Albedo - The new HPC-System at AWI in 2022

Stephan Frickenhaus, Bernadette Fritzsch, Sebastian Hinck,
Natalja Rakowsky, Malte Thoma

Computing Center, HPC & Data Processing and Storage & Server
Alfred Wegener Institute for Polar and Marine Research, Bremerhaven

Thanks a lot to

AWI Einkauf - Frank Chnelewski, Cecil Feierabend, Jörg Eilers
Legal Consulting - Thomas Haug (Castringius RA & Notare)

HPC at AWI and beyond

HPC - High Performance Computing

Tier-3 local resources at AWI: Ollie (2016-22), Albedo (2022-. . .)
For code development, testing, and small projects.
Click “HPC” in AWIID, approval by section head, and go!

Tier-2 Network of 8 centres: National High Performance Computing (NHR) with specialised focus.
DKRZ: Free within AWI share & application based
HLRN: Small tests and projects are free, application for larger projects, support by BremHLR (Lars Nerger)

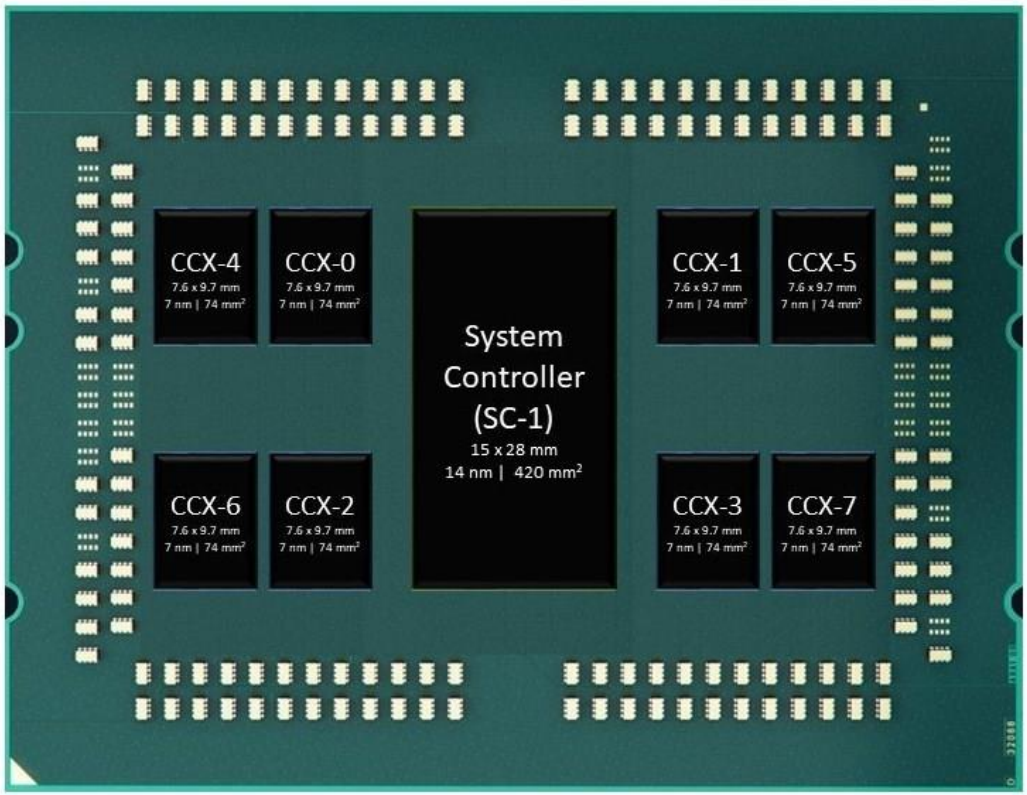
Tier-0/1 Germany: Gauss Center for Supercomputing
JSC Jülich, HLRS Stuttgart, LRZ Munich
Europe: Partnership for Advanced Computing in Europe (PRACE)

Albedo hardware - compute

- vendor: NEC
- 240 compute nodes with
  - 2x AMD Rome Epyc 7702 2GHz (3.3GHz Boost), 64 Core, cTDP reduced from 200W to 165W
  - 256 GB RAM, 500GB SSD
- 4 ”fat” nodes as above, 4TB RAM, 7.5TB SSD
- 1 GPU node with 1TB RAM, 4x NVIDIA A100/80
  More GPU nodes will follow later, after first experience
- 3 login nodes, one with GPUs 2x NVIDIA A40
- integrate test node with NEC SX-Aurora TSUBASA vector engine
- fast interconnect: HDR Infinband

Albedo hardware - compute

AMD Epyc 7702 Multi-Chip Module: 8 Core-Dies and I/O Die



Each 8 Core Die is connected to 16GiB RAM
ccNUMA - cache coherent Non Uniform Memory Access

Albedo hardware - file systems

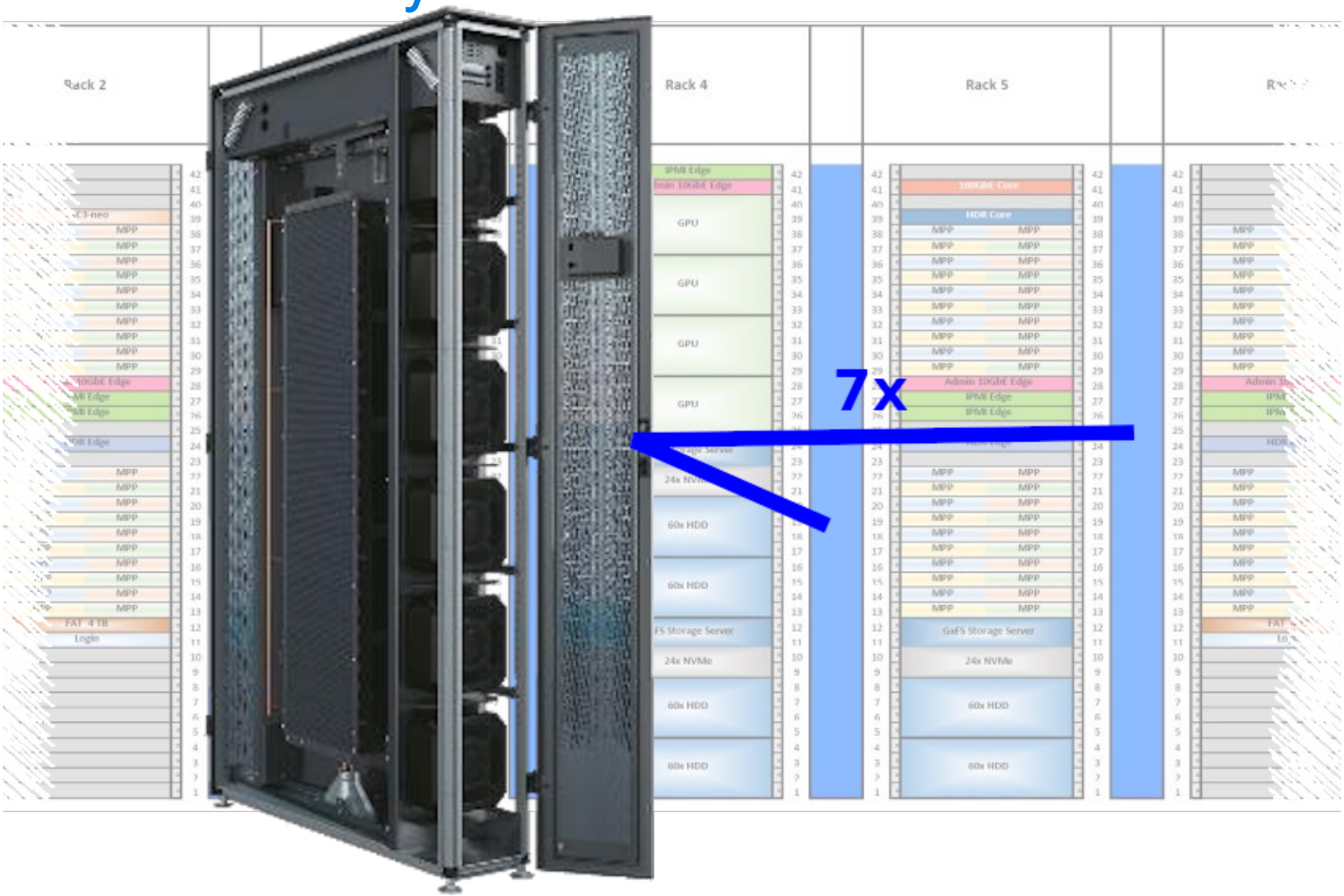
- 5 PB NEC GxFS (IBM Spectrum Scale)
  - /scratch, /home, /global
  - 220 TB as NVMe SSDs as fast cache and/or burst buffer
  - extension (capacity, bandwidth) possible
  - improved policy: small soft quota, large hard quota, grace period, project quotas e.g., for forcing data
- Dell EMC Isilon /isibhv connected to all nodes with 10GbE
  - main purpose: ease working with observation data
  - convenience: shared directories with other AWI servers
- local SSDs inside compute nodes
  - significant speed up of jobs that frequently read or write small amounts of data
  - reduce load on /scratch
  - as cache for GxFS?

Albedo racklayout (offer)

Albedo racklayout (offer)

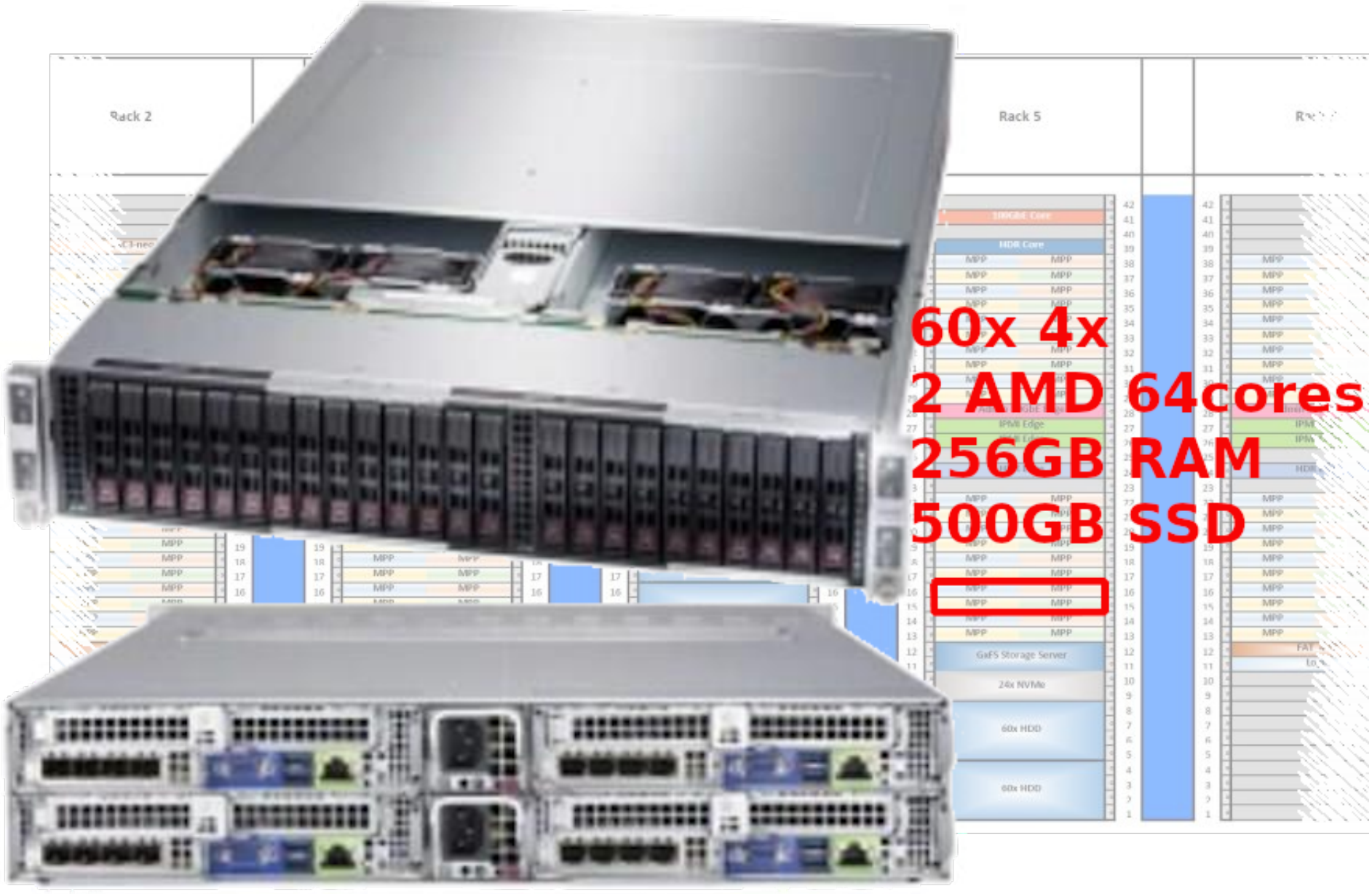
HELMHOLTZ

Albedo racklayout: sidecooler



HELMHOLTZ

Albedo racklayout: compute nodes



HELMHOLTZ

Albedo racklayout: file system

HELMHOLTZ

Albedo software

- Alma Linux (Redhat clone; CentOS with rolling release no longer an option)
- Slurm batch system
- Programming Environment
  - Intel Compiler, MPI, MKL, Profiling: Intel oneAPI (now w/o license!)
  - Gnu Compiler Collection
  - AMD Optimizing C/C++ Compiler AOCC (includes Fortran)
  - Mellanox HPC-X MPI (OpenMPI)
  - NVIDIA compiler + MPI (GPUs)
  - NEC compiler + MPI (NEC SX-Aurora vector)
  - ARM DDT parallel debugger
- Application software $\rightarrow$ whatever you need!

Albedo timeline

11/2021 procurement ended

12/2021 contract signed

≈ 05/2022 start with a few compute nodes, Gigabit network (due to a severe delivery delay of Mellanox)

- copy data from Ollie
- install software, port applications
- small projects

≈ 07/2022 switch to fast Infiniband + most compute nodes

- 30 days stability phase
- production!

2023 Evaluate, decide on extensions (GPUs, /scratch,...?)

How to buy a supercomputer

- continuously:
  - inform and seek permissions from AWI boards and directorate
  - exchange with users, with other computer centers
  - get informations from vendors on new technologies (until procurement starts)
- write proposal to Helmholtz for a medium sized investment
- Prepare all documents (together with procurement)
  - call for competition
  - *Leistungsverzeichnis* with requirements, evaluation criteria
  - benchmarks (FESOM2, HPCC, IO500)
  - general paperwork . . .

How to buy a supercomputer

Call for Competition

Vendors apply with a concept for the installation of a supercomputer and the support. We evaluate the concepts and references.
Three vendors will be admitted.

Negotiation Procedure

For Albedo: 40% price (investment + 7 years of support)
60% computing power
with fixed aspects like

- 170kW max. power consumption under “every day load”
- Fat and GPU nodes
- minimum size and speed of /scratch
- connection to /isibhv

Three rounds to fine tune the offers $\longrightarrow$ award the contract

Albedo Benchmark

FESOM2 fArc (640.000 2D nodes, 1800 steps)
runtime of time stepping must be 3min or better, for example:

ollie.awi.de: 18 nodes (2x Intel Broadwell 18 core)
lise.hlrn.de: 7 nodes (2x Intel Cascade Lake 48 core)
albedo.awi.de: 6 nodes (2x AMD Rome 64 core)

Criterion

$$\Lambda = \frac{N}{n} \cdot \frac{3600s}{\tau} = \frac{240}{8} \cdot \frac{3600s}{135s} = \mathbf{800}$$

n number of nodes for benchmark (8 because of superlinear scaling)
N number of compute nodes, τ benchmark runtime

Comparison: $\Lambda = 1$ corresponds to 1 ollie node, thus $\Lambda_{\mathrm{ollie}} = 316$