High-performance computing ecosystem in the Netherlands: Best practices and new developments

Dr. Ing. Valeriu Codreanu
Head of High-Performance Computing and Visualisation, SURF, Amsterdam

What is SURF?

Cooperative of >100 education and research institutions in the Netherlands

Working together in the SURF cooperative to fully utilise the opportunities of digitalisation

Mission: improved and more flexible education and research

Since 1984 the Dutch national HPC center (SARA – SURFsara – SURF)

Compute and data Infrastructures

- HPC systems: currently Snellius ~24PF in 2024. General purpose CPU and GPU serving many/all scientific disciplines
- HTC systems: grid, cloud as infrastructure for LHC, LOFAR, SKA, life sciences,…
- Experimental LIZA system (CDI setup in place)
- Storage: online, offline, backup

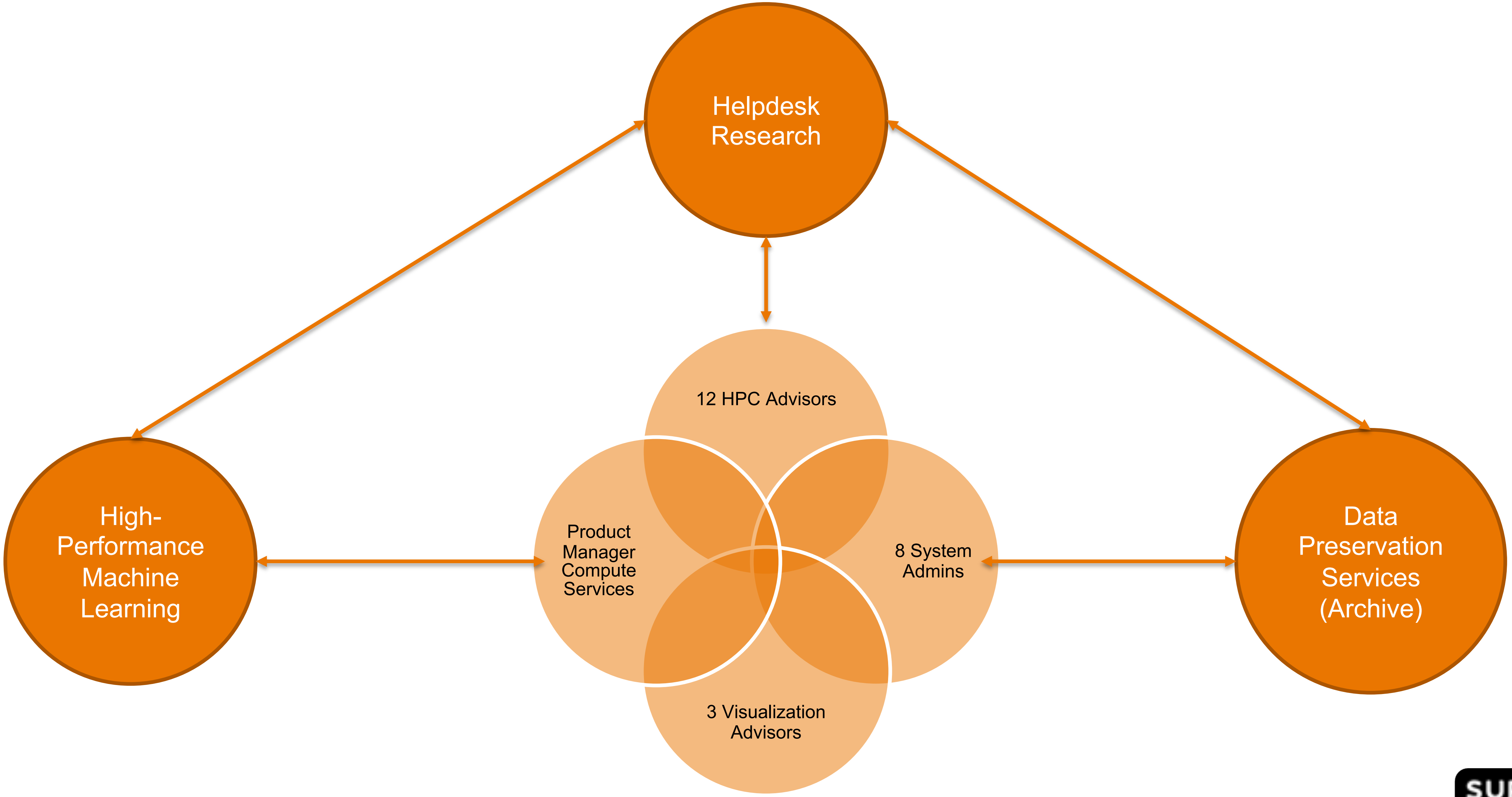
Expertise

- system administration, domain knowledge, HPC/HTC/HPDA expertise, futuring and technology watch, procurement, innovation (green, quantum, neuromorphic, …)

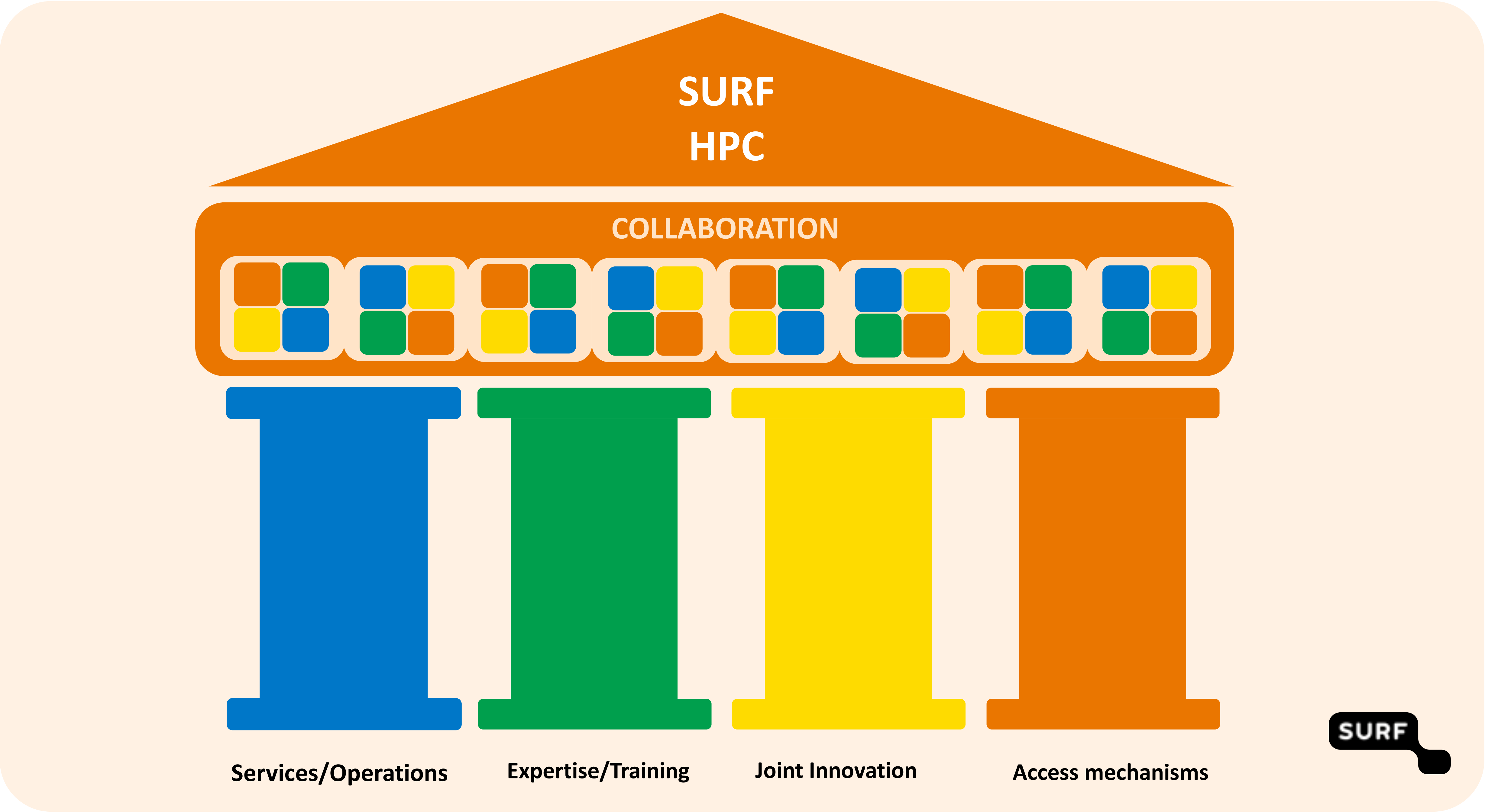
International links

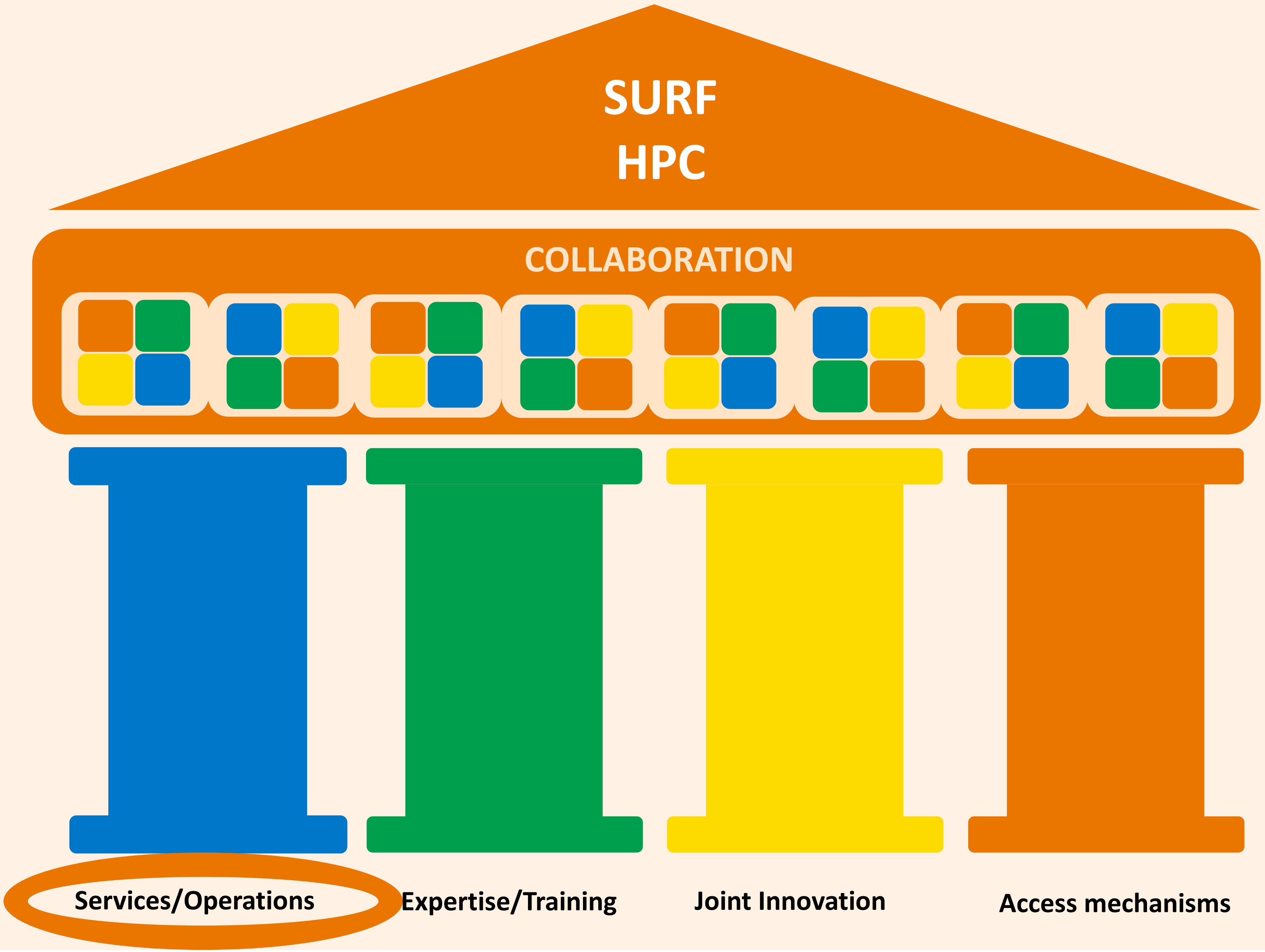
- Founding member of DEISA, PRACE, EuroHPC, EGI, EUDAT, …
- Active in many partnerships and EC projects
- Collaborations with technology vendors (Intel, AMD, NVIDIA)
- Involvement in EuroHPC pre-exascale and exascale consortiums

The HPCV team in SURF



SURF


SURF
HPC
COLLABORATION
Services/Operations
Expertise/Training
Joint Innovation
Access mechanisms
SURF


SURF
HPC
COLLABORATION
Services/Operations
Expertise/Training
Joint Innovation
Access mechanisms

SURF

Dutch National Supercomputer – History

Year	Name	System	R_{peak} flop/s	kW	flop/J
1984	–	CDC Cyber 205-611	100 M	250	400
1988		CDC Cyber 205-642	200 M	250	800
1991	YMP	Cray Y-MP4/464	1.3 G	200	6.7 k
1994	Elsa	Cray C98/4256	4 G	300	13 k
1997		Cray C916/121024	12 G	500	24 k
2000	Teras	SGI Origin 3800	1 T	300	3.4 M
2004	Teras + Aster	SGI Origin 3800 + SGI Altix 3700	3.2 T	500	6.4 M
2007	Huygens	IBM p575 Power5+	14.6 T	375	39 M
2008		IBM p575 Power6	62.6 T	540	116 M
2009		IBM p575 Power6	65 T	560	116 M
2013	Cartesius	Bull bullx B710 + R428 E3	271 T	245	1106 M
2014		+ Bull bullx B515 (NVIDIA K40m)	480 T	289	1662 M
2014		+ Bull bullx B720	1.6 P	791	1972 M
2016		+ Bull sequana X1110 + X1210	1.8 P	884	2085 M

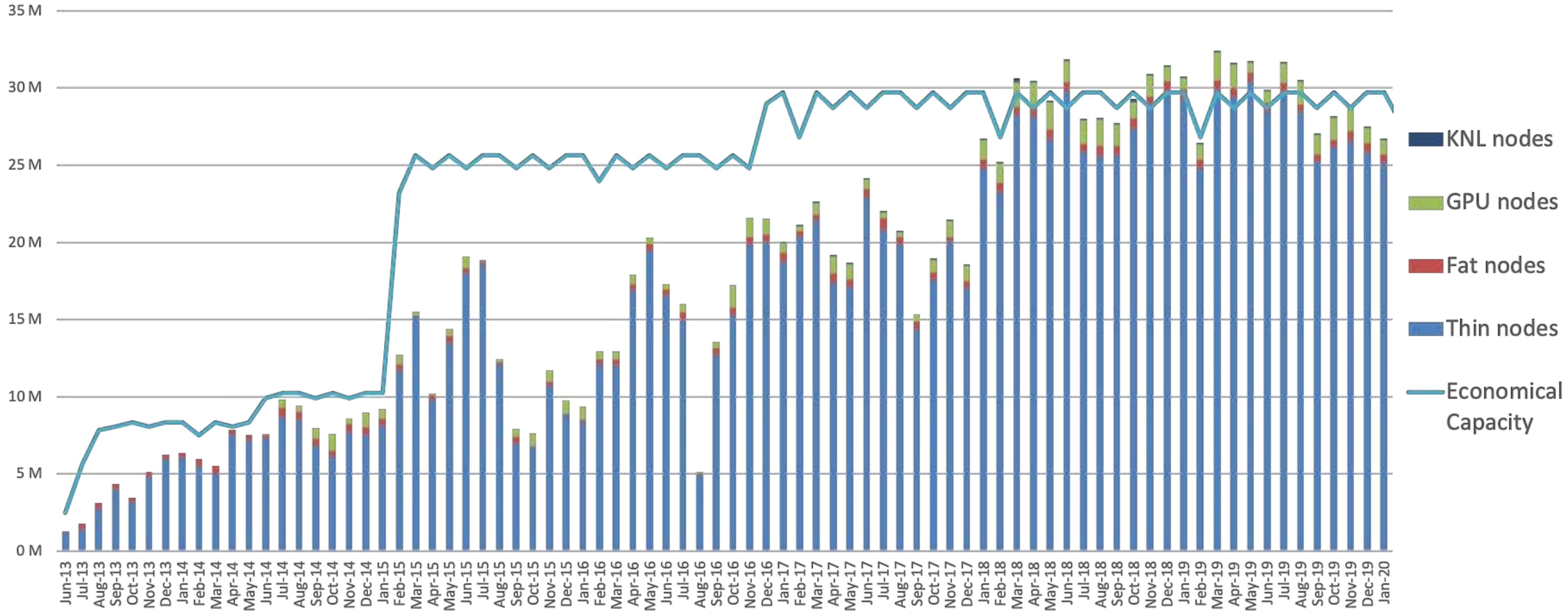
Dutch National Supercomputer – History

Year	Name	System	R_{peak} flop/s	kW	flop/J
1984	–	CDC Cyber 205-611	100 M	250	400
1988		CDC Cyber 205-642	200 M	250	800
1991					
1994					
1997					
2000					
2004					
2007					
2008					
2009					
2013					
2014					
2014					
2016					

1.5 TFLOPS. More than National supercomputer of 2000!

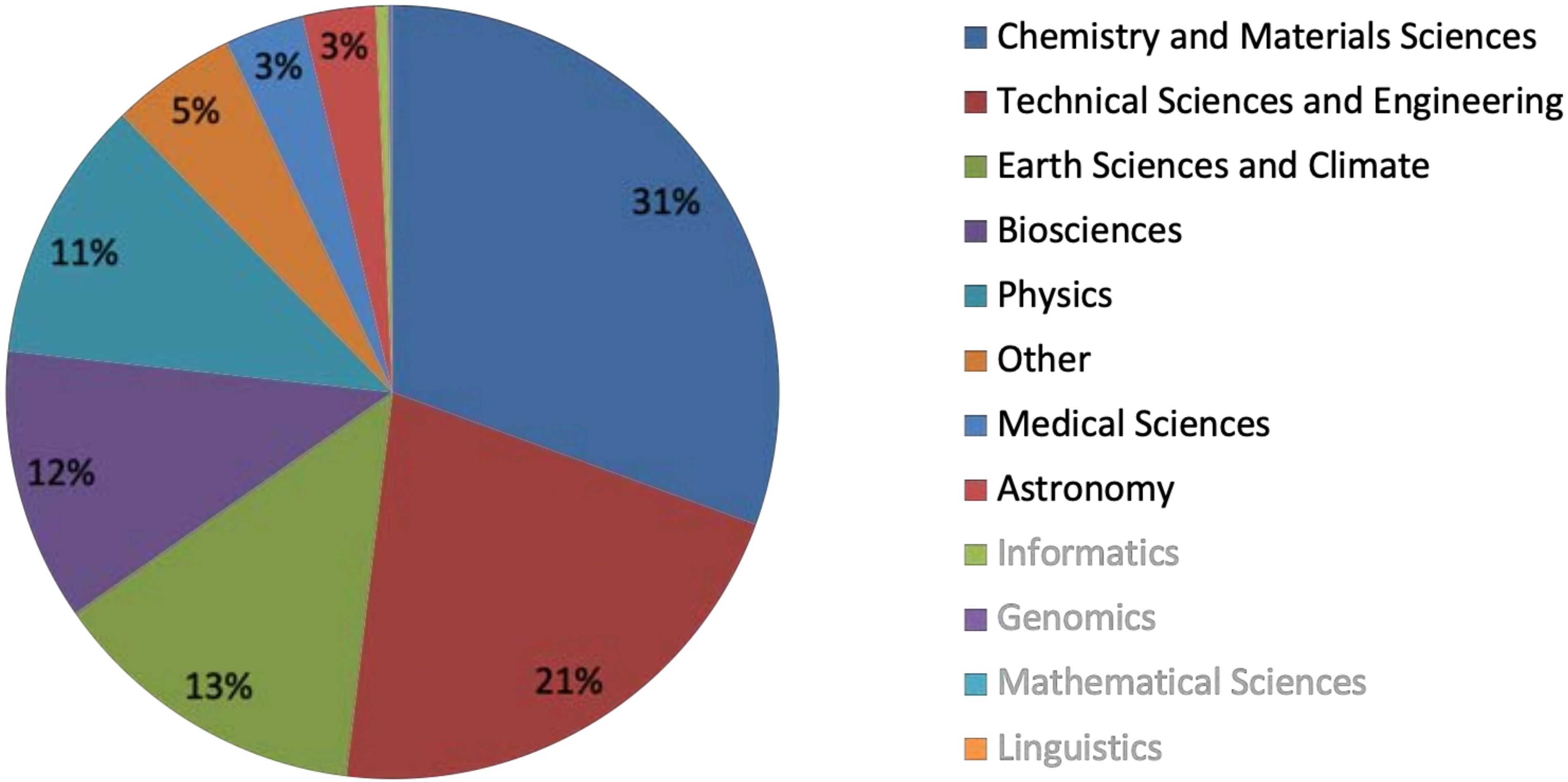
SURF

Cartesius – Usage in Core Hour per Month

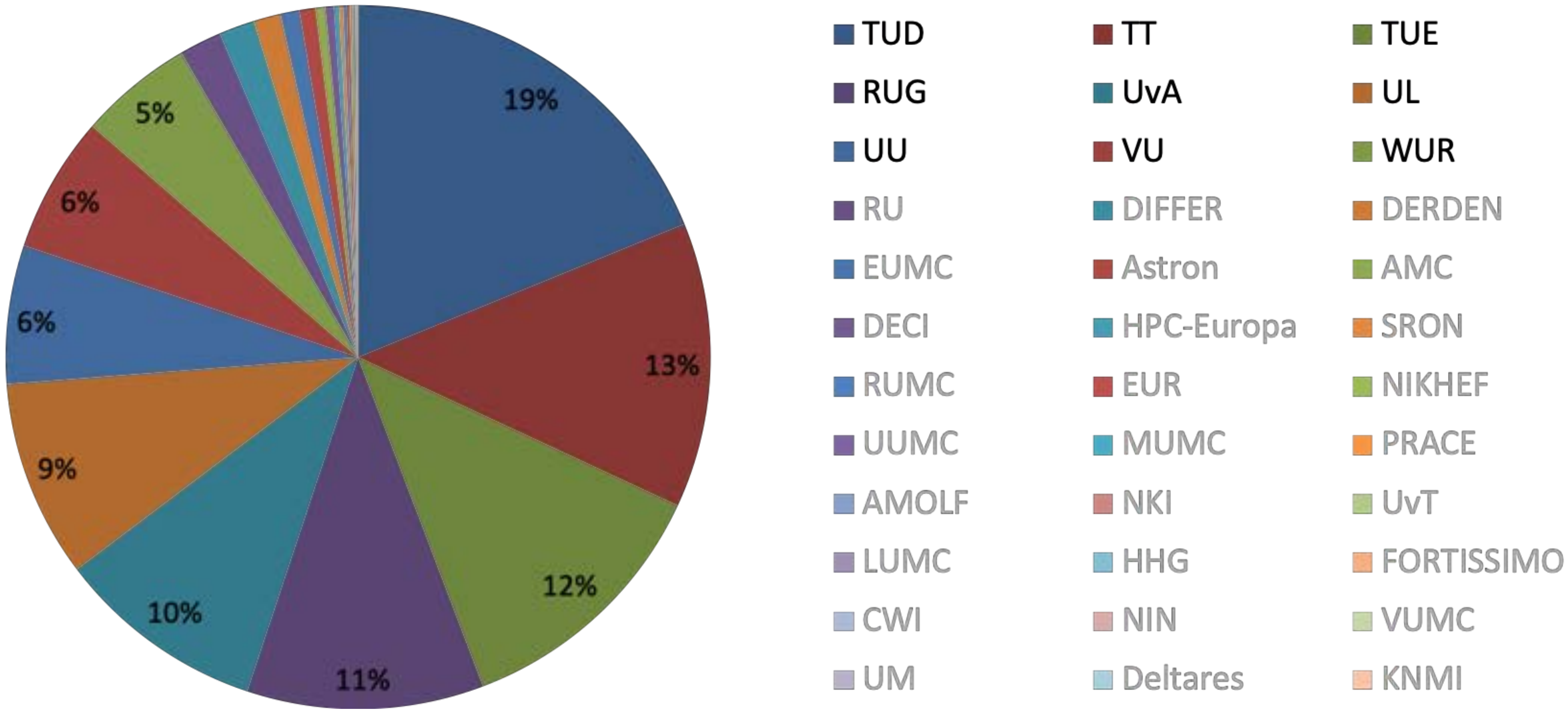


Takes time to fill a new system -> **phased growth**

Cartesius Usage 2019 – Research Area



Cartesius Usage 2019 – Affiliation



Applications on Cartesius

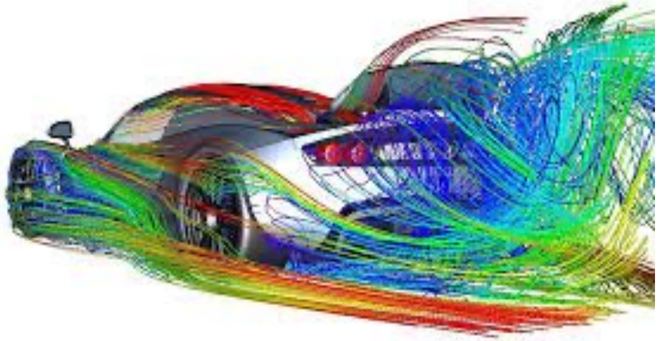
Usage 2019

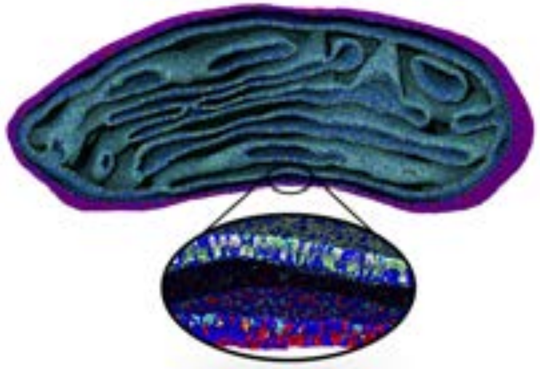
Rank	Name	Percentage
1	Other	17.05%
2	VASP	12.34%
3	Gromacs	10.83%
4	AFiD	8.02%
5	Fluent	3.71%
6	WRF	3.26%
7	CP2K	3.07%
8	ADF	2.81%
9	CESM-POP	2.74%
10	SCAN	2.49%
11	SURFSCAT	2.31%
12	CPMD	2.08%
13	Rbflow	2.01%
14	LAMMPS	1.95%
15	RASPA	1.57%
16	CFCMC	1.57%
17	OpenFOAM	1.37%
18	Orca	1.33%
19	SWAN-ADCIRC	1.27%
20	Gaussian	1.14%

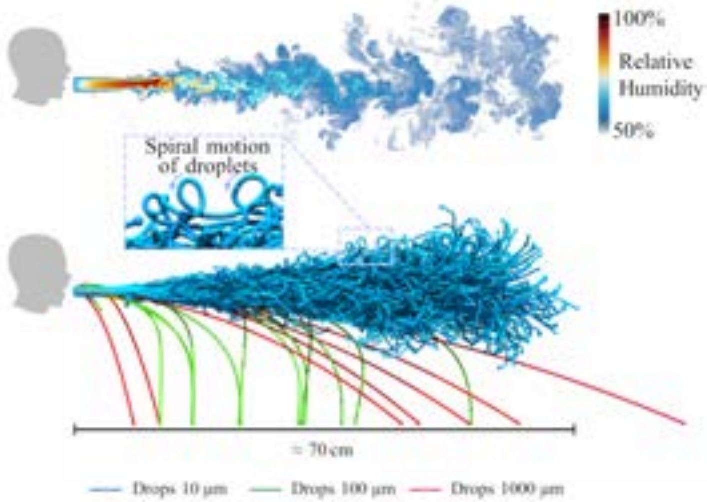
Cumulative Usage 2013 – 2019

Rank	Application	Percentage
1	VASP	14.42%
2	Other	14.17%
3	GROMACS	10.68%
4	AFiD	8.27%
5	CESM-POP	6.18%
6	PRESTO	3.10%
7	CP2K	2.89%
8	SURFSCAT	2.74%
9	SCAN	2.25%
10	ADF	2.20%
11	Fluent	2.08%
12	Gaussian	1.94%
13	CPMD	1.42%
14	LAMMPS	1.33%
15	RASPA	1.28%
16	Rbflow	1.25%
17	LB3D	1.19%
18	WRF	1.19%
19	OpenFOAM	1.07%
20	TMMC	0.97%
(…)		
80	**TensorFlow**	0.08%

SURF Applications Benchmark Suite (SABS) 2020







- Application benchmark codes selected based on
  - use
  - spread across scientific areas
  - scaling (potential)
- **Added Big Data and Machine Learning**
- Selected 8 codes that represent 45% of the workload on Cartesius (2013 – 2019)
- Included GPU benchmarks

SURF Application Benchmark Suite awarding

- Measurements:
  - Time to solution: “speed”
  - Energy to solution
- Throughput: best value for money
  - Combination of “speed” and size
  - “as is” results: main awarding
  - optimized results: awarding bonus
- Scaling: used in architecture awarding
- Relative Energy efficiency: part of Corporate Social Responsibility awarding

SURF

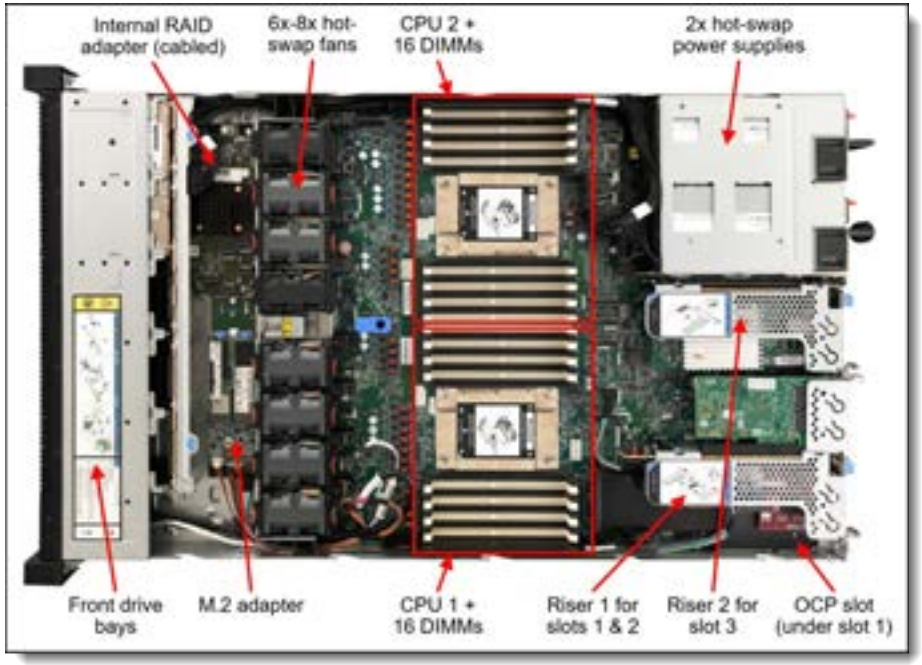
Tendering for a new system (ITT Published: July 2020)

- Competitive dialogue procedure
- Heterogeneous system, able to accommodate more than typical HPC workloads (**Unified Computing**):
  - Thin & fat CPU-only nodes
  - GPU-enhanced nodes
  - High memory nodes
  - Special storage solutions for particularly meta-data intensive applications
    - NVMe based file system
    - Truly node-local or node-dedicated file system storage for a subset of the worker nodes
  - **3-phased** system expansion to full capacity, to take advantage of specific developments in the market

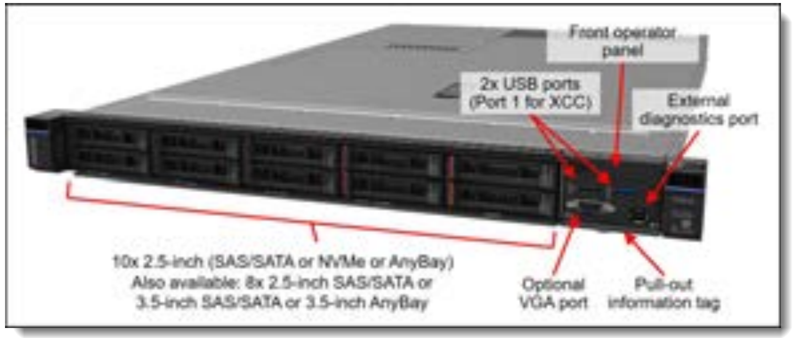
Snellius – New Dutch National Supercomputer

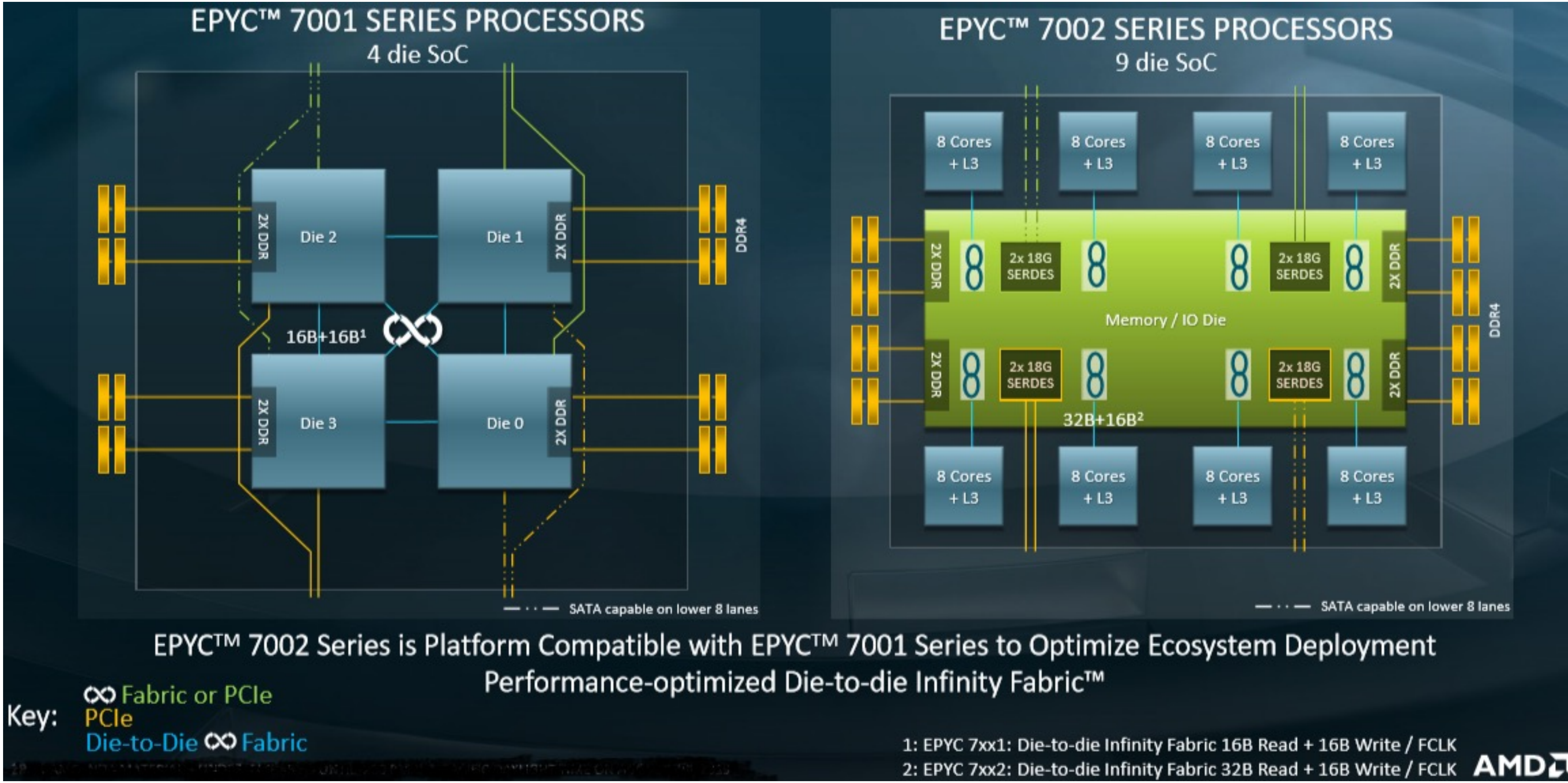
SURF

Snellius service– Phase 1 (Operational since October 2021) (1/2)



- **Phase 1 CPU nodes**
  - All nodes: 2 × 64-core 2.6 GHz **AMD 7H12** (Rome) CPUs
    - Thin nodes: 504 × Lenovo ThinkSystem SR645, 256 GB
    - Fat nodes: 72 × Lenovo ThinkSystem SR645, 1 TB, 6.4 TB NVMe
    - High memory nodes: 2 × Lenovo ThinkSystem SR665, 4 TB
    - High memory nodes: 2 × Lenovo ThinkSystem SR665, 8 TB
  - Total peak performance: 3.1 Pflop/s



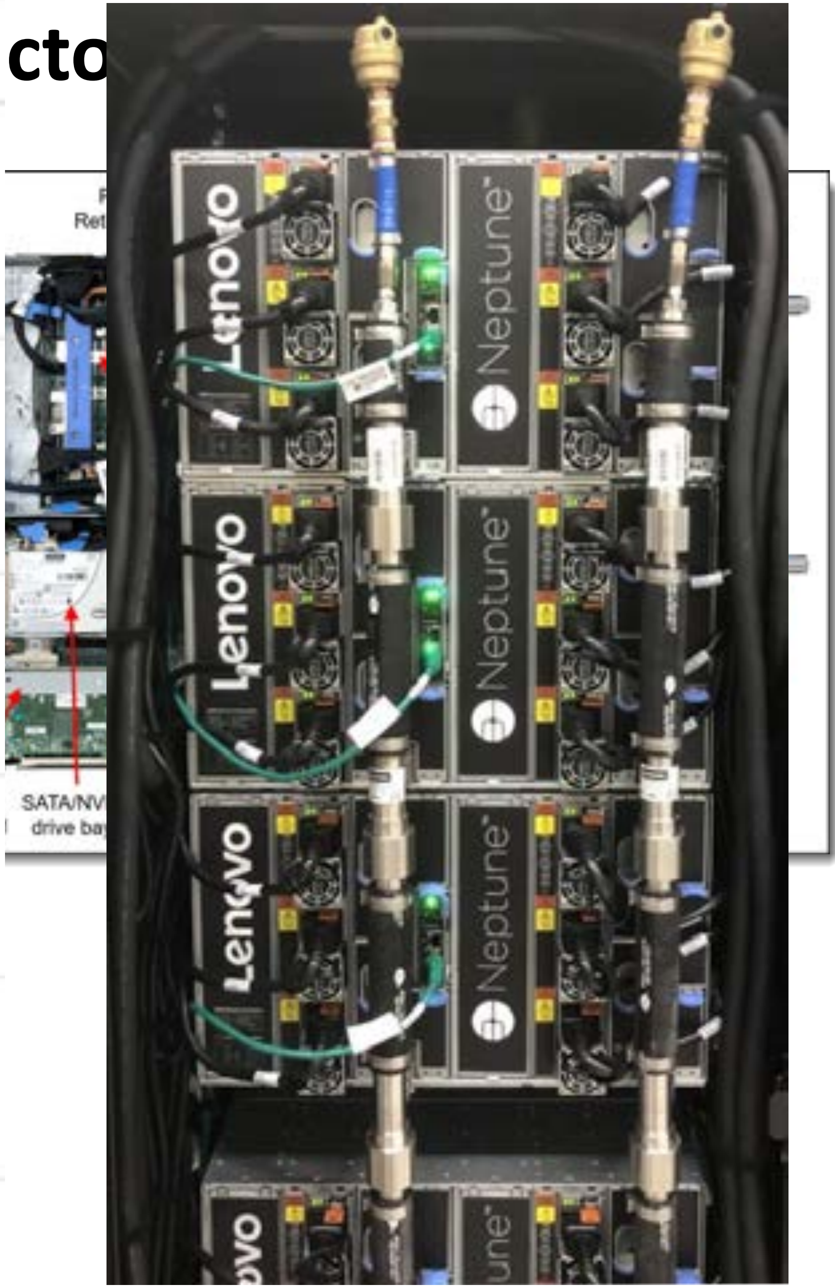

EPYC™ 7001 SERIES PROCESSORS
4 die SoC
Die 2
Die 1
Die 3
Die 0
2X DDR
2X DDR
2X DDR
2X DDR
DDR4
16B+16B¹
SATA capable on lower 8 lanes
EPYC™ 7002 SERIES PROCESSORS
9 die SoC
8 Cores + L3
8 Cores + L3
8 Cores + L3
8 Cores + L3
2X DDR
2x 18G SERDES
2x 18G SERDES
2X DDR
Memory / IO Die
2X DDR
2x 18G SERDES
2x 18G SERDES
2X DDR
32B+16B²
8 Cores + L3
8 Cores + L3
8 Cores + L3
8 Cores + L3
DDR4
SATA capable on lower 8 lanes
EPYC™ 7002 Series is Platform Compatible with EPYC™ 7001 Series to Optimize Ecosystem Deployment
Performance-optimized Die-to-die Infinity Fabric™
Key:
Fabric or PCIe
PCIe
Die-to-Die Fabric
1: EPYC 7xx1: Die-to-die Infinity Fabric 16B Read + 16B Write / FCLK
2: EPYC 7xx2: Die-to-die Infinity Fabric 32B Read + 16B Write / FCLK
AMD

Sne… cto

- Pha…

Green500 Data

Rank	TOP500 Rank	System	Cores	Rmax (TFlop/s)	Power (kW)	Power Efficiency (GFlops/watts)
1	301	**MN-3** - MN-Core Server, Xeon Platinum 8260M 24C 2.4GHz, Preferred Networks MN-Core, MN-Core DirectConnect, **Preferred Networks** Preferred Networks Japan	1,664	2,181.2	55	39.379
2	291	**SSC-21 Scalable Module** - Apollo 6500 Gen10 plus, AMD EPYC 7543 32C 2.8GHz, NVIDIA A100 80GB, Infiniband HDR200, **HPE** Samsung Electronics South Korea	16,704	2,274.1	103	33.983
3	295	**Tethys** - NVIDIA DGX A100 Liquid Cooled Prototype, AMD EPYC 7742 64C 2.25GHz, NVIDIA A100 80GB, Infiniband HDR, **Nvidia** NVIDIA Corporation United States	19,840	2,255.0	72	31.538
4	280	**Wilkes-3** - PowerEdge XE8545, AMD EPYC 7763 64C 2.45GHz, NVIDIA A100 80GB, Infiniband HDR200 dual rail, **DELL EMC** University of Cambridge United Kingdom	26,880	2,287.0	74	30.797
5	30	**HiPerGator AI** - NVIDIA DGX A100, AMD EPYC 7742 64C 2.25GHz, NVIDIA A100, Infiniband HDR, **Nvidia** University of Florida United States	138,880	17,200.0	583	29.521
6	403	**Snellius Phase 1 GPU** - ThinkSystem SD650-N V2, Xeon Platinum 8360Y 36C 2.4GHz, NVIDIA A100 SXM4 40 GB, Infiniband HDR, **Lenovo** SURF Netherlands	6,480	1,818.0	63	29.046

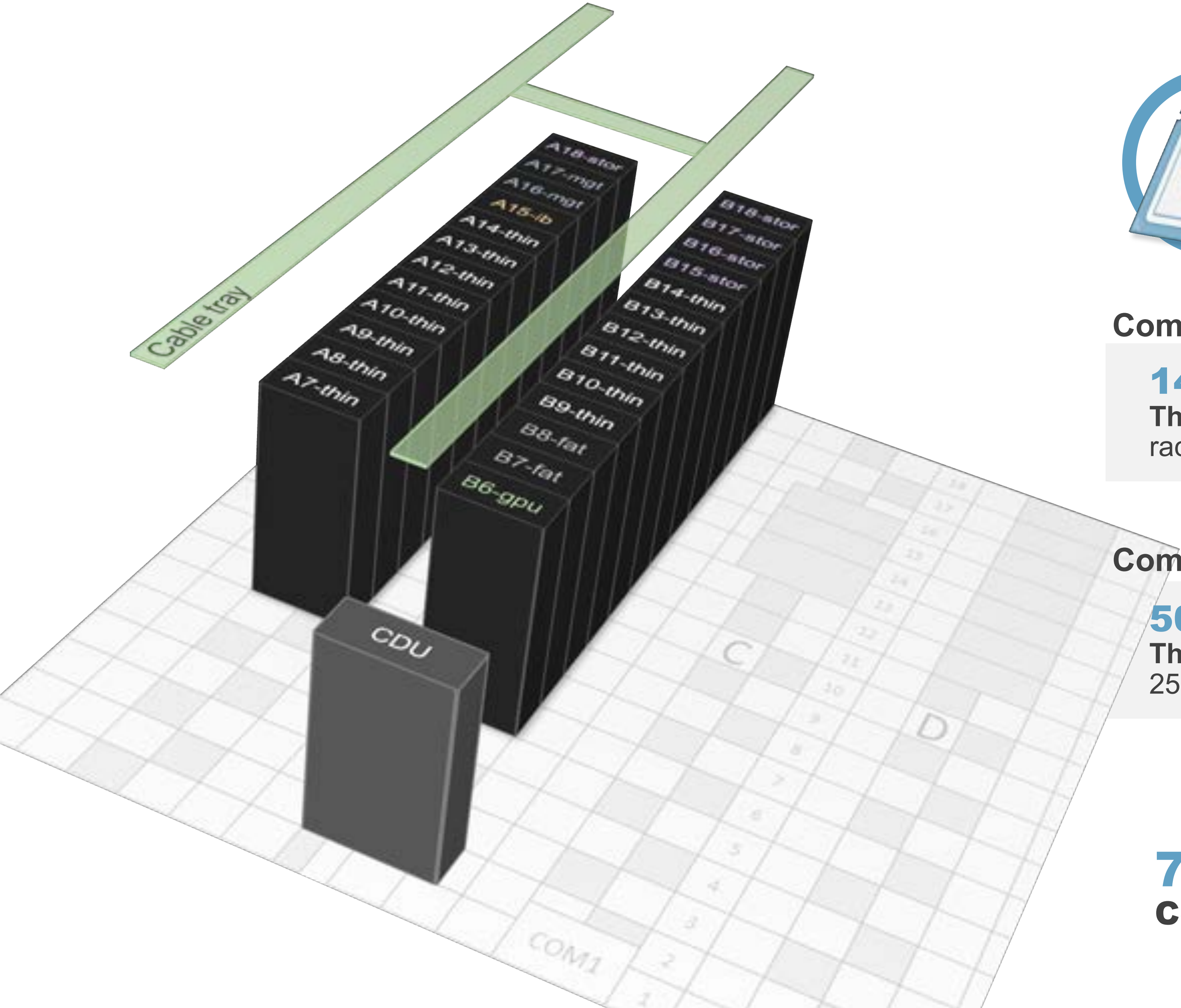


SURF

Snellius – Storage & Interconnect

- Lenovo Distributed Storage Solution (DSS-G) using IBM Spectrum Scale (formerly IBM General Parallel File System, GPFS) file systems
  - HDD based file systems
    - 4 x 2,489 TiB project file systems
    - 2,489 TiB scratch file system
  - SSD based file systems
    - 6 x 120 TiB home file systems
    - 22 TiB admin file system (Serving, among other things, supported applications and libraries, SURF maintained data sets)
  - **Completely NVMe based file system:**
    - **215 TiB project space for metadata intensive work**
- Totals
  - 12.4 PiB project and scratch file systems
  - 720 TiB home file systems

- Mellanox InfiniBand
  - HCAs
    - CPU nodes all phases: Mellanox ConnectX-6 HDR100 (100 Gbps)
    - **GPU nodes phase 1: 2 x Mellanox ConnectX-6 HDR (total: 400 Gbps)**
    - **GPU nodes phase 3: 2 x Mellanox NDR (total: 800 Gbps)**
  - Switches
    - Phase 1: HDR
    - Phase 2 and 3: NDR
- Topology:
  - **Phase 1: non-blocking 2-level fat tree**
  - **Phase 2 and 3: non-blocking 2-level fat tree**
  - Total: (pruned) fat tree



Phase 1

AMD **ROME** 7H12 64C 2.6Ghz

Compute Racks

14 Thin racks	2 Fat racks	1 GPU racks

14 **Thin** racks

2 **Fat** racks

1 **GPU** racks

Compute Nodes

504 **Thin** 256GB

72 **Fat** 1TB

4 **High Mem** 4/8TB

36 **Nvidia A100** GPU w/ IceLake CPUs

76,8K **Cores**

144 **A100** GPUs

3PF **CPU** Rpeak

2,8PF **GPU** Rpeak

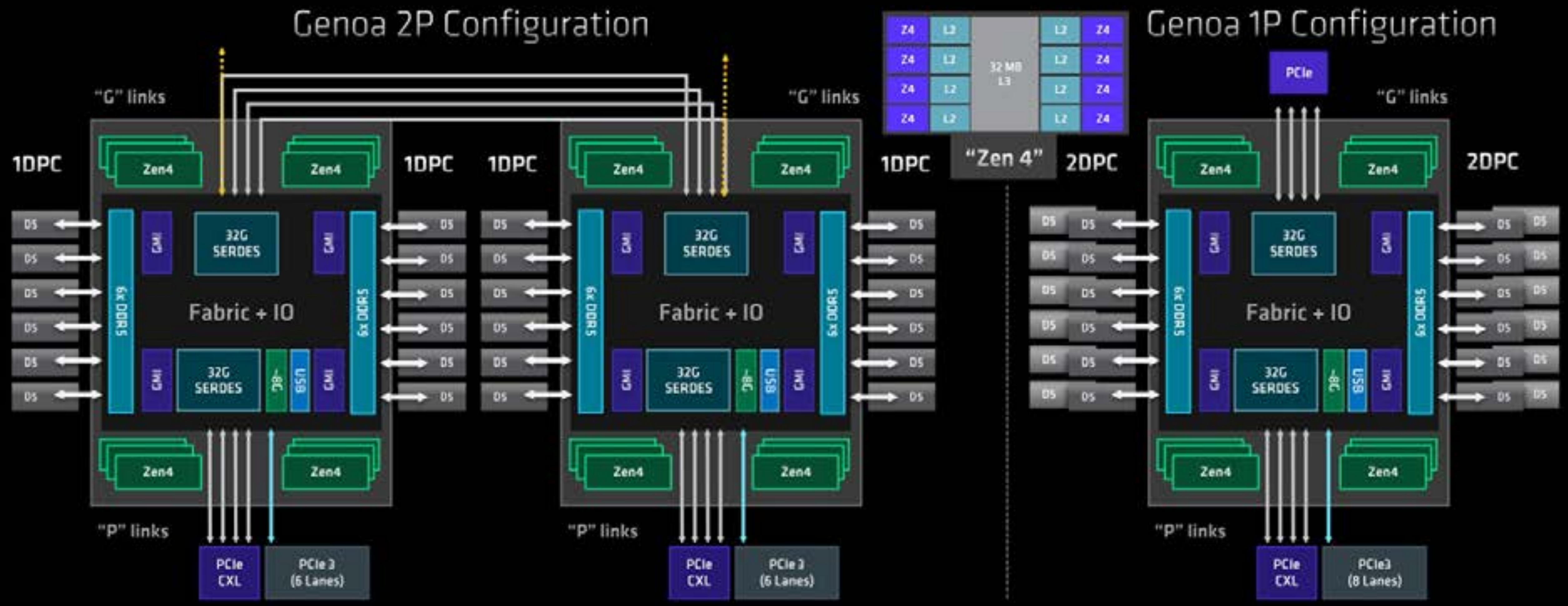
622kW **HPL** / **541kW** **Typical max (85%)** incl storage, mgt and IB

Lenovo

Snellius – Phase 2 and Phase 3

- Phase 2 will be CPU thin nodes only
  - future generation AMD EPYC “Zen 4” CPUs; 2 GB/core; DLC
  - Total peak performance: 5.6 Pflop/s
  - April 2023 delivery, July 2023 in operation
- For Phase 3 we still have three options. The choice will be made 1.5 year after the start of production of Phase 1 (around March 2023) and will be based on actual usage/demand.
  - CPU thin nodes (same as in Phase 2), 2.4 Pflop/s
  - GPU nodes, future generation NVIDIA Hopper GPUs, 10.3 Pflop/s
  - A still to be determined amount of storage
  - Q42023 delivery, Q12024 in operation
- Assuming a GPU-based Phase 3, the total peak performance will reach ~21.5 Pflop/s
- Power requirements of new HW puts pressure on datacenter design!

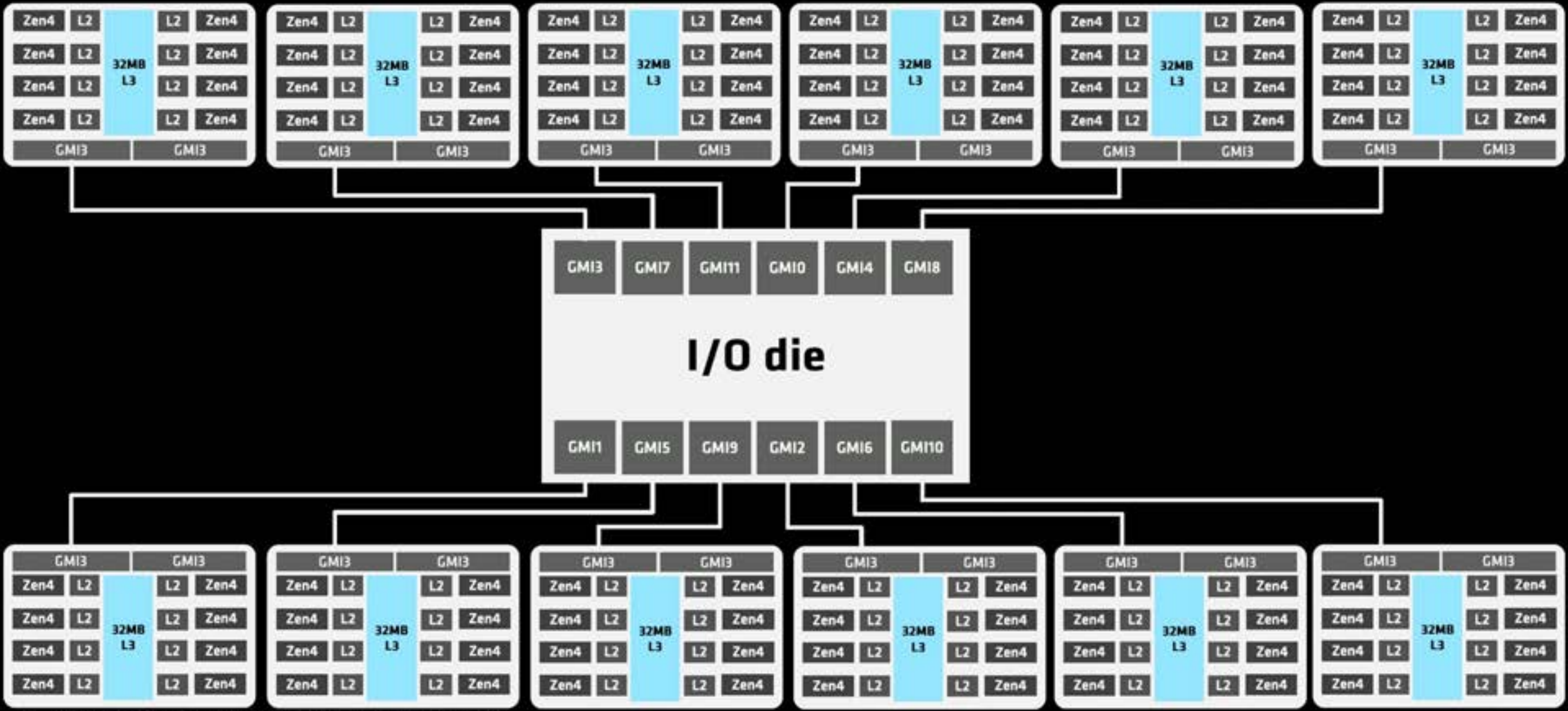
4th Gen EPYC™ SOC Platform Overview



SP5 Platform Delta vs. **SP3**

- 12ch DDR5, 2P-capable, Integrated Server Controller Hub (SCH)
- New socket: Improved SI, increased footprint (~72mm x ~75mm Pkg), 0.94x0.81mm pitch
- 128L: 32Gbps-capable SERDES; PCIe® 5 support, peak xGMI3 product speeds up-to 32Gbps; additional 6-8L PCIe3/socket
- Links: All 8 links PCIe-capable, 3Link or 4Link xGMI 2P topologies (not shown), "P" links CXL-capable

AMD EPYC™ 9004: 12 CCD Configuration



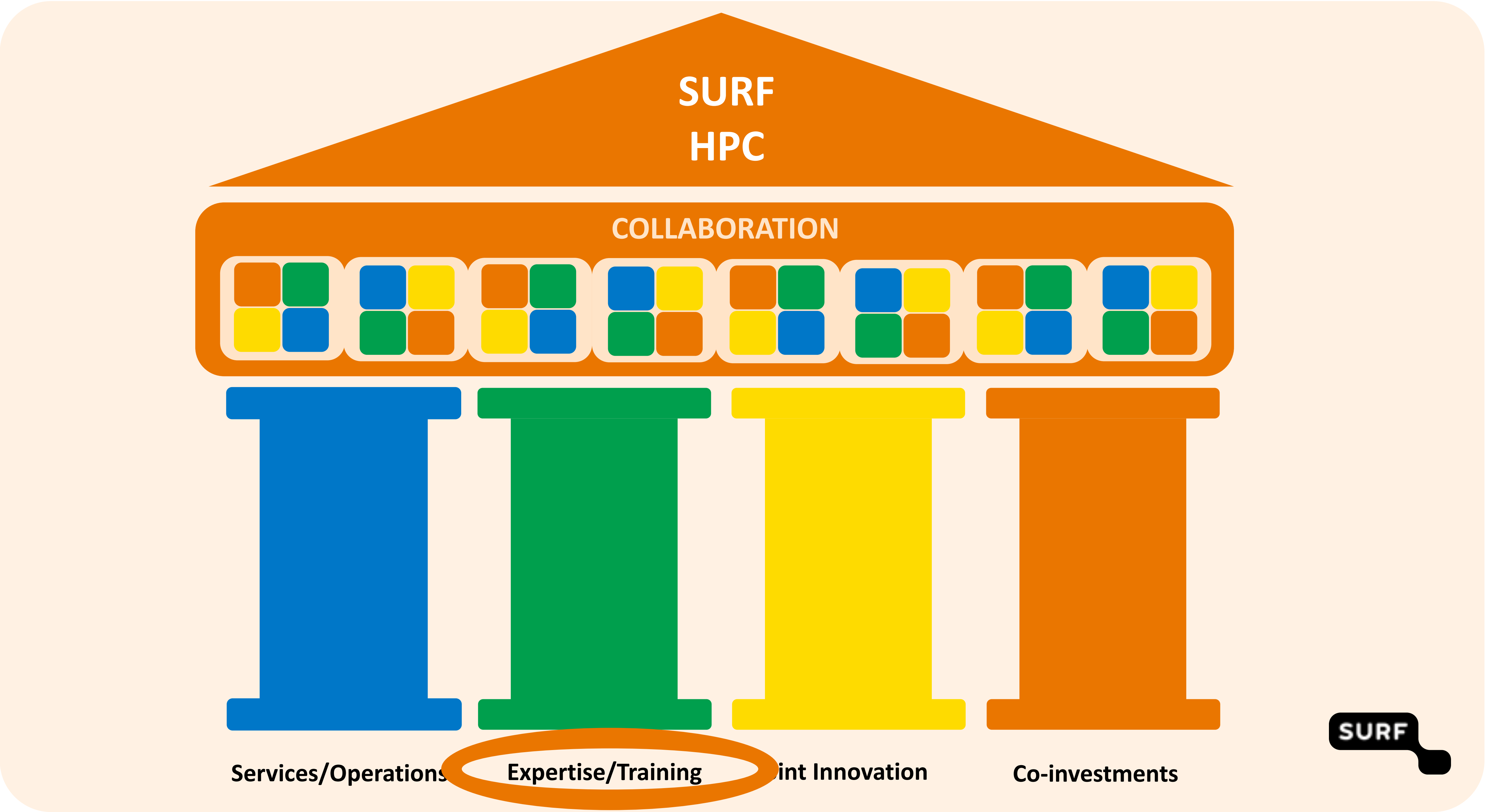
Dutch National Supercomputer

Year	Name	System	R_{peak} flop/s	kW	flop/J
1984	–	CDC Cyber 205-611	100 M	250	400
1988		CDC Cyber 205-642	200 M	250	800
1991	YMP	Cray Y-MP4/464	1.3 G	200	6.7 k
1994	Elsa	Cray C98/4256	4 G	300	13 k
1997		Cray C916/121024	12 G	500	24 k
2000	Teras	SGI Origin 3800	1 T	300	3.4 M
2004	Teras + Aster	SGI Origin 3800 + SGI Altix 3700	3.2 T	500	6.4 M
2007	Huygens	IBM p575 Power5+	14.6 T	375	39 M
2008		IBM p575 Power6	62.6 T	540	116 M
2009		IBM p575 Power6	65 T	560	116 M
2013	Cartesius	Bull bullx B710 + R428 E3	271 T	245	1106 M
2014		+ Bull bullx B515 (NVIDIA K40m)	480 T	289	1662 M
2014		+ Bull bullx B720	1.6 P	791	1972 M
2016		+ Bull sequana X1110 + X1210	1.8 P	884	2085 M
2021	**Snellius**	**Lenovo ThinkSystem SR645 + SD650-N V2**	**6.1 P**	**620**	**9836 M**
2022		**+ Lenovo ThinkSystem (CPU)**	**11.2 P**	**1200**	**9332 M**
2023		**+ Lenovo ThinkSystem (CPU or GPU)**	**13.6–21.5 P**	**1430**	**15034 M**

SURF

Snellius – Software

- Compilers
  - GCC/FOSS
  - AMD AOCC
  - Intel oneAPI (first three years)
  - NVIDIA HPC SDK (includes PGI compilers and tools)
- **Broad software stacks built using EasyBuild**
- Batch system
  - SLURM
- EAR (energy aware runtime) to monitor and manage energy consumption
  - collaboration to extend EAR support for AMD hardware
- Operating System(s)
  - Red Hat 8


SURF
HPC
COLLABORATION
Services/Operations
Expertise/Training
int Innovation
Co-investments

HPC support and consultancy

Supporting the Dutch HPC community

HPC systems usage

- 1st line Helpdesk: accounts managements, basic problems and support (Jira)
- 2nd/3rd line Helpdesk: complex problems, software installations, etc. (Jira)
- Documentation and Trainings (Confluence)
- Reporting to stakeholders (Qlik sense)

HPC systems access

- Support with NWO large applications
- Pilot projects directly at SURF

Supporting the Dutch HPC community

Promising application programme

- Specialised support for HPC workloads
- Additional support coupled to account on our systems
- Collaboration with external institutes (e.g.: eScience Centre, EU projects, local HPC)
- Topic of interest:
  - Code benchmark and profiling
  - GPU porting, and validation
  - Code performance optimization
  - Service development/Tools integrations

Contributions to EU projects/initiatives

PRACE

CompBioMed

EESSI
EUROPEAN ENVIRONMENT FOR SCIENTIFIC SOFTWARE INSTALLATIONS

EURO
CC

Training courses for research

Want to get started with our systems but lack the necessary knowledge? We regularly organize hands-on systems training courses at our offices in Utrecht and Amsterdam or at your education or research institution. You can also include the training courses in the educational programme of your institution.

Systems training

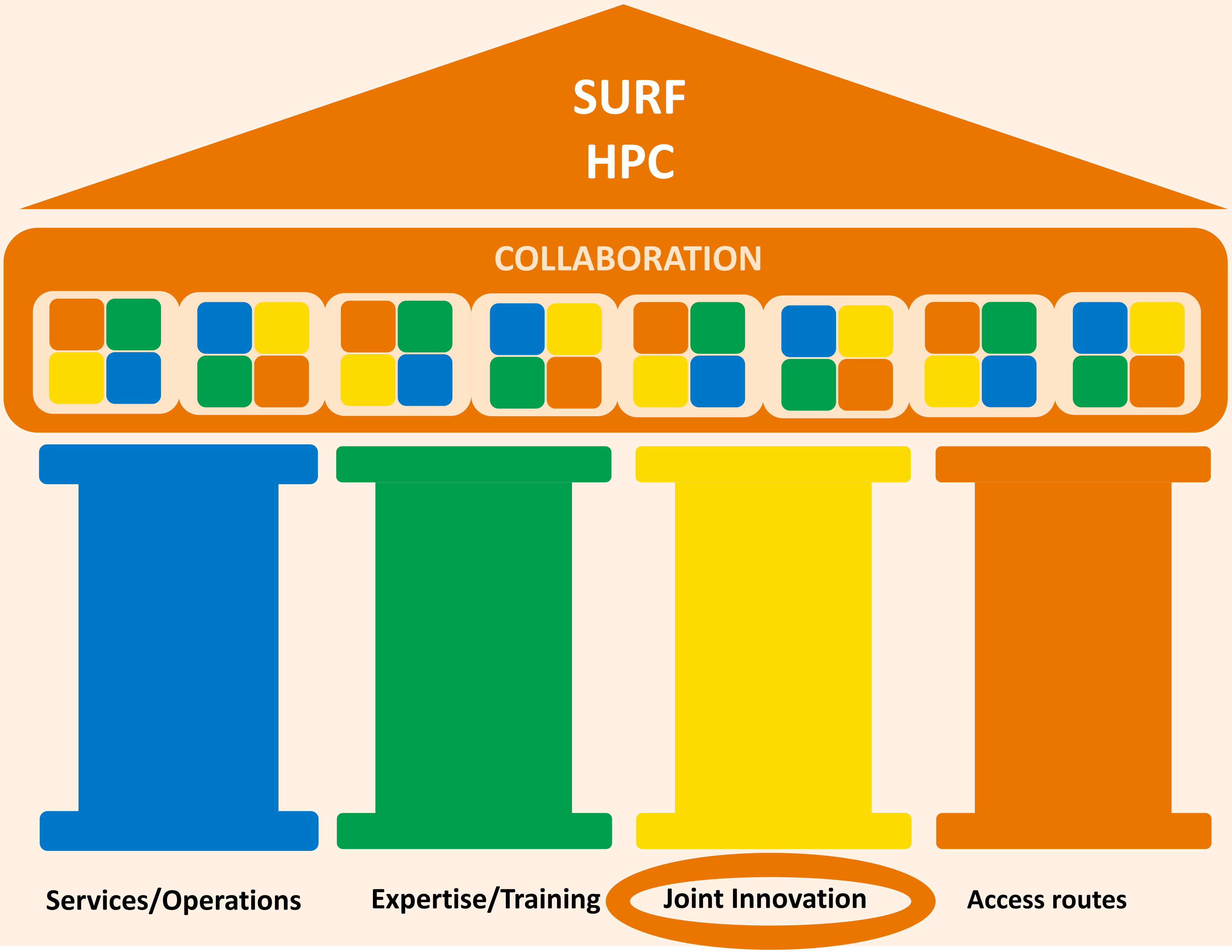
Learn how to work with our systems.

- Supercomputing
- HPC Cloud
- Data management

Technical skills

- Parallel programming
- Machine learning
- Big data
- Visualisation
- Software containers

SURF


SURF
HPC
COLLABORATION
Services/Operations
Expertise/Training
Joint Innovation
Access routes

Current and future priorities

Co-exploration and design of future applications and infrastructure

Trusted infrastructures / Responsible computing

Manageability / Reproducibility

Usable infrastructures

Sustainable infrastructures

Federated and interoperable infrastructures

Co-exploration and design of future applications and infrastructure

- Creating and professionally managing experimental testbeds (LIZA)
- Engaging with the community for mapping novel workflows
- Addressing programmability and heterogeneity
    - For modern/accelerated computing systems
- Disseminating best practices

Quantum Computing at SURF

At SURF we have a tailored development and execution environment for (hybrid) quantum simulations.

Hardware:

- **Four nodes each with 1.5 TB RAM, two NVIDIA TitanRTX GPU cards and 4 sockets CPUs with 10 cores/socket (Intel Gold Skylake) in LISA cluster computing system.**
- Computing capacity of the national supercomputer Cartesius

Software:

- An easy to use module environment with all the neccesary tools
- **Singularity container (Quantum Bundle) with several ready-to-go quantum tools**
- QuTech's Quantum Inspire Platform (more info here)

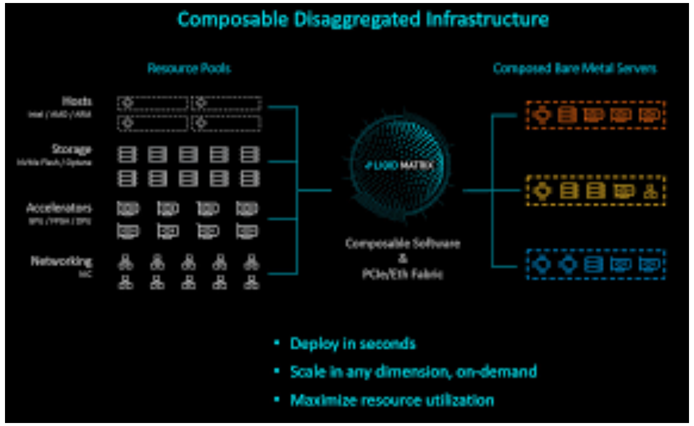
NextSilicon hardware

NextSilicon Maverick™ is a pioneering new **general-purpose coprocessor**. Its revolutionary **software-defined hardware** architecture and **dataflow approach** vastly accelerate high-performance computing (HPC). Improved utilization of silicon resources yields unprecedented performance and lower power consumption.

System on a chip

SURF and Nikhef collaborate on innovative GPU powered compute system

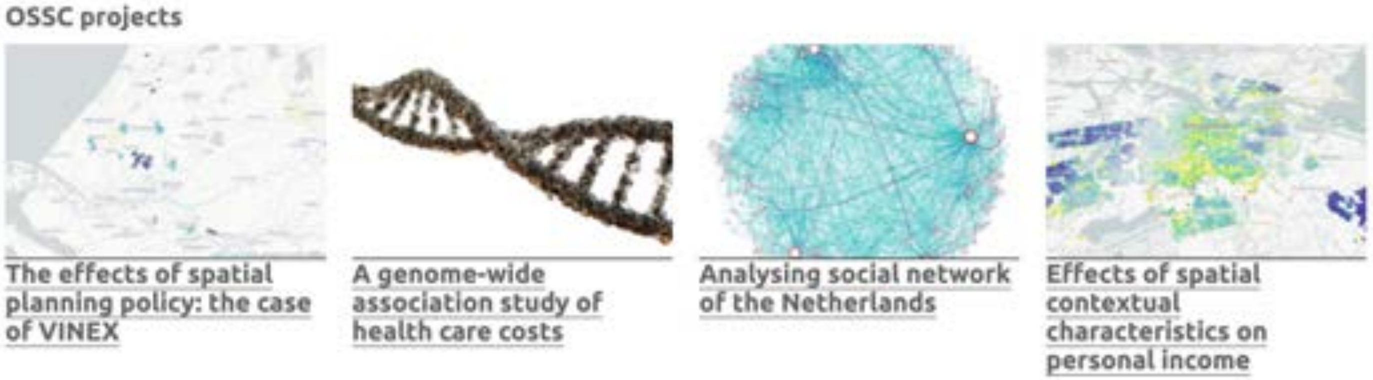
Starting this month, the SURF Open Innovation Lab (SOIL) has introduced a new, experimental compute cluster based on AMD Radeon Instinct MI50 GPU accelerators. The cluster consists of 6 servers each powered by 8 AMD GPUs. These servers are unique in the Netherlands and will be accessible to all Dutch researchers.



Trusted infrastructures / Responsible computing

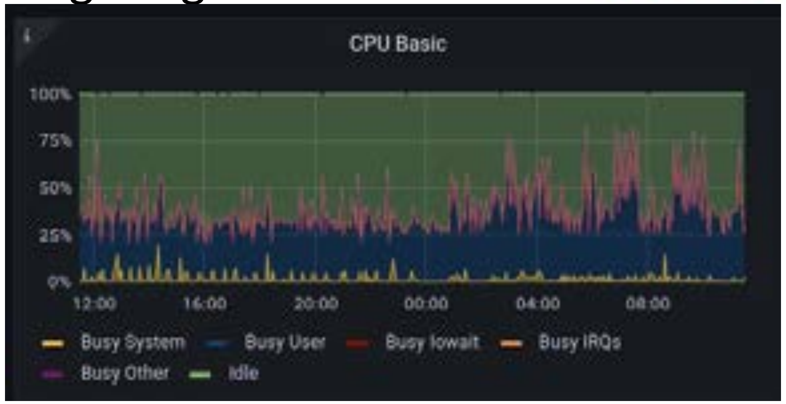
- Data sources are becoming more diverse, each with various degrees of confidentiality
- Combining data sources can unleash social sciences
  - But this requires trust
- Traditionally processing done on-prem
  - But HW requirements are increasing
- Trusted large-scale data and compute environments are required
  - Data sovereignty needs to be preserved

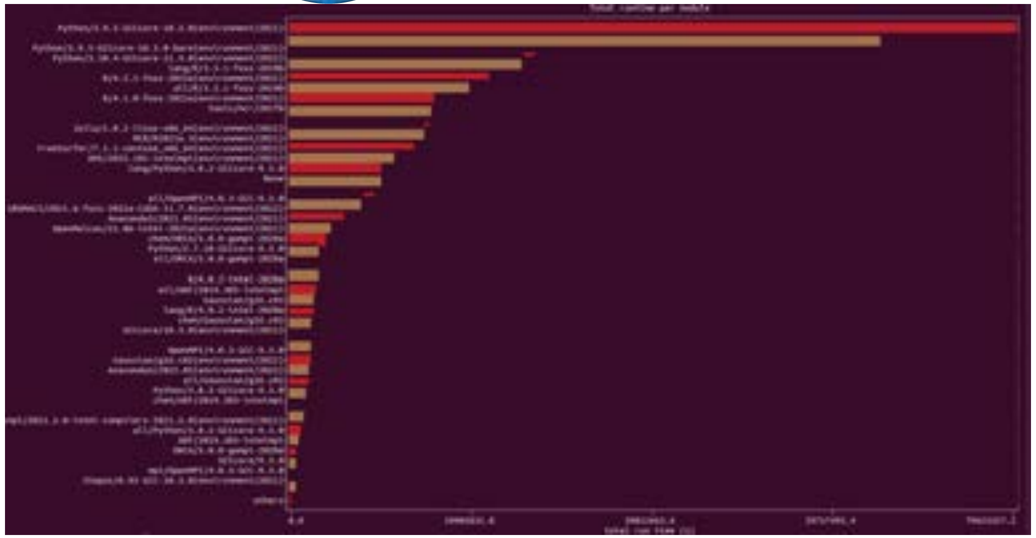
Using the ODISSEI Secure Supercomputer



Manageability / Reproducibility

- Researchers require professionally managed systems, not just HW access
- Professionally managing a heterogeneous system becoming increasingly involved
  - Complexity is only increasing
- Automation required at all levels
  - Building SW stacks in HW-agnostic way
  - Testing functionality/performance
  - Monitoring usage





Usable infrastructures

- The landscape of researchers is diversifying
  - Need more easy-to-use solutions
  - Easier/clearer access mechanisms
- HW/SW/workflow complexity is increasing
  - Need increased and diversified support/collaborations, nationally and internationally

OPEN OnDemand

Pages / SURFsara Knowledge Base / Services

JupyterHub for education

Created by Caspar van Leeuwen, last modified on Nov 19, 2020

What is JupyterHub?

JupyterHub provides a multiuser environment for Jupyter notebooks. The SURF JupyterHub service facilitates external courses, e.g. programming courses, and runs on our Lisa cluster.

Each course receives its own instance of JupyterHub. Users who log in will be provided with their own Jupyter Notebook Server, where they can create, download, upload and run notebooks. The service provides functionality for teachers to easily share notebooks, data and installations with their students.

NOTE: A Jupyter environment for researchers working on Lisa is also in development, but this documentation does not apply to that!

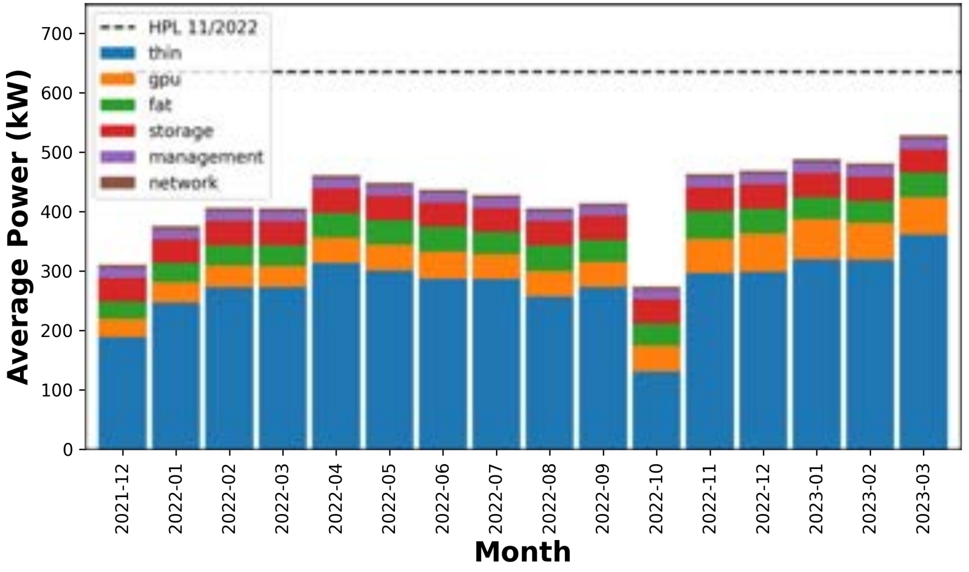
SURF

Sustainable infrastructures

EAR Energy Aware Runtime

- Energy impact of ICT is growing
  - Has financial and environmental impact
- Supporting green research
  - Developing tools to increase awareness
  - Incentivizing green computing
  - Modern datacenter technologies
- Use of accelerated technologies
  - Modernizing applications and workflows

Energy visualization prototype @ SURF



CPU

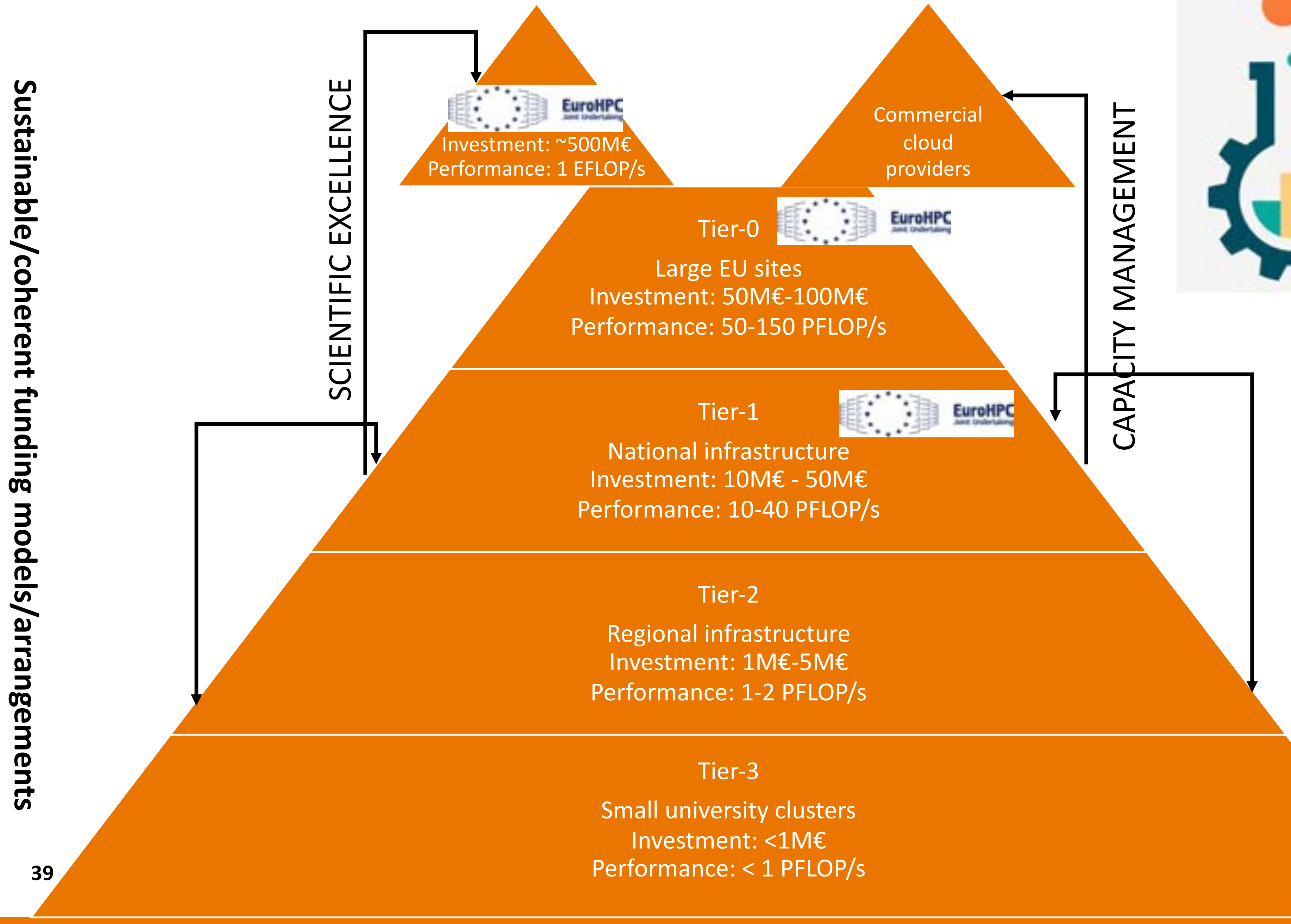
Theoretical (Rpeak)
2.13 PFLOP/s

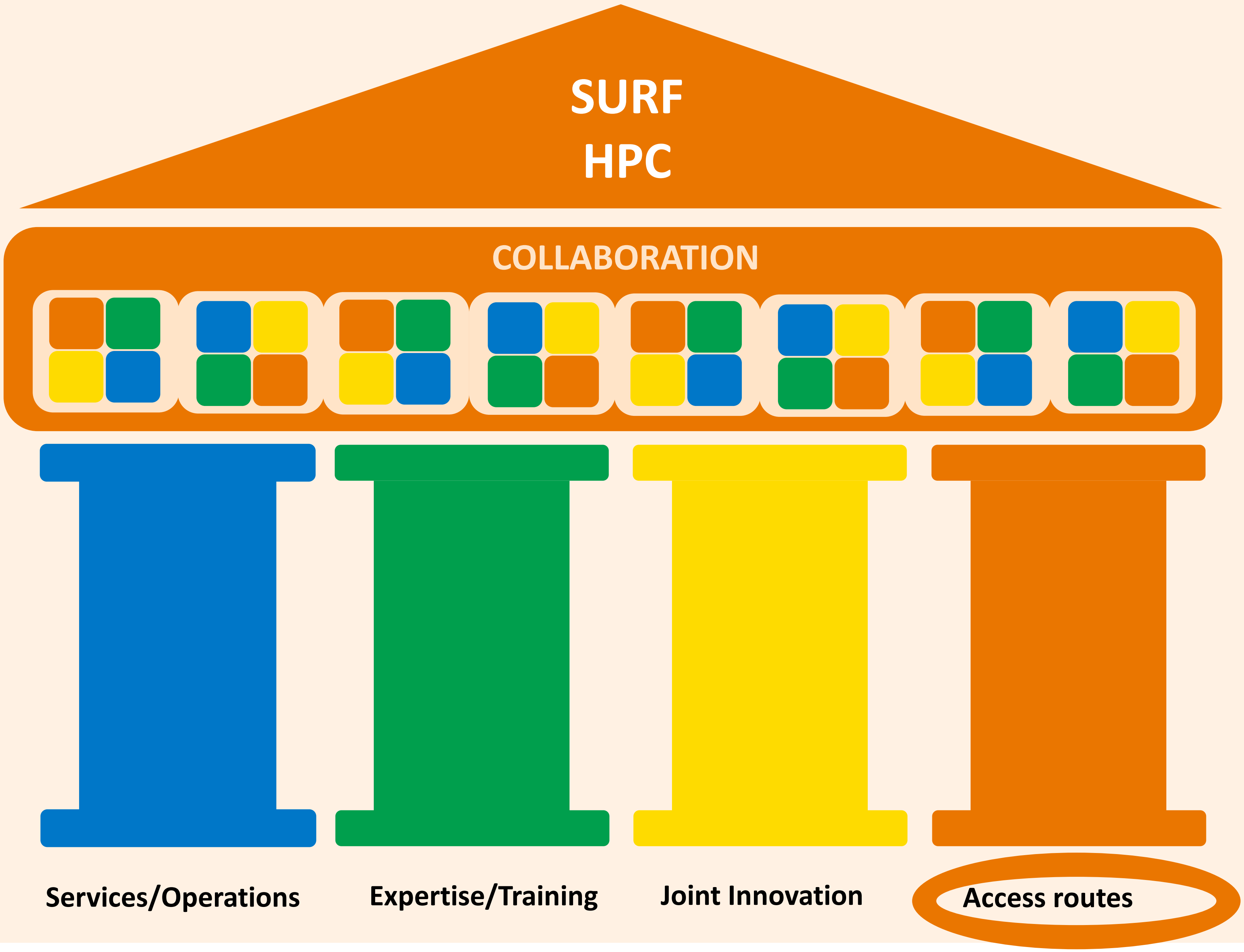
GPU

Theoretical (Rpeak)
4.34 PFLOP/s

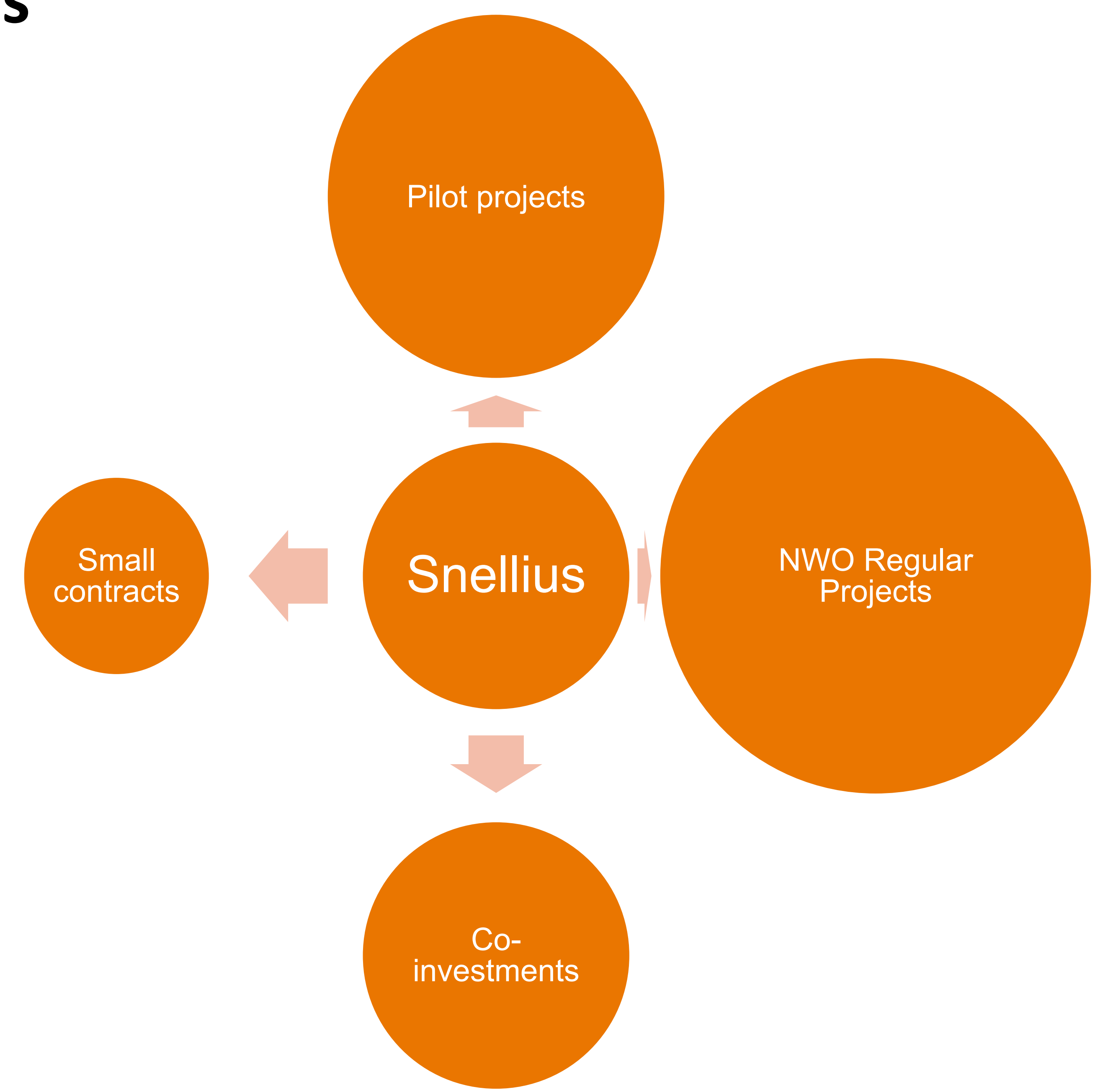
GPU x2 compute for 1/4 the energy

Federated/interoperable infrastructures




SURF
HPC
COLLABORATION
Services/Operations
Expertise/Training
Joint Innovation
Access routes

Access routes



Snellius – Co-investment access route

- 3-year commitment from participants required
  - Participating institute receives the equivalent “economical capacity” of the investment in **each** of the 3 contract years
  - Smallest participation unit: 1 CPU/GPU node
- Snellius introduced a dynamic growth model
  - Institutes that co-participate can
    - Prioritize access for given research groups -> Resource reservations
    - Access to latest HPC technology -> exploiting technology roadmaps
    - Economy of scale benefits -> Lower costs
    - Easier access to resources (no review process)
- Two co-investment rounds confirmed

Let’s grow Snellius together

Allocation Size	Variable
Resources granted	1-2 days
Application form complexity	N/A

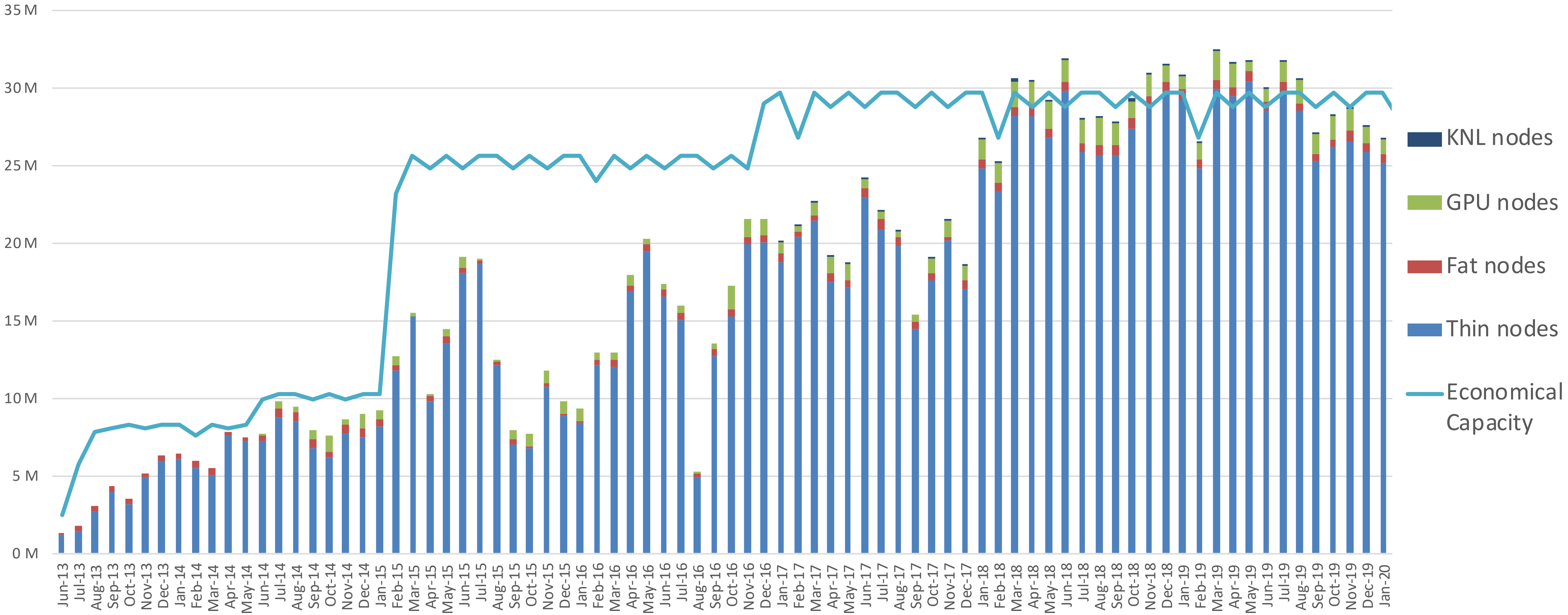
LISA in Snellius

- Two co-investment rounds confirmed
  - Phase 1A is operational Q4 2022 – CPU (21), GPU (**36**), and storage capacity (2PB)
  - Phase 2A is expected end of Q2 2023 – 1 extra Phase 2 CPU rack (72)
    - SURF also co-invests here and offers “small” SBU contracts
    - This capacity is relatively limited (a cap will be imposed)

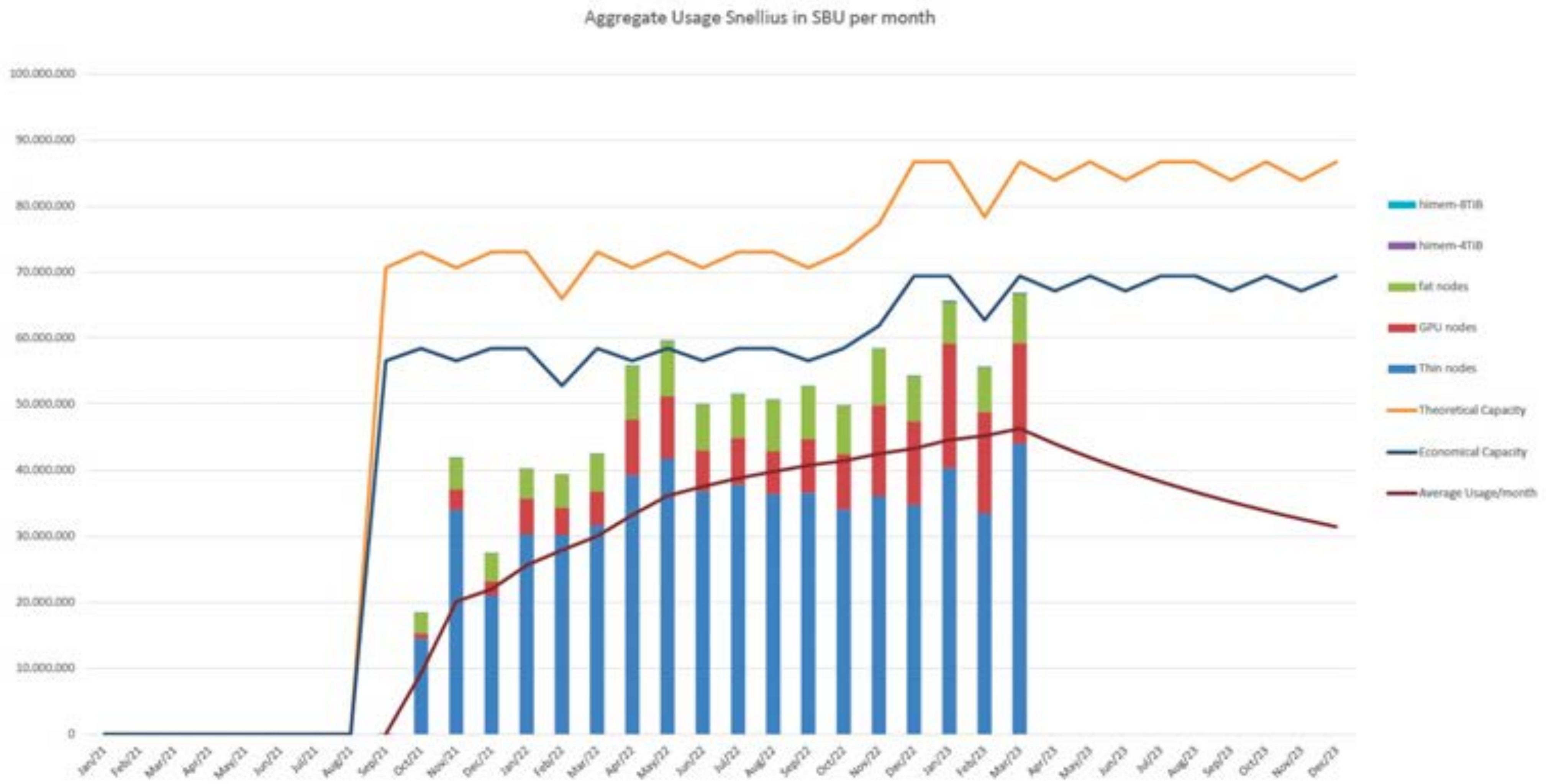
	LISA	Snelliua P2A
# cores	~5000	13824
Performance	~302 TFLOP/s	~569 TFLOP/s
Local disk/node	1.7TB	6.4TB
Memory/node	96GB DDR4	384GB DDR5
Interconnect	10Gbe	HDR100

SURF

Cartesius – Usage in Core Hour per Month



Snellius – Usage in Core Hour per Month



System at a glance

1.525 Compute nodes

230K CPU Cores

552 GPUs

>24PF Rpeak

580 + 21
AMD ROME CPU compute nodes

36 + 36
Nvidia A100 GPU compute nodes

76,8K + 2.7K Cores

144 + 144 A100 GPUs

3PF +0.15PF CPU Rpeak

2,8PF + 2,8PF GPU Rpeak

Phase 1+1A

NL-SAS, NVMe and SSD
storage

>13.3PiB + 2PiB Usable storage capacity

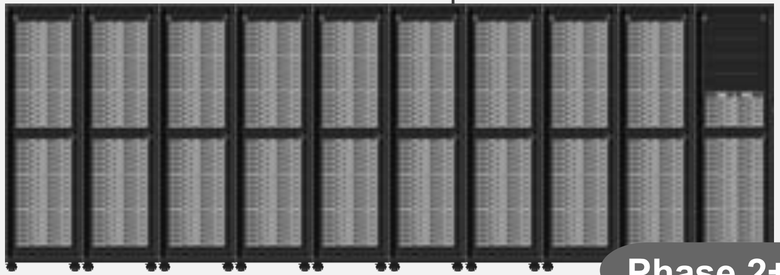
>300GiB/s Sequential perfromance

Storage

714 + 72
AMD NG CPU compute nodes

137K + 14K Cores

5,63PF + 0,5PF CPU Rpeak



Phase 2+2A

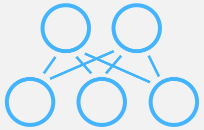
66 + ?
Nvidia Hopper GPU compute nodes

264 + ? GPUs

13PF + ? GPU Rpeak

Phase 3

HDR and NDR
Fat-Tree Mellanox infiniband fabric



High speed network

Conclusion

- The Snellius system is well supported and accepted by the community
- Emerging AI workloads, and AI/HPC workloads are growing -> GPU partition is highly used
- **Heterogeneous** (compute & storage) solutions cater Dutch science use-cases best -> **Unified computing!**
- Phased-system growth optimally exploits vendor roadmaps
- On-demand growth facilitates **HPC federation** efforts
- **Energy efficiency** and sustainability are key drivers
- **Usability, accessibility, interoperability are key challenges**
- Increased power density will be the norm -> 100kW+ racks

Thank you!