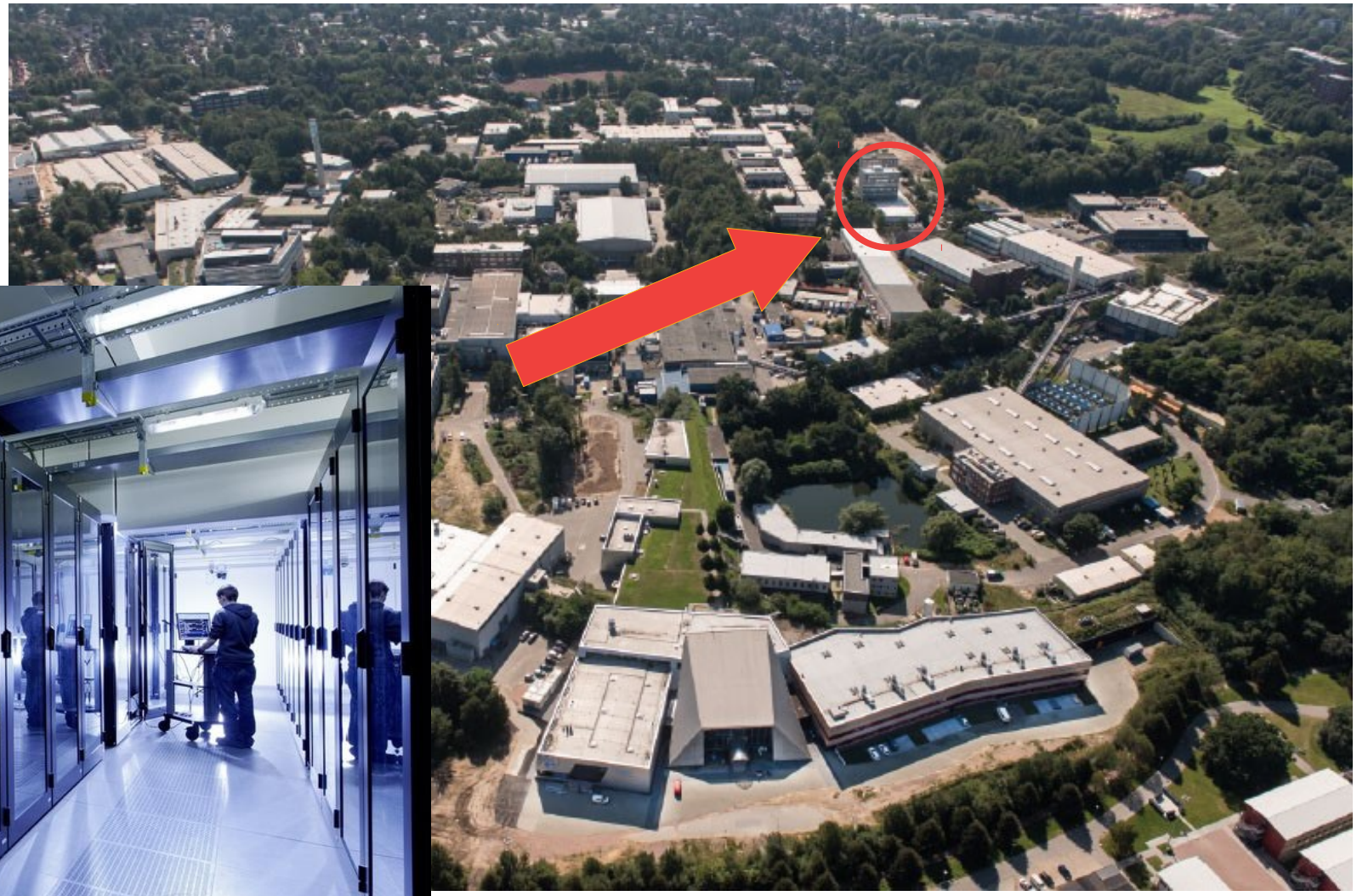
AI/ML acceleration on Maxwell

Current overview and roadmap

Yves Kemp, Tim Wetzel
AI/ML round table @ DESY
Hamburg, 03.12.2021

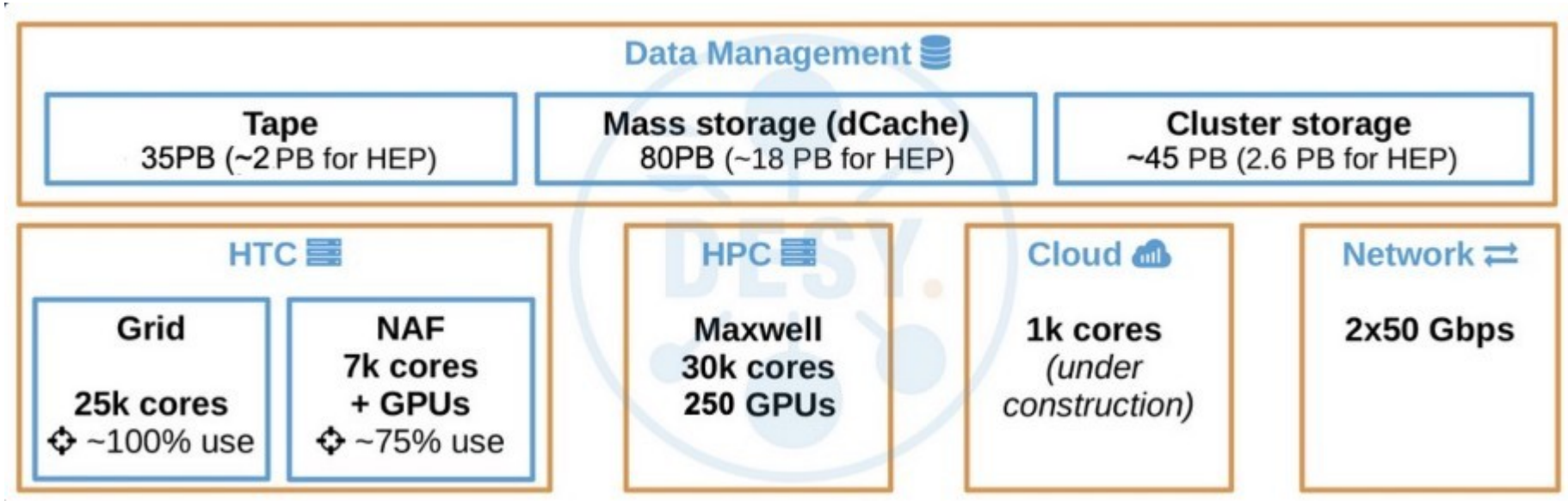
HPC-Cluster Maxwell

On DESY campus



Maxwell

Current hardware



Maxwell

- ~2.5 PFLOPs total compute power
- Extension to ~5 PFLOPs planned
- Currently difficult to get hardware due to chip shortage
- MLLab on Maxwell for testing new hardware: acquisition of AMD and Intel GPUs as soon as possible, use of external DevCloud

GPUs

Quelle: C.Voss, DESY-IT

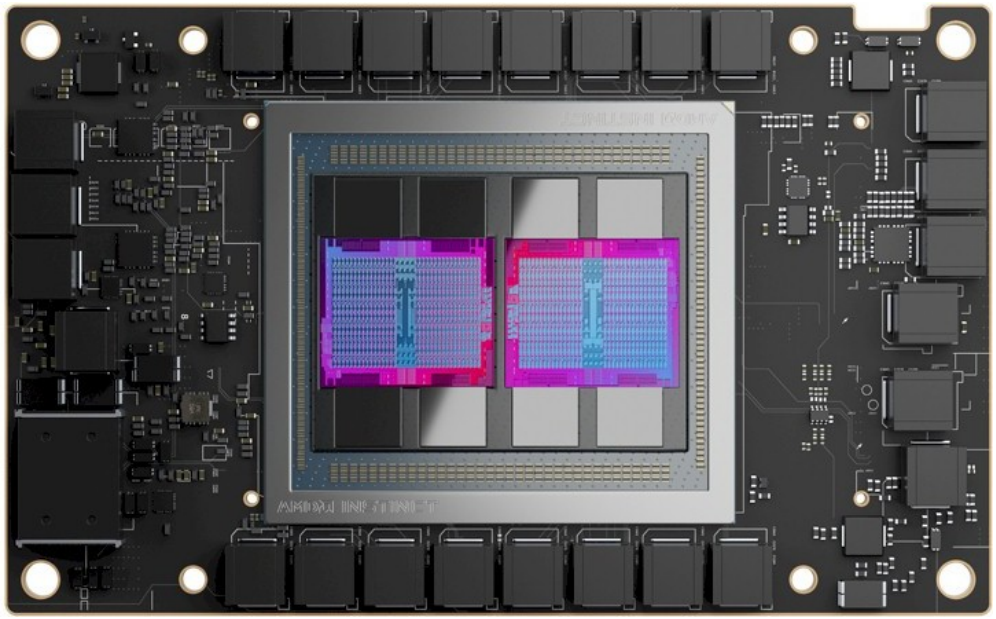
- Currently 154 GPU nodes with 292 GPUs (~1.6 PFLOPs)
- Models range from Nvidia RTX/Quadro to P100, V100, A100
- More GPUs will be purchased

AMD

Planned developments for the next generation of AMD data center GPUs

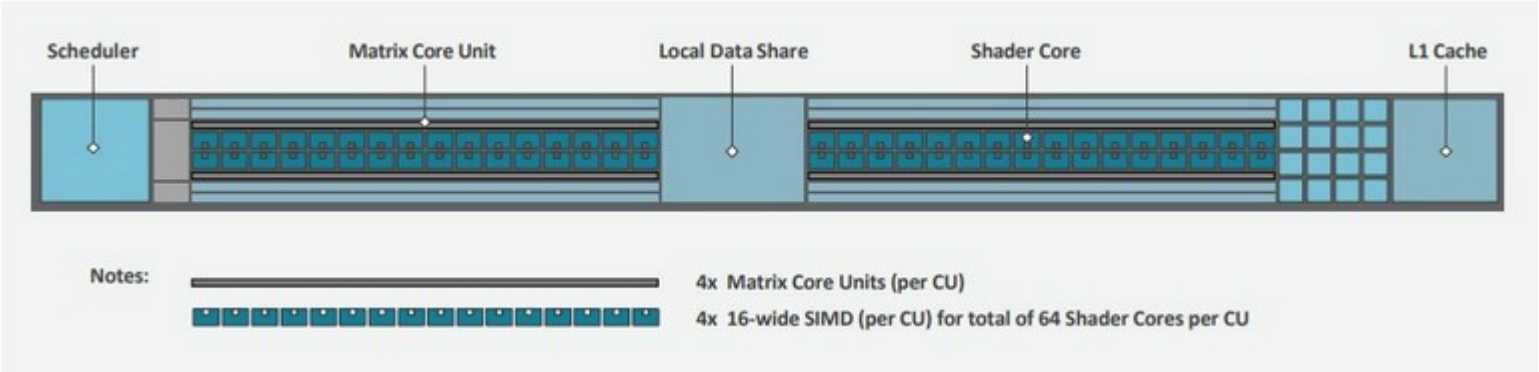
AMD Instinct MI 200 series (MI250)

- Exclusively available for newly built Exaflop HPC cluster *Frontier* in the beginning of 2022
- Publicly available in Q3/4 2022 (planned)
- Test setups eventually possible → interested?



Currently known specs

- 2 Dies per Card (CDNA2-Arch)
- ~47,9 Tflops (FP64)
- 128 GB RAM
- 560W TDP
- FP64 matrix cores, vector cores



Directly programmable via **ROCm/HIP**

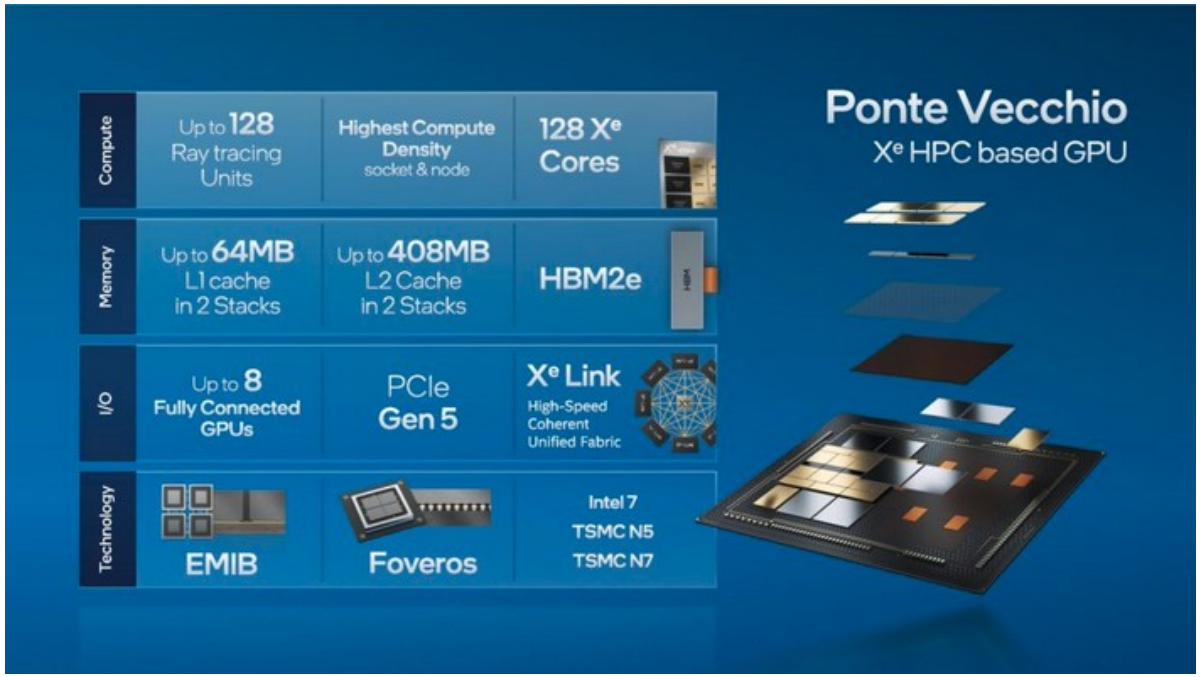
Intel

Planned developments for the next generation of Intel data center GPUs

Intel Xe GPU series (Ponte Vecchio)

- Preliminarily available for newly built Exaflop HPC clusters from beginning of 2022
- Publicly available in Q3/4 2022 (expected, more info beginning 2022)
- Test setups accessible via Intel DevCloud

  ➔ If you are interested, please let us know!



Currently known specs

- 128 Xe cores
- ~45 Tflops (FP64)
- 128 GB HBM2e RAM
- 64MB/408MB L1/L2 caches
- 560W TDP
- 1024 matrix engines, 128 ray-tracing units

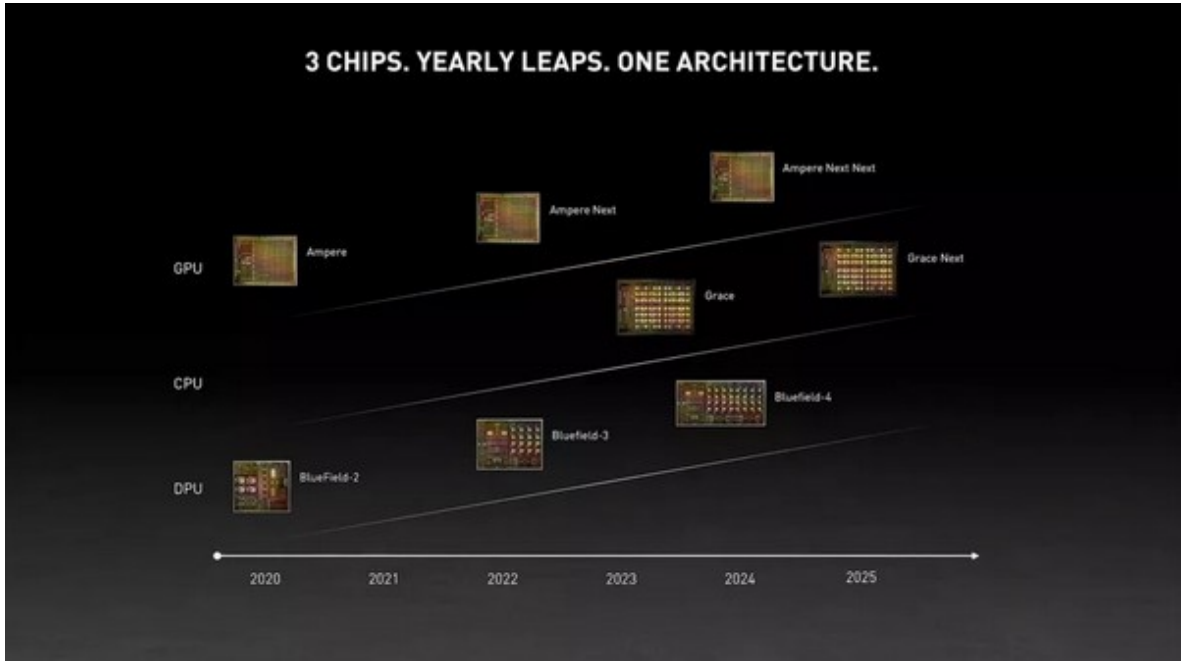
Directly programmable via Intel **OneAPI**
→ together with other Intel hardware (CPU, FPGA)

Nvidia

Planned developments for the next generation of Nvidia data center GPUs

Nvidia Ampere Next (ANext)

- Presented at GTC, release planned for mid of 2022
- Possibly 80GB RAM at first, 120GB later (~2023)
- Scaling from earlier generations:
  - P100: 4.7 TFlops (FP64)
  - V100: 7 TFlops (FP64)
  - A100: 9.7 TFLops (FP64) / 19.5 TFlops (FP64-TC)
  - ➔ ANext: >12 TFlops (FP64) / ~50 TFlops (FP64-TC) to keep up with AMD



- Cpu-only (Grace, ARM-based) announced for end of 2023
- DPU Bluefield-3 announced for 2022

no general use case as of now, only niches, e.g. data compression, reduction

ML-Frameworks

Currently supported compute architectures

PyTorch

- CUDA 10.2, 11.3
- ROCm/HIP 4.2 (beta)
- CPU

Tensorflow

- CUDA (multiple versions)
- CPU

Keras

Depends on backend

If you are interested in using a backend different from CUDA and need support, please let us know.

We would like to hear about your experiences.

Thanks!

Questions?

We are planning to purchase different test systems with non-Nvidia architectures.

We can also arrange access to testing infrastructure outside of DESY if this is wanted for evaluation purposes.
→ e.g. Intel DevCloud, Megware Test Lab, ...

Kontakt

DESY. Deutsches Elektronen-Synchrotron

www.desy.de

Tim Wetzel
IT - RIC
tim.wetzel@desy.de
040 8998-2911