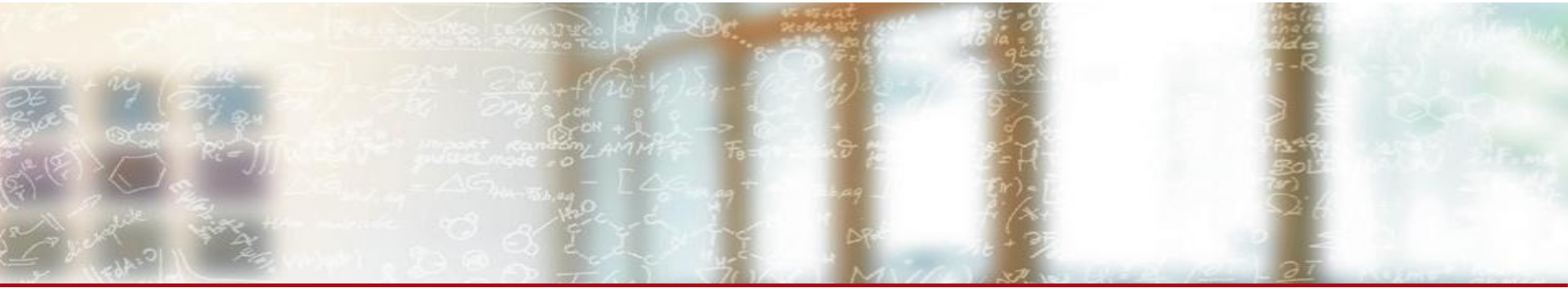
CSCS
Centro Svizzero di Calcolo Scientifico
Swiss National Supercomputing Centre

ETH zürich



Onboarding SDSC users at CSCS

Workshop

Prashanth Kanduri and Lukas Drescher
CSCS

19th October, 2023

Structure of the day

Morning

- Alps overview
- MFA access
- SSH configuration:
  - Daint login nodes and compute nodes
  - Login with VS code
- Running jobs using sbatch
- Conda environment
  - create custom jupyter-kernel
    - shared between Jupyter service, IDE and shell

Afternoon

- Running containers with Sarus on Piz Daint
  - …using NGC containers for single node and distributed deep learning
  - Large scale training on Piz Daint in MLPerf

- Outlook on Alps
  - New container engine
  - High-performance data science with RAPIDS on Clariden

CSCS
Centro Svizzero di Calcolo Scientifico
Swiss National Supercomputing Centre

ETH zürich

Quick Intro to CSCS

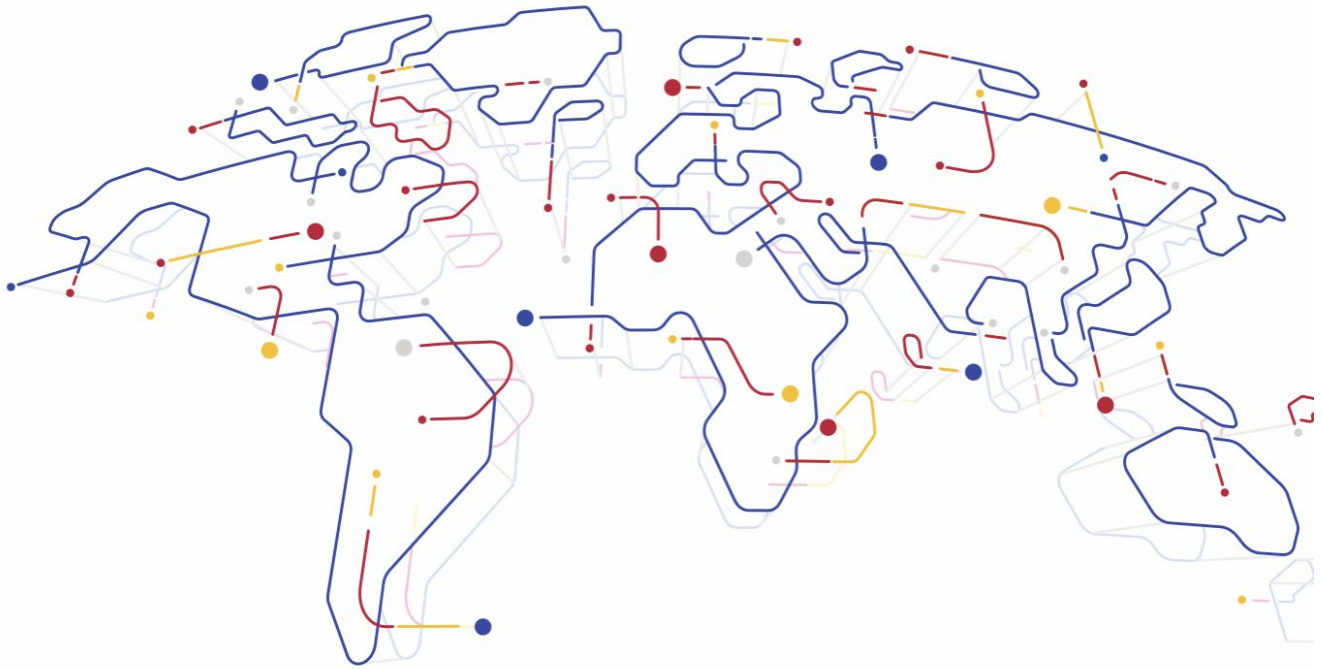
A unit of the Swiss Federal Institute of Technology, ETH Zürich

Mission

«We develop and operate a high-performance computing and data research infrastructure that supports world-class science in Switzerland»

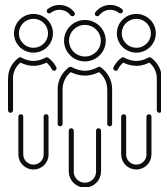
- Located in Ticino since 1991

- National and international collaborations in the research of new technologies for HPC



Some numbers

Staff

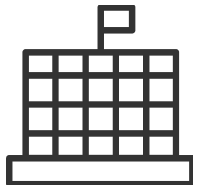


- 120 members
- 26+ nationalities
- Official language: english

Budget

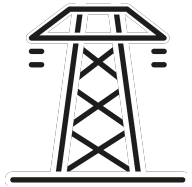
- CHF 30 Mio. operating budget
- CHF 20 Mio. IT investment

Building



- 2’600 m^2 office building
- 2’000 m^2 machine room
- «Free cooling» with lake water

Electricity

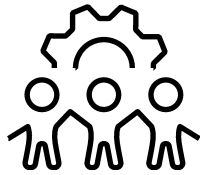


- Currently 11 MW
- Possible extension to 25 MW
- 100% hydro-electric source

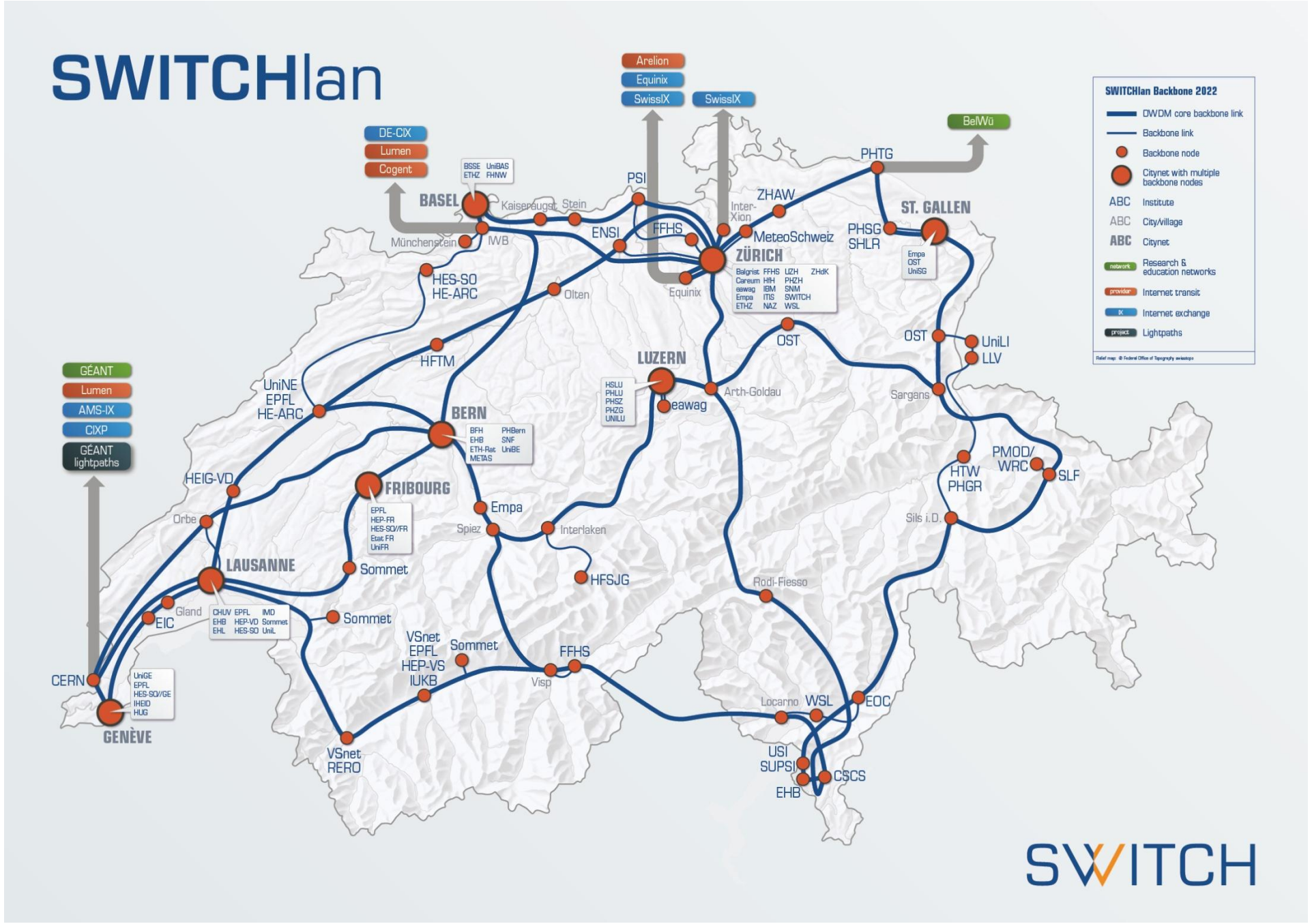
User Lab

- 2’300 users
- 130 projects

Partnerships



- MeteoSwiss, NCCR Marvel, PSI, CHIPP, Empa, ETH Zurich, CERN, USI, UZH, BlueBrain ...


SWITCHlan
Arelion
Equinix
SwissIX
SwissIX
DE-CIX
Lumen
Cogent
BelWü
BSSE UniBAS
ETHZ FHNW
BASEL
Kaiseraugst
Stein
PSI
PHTG
ZHAW
Inter-Xion
ST. GALLEN
Münchenstein
IWB
ENSI
FFHS
MeteoSchweiz
PHSG
SHLR
Empa
OST
UniSG
ZÜRICH
Balgrist FFHS LIZH ZHdK
Careum HfH PHZH
eawag IBM SNM
Empa ITIS SWITCH
ETHZ NAZ WSL
Equinix
HES-SO
HE-ARC
Olten
OST
OST
UniLI
LLV
HFTM
LUZERN
HSLU
PHLU
PHSZ
PHZG
UNILU
eawag
Arth-Goldau
Sargans
GÉANT
Lumen
AMS-IX
CIXP
GÉANT lightpaths
UniNE
EPFL
HE-ARC
BERN
BFH PHBern
EHB SNF
ETH-Rat UniBE
METAS
PMOD/
WRC
SLF
HTW
PHGR
HEIG-VD
FRIBOURG
EPFL
HEP-FR
HES-SO/FR
Etat FR
UniFR
Empa
Orbe
Spiez
Interlaken
Sils i.D.
LAUSANNE
Sommet
HFSJG
Rodi-Fiesso
Gland
EIC
CHUV EPFL IMD
EHB HEP-VD Sommet
EHL HES-SO UniL
Sommet
VSnet
EPFL
HEP-VS
IUKB
Sommet
FFHS
Visp
CERN
UniGE
EPFL
HES-SO/GE
IHEID
HUG
GENÈVE
VSnet
RERO
Locarno
WSL
EOC
USI
SUPSI
EHB
CSCS
SWITCHlan Backbone 2022
DWDM core backbone link
Backbone link
Backbone node
Citynet with multiple backbone nodes
ABC Institute
ABC City/village
ABC Citynet
network Research & education networks
provider Internet transit
IX Internet exchange
project Lightpaths
SWITCH

The Facility in Lugano

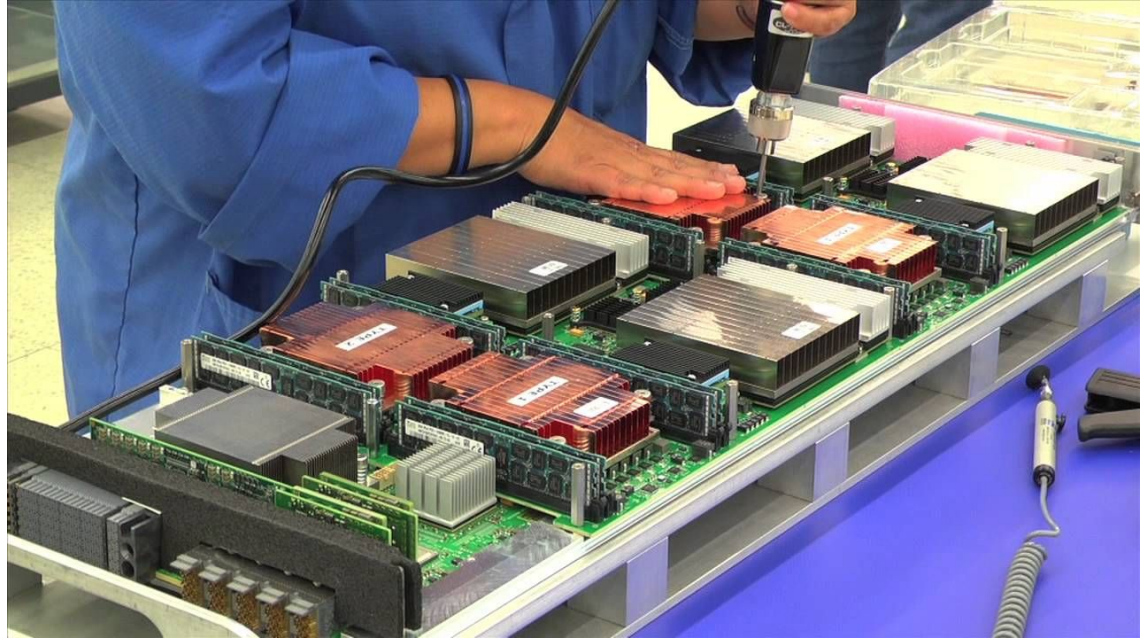
The Facility in Lugano

The Machines

Current Flagship System: Piz Daint

Model	Cray XC40/XC50
XC50 Compute Nodes	Intel® Xeon® E5-2690 v3 @ 2.60GHz (12 cores, 64GB RAM) and NVIDIA® Tesla® P100 16GB - 5704 Nodes
XC40 Compute Nodes	Two Intel® Xeon® E5-2695 v4 @ 2.10GHz (2 x 18 cores, 64/128 GB RAM) - 1813 Nodes
Login Nodes	Intel® Xeon® CPU E5-2650 v3 @ 2.30GHz (10 cores, 256 GB RAM)
Interconnect Configuration	Aries routing and communications ASIC, and Dragonfly network topology
Scratch capacity	8.8 PB

Daint-GPU Nodes



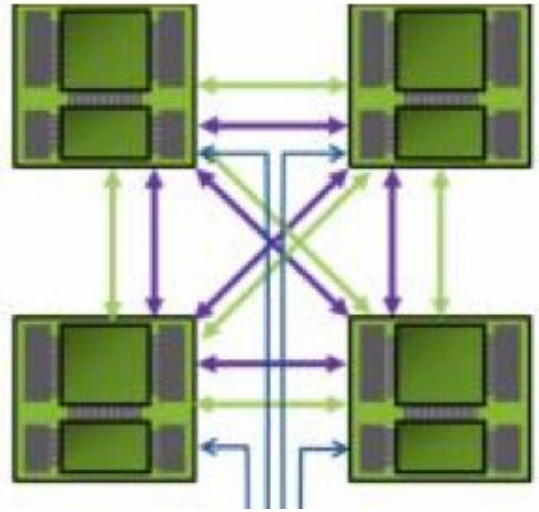
Daint-GPU node has a simple architecture

- 1 Haswell CPU socket
- 1 P100 GPU
- PCI-E connection between host-device
- 1 NIC

The ratio of 1-1 made allocating MPI ranks relatively simple:

- One rank per GPU + CPU
- Or multiple ranks sharing the GPU using CUDA MPS (multi-process service)

Alps Phase II Nodes



Grace-Hopper modules are *conceptually similar*

- 1 Grace CPU socket and one Hopper GPU per module
- Cache-coherent NVLINK connection between host and device memory
- One NIC per module

Each node will have 4 Grace-Hopper modules

- All-to-all cache-coherent memory NVLINK between all host and device memory

The one-to-one CPU to GPU ratio remains

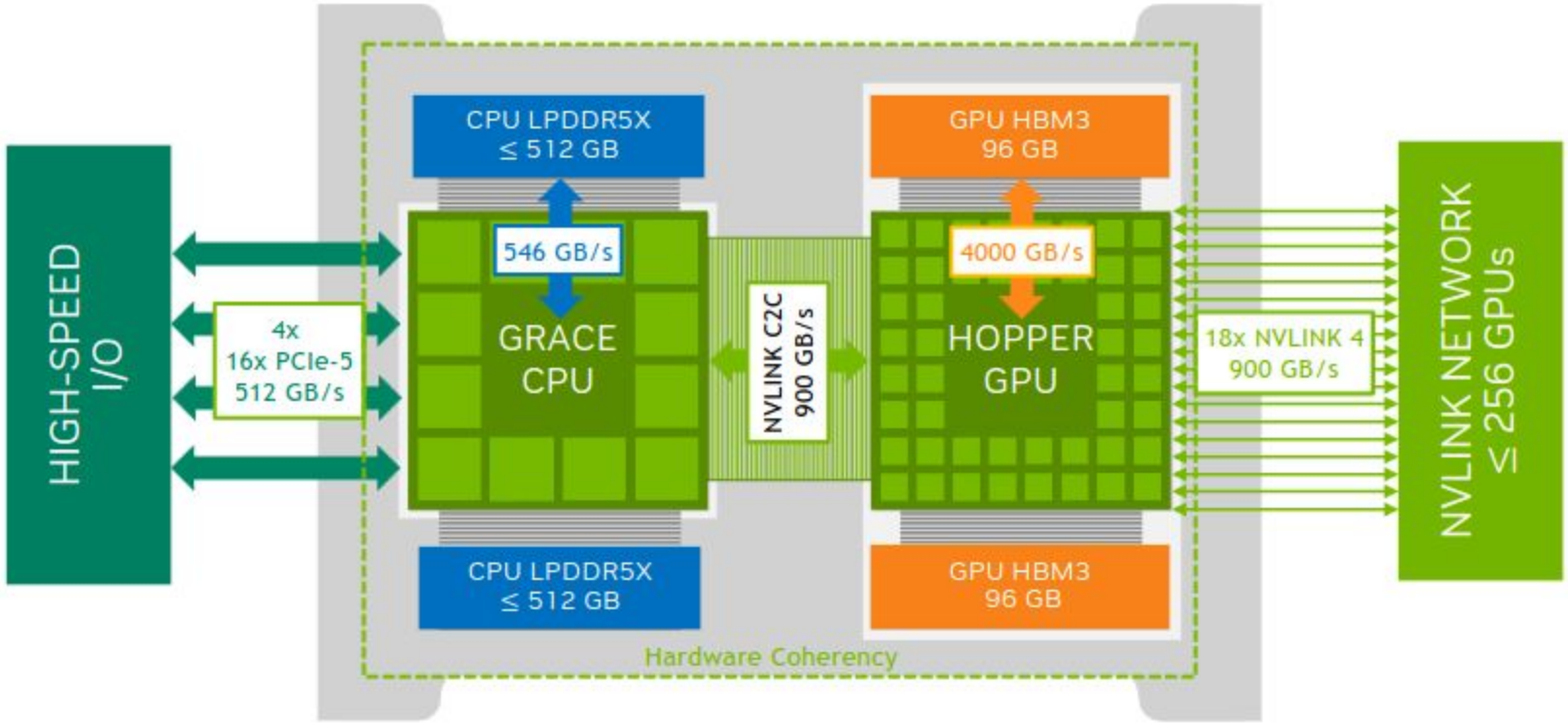
- The 4 modules on a node form an optimised communication network.

The Grace-Hopper “Super Chip”

NVIDIA are releasing are two super chips:

1. Grace-Grace: dual-socket Grace CPU with NVLINK C2C
2. Grace-Hopper: Grace CPU + Hopper GPU with NVLINK C2C

→ Alps Phase II



Bandwidth: Daint-GPU Node vs. one GH Module

Comparing the raw speeds and feeds of the CPU and GPU

GPU	P100	Hopper	Increase
Bandwidth	700 GB/s	4000 GB/s	5.7x
FP64	4.7 TFlops	34/67 TFlops	7-14x
Memory	**16 GB HBM**	**96 GB HBM3**	**6x**

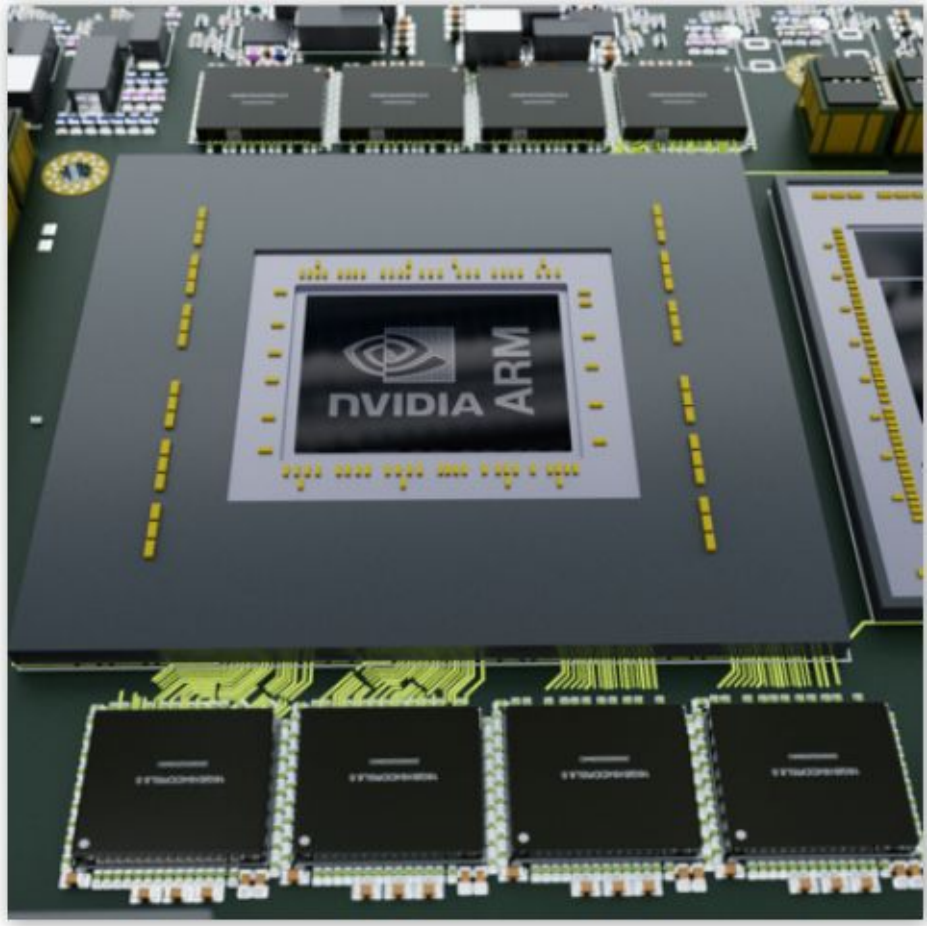
CPU	Haswell	Grace	Increase
Cores	12	72	6x
Bandwidth	**60 GB/s**	**475 GB/s**	**8x**
FP64	0.49 TFlops	> 2.5 TFlops	5x
Memory	64 GB DDR3	128 GB LPDDR	2x

Data Movement	Daint-GPU	Alps Phase II	Increase
Host-Device	**22 GB/s**	**480 GB/s**	**20x**
Device-Device on node	-	900 GB/s	-
node-node	11 GB/s	23 GB/s	2x

- The Grace-Hopper module delivers 5-10x improvement across the board
- Speedup may be lower or higher depending on the existing bottlenecks.

Grace: Server Class ARM CPU

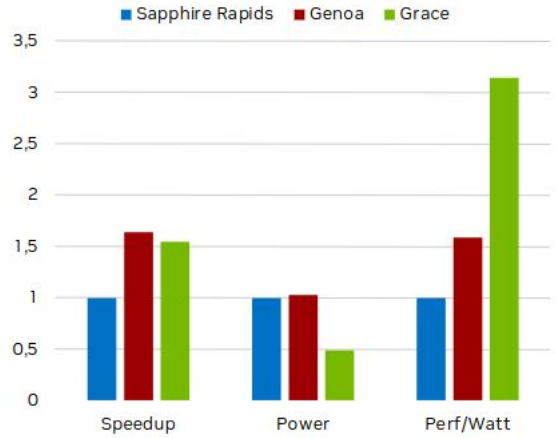
- 64bit Server Class Core and SoC
  - Arm V9.0 ISA Compliant aarch64 core (Neoverse V2 “Demeter”architecture)
  - Full SVE-2 Vector Extensions support, inclusive of NEON instructions
  - Supports 48-bit Virtual and 48-bit Physical address space
- Implemented on 5nm Process technology
- Balanced architecture between Single Core Perf, Core count, Memory and IO subsystems

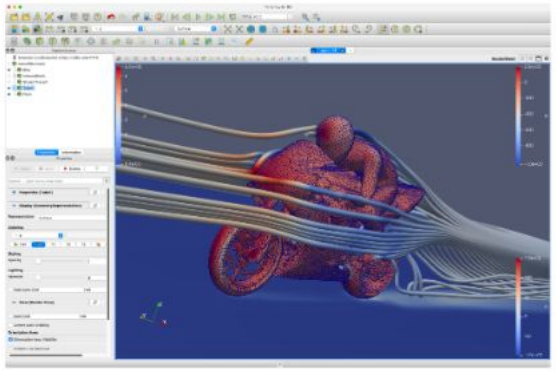


Grace Performance

OPENFOAM 2206

MotorBike 5M





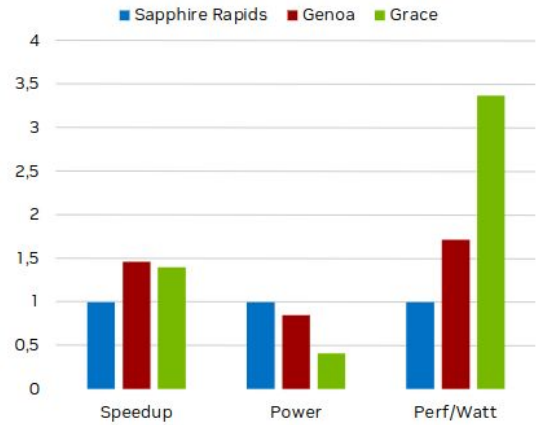
Sapphire Rapids: Intel Xeon Platinum 8470Q, 52c @ 2.1GHz - 3.8GHz
Genoa: AMD EPYC 9654, 96c @ 1.5GHz – 3.7GHz
Grace: NVIDIA Engineering Sample, 72c @ 3.2GHz
Best single socket time using best compilers (GCC, ICC, AOCC) and best rank/thread decomposition

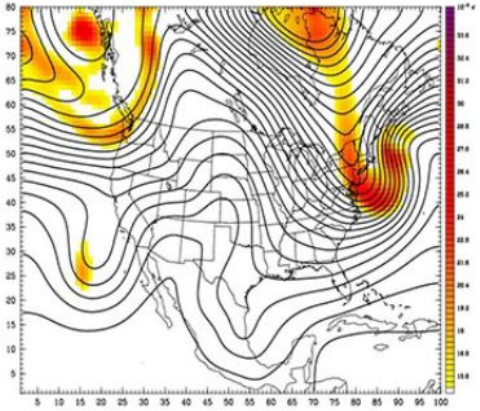
- Grace will be competitive with x86 HPC CPU architectures.
- Each GPU will have a full CPU socket – workloads that can “reverse offload” or have CPU-intensive components will benefit.

Results based on early engineering samples of Grace

WRF 4.4.2

CONUS12, 24hr simulation time



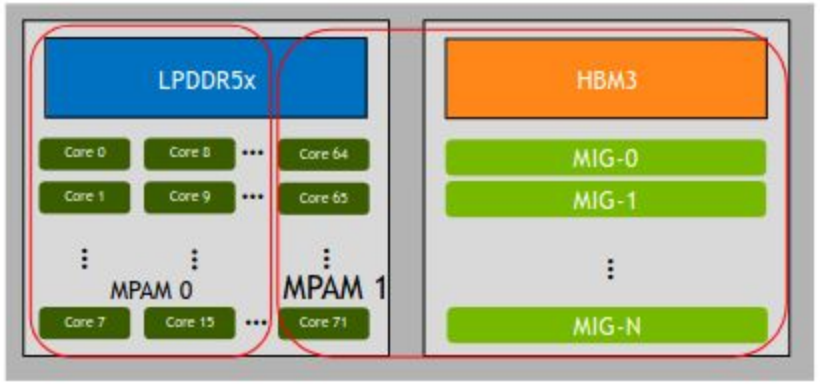


Sapphire Rapids: Intel Xeon Platinum 8470Q, 52c @ 2.1GHz - 3.8GHz
Genoa: AMD EPYC 9654, 96c @ 1.5GHz – 3.7GHz
Grace: NVIDIA Engineering Sample, 72c @ 3.2GHz
Best single socket time using best compilers (GCC, ICC, AOCC) and best rank/thread decomposition

Flexible distribution of resources

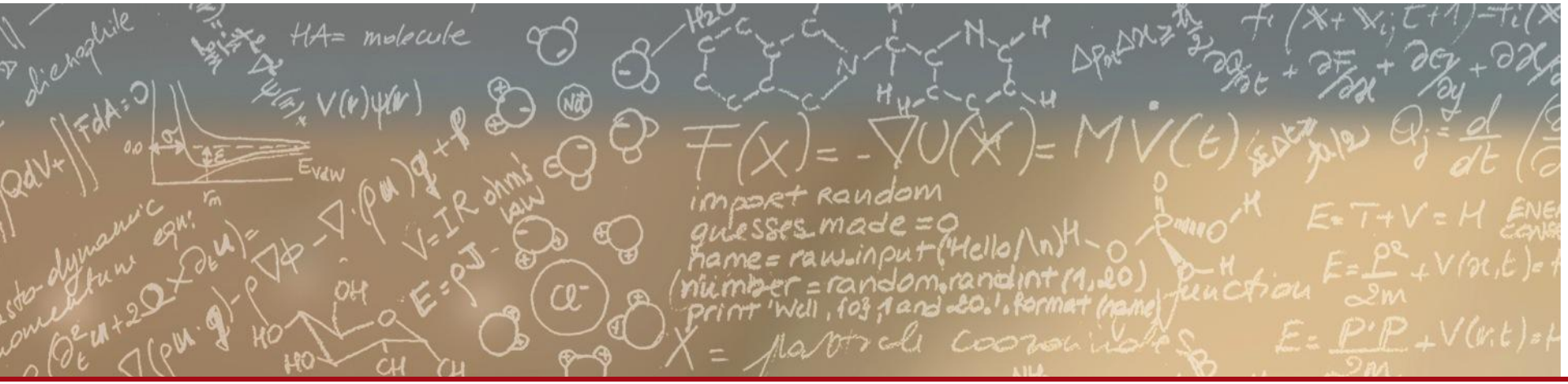
Grace-Hopper supports flexible allocation of CPU and GPU resources over multiple jobs and tasks

One Job exclusive to the node



Job A - 64 Grace CPU MPAM
Job B - 8 Grace Cores MPAM + Hopper GPU

Job A - 8 Grace Cores MPAM + Hopper MIG
Job B - 8 Grace Cores MPAM + Hopper MIG
Etc.



Thank you for your attention.