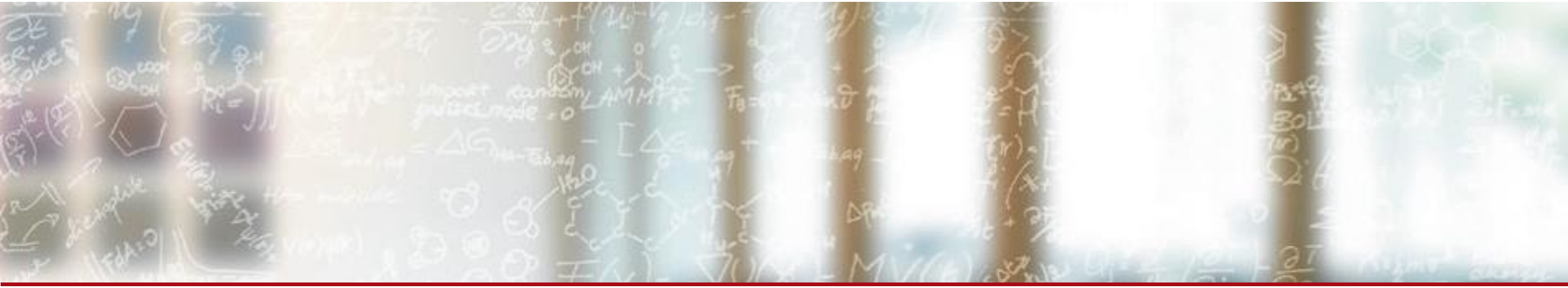
CSCS
Centro Svizzero di Calcolo Scientifico
Swiss National Supercomputing Centre

ETH zürich

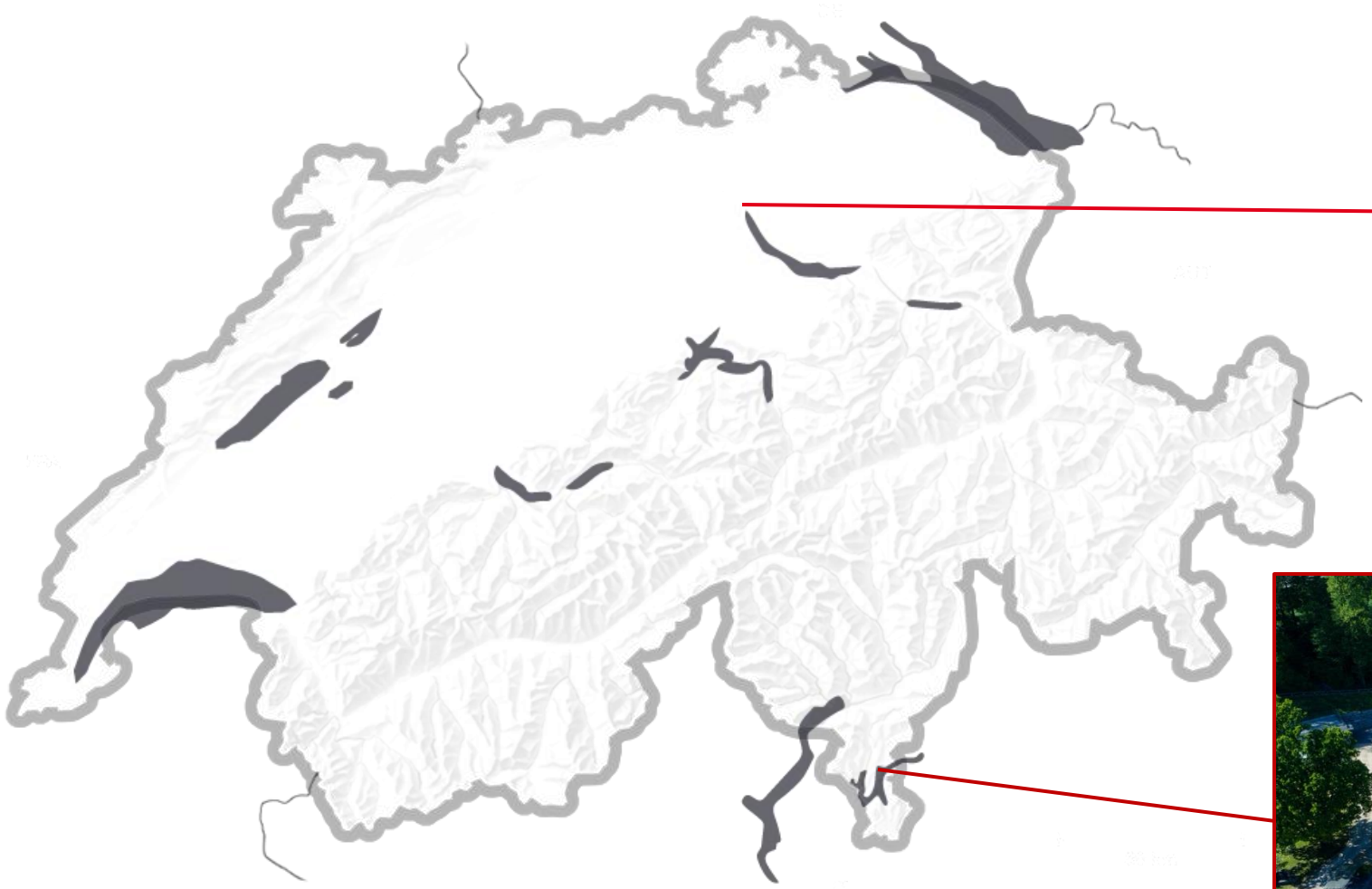


CSCS and Alps: the future is present

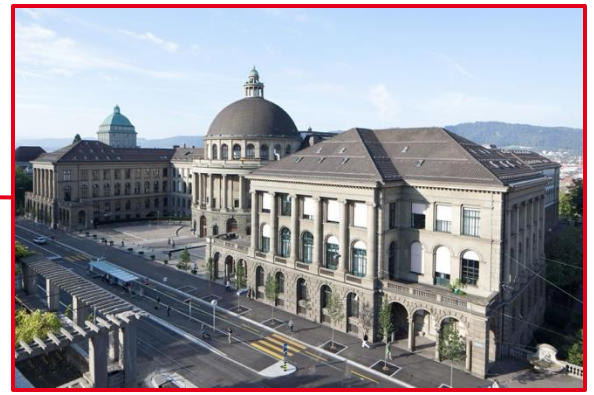
Dr. Riccardo Di Maria

The Swiss National Supercomputing Centre (CSCS), ETH Zurich

The *Swiss National Supercomputing Centre,* located in Lugano,
is a unit of the *Swiss Federal Institute of Technology in Zurich* (ETH Zurich)



ETH Zurich

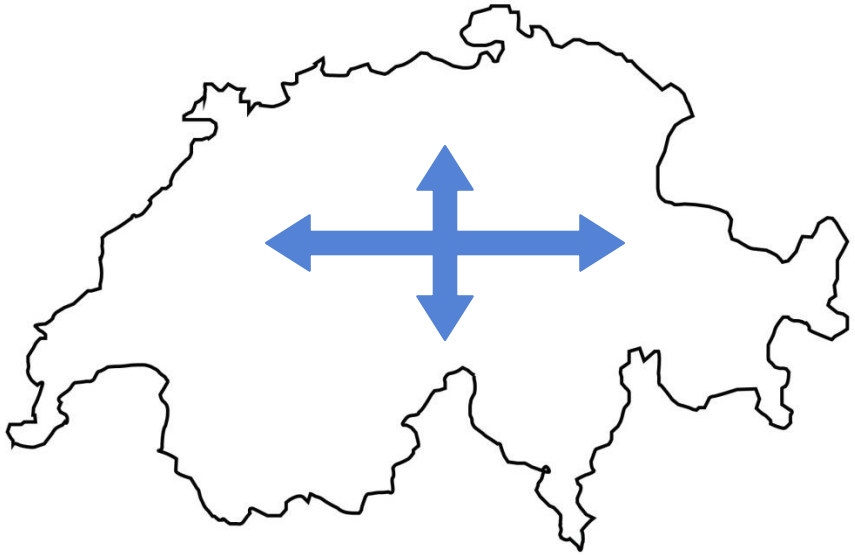


CSCS Lugano

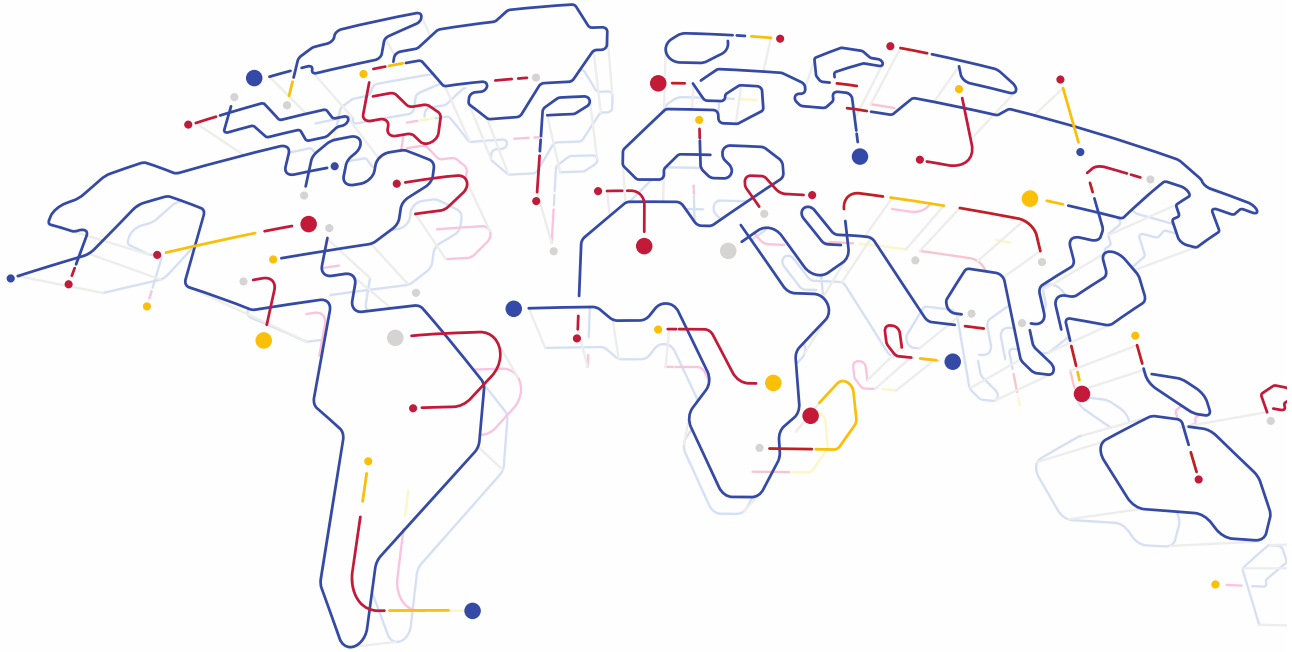
Mission

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

- Located in Ticino since 1991

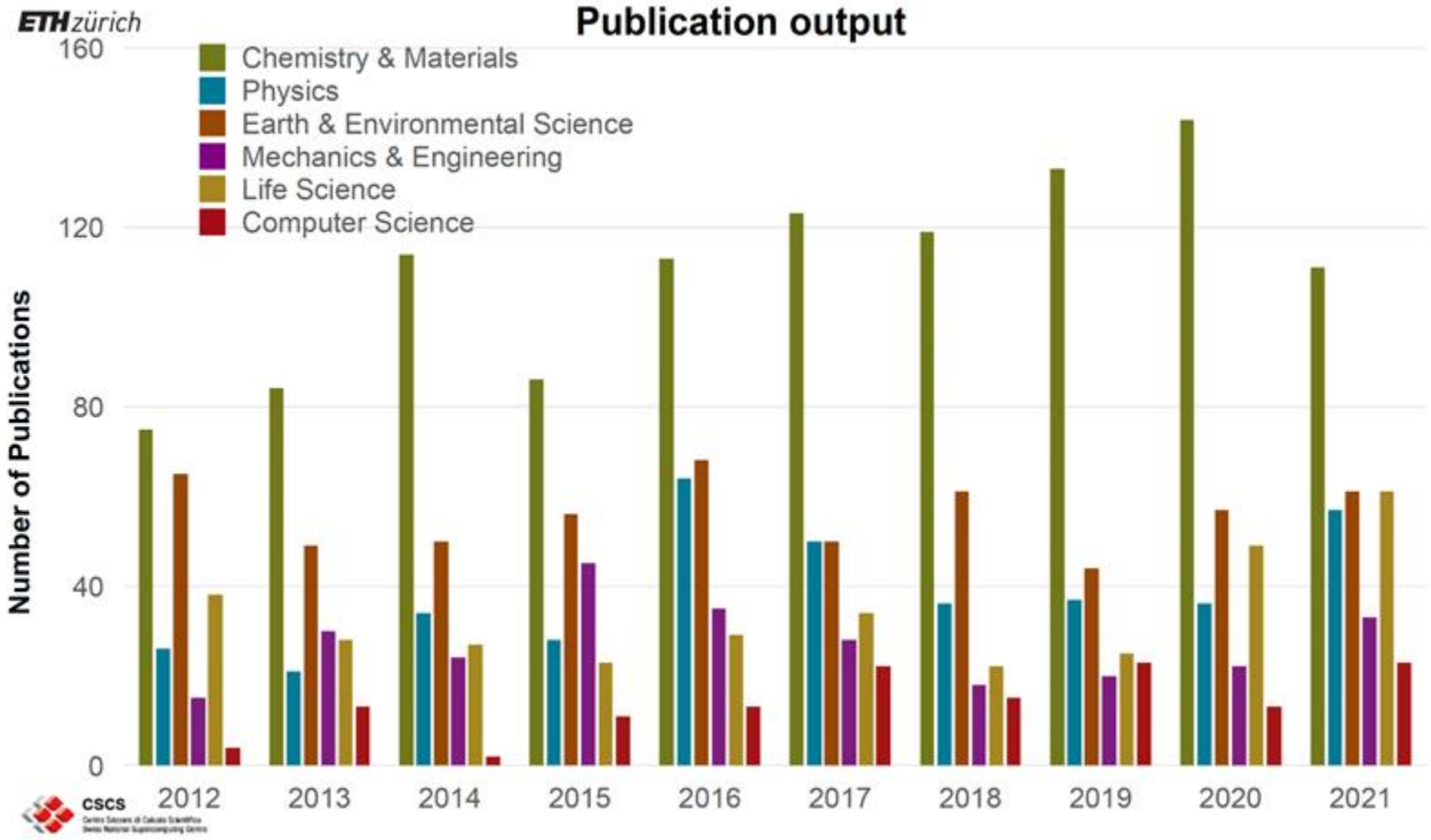


- National and international collaborations in the research of new technologies for HPC



User Lab Program, PASC, and Partnerships

The **User Lab Program** and **PASC** (Platform for Advanced Scientific Computing) initiative provide access to resources and knowhow based on a *peer review process* and are funded through the HPCN initiative by the Swiss Government

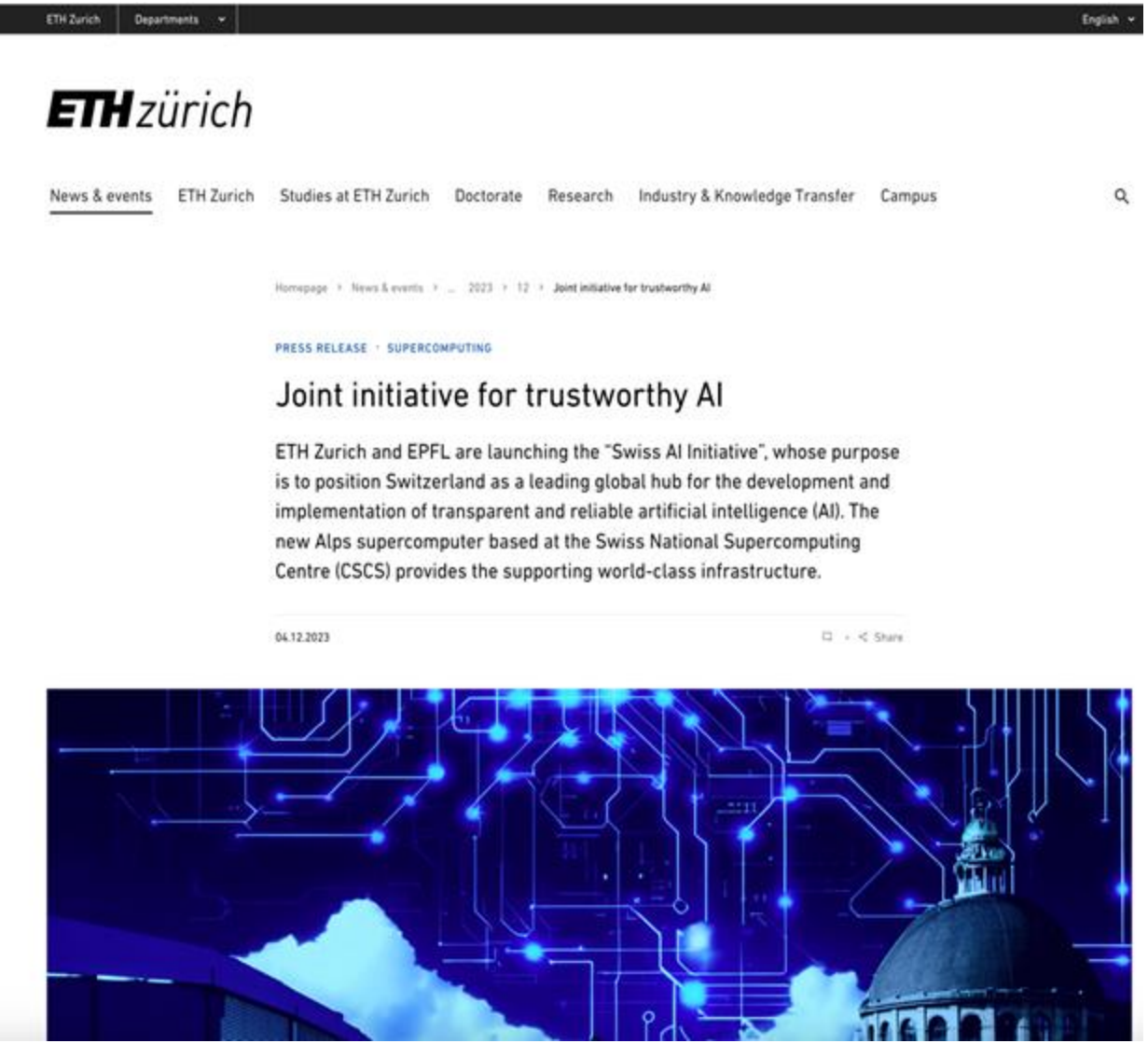


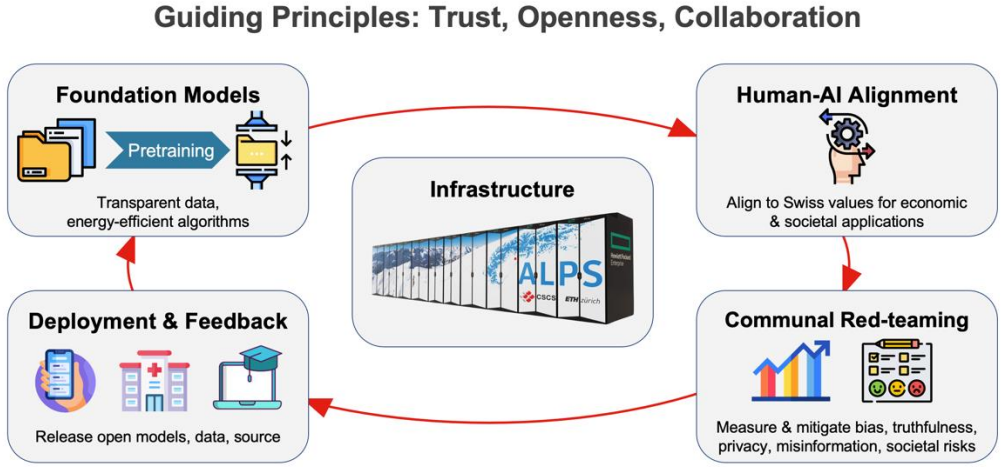
Partnerships provide access to services at CSCS based on *external funding*, and involve collaboration, exchange of knowhow, provisioning of resources

- MeteoCH
- PSI
- CHiPP
- CTA / SKA
- HBP
- UZH
- EMPA
- Euler
- BBP
- …

ETH zürich

The Swiss AI initiative





Develop capabilities, know-how & talent to build trustworthy Generative AI aligned with Swiss values

Make these resources available for the benefit of Swiss society and global actors

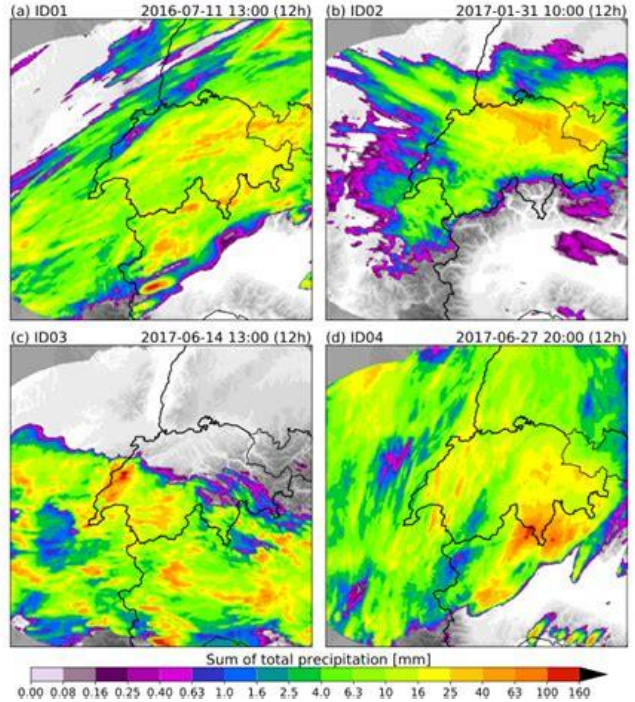
- Uniting researchers in Switzerland to tackle challenges out-of-reach of single research groups
- 70 professors, > 500 PhD and postdocs
- Lead by the ETHZ and EPFL AI centers
- Involving many of the Swiss universities and universities of applied sciences

https://www.swiss-ai.org/

Case Studies

MeteoSwiss - Numerical Weather Forecasts

- MeteoSwiss computes its daily weather forecasts on two versatile, software-defined clusters (Tasna and Balfrin) at CSCS
  - geo-distributed computing: Lugano (AlpsM) and Lausanne (AlpsE)
  - future weather conditions and possible natural hazards
- Essential information for air traffic control services and disaster mitigation (e.g. radioactive leak)
- Production and R&D seamlessly migrated to Alps in 2024
  - AI/ML based on Kubernetes
- One of the important partners for the strategic pillar of Weather and Climate at CSCS
  - ETHZ Exclaim
  - ECMWF



MeteoSwiss

HPC Platform on Alps

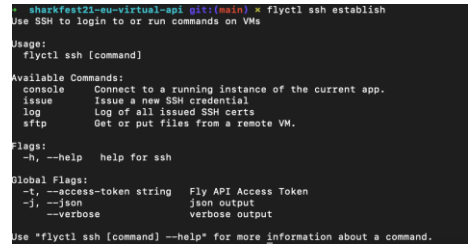
- The CSCS' HPC Platform provides a generic environment to run scientific codes at scale
  - focus on performance and scalability
  - aimed at thousands of users from all scientific domains
  - high performance hardware and libraries
  - enables automated workflows via REST APIs and containers
  - features interactive computing, data-movers, compilers, debuggers, job schedulers, parallel file systems, etc. for the convenience of all users

jupyterhub

slurm workload manager

UENV

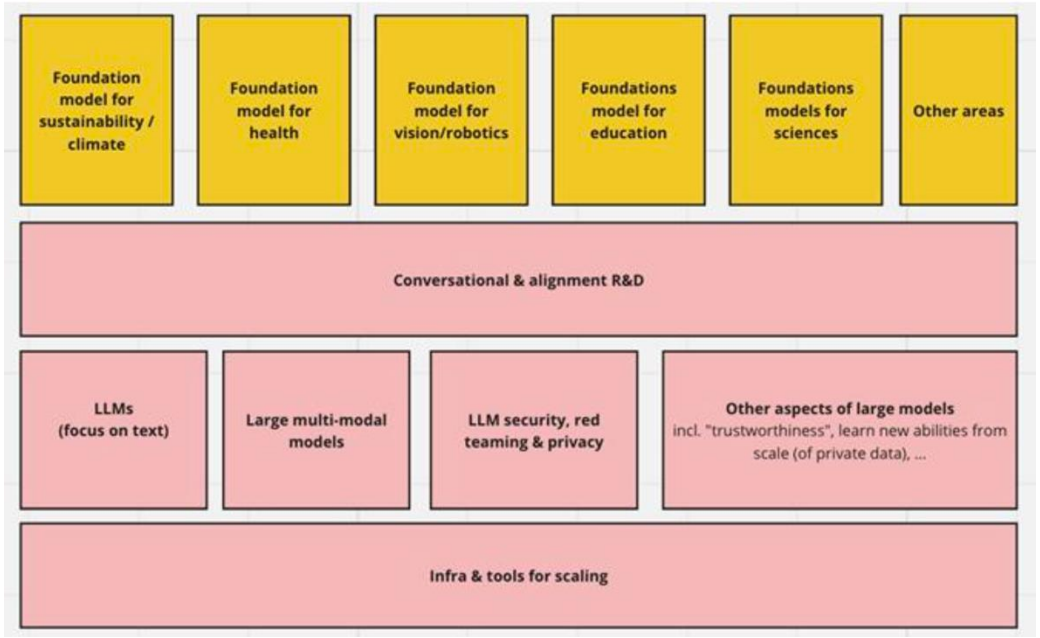
FIRECREST



SARUS

Machine Learning Platform on Alps

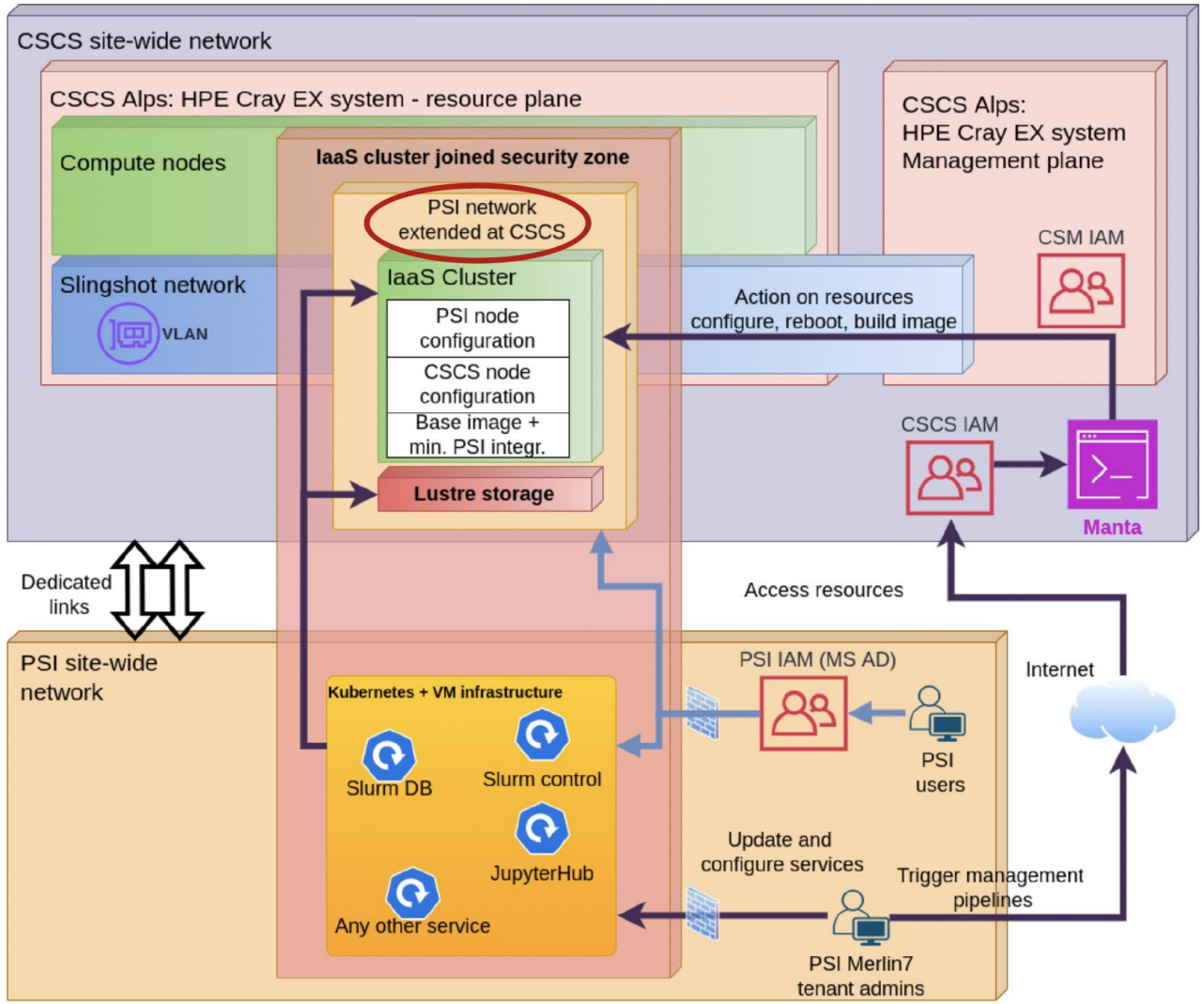
- Alps provides a dedicated Machine Learning Platform leveraging thousands of Grace-Hopper GPUs
- Alps is the underlying infrastructure for the Swiss AI Initiative



ETH Zurich and EPF Lausanne, teaming up with major universities in Switzerland

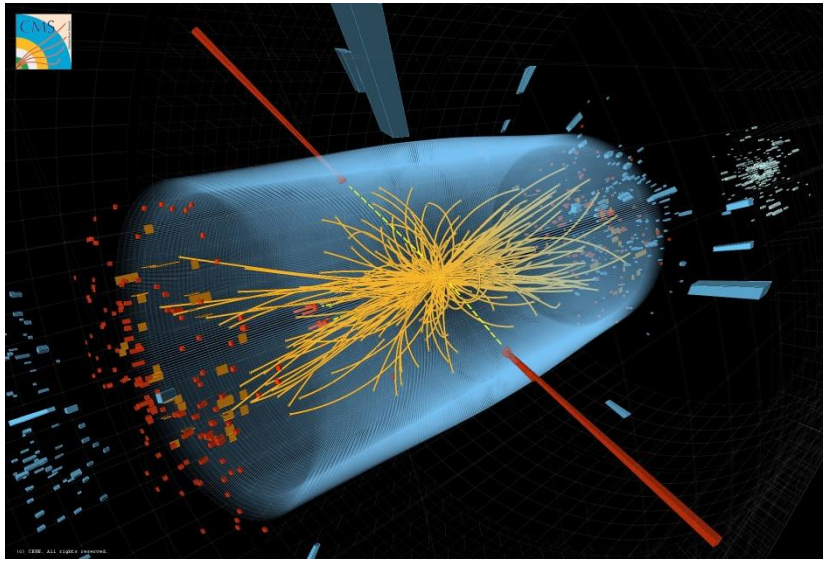
The Paul Scherrer Institute (PSI) – Infrastructure-as-a-Service

- The **Paul Scherrer Institute** is the largest research centre for natural and engineering sciences within Switzerland
  - world-class research in three main subject areas:
    - matter and material
    - energy and environment
    - human health
- PSI operates large scale research facilities (e.g. SwissFEL) and conducts research that produces a large amount of data
  - >10PB of archived data at CSCS
- Currently migrating their dedicated HPC cluster to Alps
  - around 100 nodes currently at PSI
  - including Kubernetes-based management of clusters
  - working on geo-redundancy of the CSCS archive
- CSCS provides Infrastructure-as-a-Service to PSI
  - 30PB of archived data at CSCS
  - Geo-redundant tape archive at PSI
  - Deploying 100-node HPC Cluster at CSCS with multi-tenancy on Alps



CHiPP - Analysis of Data from the Large Hadron Collider (LHC)

Swiss Institute of Particle Physics



- Swiss particle physics community in the context of the LHC at CERN
  - help understanding the building blocks of our Universe by particle collisions
- On behalf of CHiPP, CSCS operates a mid-size Tier-2 grid site for three of the four experiments: ATLAS, CMS, and LHCb
- Grid site (*currently*) fully running on HPC resources
  - main services “Kubernetes-ised”

Shared HPC resources for Universities and Research Institutes

- The University of Zurich moved their local HPC service into CSCS' shared resources back in 2014 on Piz Daint
  - migrated into Alps in 2021
- Resources are shared among 15 different groups inside UZH
  - UZH local Scientific IT service (S3IT) distributes the resources and supports their scientists with their research
- Classic HPC
  - ~200 multi-core nodes

University of Zurich[UZH]

*Similar model followed by **Empa** (Swiss Federal Laboratories for Materials Science and Technology), **Eawag** (Swiss Federal Institute of Aquatic Science and Technology), **USI** (Università della Svizzera Italiana), **FHGR Chur** (University of Applied Sciences of the Grisons), **SDSC** (Swiss Data Science Center) and others*

Observatories – Square Kilometer Array & Cherenkov Telescope Array

SKACH

cta cherenkov telescope array

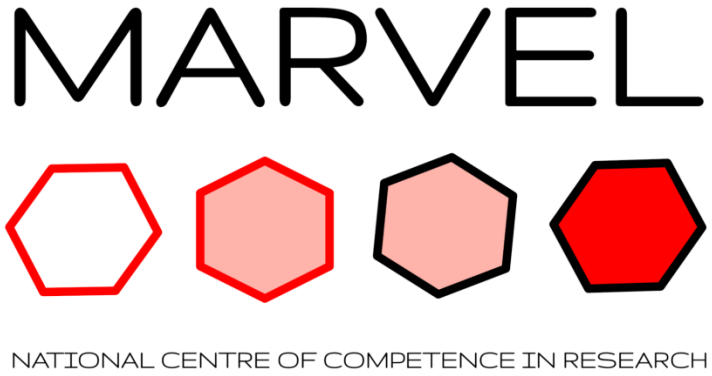
- Shaping the future of Astronomy at Petabyte-to-Exabyte scales is of strategic importance for Switzerland to support research, education and innovation
- **Cherenkov Telescope Array (CTA)** and the **Square Kilometer Array (SKA)** Observatories are at the forefront of this new Big Data astronomy
  - ~1 Exabyte per year, to be distributed around the globe

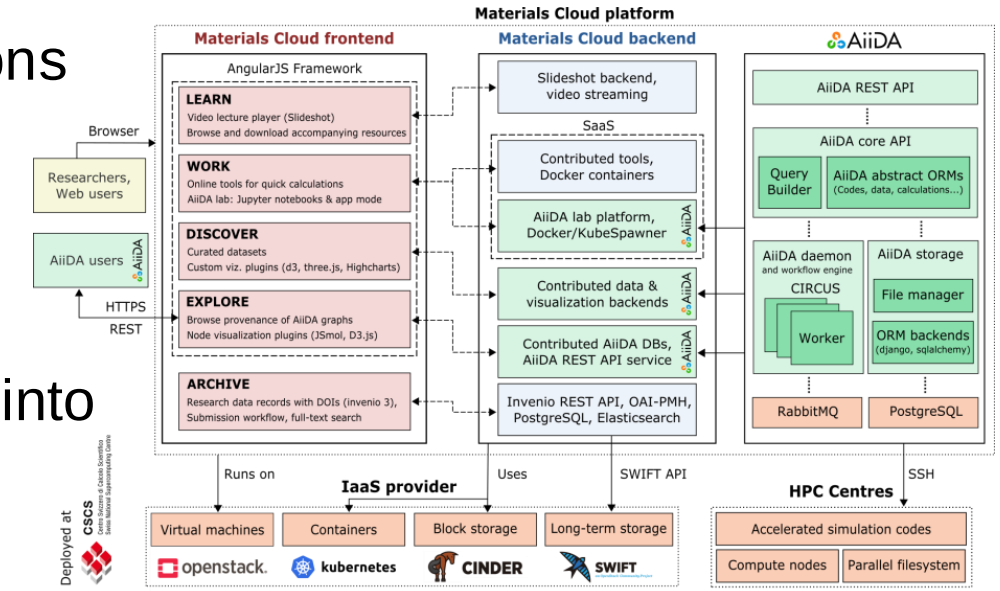
MARVEL – Compute and Long-Term Storage Resources

- **MARVEL** (National Centre of Competence in Research - NCCR) aims to radically transform and accelerate the design and discovery of novel materials

- CSCS supports MARVEL by providing compute and long-term storage resources, and the two organizations work closely together in defining the **Materials Platform** of the future

- Materials Cloud uses OpenStack and is now moving into Kubernetes





EXCLAIM : Cloud-Resolving Weather and Climate

High-level implementation plan

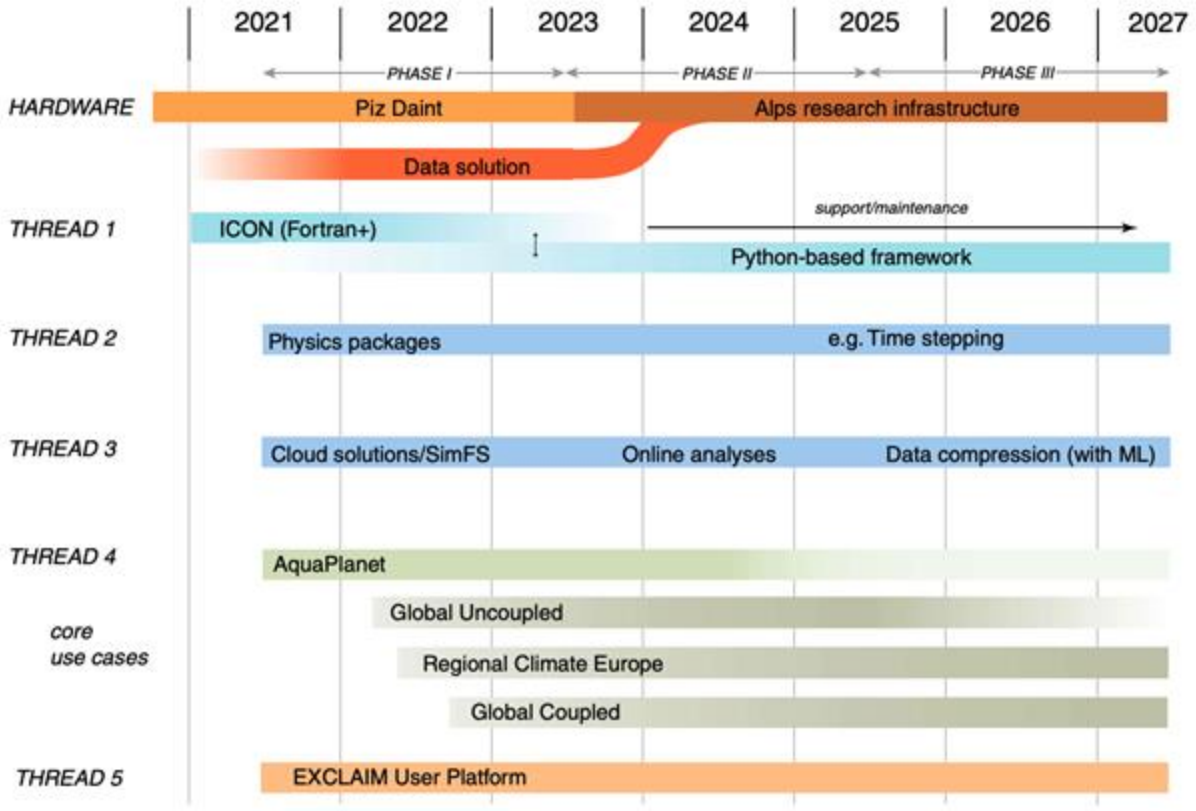
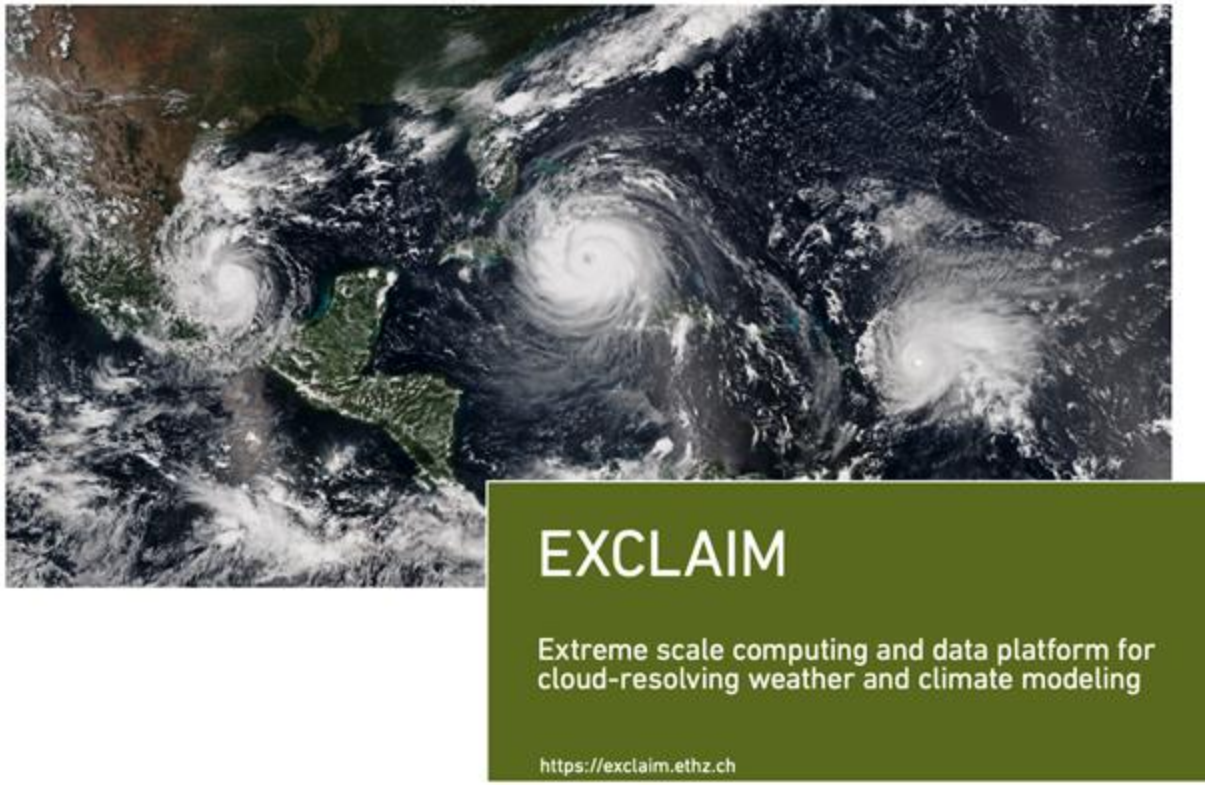


Figure 2: High-level road-map of the EXCLAIM project showing the five key development threads

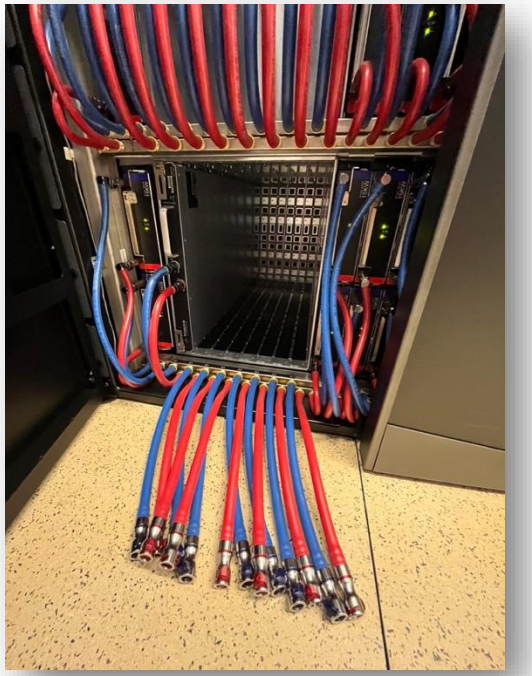
A community united around an ambitious roadmap that jointly decides on the milestone simulations, shares data for analysis, and develops a platform with dedicated functionality in a specialized platform. Model for how we see communities develop their HPC roadmaps in the future.



The Research Infrastructure: Alps

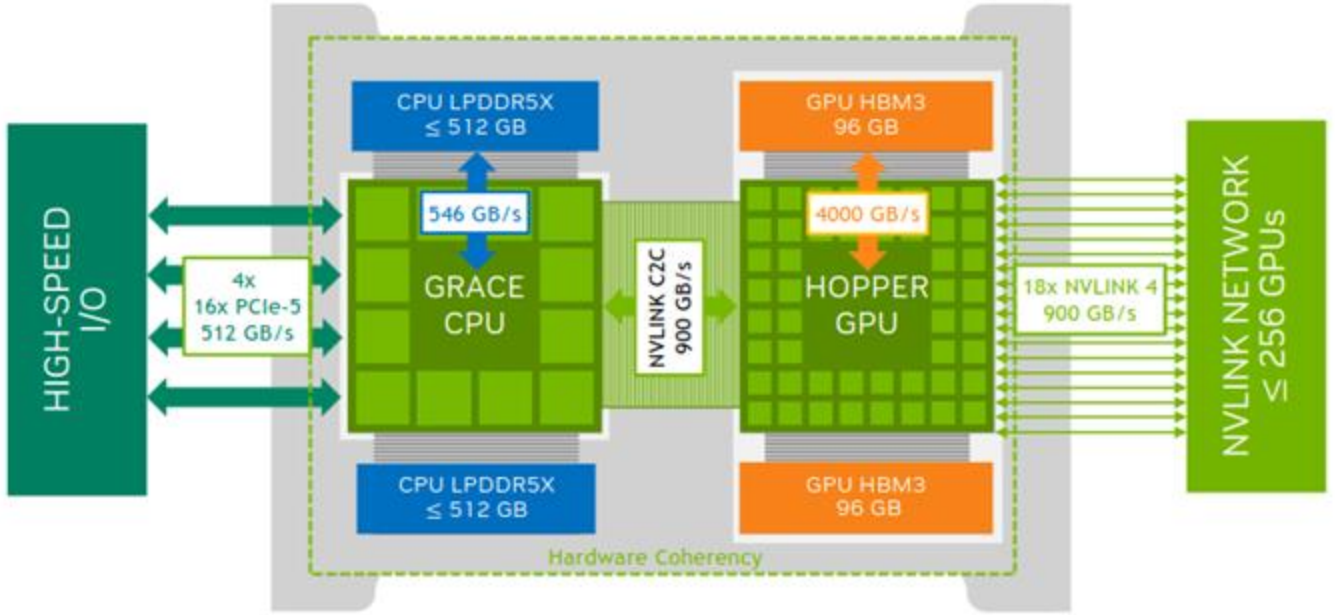
Alps

- Alps is an HPE Cray EX supercomputer meant to be our new flagship infrastructure
- Multi-phase installation started in 2020
- Multiple **geo-distributed** infrastructure
- Specification
  - 1024 AMD Rome (256/512GB RAM) nodes
  - 48 AMD Milan nodes
  - 144 NVIDIA A100 GPU nodes
  - 24 AMD MI250X GPU nodes
  - 128 AMD MI300 GPU nodes
  - ***2688*** GraceHopper nodes (4xGH200 → ***10752***)
  - Slingshot network
  - 2 availability zones (HA, non-HA)
  - 100% liquid cooled



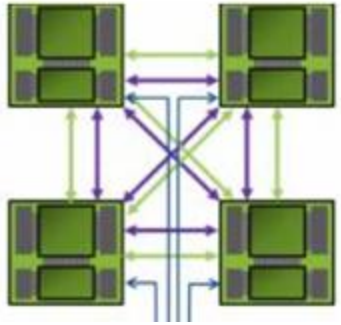
water cooled blades

Alps Phase II:
the Grace-Hopper “Super Chip”

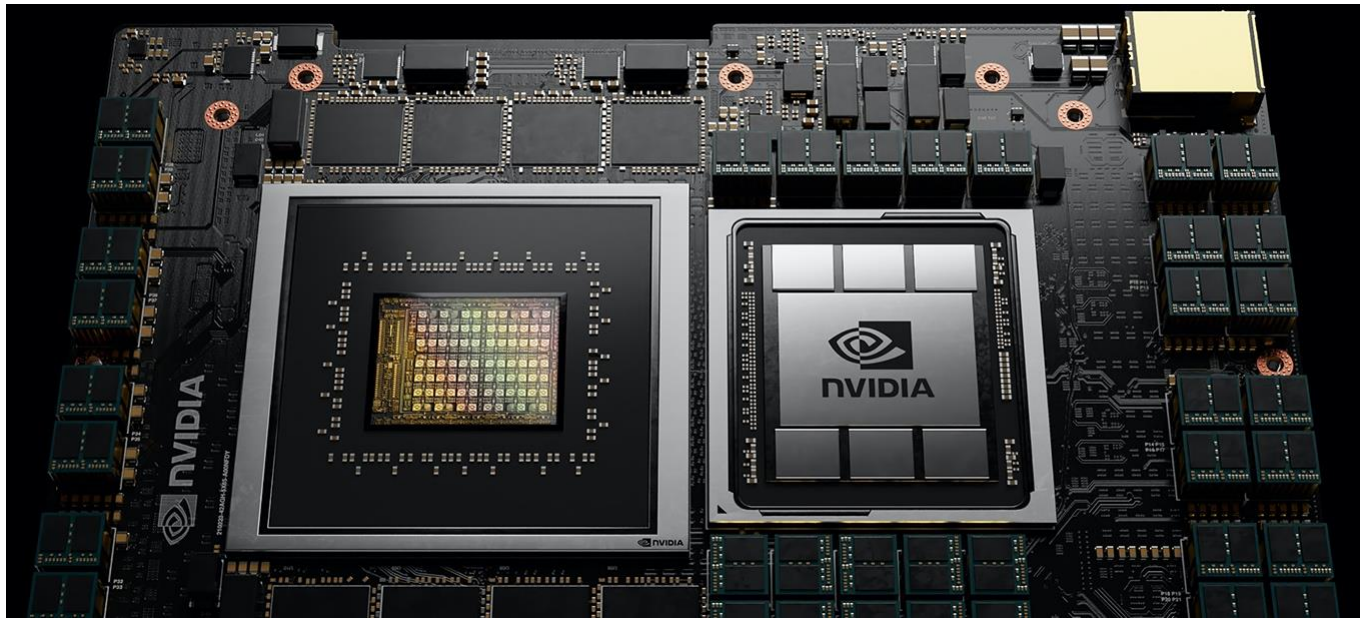


Each node will have 4 Grace-Hopper modules

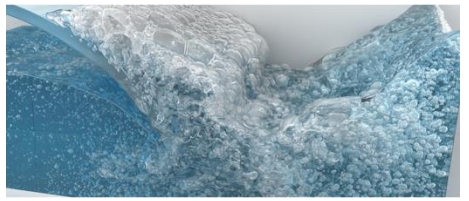
- 1 Grace CPU socket and
  1 Hopper GPU per module
- all-to-all cache-coherent memory NVLINK
  between all host and device memory



- one NIC per module



HPC and Cloud Convergence



- Science and engineering requires more and more computer-assisted experiments
  - simulation of physical phenomena and behaviours
  - Digital Twins
  - design engineering products
  - AI/ML statistic solutions

1. HPC offers high-performance compute and data access
   - improves Time to Solution
   - manage efficiently data to compute
   - bare-metal performance, fixed amount of resources

1. + 2. = ?

2. Cloud offers high flexibility for business needs
   - XaaS – business logic as a service
   - economy of scale – oversubscription of resources
   - virtualized resources, scalable to infinity (and beyond)

Achieve Best of HPC and Cloud

- Performance and flexibility
  - container as a virtualization layer with OCI hooks
    - keep OS near bare metal – accelerators and high-speed network drivers
    - bring low-level libraries in the container
  - own user environment
    - decouple HPC programming environments from underlying layers
    - UE can potentially become “just” an artifact mounted in the container

- Separation of concerns with layers
  - Platforms
    - provisioning of services with Kubernetes and/or Nomad
    - container as an abstraction layer for compute nodes
  - **Infrastructure as code**
    - APIs and configuration management
    - multi-tenancy: exclusive compute, network, and storage segregation
    - update components independently and minimise downtime across tenants

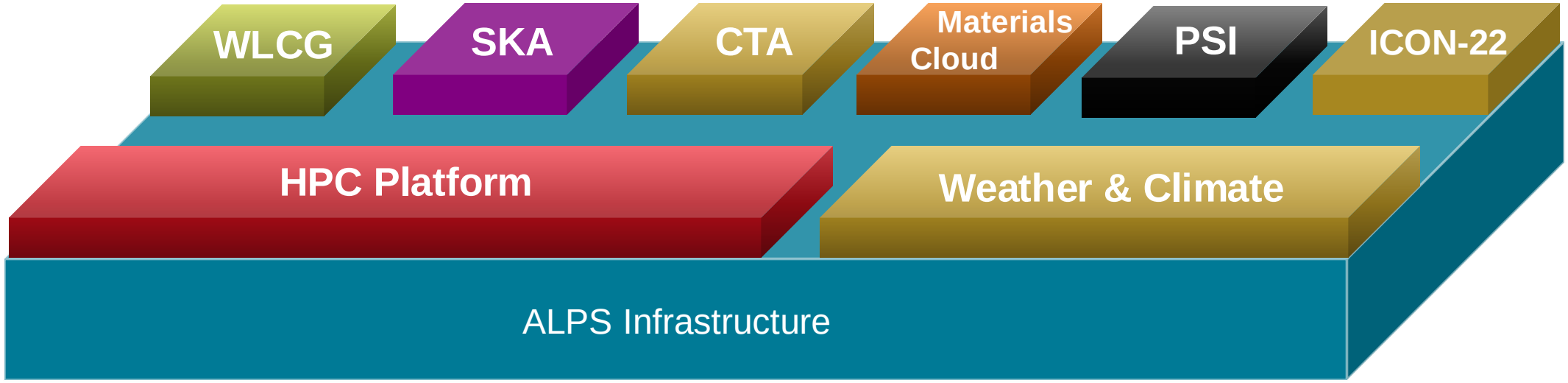
vCluster
versatile software-defined cluster

- HPC business logic
  - web-facing API to access HPC resources (submit jobs, move data)
  - web gateway

Cloud-native Supercomputing and Big Data

vCluster: versatile software-defined cluster

- custom user environments
- manage platform services, provisioning of clusters
- possibility of network and storage isolation



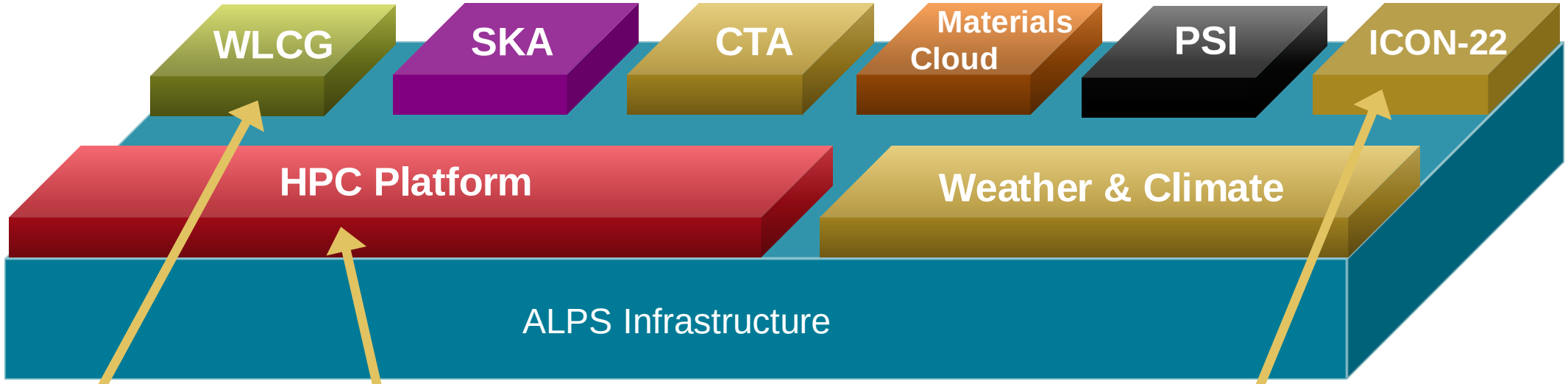
User environments management

Platforms and services management

Infrastructure as Code

A layered, versatile research infrastructure, providing the environment for various communities to excel and innovate in their fields

Cloud-native Supercomputing and Big Data



Piz Daint
User Lab and WLCG

MeteoSwiss
Arolla/Tsa
COSMO/ModInterim

Service Orchestration

a. Full service on HPC
   - security challenges
     - VLANs help
     - ad-hoc configurations between management and managed plane
   - inefficiency on costly resources
   - additional ”virtualisation” layer
     → complexity (e.g. network)

b. Adjacent/front-end services orchestrated within
   - compute nodes of Alps using NOMAD
   - dedicated Kubernetes clusters
     - efficient use of HPC resources
     - necessity of middleware/interface between user and compute

- Orchestration and maintenance driven by use-case needs

- Advantages
  - decoupling from the infrastructure
  - declarative configuration
  - reusage of code
  - load balancing
  - automated rollouts and rollbacks
  - self-healing
  - secret management
  - observability and traffic management
  - **disaster recovery management and one-button deployment**

- Challenges
  - additional “moving parts” and complexity layers
    - networking: Cilium vs. Calico, service mesh
  - security
    - additional configuration

- Riccardo Di Maria *“The WLCG Journey at CSCS: from Piz Daint to Alps”*, HEPiX 2023, Taipei, Taiwan

Summary and Conclusions

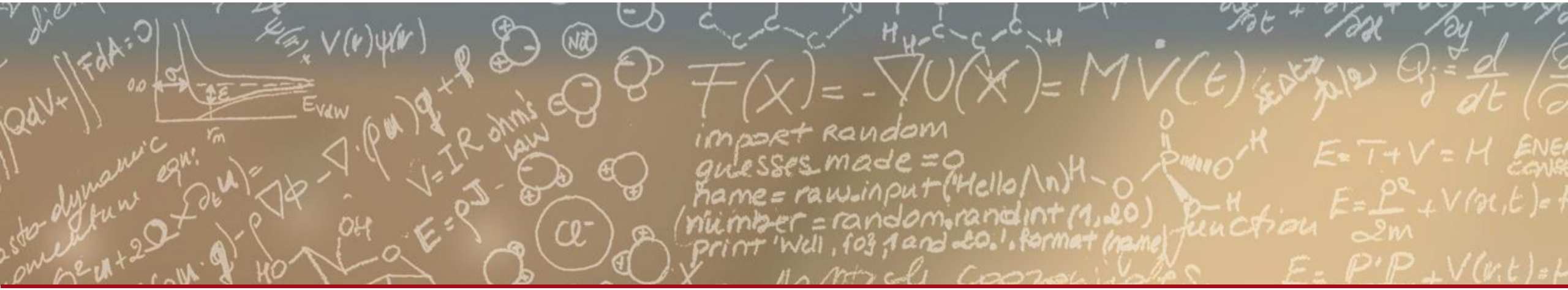
- CSCS mission is to enable world-class science in Switzerland
- Alps towards improving standard HPC through cloud-native approach
- Infrastructure is key to modern AI and foundational models, and the Swiss AI initiative unites and will generate top talent in AI
- Multiple geo-distributed infrastructure
- Heterogeneous infrastructure
  → NVIDIA and AMD GPU, AMD CPU, GH200
- Infrastructure as Code based implementation of Alps vClusters
  - scale the infrastructure dynamically and according to the changing requirements of the partners
- Science as a Service concept with innovative resource access

TOP 500 The List.

Rank	System	Cores	Rmax (PFlop/s)	Rpeak (PFlop/s)	Power (kW)
1	**El Capitan** - HPE Cray EX255a, AMD 4th Gen EPYC 24C 1.8GHz, AMD Instinct MI300A, Slingshot-11, TOSS, HPE DOE/NNSA/LLNL United States	11,039,616	1,742.00	2,746.38	29,581
2	**Frontier** - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, HPE Cray OS, HPE DOE/SC/Oak Ridge National Laboratory United States	9,066,176	1,353.00	2,055.72	24,607
3	**Aurora** - HPE Cray EX - Intel Exascale Compute Blade, Xeon CPU Max 9470 52C 2.4GHz, Intel Data Center GPU Max, Slingshot-11, Intel DOE/SC/Argonne National Laboratory United States	9,264,128	1,012.00	1,980.01	38,698
4	**Eagle** - Microsoft NDv5, Xeon Platinum 8480C 48C 2GHz, NVIDIA H100, NVIDIA Infiniband NDR, Microsoft Azure Microsoft Azure United States	2,073,600	561.20	846.84	
5	**HPC6** - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, RHEL 8.9, HPE Eni S.p.A. Italy	3,143,520	477.90	606.97	8,461
6	**Supercomputer Fugaku** - Supercomputer Fugaku, A64FX 48C 2.2GHz, Tofu interconnect D, Fujitsu RIKEN Center for Computational Science Japan	7,630,848	442.01	537.21	29,899
7	**Alps** - HPE Cray EX254n, NVIDIA Grace 72C 3.1GHz, NVIDIA GH200 Superchip, Slingshot-11, HPE Cray OS, HPE Swiss National Supercomputing Centre (CSCS) Switzerland	2,121,600	434.90	574.84	7,124

CSCS
Centro Svizzero di Calcolo Scientifico
Swiss National Supercomputing Centre

ETH zürich



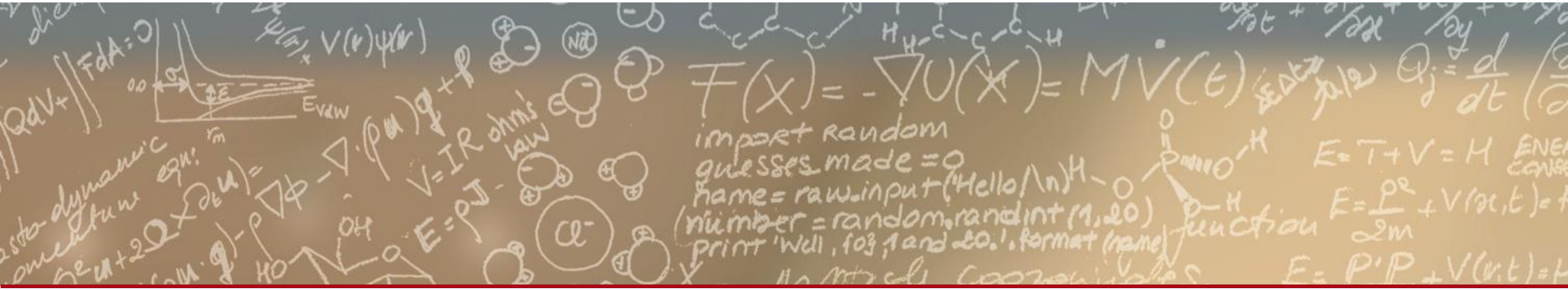
Thank you for your attention!

Acknowledgements

Presentation Material: J. VandeVondele, P. Fernandez, M. Martinasso, M. Klein

Contact: riccardo.dimaria@cscs.ch

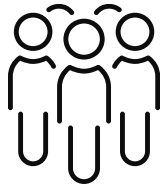
cscsch



Backup Slides

Some numbers

Staff

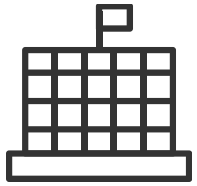


- 120 collaborators
- 25 nationalities
- Official language: English

Budget

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

Building



- 2’600 m^2 office buildings
- 2000 m^2 machine room
- «Free cooling» with lake water

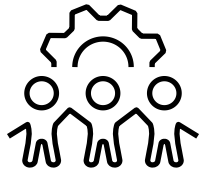
Electricity

- Currently 11 MW
- Possible extension to 25 MW
- 100% hydroelectrical source

User Lab

- 2300 users
- 130 projects

Third-party



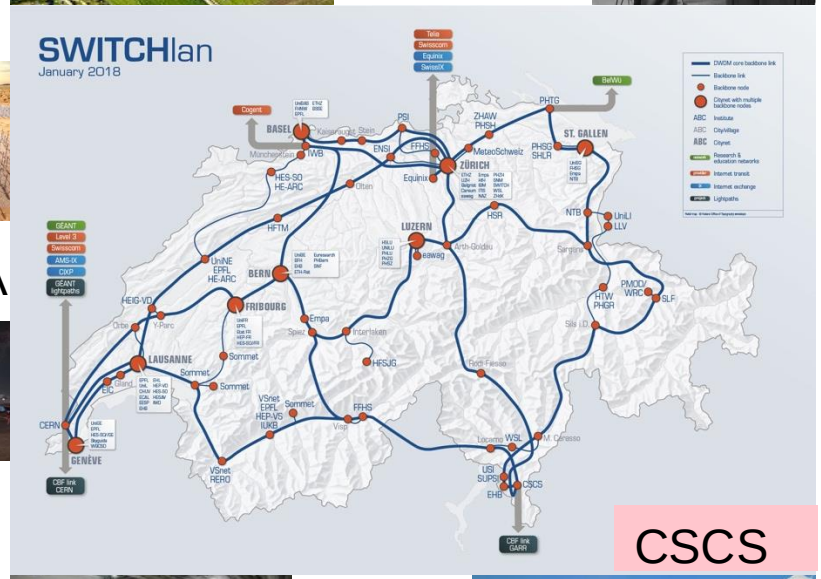
- MeteoSwiss, NCCR Marvel, PSI, CHiPP, Empa, ETH Zurich, SDSC, USI, UZH, BlueBrain ...

A RI connected to experiment, computational science, and the world

PSI

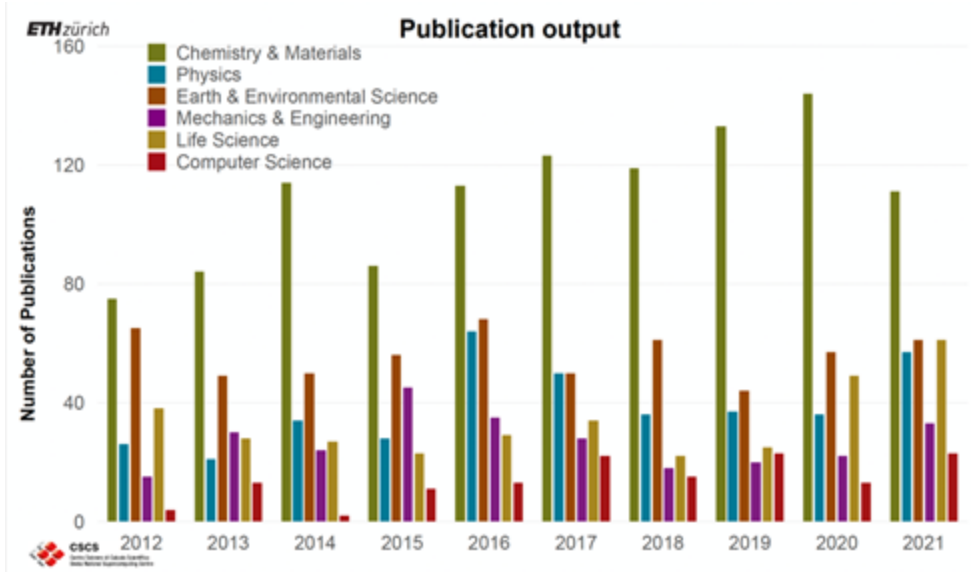
MCH

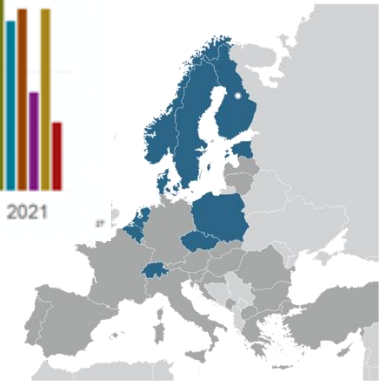
SKA / CTA



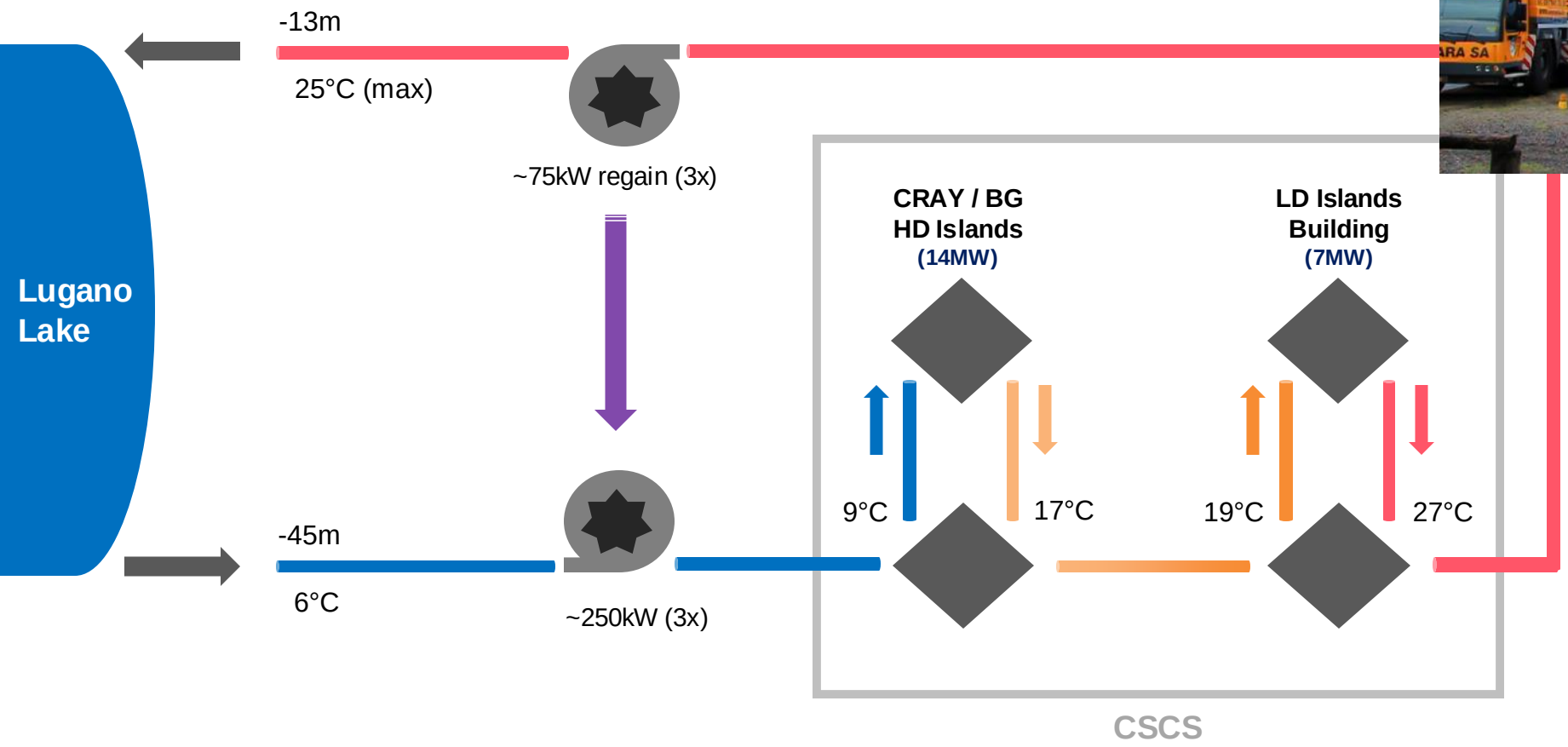
CSCS

CERN



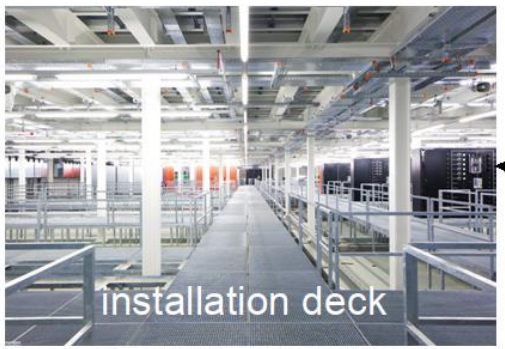


CSCS Facility – Water Cooling System

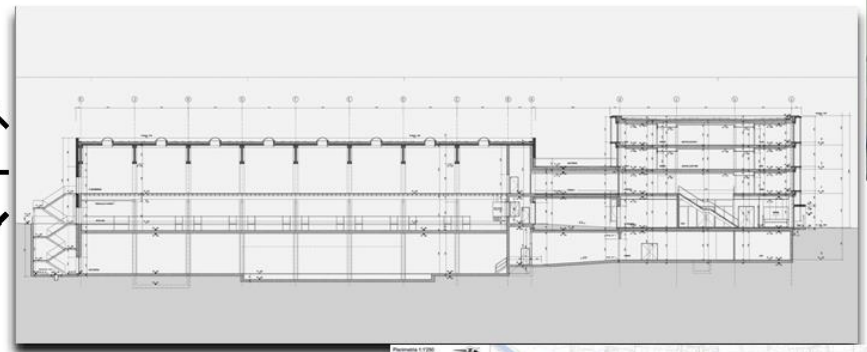


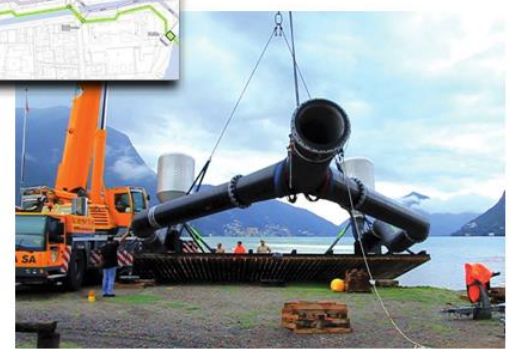
- pipeline length: 2.8 km
- height: 30 m
- max. flow rate: 760 l/s

We need flexible infrastructure, since we can only predict that technology will change



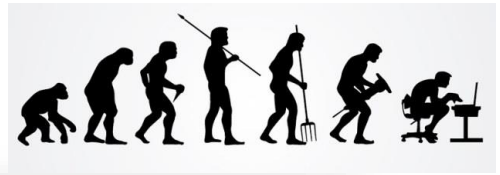
Power/cooling: 12 MW
(upgradable to 25MW)
Current (2018) load ~3.5 MW
Current power use efficiency: <1.2

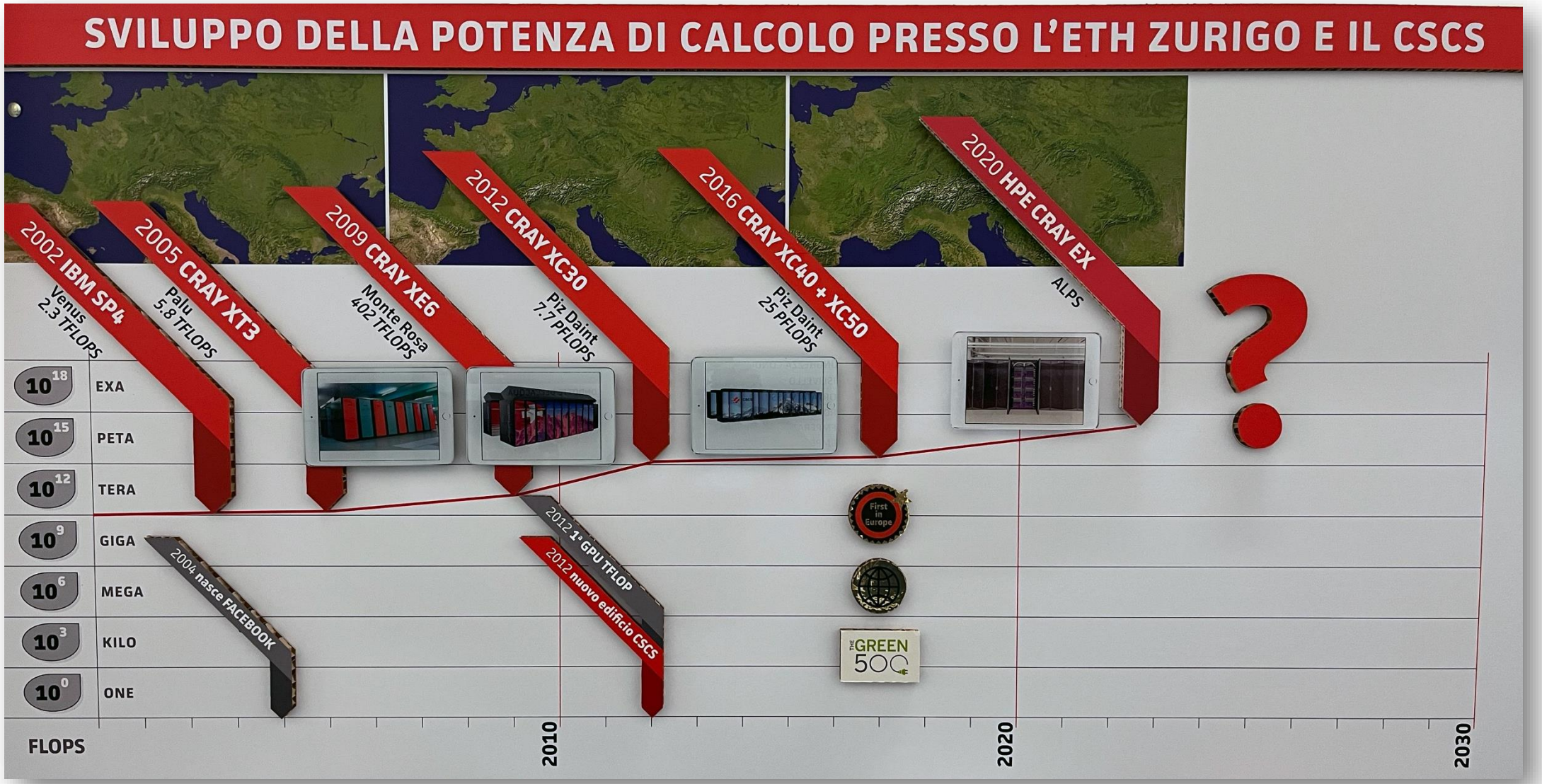




ETH Board: CHF 62.5 million for the building
Canton Ticino: CHF 5 million for lake water cooling
City of Lugano: donated the land
ETH Zurich: CHF 12.5 million for an extension **+ CHF ~10 million extension**

Evolution




SVILUPPO DELLA POTENZA DI CALCOLO PRESSO L'ETH ZURIGO E IL CSCS
2002 IBM SP4
Venus
2.3 TFLOPS
2005 CRAY XT3
Palu
5.8 TFLOPS
2009 CRAY XE6
Monte Rosa
402 TFLOPS
2012 CRAY XC30
Piz Daint
7.7 PFLOPS
2016 CRAY XC40 + XC50
Piz Daint
25 PFLOPS
2020 HPE CRAY EX
ALPS
?
10^18 EXA
10^15 PETA
10^12 TERA
10^9 GIGA
10^6 MEGA
10^3 KILO
10^0 ONE
2004 nasce FACEBOOK
2012 1° GPU TFLOP
2012 nuovo edificio CSCS
First in Europe
THE GREEN 500
FLOPS
2010
2020
2030