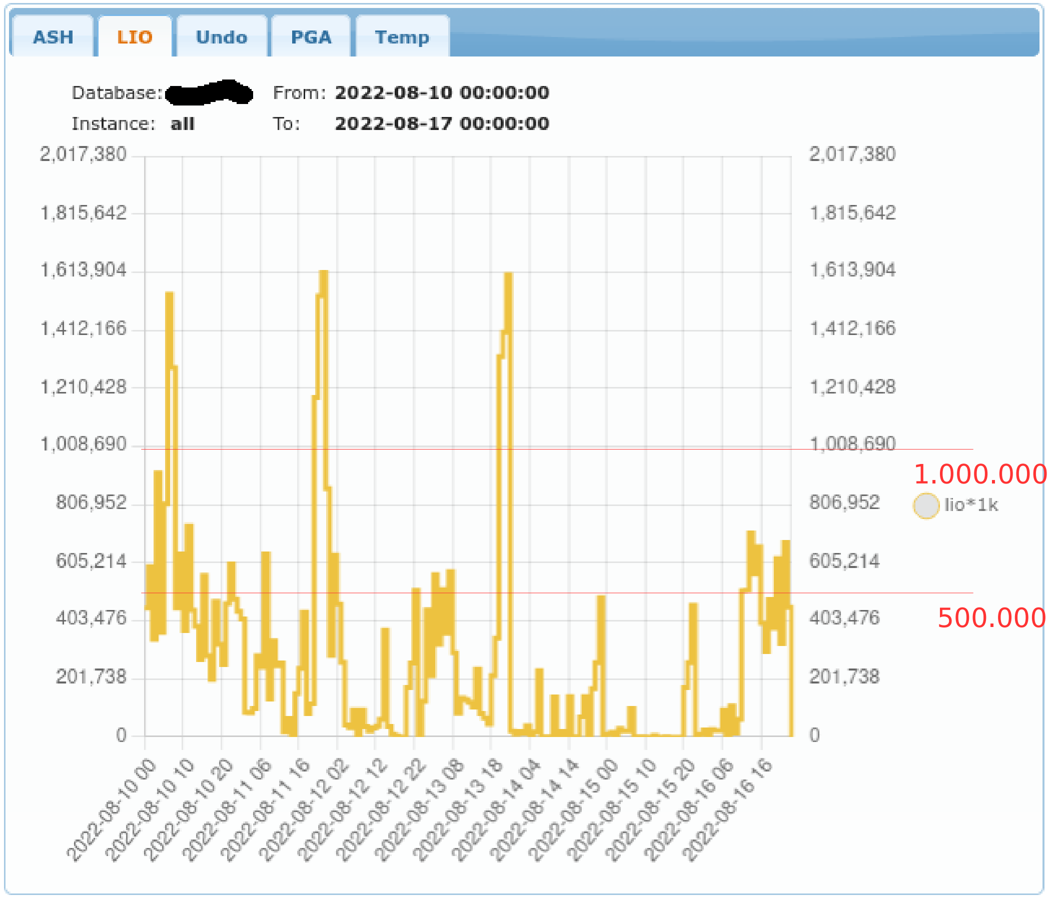
Type	Characteristics
FC (Fibre Channel)	low throughput, expensive
iSCSI (tcp)	increased latency
CEPH	scalable, featureful, but slow
NVMEoF (rdma)	high througput, low latency, no features at all, perfect for ASM

NVMEoF Configuration

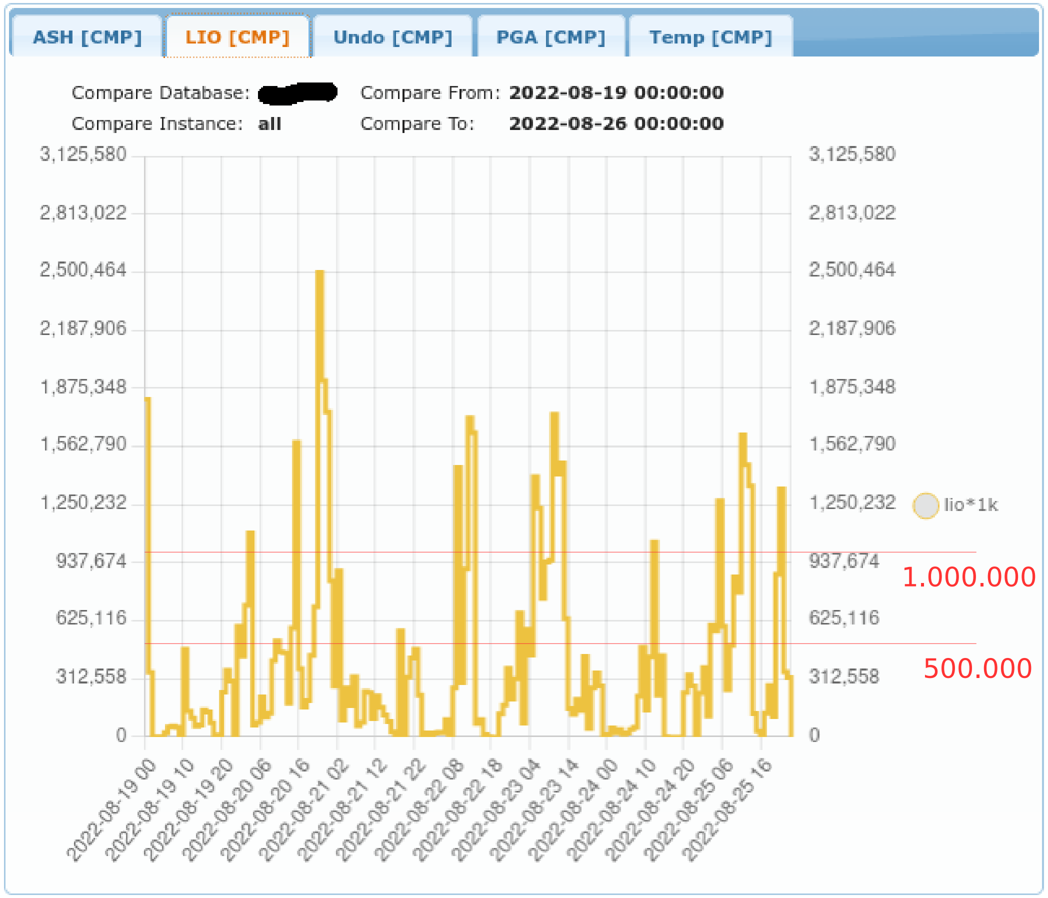
```
# storage
mkdir /sys/kernel/config/nvmet/subsystems/aba1
echo 1 > /sys/kernel/config/nvmet/subsystems/aba1/attr_allow_any_host
# echo 1 > /sys/kernel/config/nvmet/subsystems/aba1/attr_offload  # offloading is not stable
echo aba1 > /sys/kernel/config/nvmet/subsystems/aba1/attr_model

mkdir /sys/kernel/config/nvmet/subsystems/aba1/namespaces/1
echo -n /dev/nvme0n1 > /sys/kernel/config/nvmet/subsystems/aba1/namespaces/1/device_path
echo 1 > /sys/kernel/config/nvmet/subsystems/aba1/namespaces/1/enable

mkdir /sys/kernel/config/nvmet/ports/1
echo 4420 > /sys/kernel/config/nvmet/ports/1/addr_trsvcid
echo 192.168.250.1 > /sys/kernel/config/nvmet/ports/1/addr_traddr
echo "rdma" > /sys/kernel/config/nvmet/ports/1/addr_trtype
echo "ipv4" > /sys/kernel/config/nvmet/ports/1/addr_adrfam
ln -s /sys/kernel/config/nvmet/subsystems/aba1/ /sys/kernel/config/nvmet/ports/1/subsystems/aba1
```

```
# server
modprobe nvme-rdma
nvme discover -t rdma -a 192.168.250.1 -s 4420
nvme connect -t rdma -n aba1 -a 192.168.250.1 -s 4420
```

Migration: Exadata → Server



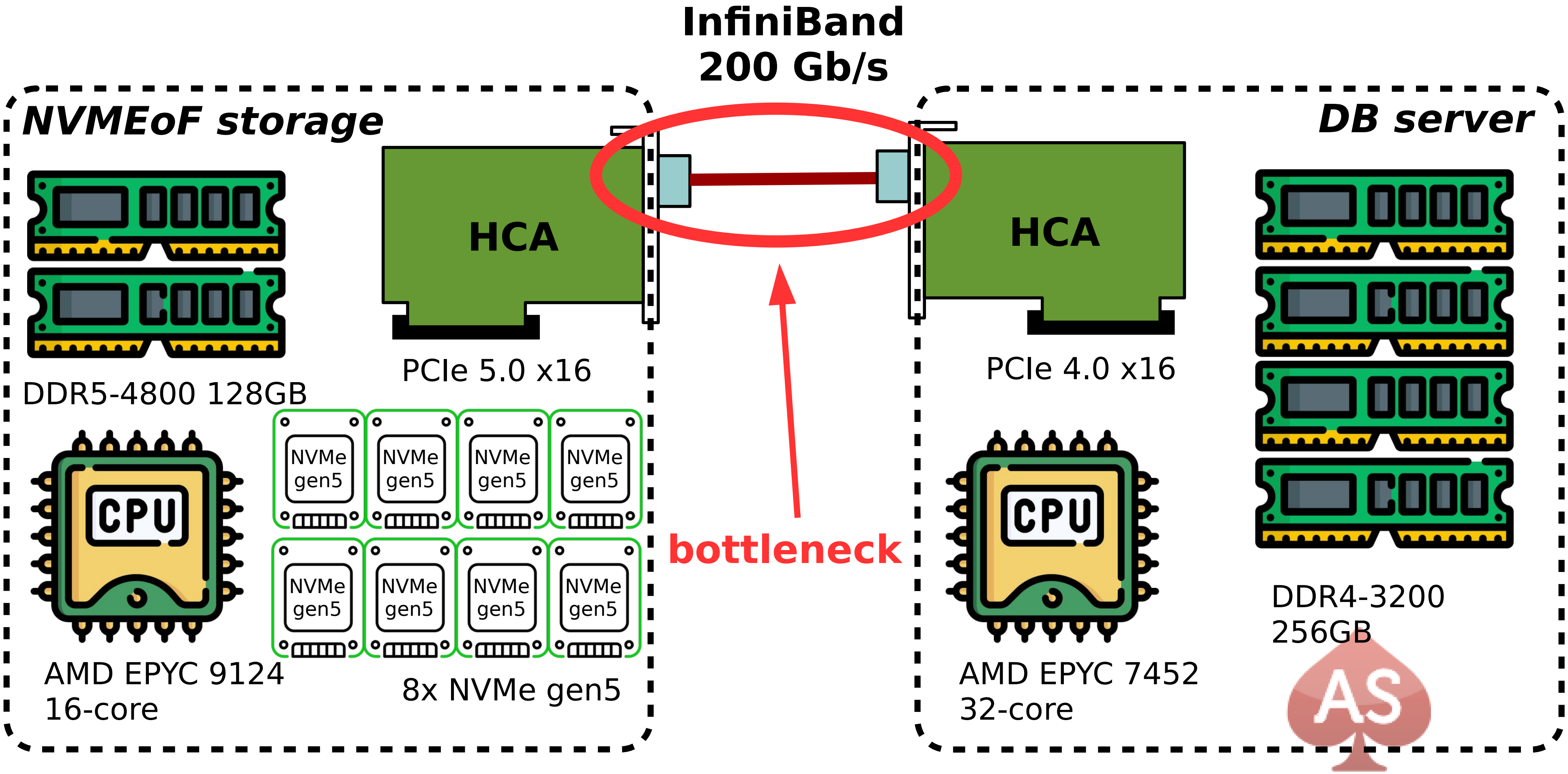
Database: [redacted] From: **2022-08-10 00:00:00**
Instance: **all** To: **2022-08-17 00:00:00**



Compare Database: [redacted] Compare From: **2022-08-19 00:00:00**
Compare Instance: **all** Compare To: **2022-08-26 00:00:00**

Abakus
As na disku.

Test System



InfiniBand
200 Gb/s
NVMEoF storage
HCA
DDR5-4800 128GB
PCIe 5.0 x16
CPU
NVMe gen5
NVMe gen5
NVMe gen5
NVMe gen5
NVMe gen5
NVMe gen5
NVMe gen5
NVMe gen5
AMD EPYC 9124
16-core
8x NVMe gen5
bottleneck
DB server
HCA
PCIe 4.0 x16
CPU
DDR4-3200
256GB
AMD EPYC 7452
32-core

Sequential Read/Write On the Storage Server

```
# fio nvme-seq-read.fio
...
Jobs: 10 (f=10): [R(10)][100.0%][r=54.1GiB/s][r=222k IOPS][eta 00m:00s]
...
```

```
# fio nvme-seq-write.fio
...
Jobs: 10 (f=0): [f(10)][100.0%][w=48.6GiB/s][w=199k IOPS][eta 00m:00s]
...
```

```
# cat nvme-seq-read.fio
[global]
name=nvme-seq-read
time_based
ramp_time=5
runtime=30
readwrite=read
bs=256k
ioengine=libaio
direct=1
numjobs=10
iodepth=32
group_reporting=1

[nvme]
filesize=10G
filename=test
```

```
# cat nvme-seq-write.fio
[global]
name=nvme-seq-read
time_based
ramp_time=5
runtime=30
readwrite=write
bs=256k
ioengine=libaio
direct=1
numjobs=10
iodepth=32
group_reporting=1

[nvme]
filesize=10G
filename=test
```

Read/Write 4k IOPS On the Storage Server

```
# fio nvme-rand-read.fio
...
Jobs: 16 (f=1): [f(6),r(1),f(9)][100.0%][r=11.5GiB/s][r=3009k IOPS][eta 00m:00s]
...
```

```
# fio nvme-rand-write.fio
...
Jobs: 16 (f=0): [f(16)][100.0%][w=7802MiB/s][w=1997k IOPS][eta 00m:00s]
...
```

```
# cat nvme-rand-read.fio
[global]
name=nvme-rand-read
time_based
ramp_time=5
runtime=30
readwrite=randread
random_generator=lfsr
bs=4k
ioengine=libaio
direct=1
numjobs=16
iodepth=16
group_reporting=1

[nvme]
filesize=10G
filename=test
```

```
# cat nvme-rand-write.fio
[global]
name=nvme-rand-read
time_based
ramp_time=5
runtime=30
readwrite=randwrite
random_generator=lfsr
bs=4k
ioengine=libaio
direct=1
numjobs=16
iodepth=16
group_reporting=1

[nvme]
filesize=10G
filename=test
```

Sequential Read/Write On the DB Server

```
# fio nvme-seq-read.fio
...
Jobs: 10 (f=10): [R(10)][100.0%][r=16.9GiB/s][r=69.2k IOPS][eta 00m:00s]
...
```

```
# fio nvme-seq-write.fio
...
Jobs: 10 (f=10): [W(10)][100.0%][w=19.2GiB/s][w=78.8k IOPS][eta 00m:00s]
...
```

```
# cat nvme-seq-read.fio
[global]
name=nvme-seq-read
time_based
ramp_time=5
runtime=30
readwrite=read
bs=256k
ioengine=libaio
direct=1
numjobs=10
iodepth=32
group_reporting=1

[nvme]
filesize=10G
filename=test
```

```
# cat nvme-seq-write.fio
[global]
name=nvme-seq-read
time_based
ramp_time=5
runtime=30
readwrite=write
bs=256k
ioengine=libaio
direct=1
numjobs=10
iodepth=32
group_reporting=1

[nvme]
filesize=10G
filename=test
```

Read/Write 4k IOPS On the DB Server

```
# fio nvme-rand-read.fio
...
Jobs: 16 (f=16): [r(16)][100.0%][r=6416MiB/s][r=1642k IOPS][eta 00m:00s]
...


# fio nvme-rand-write.fio
...
Jobs: 16 (f=16): [w(16)][100.0%][w=4654MiB/s][w=1191k IOPS][eta 00m:00s]
...
```

```
# cat nvme-rand-read.fio
[global]
name=nvme-rand-read
time_based
ramp_time=5
runtime=30
readwrite=randread
random_generator=lfsr
bs=4k
ioengine=libaio
direct=1
numjobs=16
iodepth=16
group_reporting=1

[nvme]
filesize=10G
filename=test
```

```
# cat nvme-rand-write.fio
[global]
name=nvme-rand-read
time_based
ramp_time=5
runtime=30
readwrite=randwrite
random_generator=lfsr
bs=4k
ioengine=libaio
direct=1
numjobs=16
iodepth=16
group_reporting=1

[nvme]
filesize=10G
filename=test
```

DBMS_RESOURCE_MANAGER.calibrate_io

```
SQL> SET SERVEROUTPUT ON
SQL> DECLARE
  l_latency  PLS_INTEGER;
  l_iops     PLS_INTEGER;
  l_mbps     PLS_INTEGER;
BEGIN
   DBMS_RESOURCE_MANAGER.calibrate_io (num_physical_disks => 8,
   max_latency        => 1,
   max_iops           => l_iops,
   max_mbps           => l_mbps,
   actual_latency     => l_latency);
END;
/

max_iops = 1977404
latency  = .04
max_mbps = 16045

PL/SQL procedure successfully completed.
```

Simple Select on a Large Table

```
SQL> select bytes/1024/1024/1024 gb from user_segments where segment_name = 'TBL_DISPLAY_BIG';

        GB
----------
 624.05365


SQL> set timing on
SQL> alter system flush buffer_cache;

SQL> select /*+ full(t) parallel(16) */ count(*) from tbl_display_big t;

  COUNT(*)
----------
 689078056

Elapsed: 00:00:33.05
```

Thank You

mag. Sergej Rožman
ABAKUS plus d.o.o.
Ljubljanska c. 24a, Kranj, Slovenija

e-mail: sergej.rozman@abakus.si

ANDRITZ Blubit CENTROSINERGIJA EKDIS Ekspresno. Ekonomično. Hidria jata emona LJUBLJANA

DRINA OSIGURANJE GENERALI GROUP GENERALI Zavarovalnica MILENIJUM® OSIGURANJE NLB Vita Življenjska zavarovalnica PRVA triglav