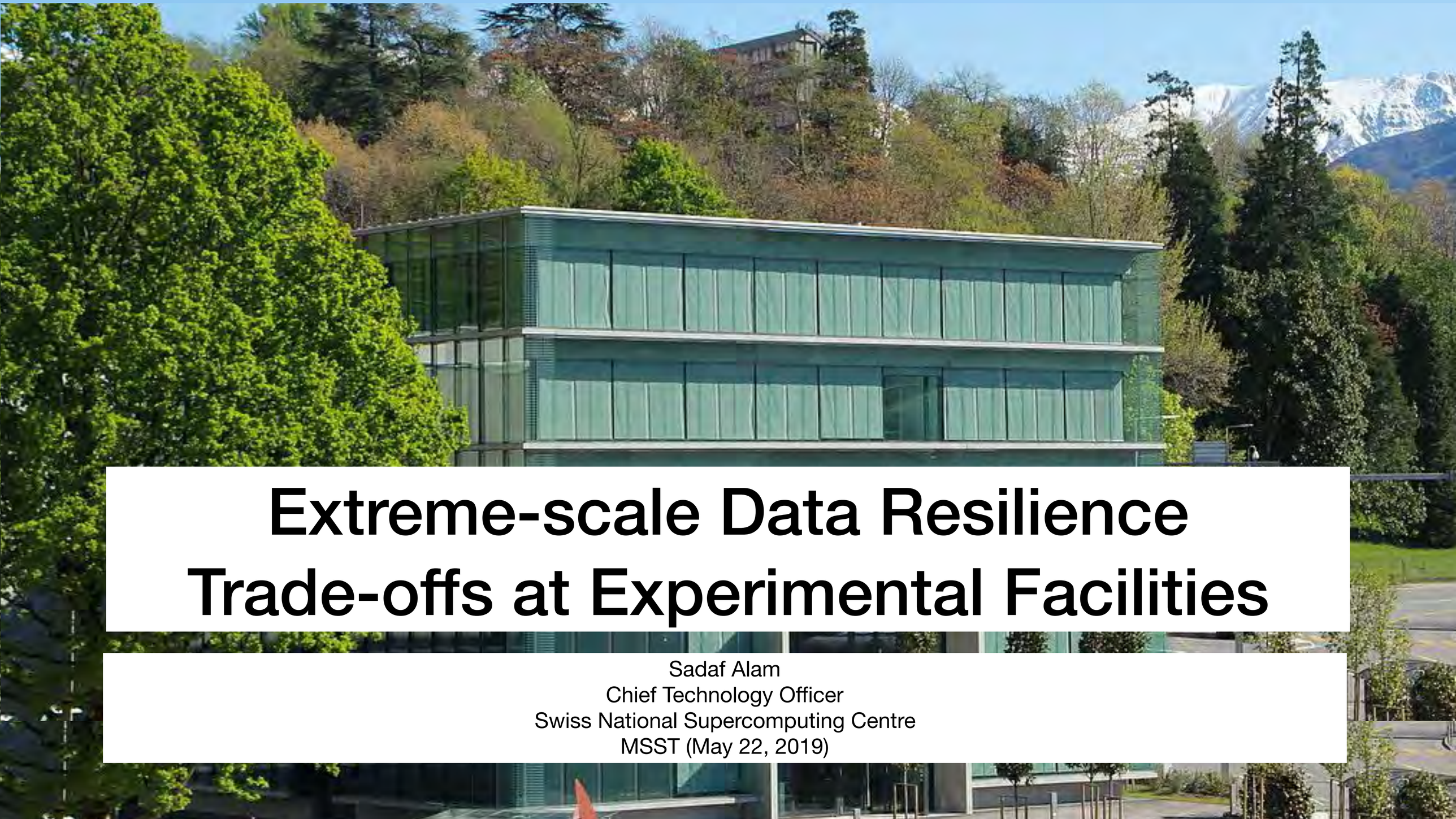
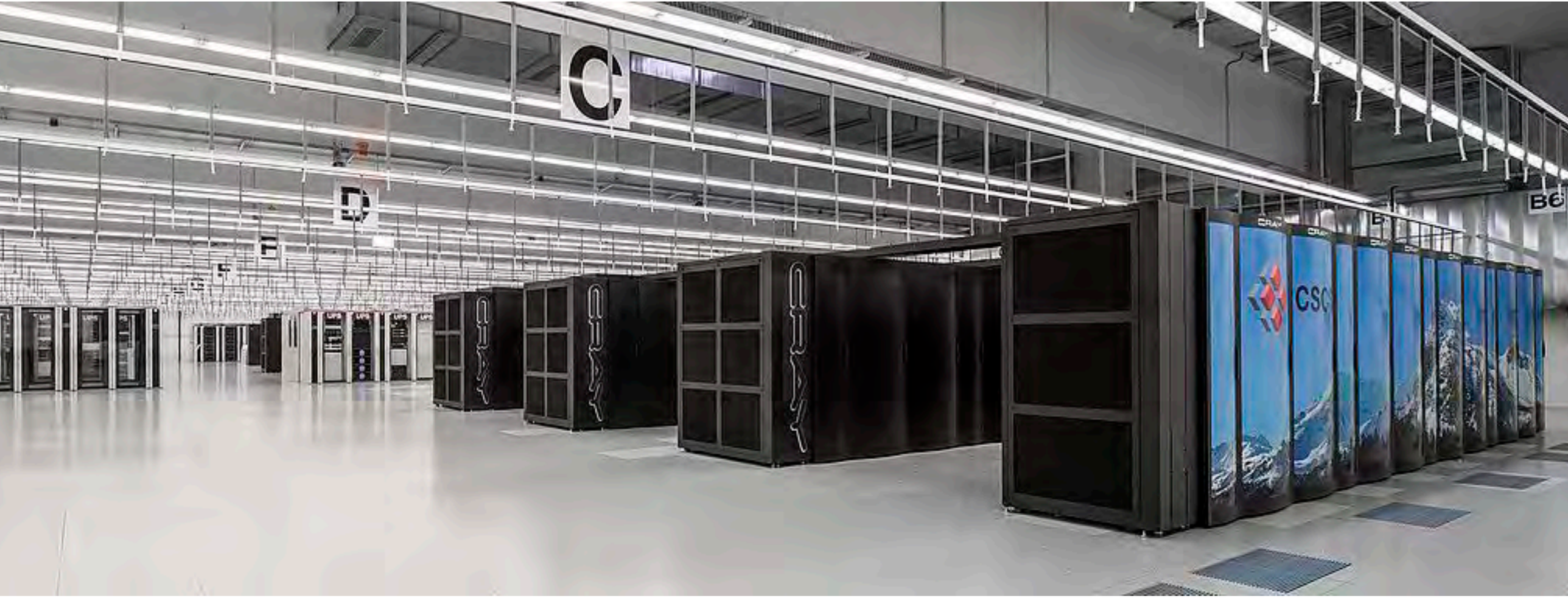


Extreme-scale Data Resilience Trade-offs at Experimental Facilities

Sadaf Alam
Chief Technology Officer
Swiss National Supercomputing Centre
MSST (May 22, 2019)

Outline

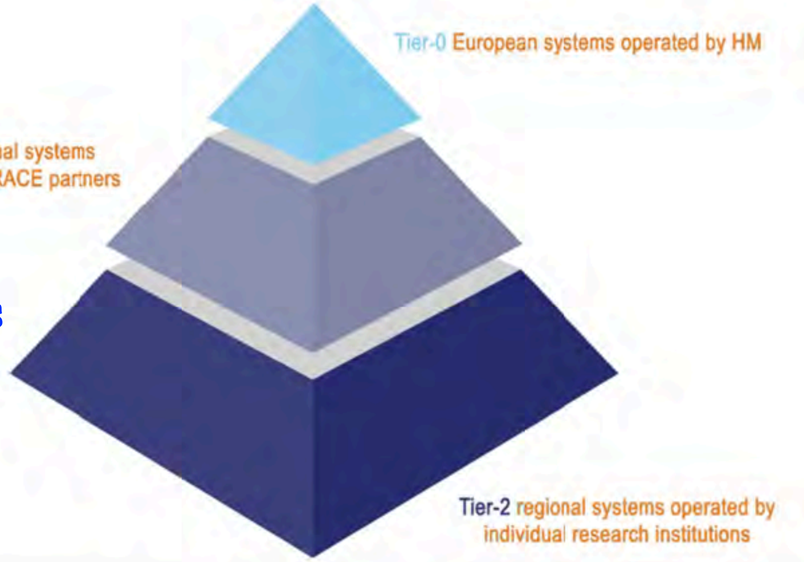
- Background
  - Users, customers and services
  - Co-design, consolidate and converge
- Resiliency in the context of experimental facilities workflows
  - Data-driven online and offline workflows
  - Data in motion and data at rest parameters
- Future: Co-designed HPC & cloud services for federated, data-driven workflows



Background

Diverse Users & Customers

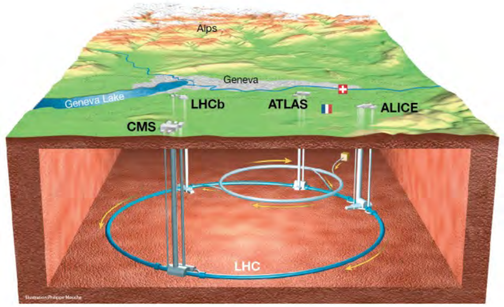
- R&D HPC services
  - Leadership scale: PRACE (Partnership for Advanced Computing in Europe)
  - Swiss & international researchers: user program
  - Customers with shares
- National services
  - Weather forecasting (MeteoSwiss)
  - CHIPP (WLCG Tier-2)
  - PSI PetaByte archive
- Federated HPC and cloud services
  - European e-Infrastructure



Supercomputing & HPC cluster workflows

Time-critical HPC workflows

Extreme-scale, data-driven HPC workflows

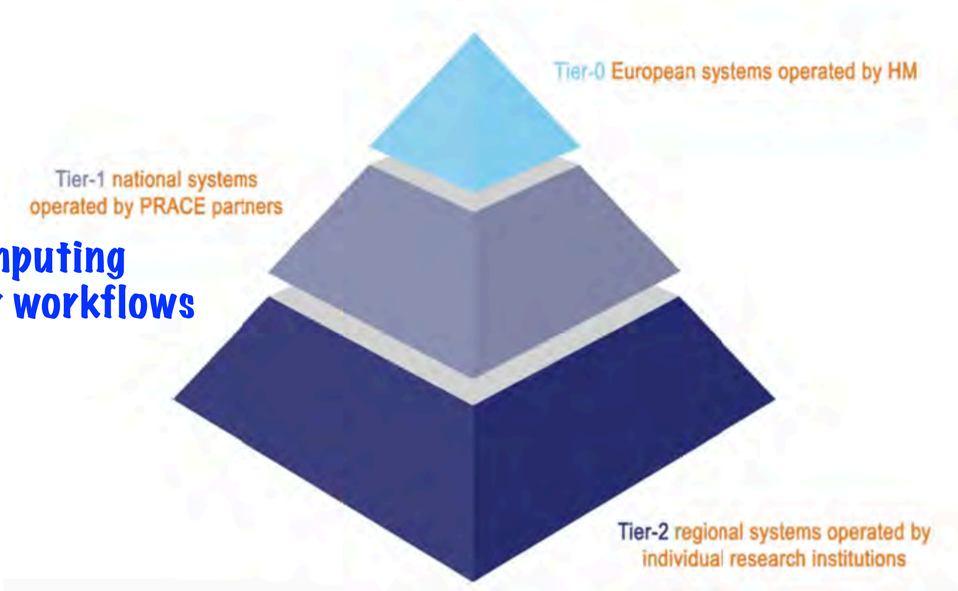


CSCS
Centro Svizzero di Calcolo Scientifico
Swiss National Supercomputing Centre

Diverse Requirements & Usages

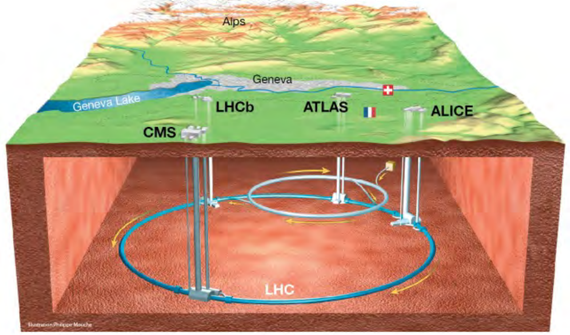
- R&D HPC services
  - Varying job sizes (full system 5000+ CPU & GPU to single core and even hyper thread for WLCG)
  - 1000s of users, 100s of applications, 10s of workflows
  - Batch & interactive batch, automated with custom middleware
  - Varying storage requirements (latency, bandwidth, ops sensitivity)
- National services
  - Different SLAs
  - Service catalog & contracts
- Federated HPC and cloud services
  - Stay tuned …

Supercomputing & HPC cluster workflows



Time-critical HPC workflows

Extreme-scale, data-driven HPC workflows



CSCS
Centro Svizzero di Calcolo Scientifico
Swiss National Supercomputing Centre

Co-design, consolidate & converge

CSCS
Centro Svizzero di Calcolo Scientifico
Swiss National Supercomputing Centre

University of Zurich

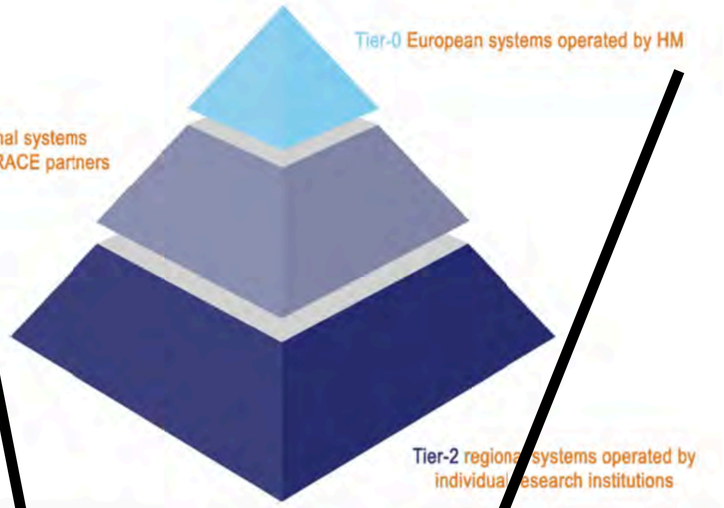
MARVEL

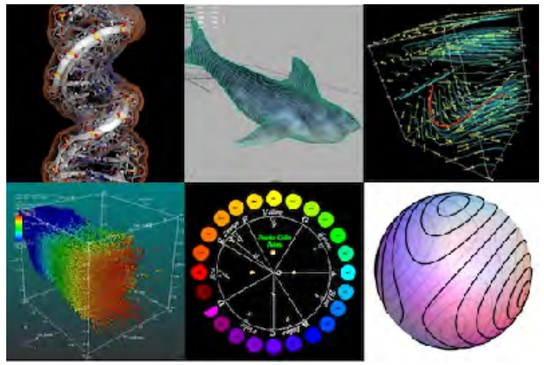
PAUL SCHERRER INSTITUT
PSI

PRACE

CHIP
Swiss Institute of Particle Physics

FENIX
RESEARCH INFRASTRUCTURE





theano

ML

Spark

TensorFlow

neon

jupyterhub

Shared , bare-metal compute & storage resources

CSCS
Centro Svizzero di Calcolo Scientifico
Swiss National Supercomputing Centre

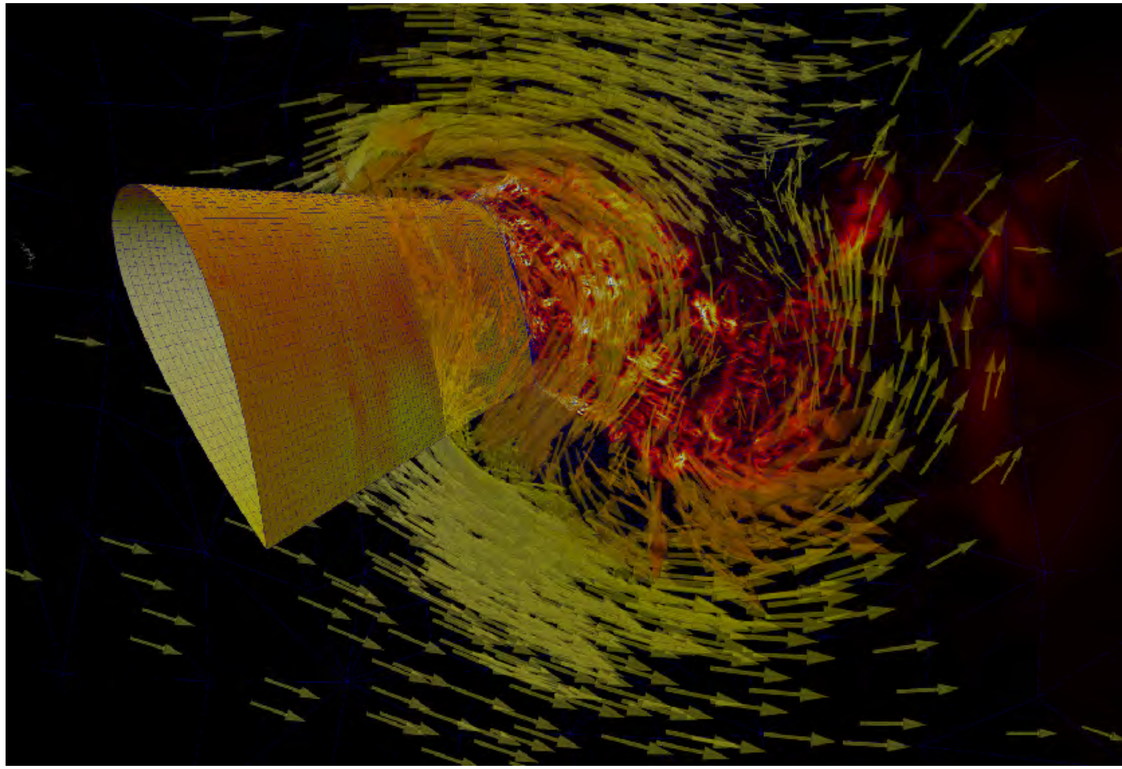
Highlights I (Users)

SIMULATING EXTREME AERODYNAMICS

- Reducing aircraft CO_2 emissions and noise. In 2016, aircraft worldwide carried 3.8 billion passengers while emitting around 700 million tons of CO2.

- Gordon Bell finalist: Researchers at Imperial College in London have used “Piz Daint” to simulate with unprecedented accuracy the flow over an aerofoil in deep stall.

- Open source platform for accelerators called PyFR (for performing high-order flux reconstruction simulations)



High-order accurate simulation of turbulent flow over a NACA0021 aerofoil in deep stall using PyFR on Piz Daint. (Image: Peter Vincent)

Highlights II (Users)

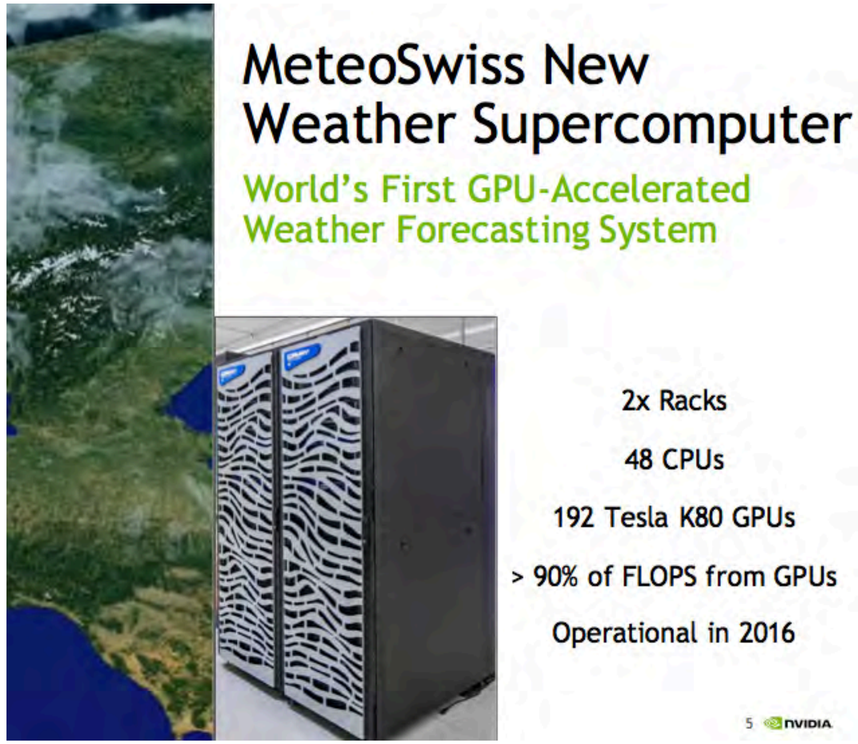
ECONOMISTS USING EFFICIENT HIGH-PERFORMANCE COMPUTING METHOD

- What-ifs scenarios for public financing models, e.g. pension models
- High dimensional modelling
  - approximating the high-dimensional functions
  - solving system of linear equations for million of grid point
- Nested models
  - combine sparse grids with a high-dimensional model reduction framework
- Hierarchical parallelism in application

Macroeconomic models, designed to study for example monetary and fiscal policy on a global scale, are extremely complex with a large and intricate formal structure. Therefore, economists are using more and more high-performance computing to try and tackle these models. (Image: William Potter, Shutterstock.com)

Highlights I (Customer: MeteoSwiss)

- MeteoSwiss mission: Acting on behalf of the Federal Government, MeteoSwiss provides various weather and climate services for the protection and benefit of Switzerland
- 40x improvement over previous generation system (2015)
  - With same CapEx and reduced OpEx
- Multi-year investment into the development and acceleration of COSMO application
- 24/7 operation with strict SLAs

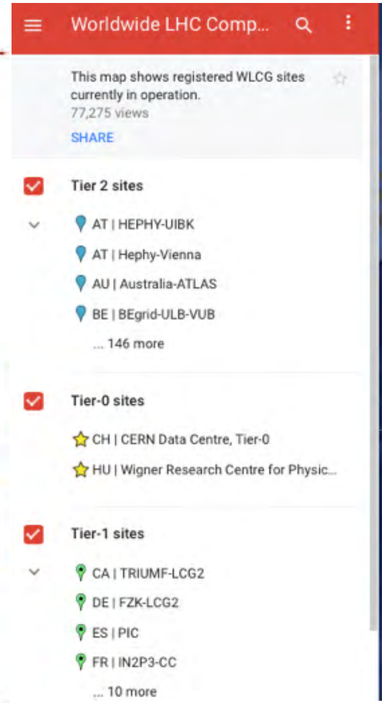


Highlights II (Customer: LHC on Cray)

The LHConCRAY project at CSCS

- **Consolidation project to run LHC jobs on Piz Daint**
  - Partners: CSCS, CHIPP (*Swiss Institute of Particle Physics* - ATLAS, CMS, LHCb)
  - Started ~2 year ago with preliminary studies on a Cray TDS
  - **Started production in April 2017 on Piz Daint**: 25 Cray nodes/1600 cores (ATLAS:CMS:LHCb - 40:40:20)
  - Operated in parallel with Phoenix
  - **The goal is to run ALL VO workloads without changes to the experiments' workflows**
- **Normal workflow:**
  - **Plugs transparently in to the experiments' WMSs**

- **Roadmap**
  - Measure performance in the production environment
  - Produce a cost study (until Dec. 2017)
  - Decision due: **migrate to the Cray or revert to invest on Phoenix**



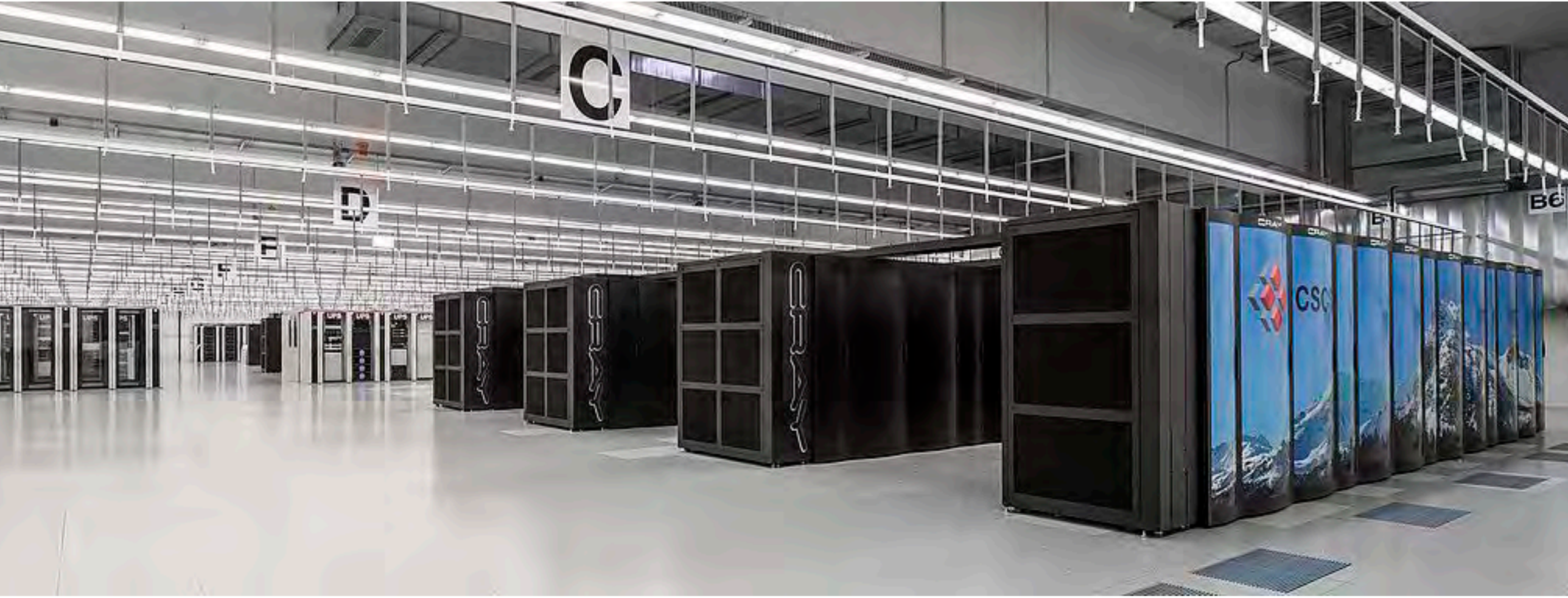
"PIZ DAINT" TAKES ON TIER 2 FUNCTION IN WORLDWIDE LHC COMPUTING GRID
April 1, 2019
"Piz Daint" supercomputer will handle part of the analysis of data generated by the experiments conducted at the Large Hadron Collider (LHC). This new development was enabled by the close collaboration between the Swiss National Supercomputing Centre (CSCS) and the Swiss Institute of Particle Physics (CHIPP). In the past, CSCS relied on the "Phoenix" dedicated cluster for the LHC experiments.

CSCS
Centro Svizzero di Calcolo Scientifico
Swiss National Supercomputing Centre

Summary: Mission, Infrastructure & Services

- CSCS develops and operates cutting-edge high-performance computing systems as an essential service facility for Swiss researchers (https://www.cscs.ch)
- High Performance Computing, Networking and Data Infrastructure
  - Piz Daint supercomputing platform
    - 5000+ Nvidia P100 + Intel E5-2690 v3 nodes
    - 1500+ dual-socket Intel E5-2695 v4 nodes
    - Single network fabric (10s of Terbytes/s bandwidth)
    - High bandwidth multi-Petabytes of scratch (lustre)
  - Storage Systems including SpectrumScale (10s of PetaBytes online & offline storage)
- Services
  - Computing services
  - Data services
  - Cloud services

CSCS
Centro Svizzero di Calcolo Scientifico
Swiss National Supercomputing Centre



Resiliency in the context of experimental facilities workflows

CSCS
Centro Svizzero di Calcolo Scientifico
Swiss National Supercomputing Centre

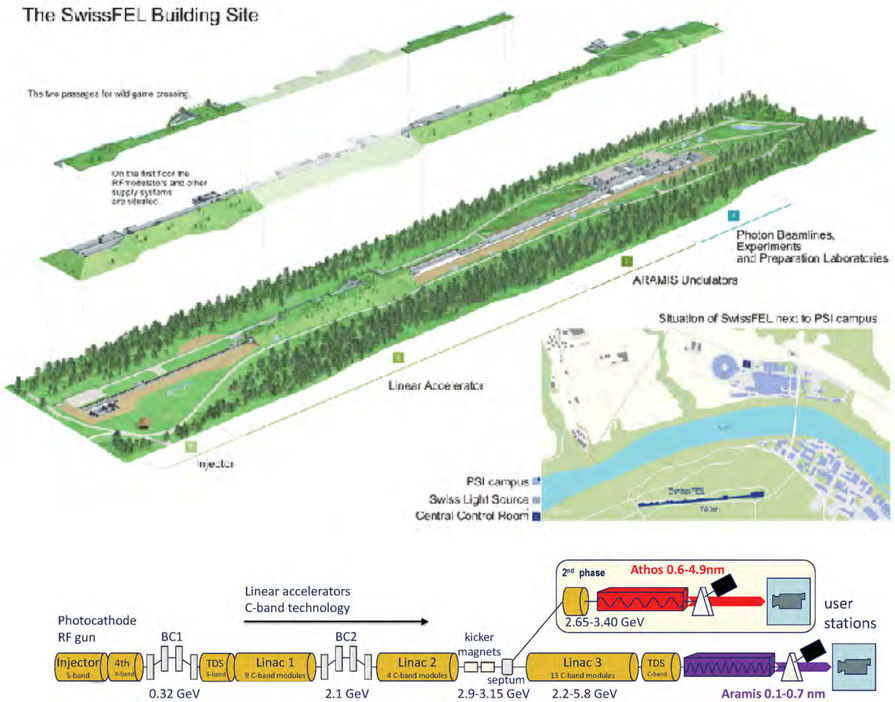
PSI Introduction (https://www.psi.ch)

The Paul Scherrer Institute, PSI, is the largest research institute for natural and engineering sciences within Switzerland. We perform world-class research in three main subject areas: Matter and Material; Energy and the Environment; and Human Health. By conducting fundamental and applied research, we work on long-term solutions for major challenges facing society, industry and science.

Labs & User services
- Research Divisions and Labs
- Facilities and Instruments
- PSI User Laboratories

Visitors
- Public Events at PSI
- Visitor Centre psi forum
- The iLab School Laboratory (in German)

Industry
- Technology Transfer
- Expertise and Services
- PSI working with companies



The SwissFEL is a X-ray free-electron laser (the FEL in its name stands for Free Electron Laser), which will deliver extremely short and intense flashes of X-ray radiation of laser quality. The flashes will be only 1 to 60 femtoseconds in duration (1 femtosecond = 0,000 000 000 000 001 second). These properties will enable novel insights to be gained into the structure and dynamics of matter illuminated by the X-ray flashes

Data Catalog and Archiving @ PSI

- https://www.psi.ch/photon-science-data-services/data-catalog-and-archive
- Data sets with PIDs
- Petabyte Archive System @ CSCS
  - packaging, archiving and retrieving the datasets within a tape based long-term storage system
- Necessary publication workflows to make this data publicly available
- PSI data policy which is compatible with the FAIR principles

PSI-CSCS PetaByte Archive Initiative

CSCS WILL STORE PETABYTE DATA FOR THE PAUL SCHERRER INSTITUTE

In the future, research data collected by the large-scale research facilities at the Paul Scherrer Institute (PSI) in Villigen will be archived at the Swiss National Supercomputing Centre (CSCS) in Lugano. Collaboration between PSI and CSCS enabled major improvements to the data transfer and storage process.

Highlights:

Archival storage for the new SwissFEL X-ray laser and Swiss Lightsource (SLS)

A total of **10 to 20 petabytes** of data is produced every year

A dedicated redundant network connection between PSI and CSCS, **10 Gbps**

CSCS tape library current storage capacity is **120 petabytes**, can be extended to 2,000 petabytes

By 2022, PSI will transfer around **85 petabytes** of data to CSCS for archiving. Around 35 petabytes come from SwissFEL experiments, and 40 come from SLS.

Problem Statement

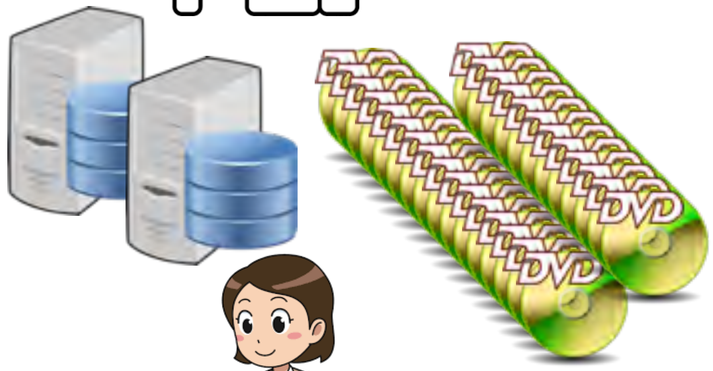
Before upgrade

Sometime before day n
User applies for beam time

Day n + couple of days/weeks
User @ PSI collects and processes data
Complete output stored on user media

After upgrade

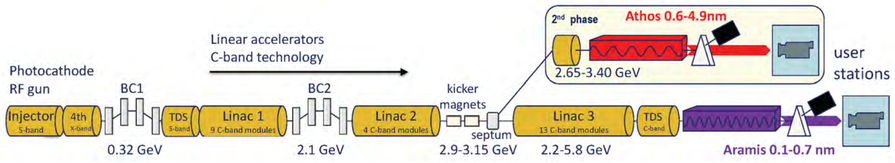
Sometime before day n
User applies for beam time



Day n + couple of days/weeks
User @ PSI collects and processes data
Complete output archived at CSCS

CSCS
Centro Svizzero di Calcolo Scientifico
Swiss National Supercomputing Centre

PSI Online Workflow (s)



Local network

Local network

Swiss A&R network

Realtime
Compression
Data Transfer
Tightly coupled & resilient

PAUL SCHERRER INSTITUT

Staging and preparation for archiving (PSI service)

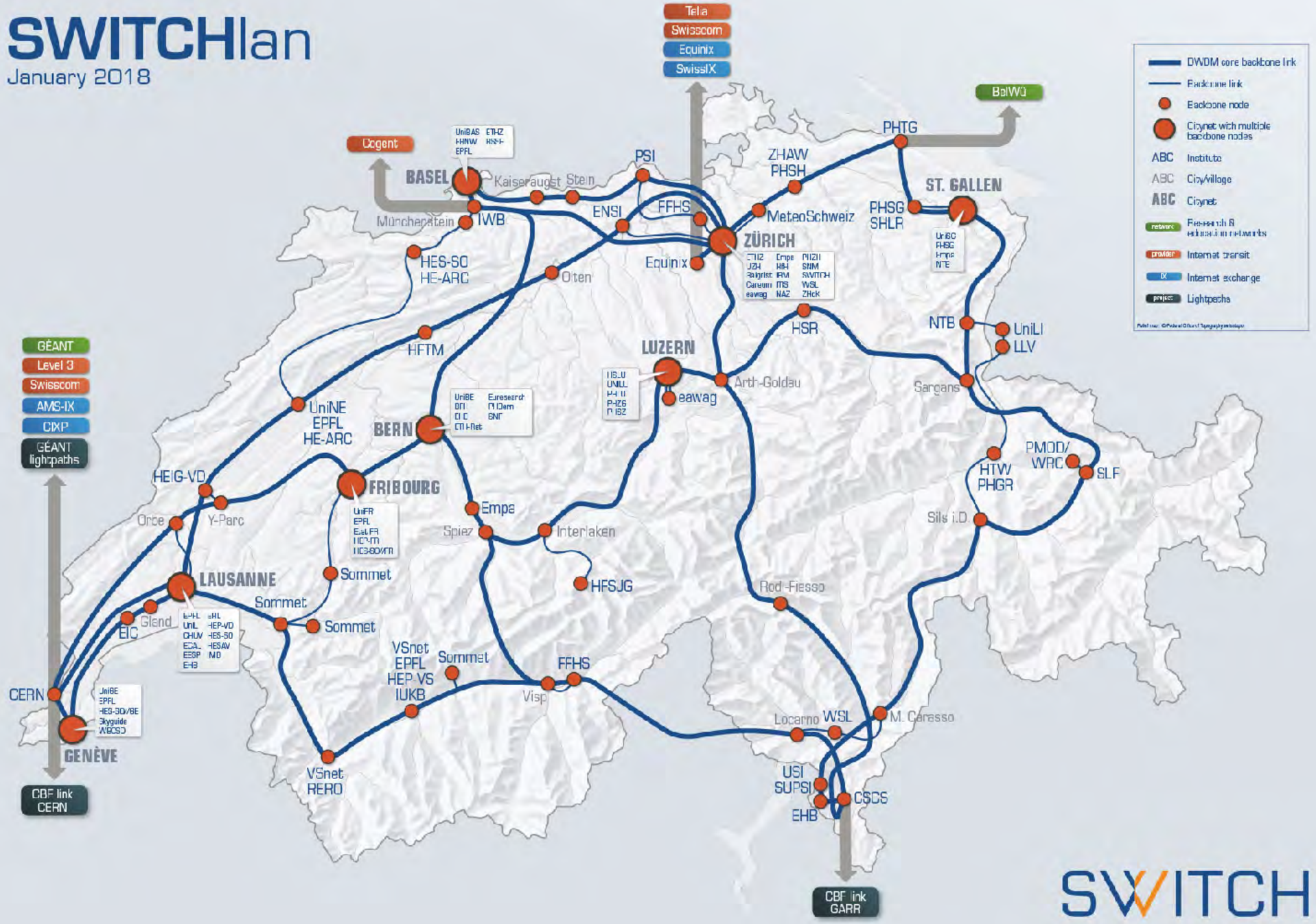
PAUL SCHERRER INSTITUT

Selected data processing by user @ PSI (PSI service)

PAUL SCHERRER INSTITUT

Archived data at CSCS (Data at rest)

CSCS
Centro Svizzero di Calcolo Scientifico
Swiss National Supercomputing Centre


SWITCHlan
January 2018
Tata
Swisscom
Equinix
SwissIX
Cogent
BelWü
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
BASEL
Kaiseraugst
Stein
PSI
ZHAW
PHSH
PHTG
ST. GALLEN
PHSG
SHLR
ENSI
FFHS
MeteoSchweiz
ZÜRICH
Equinix
Münchenstein
IWB
HES-SO
HE-ARC
Olten
HSR
NTB
UniLI
LLV
Sargans
HFTM
LUZERN
eawag
Arth-Goldau
UniNE
EPFL
HE-ARC
BERN
HEIG-VD
Orbe
Y-Parc
FRIBOURG
Empa
Spiez
Interlaken
HTW
PHGR
PMOD/
WRC
SLF
Sils i.D.
LAUSANNE
Sommet
HFSJG
Rod-Fiesso
Glend
EIC
VSnet
EPFL
HEP-VS
IUKB
FFHS
Visp
Locarno
WSL
M. Carasso
USI
SUPSI
EHB
CSCS
VSnet
RERO
CERN
GENÈVE
GÉANT
Level 3
Swisscom
AMS-IX
CIXP
GÉANT lightpaths
CBF link CERN
CBF link GARR
SWITCH

PSI Offline Workflow (s)

Data access & analysis portal

PAUL SCHERRER INSTITUT

PSI

Accesses PSI service for archival data processing

Data access service (PSI)

Workflow service (PSI)

Archived data at CSCS (Data in motion)

Data unpacking service (PSI)

Data mover service (CSCS)

Job submission service (CSCS)

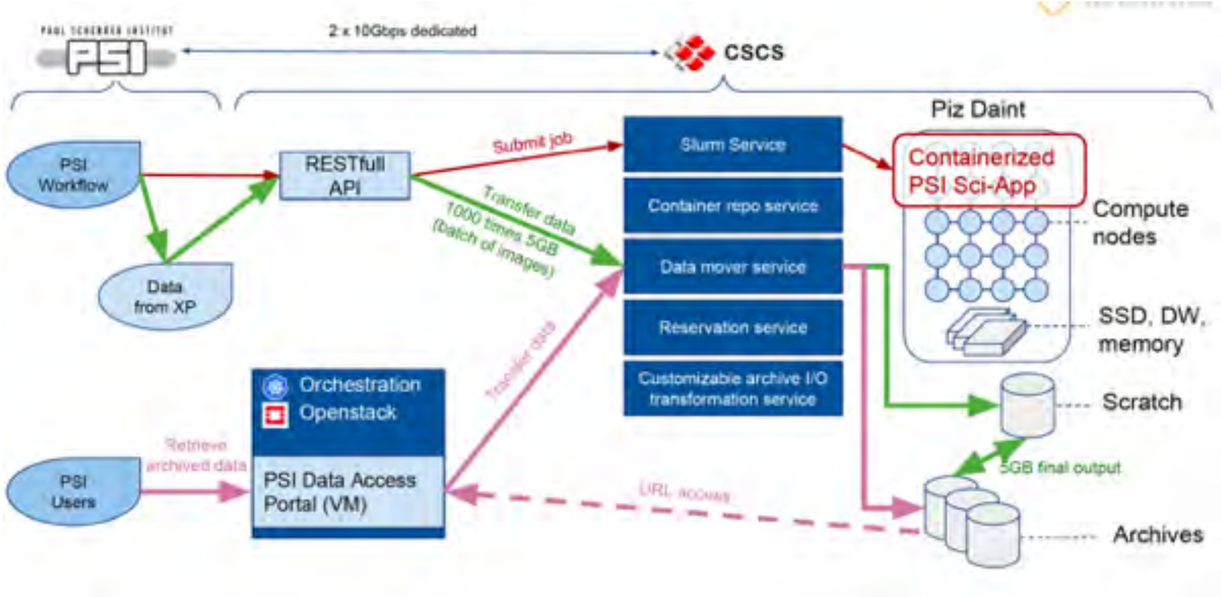
CSCS

Centro Svizzero di Calcolo Scientifico

Swiss National Supercomputing Centre

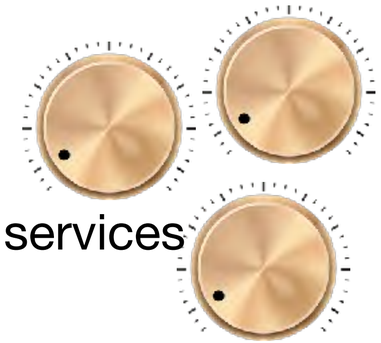
Resilience

- Full, multi-level redundancy, over-provisioning & failover not an option at scale …
  - Especially for government funded research programs
- Use case driven approach
  - Functionality tradeoffs
  - Performance tradeoffs
- Partial and programmable redundancy
  - To manage functionality & performance tradeoffs through virtualisation

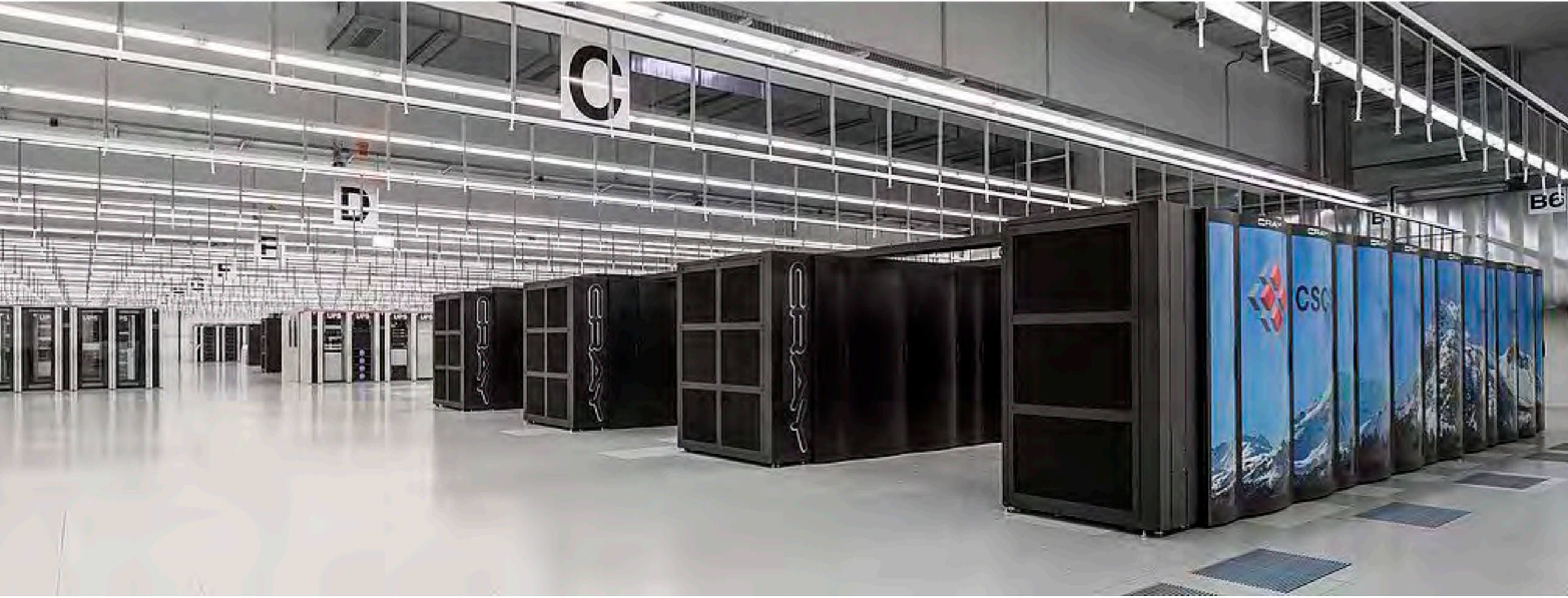


Co-designing Resilient Solutions

- Data at rest (few CSCS systems and services, mainly storage processing)
  - Functionality: network resilience (fixed CapEx/OpEx), storage system failures (programmable with extra CHF or local buffering @ PSI), data corruption (fixed CapEx/OpEx), …
  - Performance: network resilience (fixed CapEx/OpEx), regression @ CSCS (programmable/tuneable with extra CHF or local buffering), …



- Data in motion (several CSCS HPC, storage and cloud systems and services)
  - Functionality: HPC systems failure (tough), site-wide storage failure (tough), cloud services (programmable, failover to private or public cloud), …
  - Performance: HPC systems regression (programmable with extra CHF or wait or tolerate slowdown), site-wide storage regression (programmable with extra CHF or wait or tolerate slowdown), cloud services regression (really?), …



Future: Co-designed HPC & cloud services for federated, data-driven workflows

CSCS
Centro Svizzero di Calcolo Scientifico
Swiss National Supercomputing Centre

EU Fenix Research Infrastructure

Functional resilience through federation (technical and business solutions)
Performance resilience is still work in progress …
… for nationally funded programs

CSCS
Centro Svizzero di Calcolo Scientifico
Swiss National Supercomputing Centre

Use case & cost-performance driven approach
X-as-a-Service oriented infrastructure for HPC

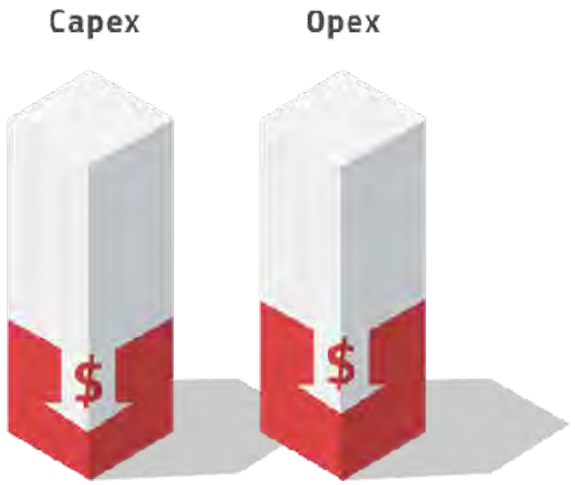
Performance

Empowering users & customers

ssh, sbatch, scp, …

—>

IaaS, PaaS, SaaS



Functionality

Invitation to SC19 Workshop (SuperCompCloud: Workshop on Interoperability of Supercomputing and Cloud Technologies)
November 18, 2019
Denver, CO, USA