NHR@FAU upcoming installations and future plans

HPC-Café 19.12.2023

Prof. Dr. Gerhard Wellein, Dr. Thomas Zeiser
Zentrum für Nationales Hochleistungsrechnen Erlangen (NHR@FAU)

Agenda

- Hardware: Status 12/2023
- Hardware upgrades 2024/2025 (NHR/FAU)
- Hardware updates
  - Decomissioning of MEGGIE (FAU)
  - New systems in test cluster (NHR/FAU)

Terms:

- NHR → NHR-funds → accessible through NHR-application process („Tier-2“)
- FAU → FAU-funds (91b) → FAU/RRZE access only („Tier-3“)

NHR FAU

FAU Friedrich-Alexander-Universität Erlangen-Nürnberg

NHR@FAU – Hardware: Status

NHR@FAU: HPC Infrastructure – 12/2023

- **HPC systems** accessible through **NHR+FAU**
  - Fritz: CPU-Cluster (11/2023: TOP500, rank 213) with 77,000+ cores
  - Alex: GPU-Cluster (11/2023: TOP500, rank 187) with 304 (+352) NVIDIA GPUs

- **Central** (online/offline) **storage**: 8 PB / >3300 tapes possible (**NHR+FAU**)
  - HOME, WORK, VAULT (ARCHIVE)
- **Prototype NVMe storage** (1 PB) → Alex → request test access via ticket

- HPC systems for **FAU** (local access only)
  - Approx. 17,000 cores (meggie, woody, tinyFAT)
  - Approx. 150 NVIDIA GPUs (tinyGPU)

NHR FAU

FAU Friedrich-Alexander-Universität Erlangen-Nürnberg

NHR@FAU – Hardware: 2024 / 25 upgrades

Procurement process

- Wettbewerblicher Dialog mit Teilnahmewettbewerb
- Ausschreibungsteam + Auswahlausschuss (HPC-Cafe 25.4.2023)

- Zeitplan des Verfahrens bis Zuschlag
  - 16.4. Veröffentlichung Teilnahmewettbewerb
  - 15.5. Ende Teilnahmewettbewerb
  - Anfang Juni: Dialogrunde
  - Mitte Juni: Finale Ausschreibungsunterlagen → Erste Angebote
  - Mitte Juli: Dialogrunde
  - Anfang September: Finale Angebote
  - Anfang November: Bekanntgabe Zuschlag
  - Dezember: Vertrag

MEGWARE

Configuration

- **Funding**: NHR + StMWK/FAU → ~10 M€ (incl. VAT)
  + HS Hof & KU Eichstätt

- **Components**
  - CPU module (Fritz.2) → FP64 Peak: ca. 7 PF/s
  - GPU module (Alex.2) → FP64 Peak: ca. 10 PF/s
  - Shared NDR-IB network infrastructure
  - Shared parallel filesystem

- Potential extension for FAU by “Forschungsgrossgeräteantrag“

CPU module Fritz.2

- Compute nodes: **312** (**264** + **21**+ **24**+ 3)
  - Each node with 2 CPUs
    - **>128 core AMD „Turin dense“** → >3.5x vs. Fritz (ICL)
    - 12 Channel DDR5 (ca. 450 GB/s) → ~3x vs. Fritz (ICL)
  - 288 nodes: 768 GB main memory
  - 24 nodes: 2,1 TB main memory

- NDR Infiniband
  - 200 Gbit/s uplink per node → 2x vs. Fritz (ICL)
  - NDR island size of at least **24** nodes

- Installation: Q1/Q2 – 2025

Warm Water Cooling

GPU module Alex.2

- GPGPU nodes: **43** (**20** + **20**+ 2+1)
  - Each node with
    - 2 x 64 core AMD
    - 768 GB main memory (host)
    - 1 x NVMe **15** TB (6 TB for user data)
    - **4 x NVIDIA H100 SXM5** (94 GB HBM2e)
- NDR Infiniband
  - 2 * 200 Gbit/s uplink per node
- Installation: Q3/Q4 – 2024

Warm Water Cooling

Parallel Filesystem & IB network

Parallel Filesystem (shared by CPU & GPU)

- Lustre
- Capacity: 3 PB
- Technology: HDDs
- Peak Bandwidth: 20 GB/s
- Uplinks im NDR IB: min. 400 Gbit/s

NHR FAU

FAU Friedrich-Alexander-Universität Erlangen-Nürnberg

NHR@FAU – Hardware Updates

MEGGIE decommissioning

Test cluster

„Meggie” cluster from 2016 – end of operation: 31.3.2024

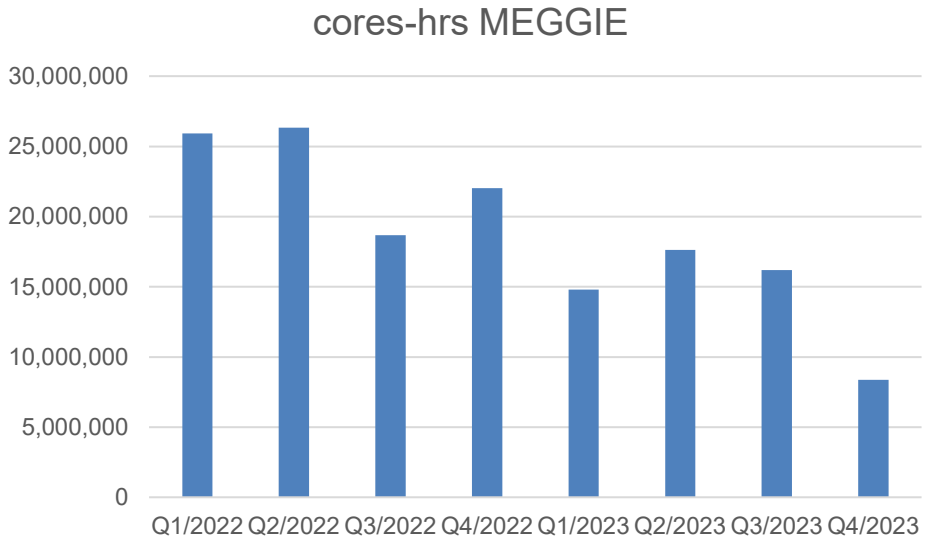
- 728 Compute nodes (14.560 cores)
  - 2 Intel Xeon E5-2630 v4 (Broadwell) 2.2 GHz (10 cores)
  - 20 cores/ node + SMT cores
  - 64 GB main memory
  - NO local disks
- Intel OmniPath network: Up to 100 Gbit/s

- Peak Performance: R_{peak} = 0.5 PF/s
- #346@TOP500 – Nov. 2016 R_{max} = 0.48 PF/s
- Price: € 2,5 Mio.
- **Power consumption [KW]: 100 - 200**

FAU users only

End of user operation: 31.3.2024

„Meggie” cluster from 2016 – end of operation: 31.3.2024



„Meggie" cluster from 2016 – end of operation: 31.3.2024

- **FAU users only – End of user operation: 31.3.2024**
- Alternatives for users
  - Woody (low / moderate parallelism)
  - Fritz (at least one node with 72 cores must be used!)

	#nodes	Node conf.	Typical job sizes
Woody	176 Intel E3-1240 v5/6 (704 cores)	1 * 4 c (8360Y) 32 GB	1 node
	110 Intel ICL (3,520 cores)	2 * 16 c (6326) 256 GB	1 node

#nodes	Node conf.	Storage	Typical job sizes	Peak (FP64)
992 Intel ICL (71,424 cores)	2 * 36 c (8360Y) 256 GB 1 x HDR100	Shared PFS • 3 PB • >20 GB/s	1 – 64 nodes	5.9 PF/s (4.1 PF/s)
64 Intel SPR (6,656 cores)	2 * 52 c (8470) **1 TB / 2 TB** 1 x HDR100		1 – 4 nodes	n.y.a.

- Need support for code porting / optimization?? hpc-support@fau.de

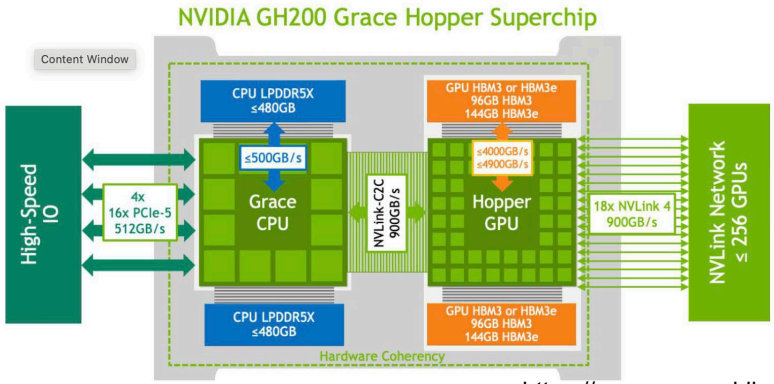
New systems in Testcluster – Nvidia GraceHopper

- Combined chip with CPU (Grace) and GPU (**H100**)
- 72 CPU cores (ARM Neoverse V2, 4x SVE128) @ 3.6 GHz
- 480 GB LPDDR5X & 96 GB HBM memory
- Strong NVLink interconnect between CPU & GPU

System:
`gracehop1`

Product Specifications

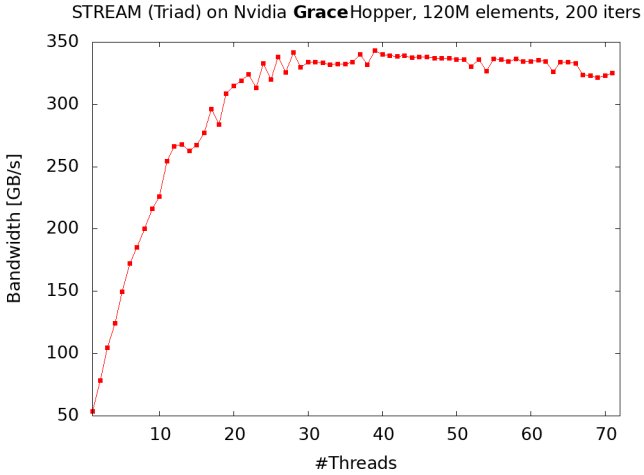
Grace CPU	Feature
CPU core count	72 Arm Neoverse V2 cores
L1 cache	64KB i-cache + 64KB d-cache
L2 cache	1MB per core
L3 cache	117MB
LPDDR5X size	Up to 480GB
Memory bandwidth	Up to 512GB/s
PCIe links	Up to 4x PCIe x16 (Gen5)

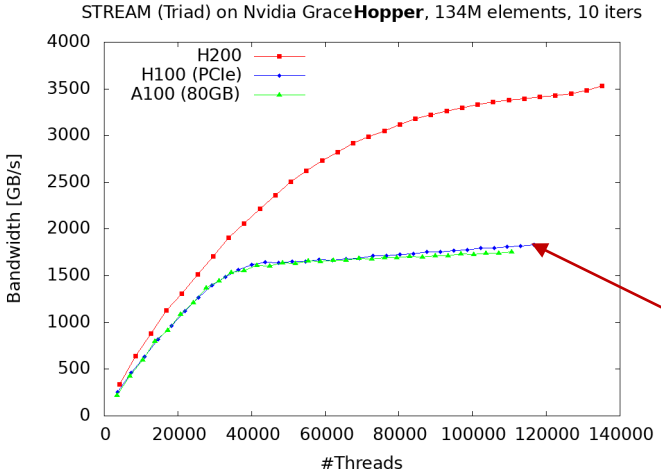


Hopper H100 GPU	Feature
FP64	34 teraFLOPS
FP64 Tensor Core	67 teraFLOPS
FP32	67 teraFLOPS
TF32 Tensor Core	989 teraFLOPS* | 494 teraFLOPS
BFLOAT16 Tensor Core	1,979 teraFLOPS* | 990 teraFLOPS
FP16 Tensor Core	1,979 teraFLOPS* | 990 teraFLOPS
FP8 Tensor Core	3,958 teraFLOPS* | 1,979 teraFLOPS
INT8 Tensor Core	3,958 TOPS* | 1,979 TOPS
High-bandwidth memory (HBM) size	Up to 96GB | 144GB HBM3e
Memory bandwidth	Up to 4TB/s | Up to 4.9TB/s HBM3e
NVIDIA NVLink-C2C CPU-to-GPU bandwidth	900 GB/s bidirectional
Module thermal design power (TDP)	Programmable from 450W to 1000W (CPU + GPU + memory)

https://resources.nvidia.com/en-us-grace-cpu/grace-hopper-superchip

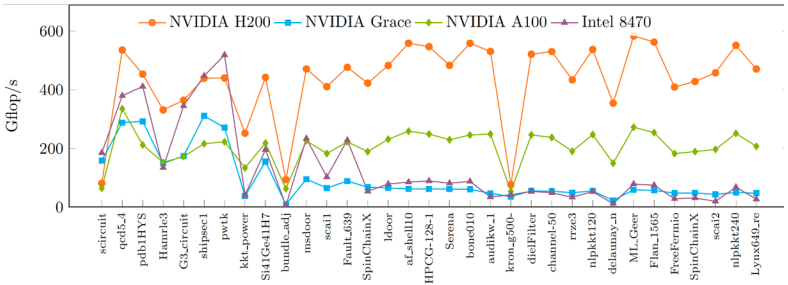
New systems in Testcluster – Nvidia GraceHopper





System:
gracehop1

H100 (PCIe)
A100



Sparse Matrix Vector Multiplication:
H100 vs. A100 vs. GRACE vs. SPR

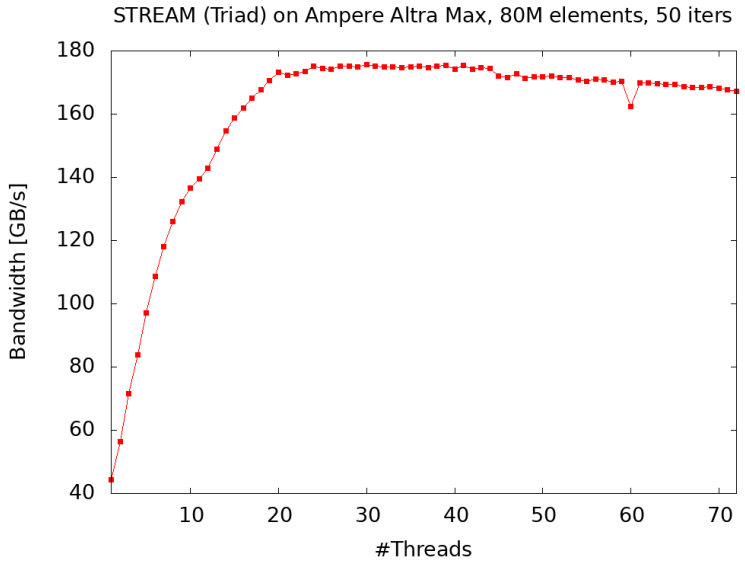
New systems in Testcluster – Ampere Altra MAX

- 2 x 128 cores ARMv8 chip (ARM Neoverse N1) @ 3 GHz
- 512GB DDR memory
- No SVE, only NEON

System:
`lukewarm`

- Only 16 Mbyte L3 (SLC) per socket (not visible for OS)

- Peak FP rate: 2.5 TF/s DP



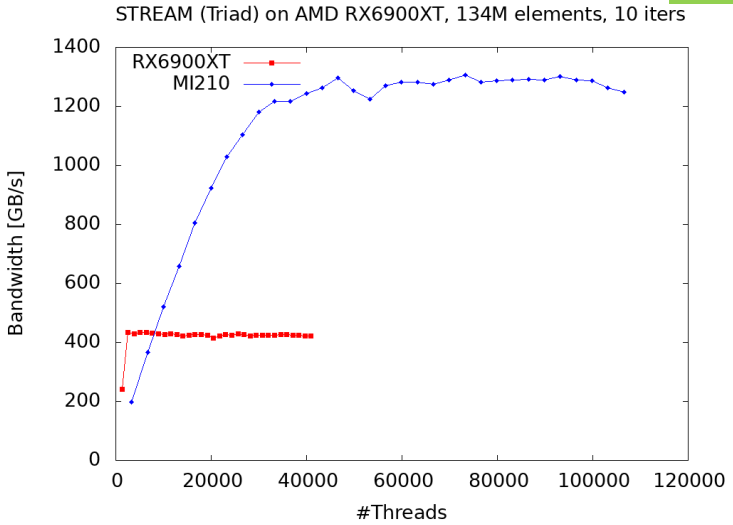
New systems in Testcluster – AMD RX6900XT

- RDNA2 architecture with 80 compute units
- Gaming GPU with **128MB cache** and 16GB GDDR6 memory

- Low DRAM bandwidth (up to 512 GB/s)
- Low FP64 (~ 1.44 TFlop/s)

- Medium FP32 (23 TFlop/s)
- Cache bandwidth: 2 TB/s

System:
`eurayle`

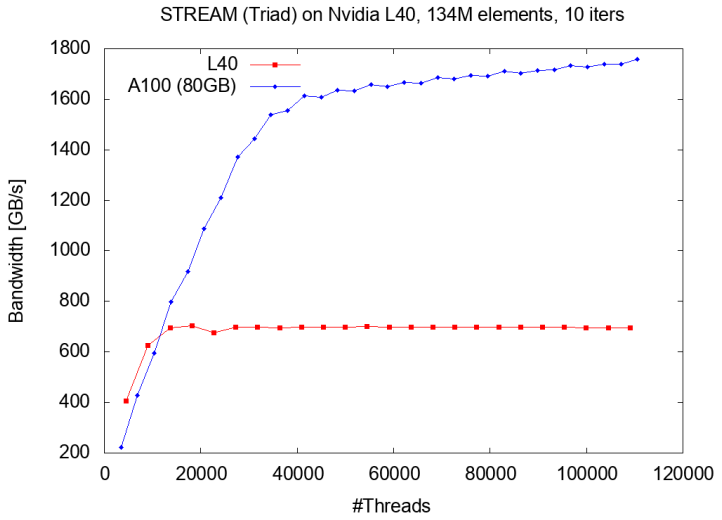


New systems in Testcluster – Nvidia L40

- Ada Architecture with 142 Streaming Multiprocessors
- „Gaming“ GPU with 92 MB cache
- 48 GB GDDR6 Memory
- Low DRAM bandwidth (up to 864 GB/s)
- Low FP64 (up to 1.4 TFlop/s)
- High FP32 (90 TFlop/s)
- Cache bandwidth: 5.3 TB/s

- Strong performance on some GROMACS simulations

System: genoa2



FAU Friedrich-Alexander-Universität Erlangen-Nürnberg

FROHE

Weihnachten

UND EIN GESUNDES
JAHR 2024

NHR FAU

„Die Zukunft gehört denen,
die die Möglichkeiten erkennen,
bevor sie offensichtlich werden."

- Oscar Wilde -