PAUL SCHERRER INSTITUT
PSI

Alun Ashton ; CaSIT Work package lead

PSI/SLS: On going project or future plan for IT transformation

19 September Soleil Visit to PSI

PAUL SCHERRER INSTITUT PSI

SLS vs SLS2.0 Network Planning

SLS Services - Overview

Core 1/10Gb/s

Machine network
- Cu 1Gb/s

BeamLine network

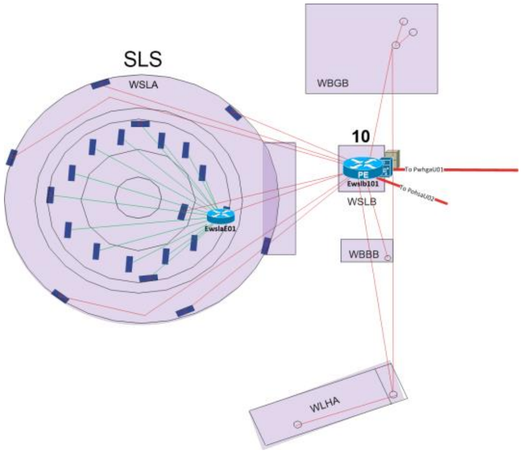
- Cu 1Gb/s, few 10Gb/s

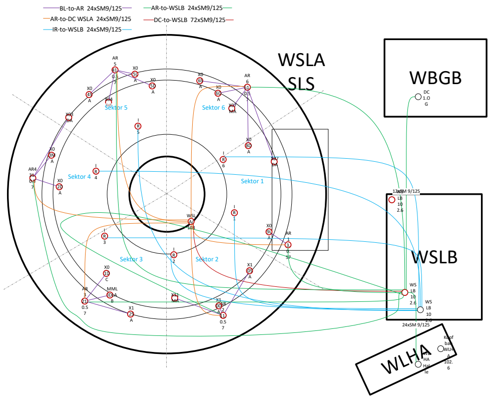
WLAN (not in Tunnel)

- corp
- guest/eduroam

Cabling

- Fibre Multimode
- Fibre Singlemode





SLS2.0 Services - Overview

- Core 100Gb/s
- Machine network
- 1/10/25/100Gb/s
- BeamLine network
- Cu 1/10Gb/s
- Fibre 10/25/100Gb/s

WLAN in Tunnel as well

- corp/infra
- guest/eduroam

Cabling

- Fibre Singlemode

Thanks to: Stefan Fries

SLS2.0 Network Planning

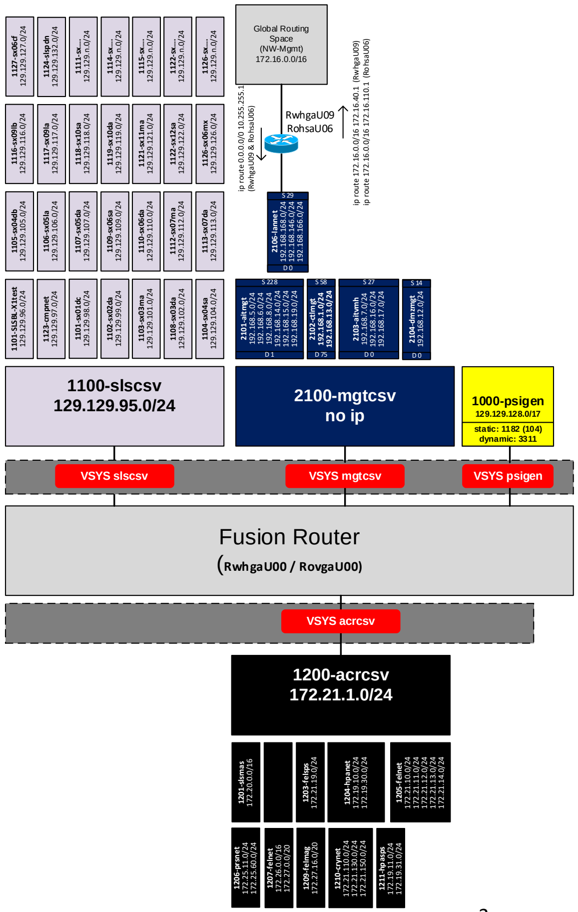
Zone concept will stay as it is today within SLS

Accelerator:

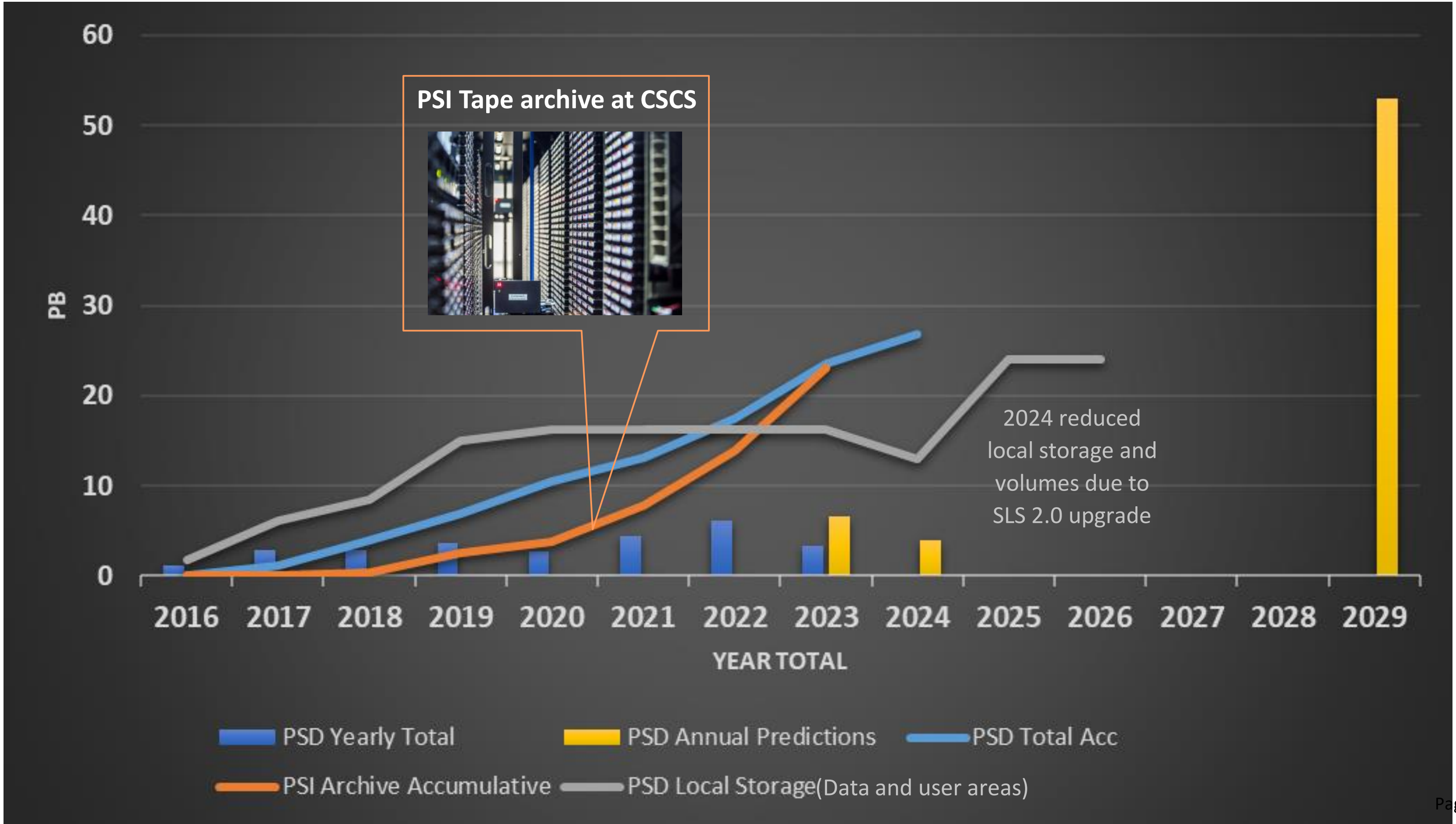
- EPICS machine Network will subneted and therefore built with /24 networks, instead of /16 as in SLS today,
- Beside EPICS Machine Network we will have noneEPICS machine network (as in SwissFEL)

Beamline

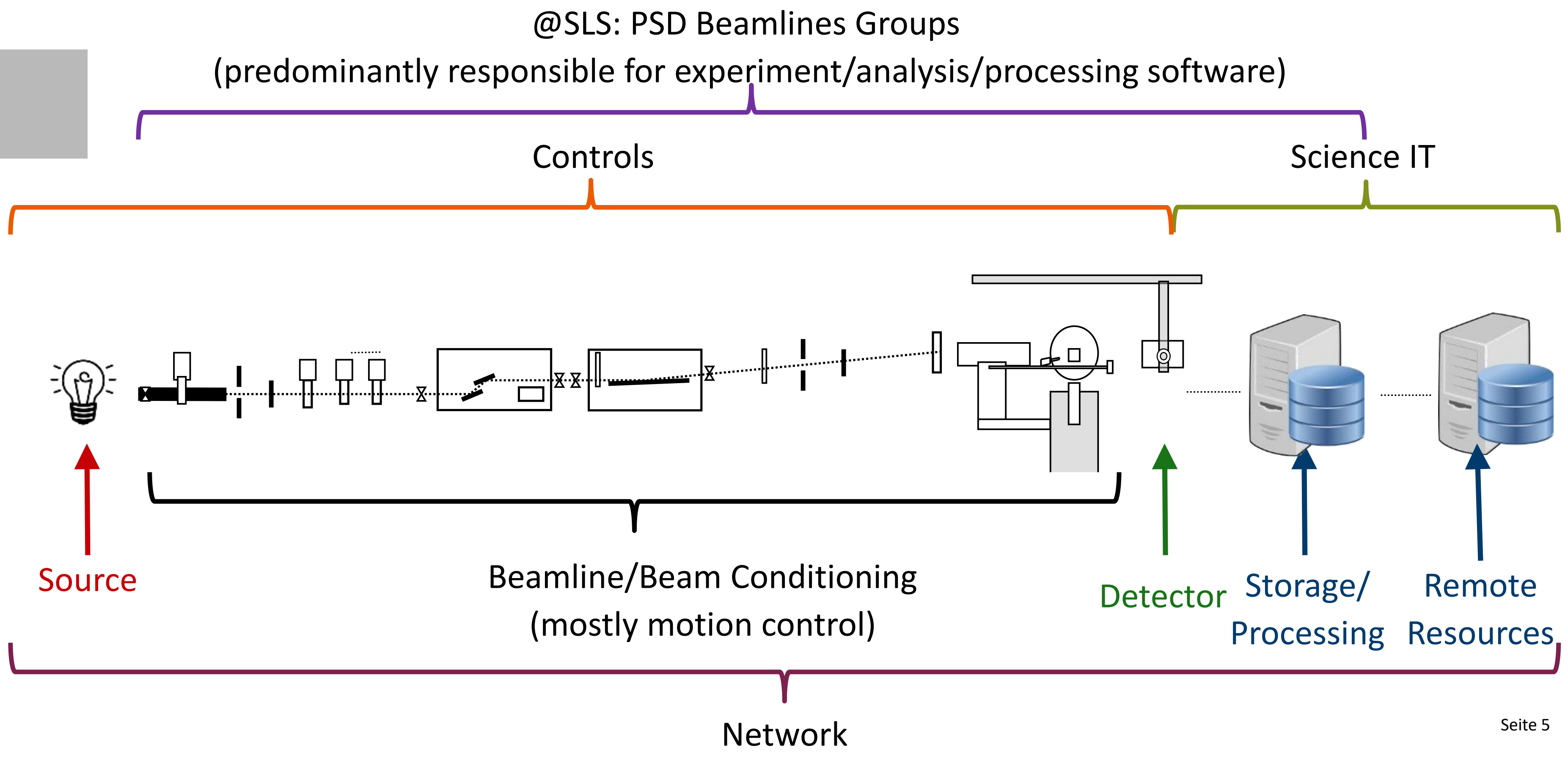
- Each Beamline will have it's own network. Beamline networks do not see each other.
- There will be a shared network between all beamline, Common Service VLAN network
- Detector Ethernet switches/networks are covered by Science IT
- Networking boundaries between accelerator and beamlines need to be agreed



Photon Science (PSD) and PSI Data Volumes



Experiment Schematic and Responsibilities



PAUL SCHERRER INSTITUT PSI

Andrej Babic :: Software Engineer :: Paul Scherrer Institute

Standard DAQ

22.5. Brugg

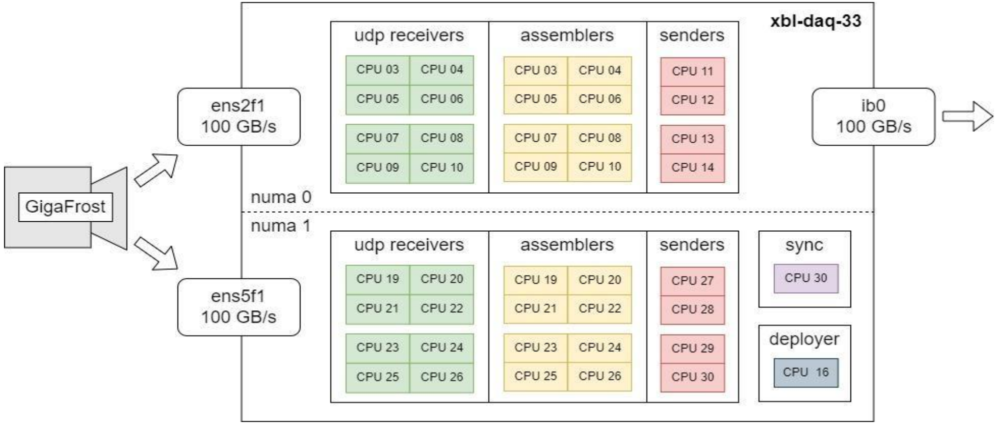
Standard platform for detectors, STD_DAQ

- Support all PSI facilities with a common solution
- Consolidation of DAQ operational knowledge
- Modular design to accommodate individual needs:
  - Different detectors
  - SwissFEL and SLS usage patterns
  - Beamline specific requirements

Thanks to Andrej Babic, Controls

Components

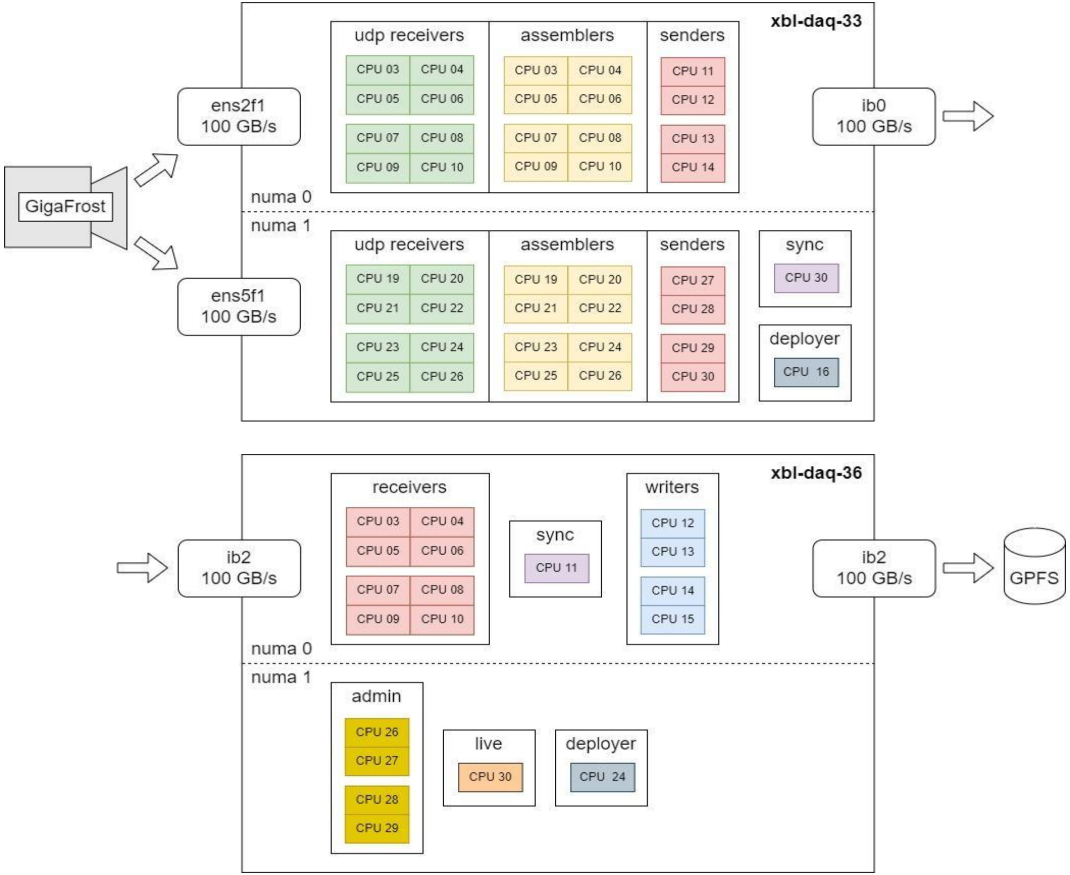
- Detector readout
  - Udp receiver
  - Image assembler
  - Synchronizer
  - Buffer sender/writer



Components

- Detector readout
  - Udp receiver
  - Image assembler
  - Synchronizer
  - Buffer sender/writer

- Stream processors
  - Buffer receiver/reader
  - Detector writer
  - Admin interface
  - Live streamer
  - Configuration deployer



PAUL SCHERRER INSTITUT

Deployment

- Reproducible builds
- Single action deployment
- Audit trail

- Addresses all but the lowest and highest data rates from beamlines.

PAUL SCHERRER INSTITUT

Filip Leonarski :: Beamlines Data Scientist :: MX Data Group

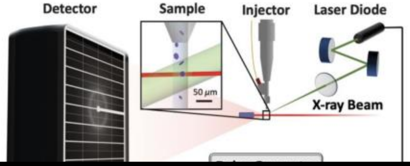
Jungfraujoch image acquisition and analysis system

Data acquisition, reduction and online processing workshop
Brugg, May 22nd, 2023

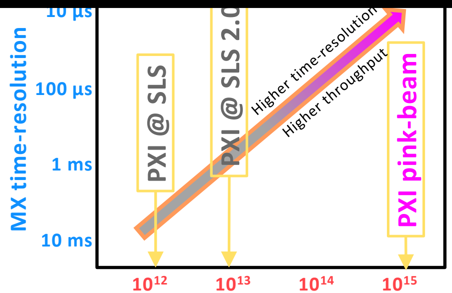
PAUL SCHERRER INSTITUT

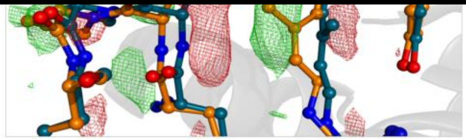
Time-resolved serial synchrotron crystallography at SLS 2.0

- **Serial crystallography** solves protein structures with diffraction images from thousands of crystals
- **PXI-VESPA**: A Versatile End-station for Scattering Pink-beam Applications



Last week we have collected protein dynamics data with
ns laser and JUNGFRAU storage cells (16 images x 100 µs) at PXI:
proof-of-concept for 10+ kHz detector for
time-resolved MX @ SLS 2.0

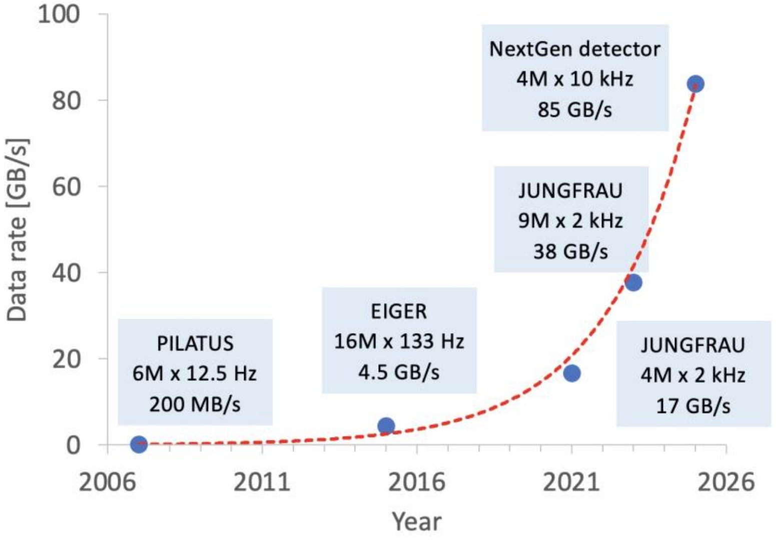




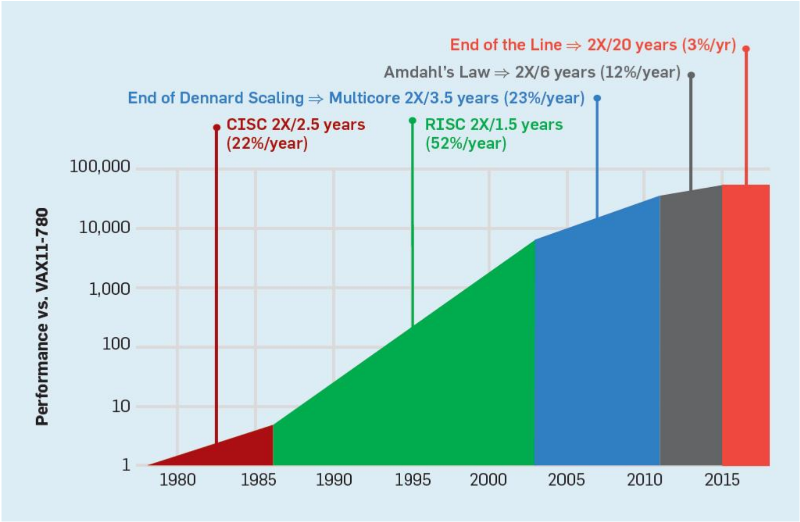
F. Leonarski, J. Nan, …, F. Dworkowski (submitted)
«Kilohertz Serial Crystallography with the JUNGFRAU Detector at a 4^{th} Generation Synchrotron Source»

Need sustainable DAQ for increasing data rates

MX detector data rates @ SLS double every two years

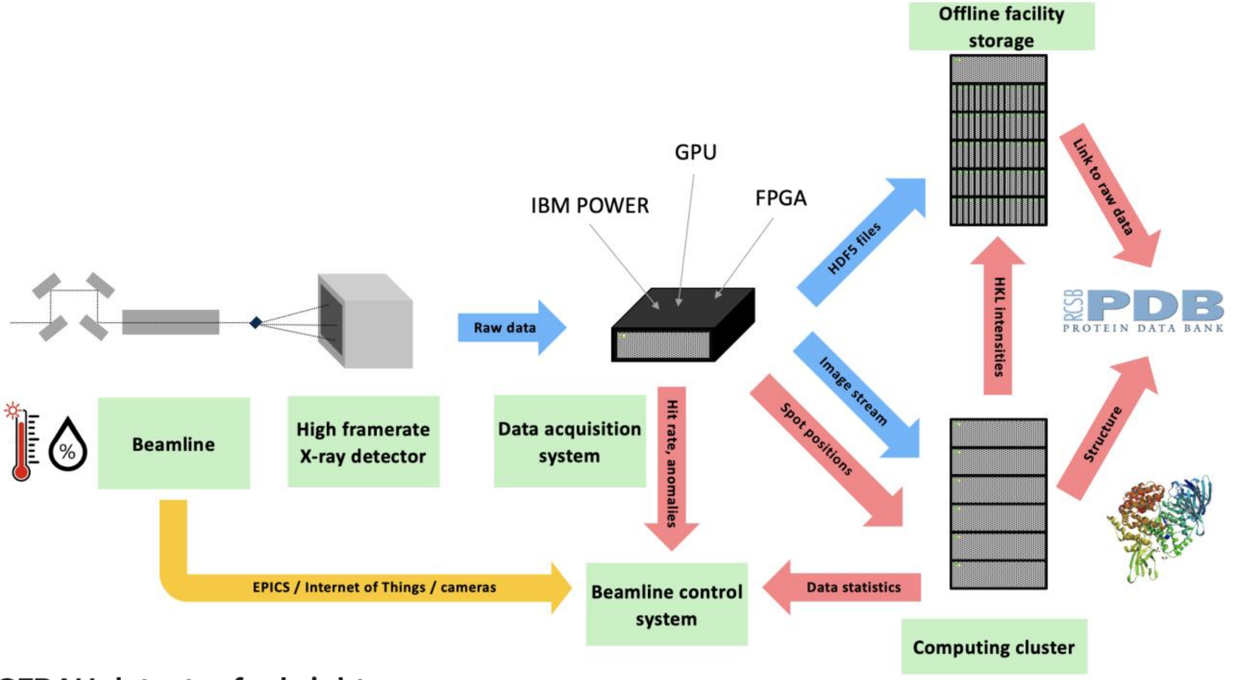


CPU performance is no match for such growth



Hennessy & Patterson
doi:10.1145/3282307

Data pipeline for crystallography beamline



JUNGFRAU detector for brighter x-ray sources: Solutions for IT and data science challenges in macromolecular crystallography

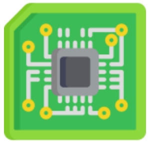
Cite as: Struct. Dyn. **7**, 014305 (2020); doi: 10.1063/1.5143480
Submitted: 27 December 2019 · Accepted: 4 February 2020 ·
Published Online: 26 February 2020

View Online Export Citation CrossMark

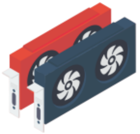
Filip Leonarski,[a] Aldo Mozzanica, Martin Brückner, Carlos Lopez-Cuenca, Sophie Redford, Leonardo Sala, Andrej Babic, Heinrich Billich,[b] Oliver Bunk, Bernd Schmitt, and Meitian Wang

Jungfraujoch: hardware-accelerated platform

“Black box” design
Like DECTRIS Detector Control Unit: all-in-one
Optimized for MX science case



x86 server (2023)
Possible performance up to 40 GB/s



HW and SW platform
Data acquisition on FPGA
Image analysis on GPU
Compression on CPU

JOURNAL OF SYNCHROTRON RADIATION

ISSN 1600-5775

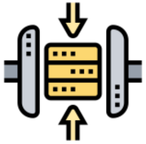
Jungfraujoch: hardware-accelerated data-acquisition system for kilohertz pixel-array X-ray detectors

Filip Leonarski,[a*] Martin Brückner,[a] Carlos Lopez-Cuenca,[a] Aldo Mozzanica,[a] Hans-Christian Stadler,[b] Zdeněk Matěj,[c] Alexandre Castellane,[d] Bruno Mesnet,[d] Justyna Aleksandra Wojdyla,[a] Bernd Schmitt[a] and Meitian Wang[a]

Received 23 June 2022
Accepted 24 October 2022

Complementary projects
Innosuisse
RED-ML
Open Research Data

Simple deployment of JUNGFRAU for MX beamlines: tested at SLS (CH), MAX IV (SE) and KEK (JP)



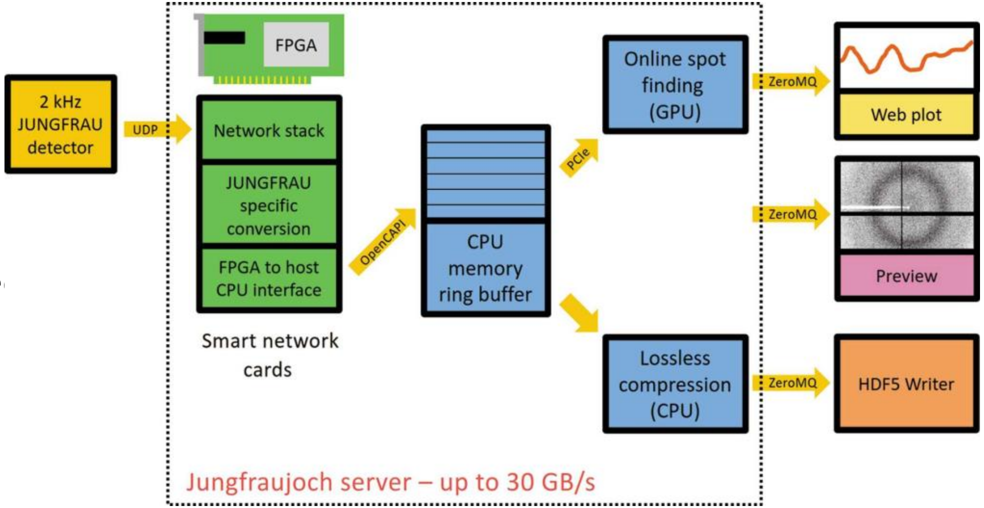
Community accepted interfaces for file writing and streaming

Jungfraujoch: architecture

- Jungfraujoch <-> JUNGFRAU
  - Control (via slsDetectorPackage)
  - Receiving UDP stream

- ZeroMQ stream output:
  - CBOR encoding (DECTRIS Stream2)
  - Image: raw or photon count, compresse
  - Optional pixel binning
  - Real-time analysis results
- Stream to GPFS node for NeXus writer

- Visualize images: DECTRIS Albula or Adxv
- Configuration and analysis result
  - gRPC or REST
  - Web frontend



Jungfraujoch: platform

- **FPGA as smart network card**
  - Offload data acquisition
  - JUNGFRAU conversion to photon counts before data arrive in host memory
  - High effort in development
- **IBM POWER9**
  - Low effort for FPGA integration
  - No suitable server with POWER10
  - From 2023 POWER9 support deprecated in Jungfraujoch
- **x86 servers**
  - Wide availability of hardware
  - Improvements with new CPU generations
  - Extra effort to develop and maintain kernel driver

Xilinx Alveo U55C

IBM POWER9

Jungfraujoch: platform

- **FPGA as smart network card**
  - Offload data acquisition
  - JUNGFRAU conversion to photon counts before data arrive in host memory
  - High eff
- **IBM POW**
  - Low eff
  - No suita
  - From 20
- **x86 servers**
  - Wide availability of hardware
  - Improvements with new CPU generations
  - Extra effort to develop and maintain kernel driver

FPGA is not the only way;
User-space (Mellanox Raw Ethernet) and Linux sockets were recently implemented in Jungfraujoch as well

IBM POWER9

PAUL SCHERRER INSTITUT

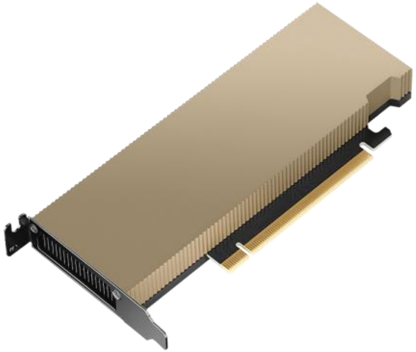
Jungfraujoch: real-time data analysis on GPU

- Real-time analysis for MX:
  - Spot finding
  - Indexing
  - Radial integration / background estimation

- Real-time analysis requires balance between precision and execution time

- GPU benefits:
  - Fast for computation
  - Extends memory bandwidth

- Inference grade GPUs are cost-effective

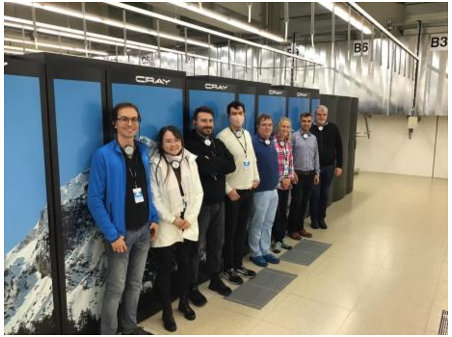


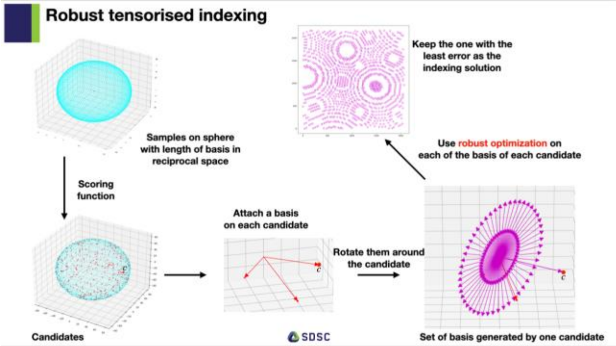
Nvidia L4

Nvidia L40

Reduction of high volume experimental data using machine learning (RED-ML)

- Funded by the Swiss Data Science Center
- Realized by Science IT, SDSC, CSCS and MX Group
- Main outcome: fast indexing algorithm for serial crystallography running on GPUs
- Solution possible in **500 µs**
  (CPU based algorithms require ~100 ms)
- CrystFEL integration: tested on Piz Daint supercomputer



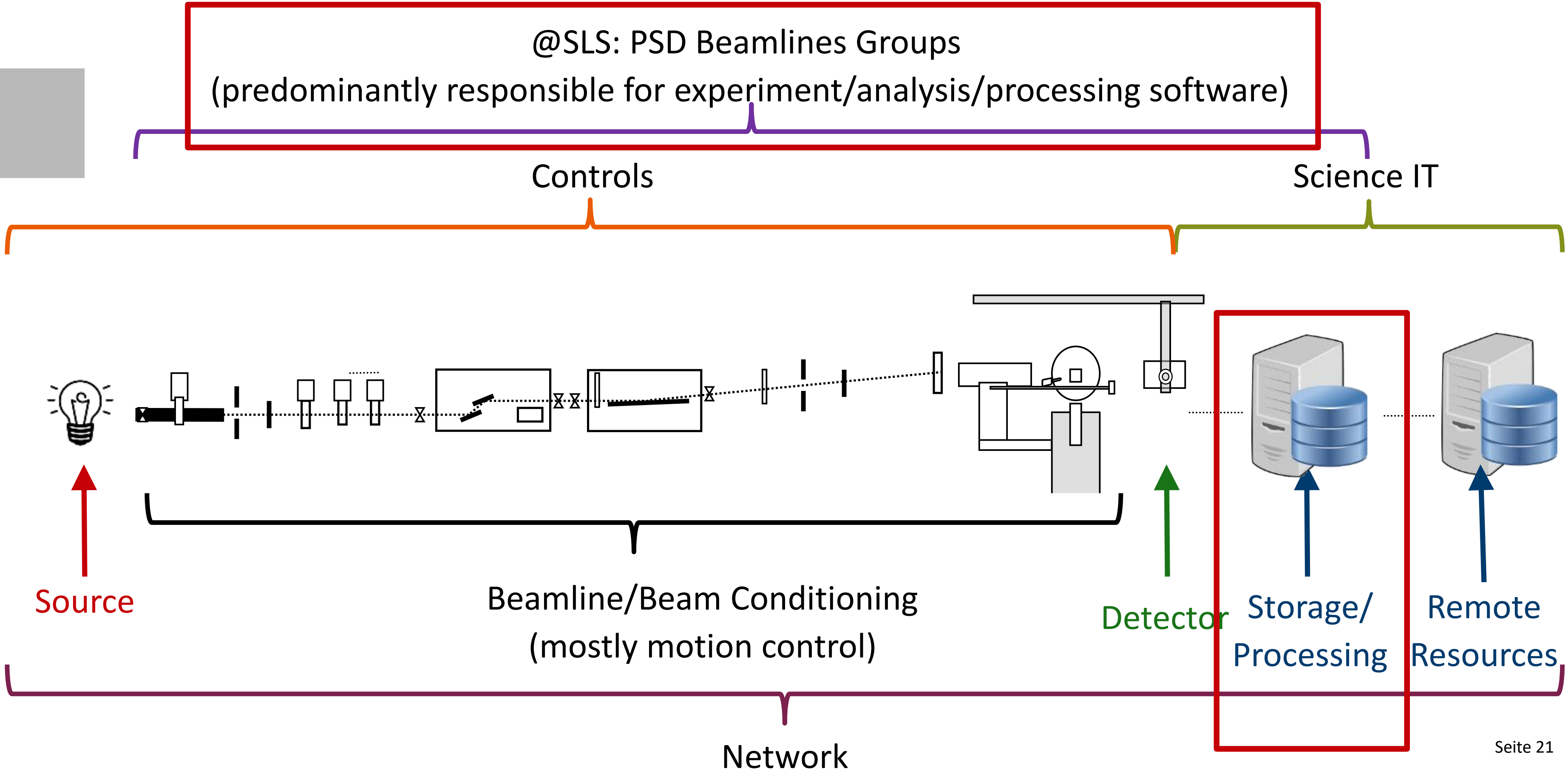


Acknowledgements:
A. Ashton, G. Assmann, L. Barba, B. Béjar,
P. Gasparotto, M. Janousch, T. Koka,
H. Mendoça, H.-C. Stadler

SDSC

CSCS
Centro Svizzero di Calcolo Scientifico
Swiss National Supercomputing Centre

Experiment Schematic and Responsibilities

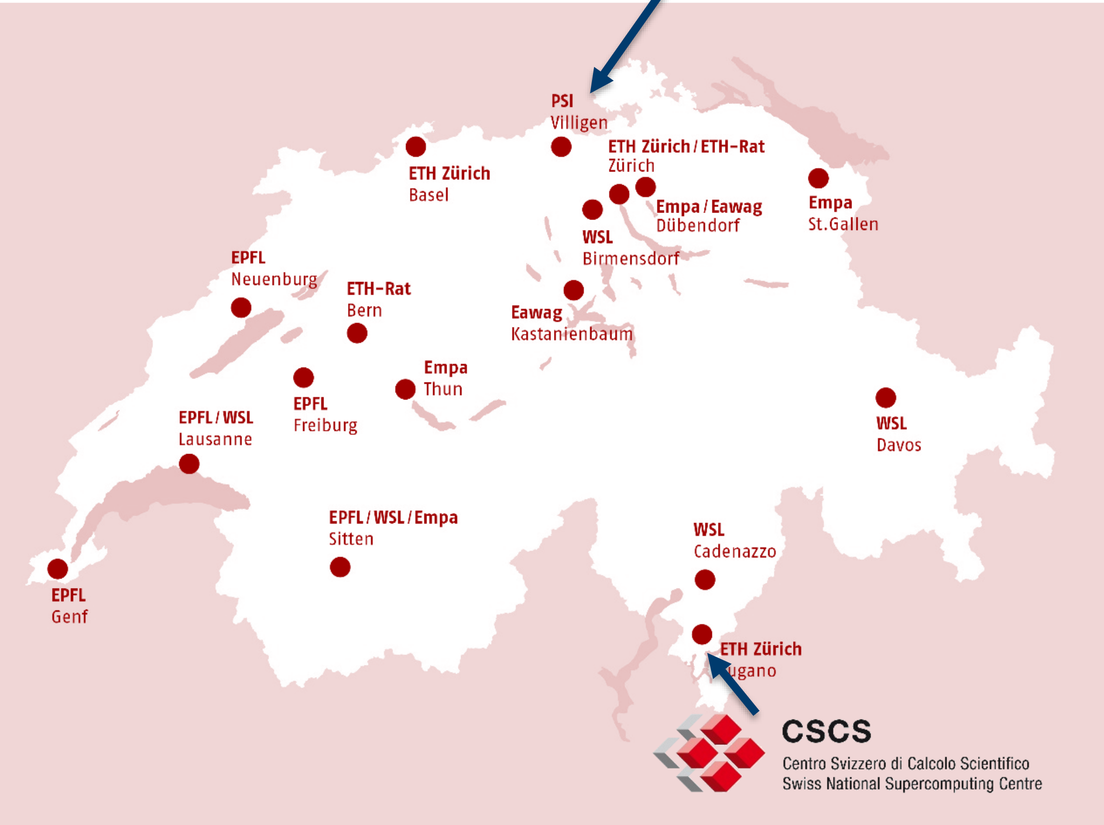


PAUL SCHERRER INSTITUT

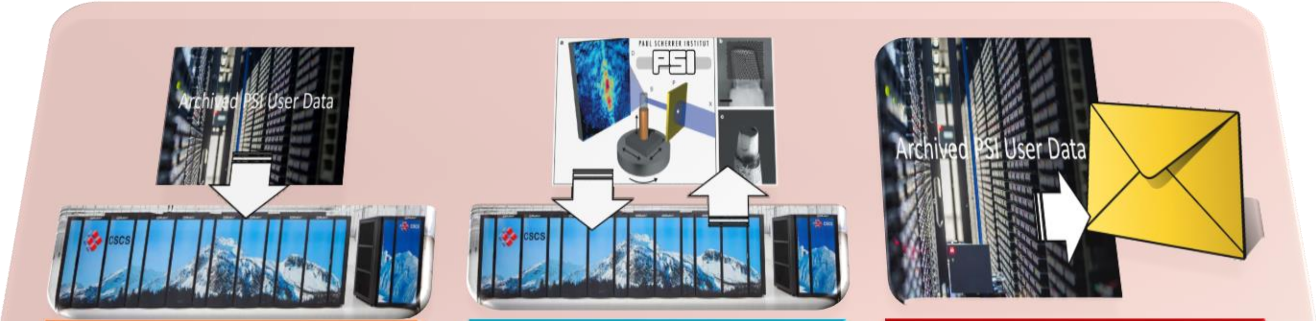
Remote resources, why?

The ETH Domain

ETH BOARD | ETH zürich | EPFL | PAUL SCHERRER INSTITUT PSI | WSL | Empa Materials Science and Technology | eawag aquatic research



SELVEDAS (2019-2022) Targeted Use cases



Use case 1

Analysis of archived data

Use case 2

Fast experiment feedback

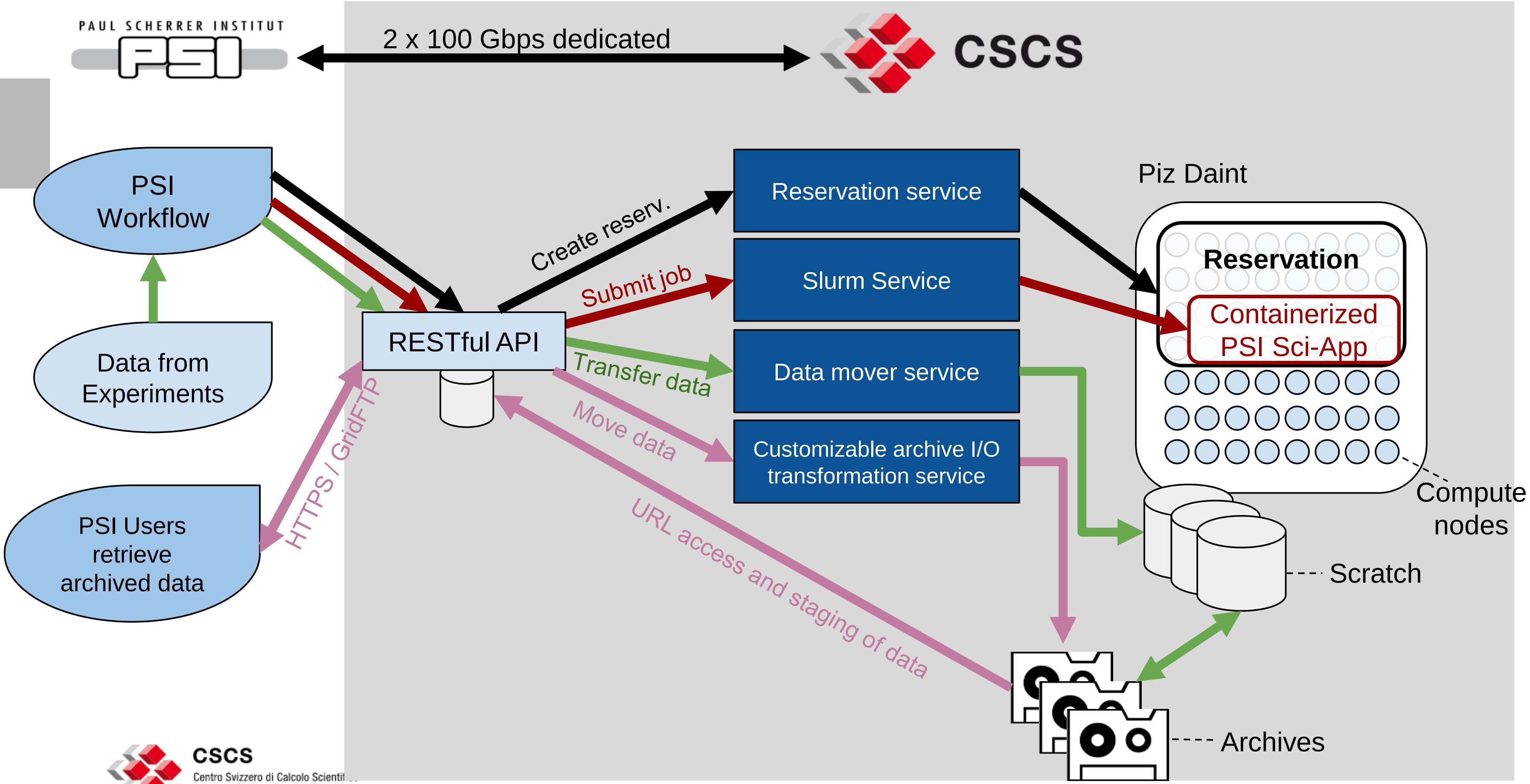
Use case3

PSI user data direct from CSCS

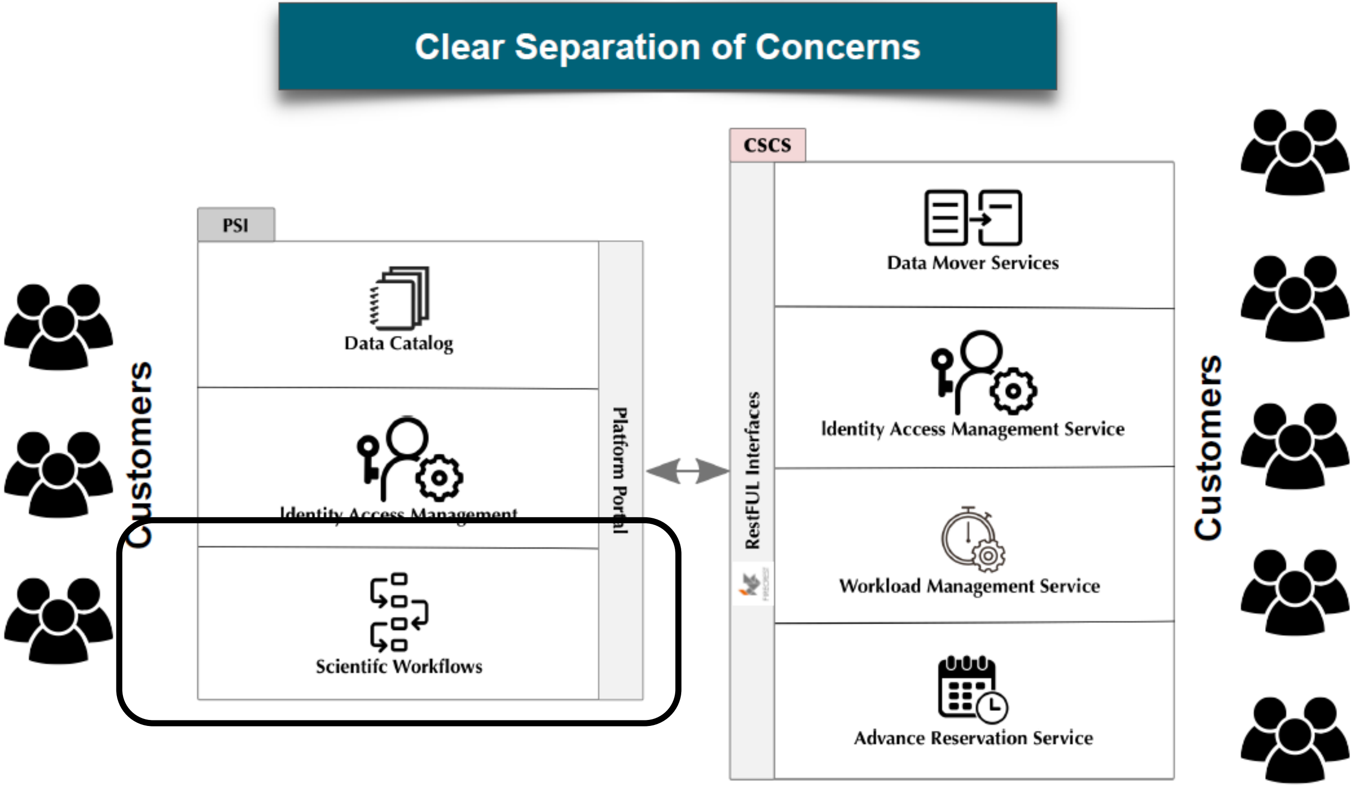
Using HPC resources (e.g. RA / CSCS)

1. Standard: login via ssh / NX / Graphical protocol, plus job submission (e.g Slurm)
2. Interactive: Jupyter notebooks (Jupyterhub running slurm jobs in background)
3. “SELVEDAS”: running containers over Slurm triggered by remote API call
4. Virtual Cluster: creation of a dedicated OpenStack cluster, composed of various VMs

The communication architecture



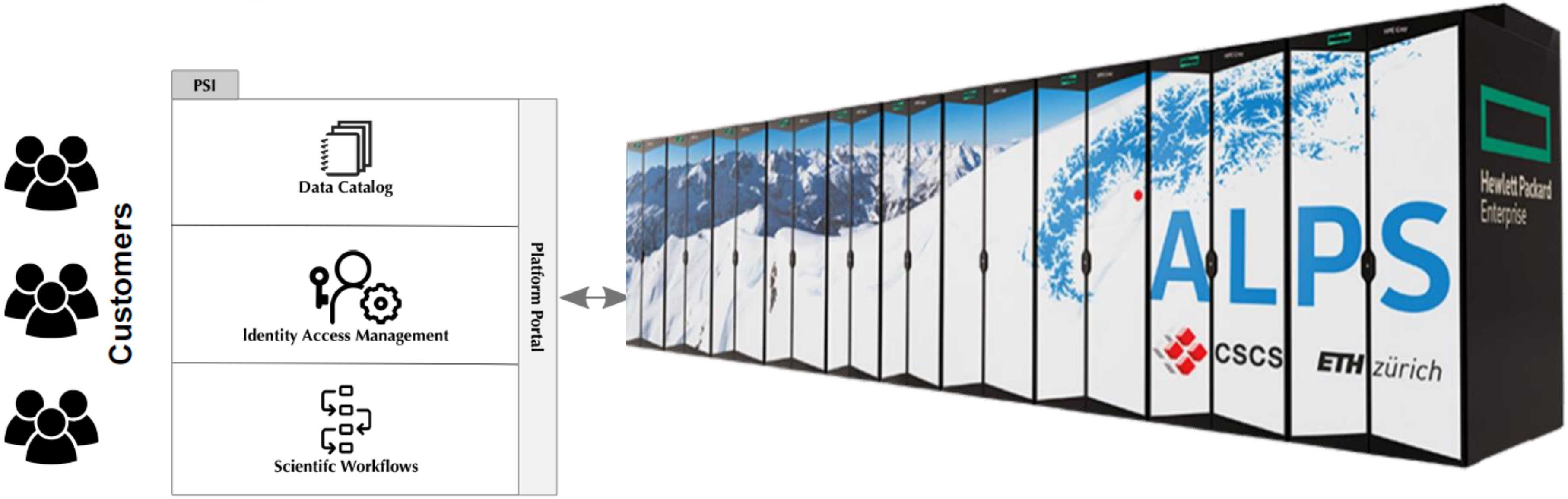
SELVEDAS



Clear Separation of Concerns
PSI
Data Catalog
Identity Access Management
Scientifc Workflows
Platform Portal
Customers
CSCS
RestFUL Interfaces
Data Mover Services
Identity Access Management Service
Workload Management Service
Advance Reservation Service
Customers

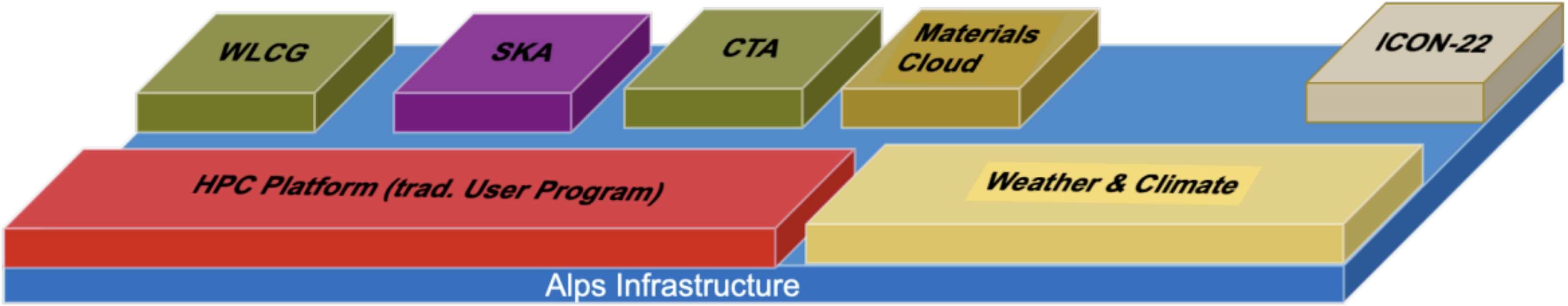
HPC for the experiment/research lifecycle: offline

Clear Separation of Concerns



CSCS and Alps background

- Background Article in Technical Publication April 2023. https://www.hpcwire.com/2023/04/05/into-the-alps-what-exactly-is-the-new-swiss-supercomputer-infrastructure/



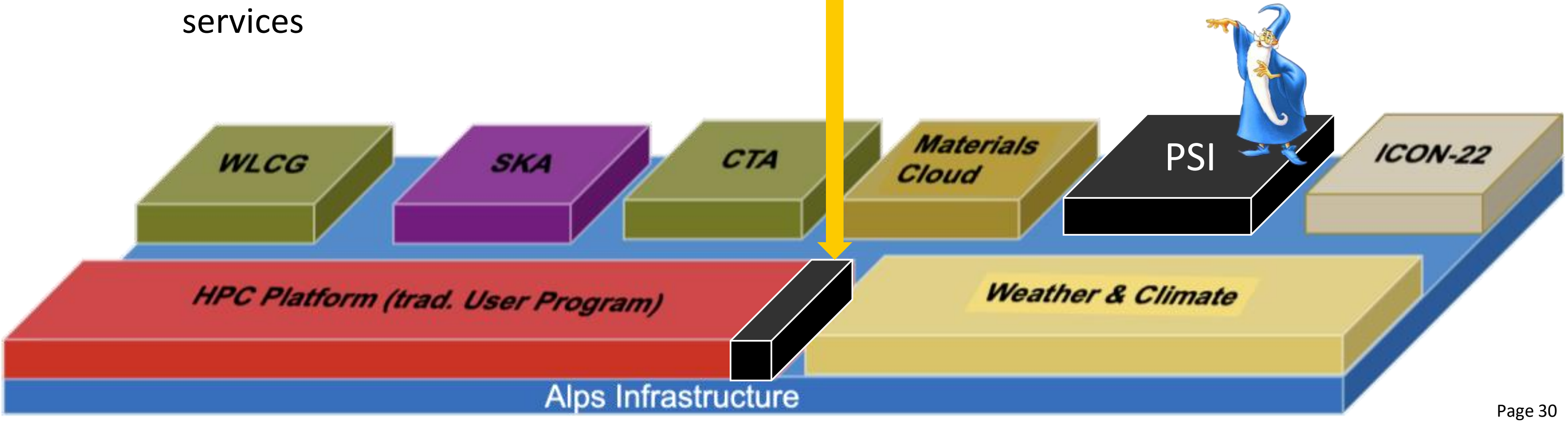
- *"…. most of Alps' computing power would come from Nvidia's novel Grace Hopper Superchips, each of which contains an Arm-based Nvidia Grace CPU and an Nvidia Hopper GPU."*

Very technical publication http://dx.doi.org/10.1177/10943420231167811

PAUL SCHERRER INSTITUT

Mapping the PSI use cases and requirements

1. **Large-scale resources**: Buy-in share in Alps Userlab (currently on Piz Daint)
   - For jobs needing *large number of parallel CPUs* over many nodes
   - ~15 projects per year, power users
2. **Mid-scale resources**: Merlin7 on Alps
   - For *high throughput* jobs (1 CPU) and smaller *parallel* jobs.
   - Need resource able to efficiently collocate several jobs per node or across a few nodes
   - Main part of PSI users (200 users, 80 groups per year), best fully integrated with PSI services



What were the PSI expectations for mid-scale

- HW costs and operation costs similar to PSI (ideally cheaper)
- CAPEX->OPEX, potential for yearly investment

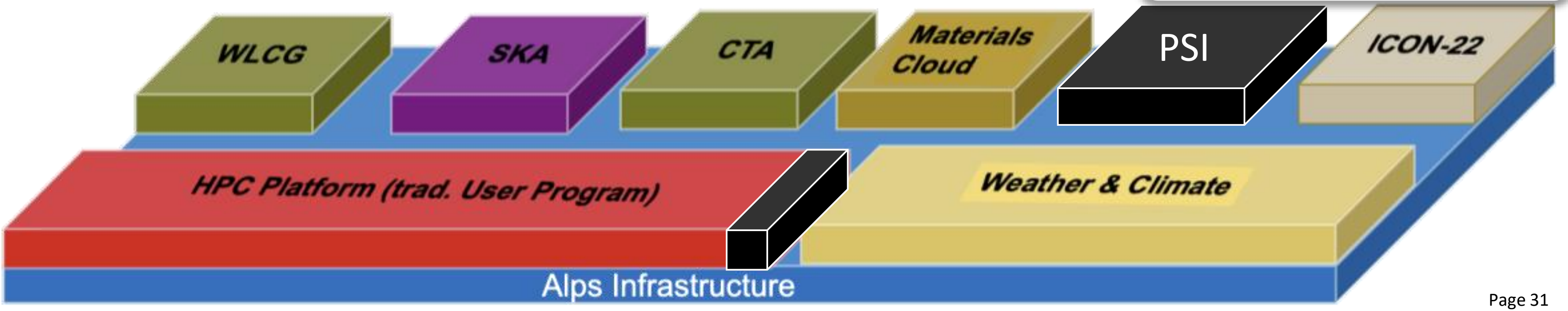
Negotiable and can evolve with time

- Elasticity, ability to expand or contract on demand
- Scaling, ability to grow baseline resources year on year
- Specialised computing requirements

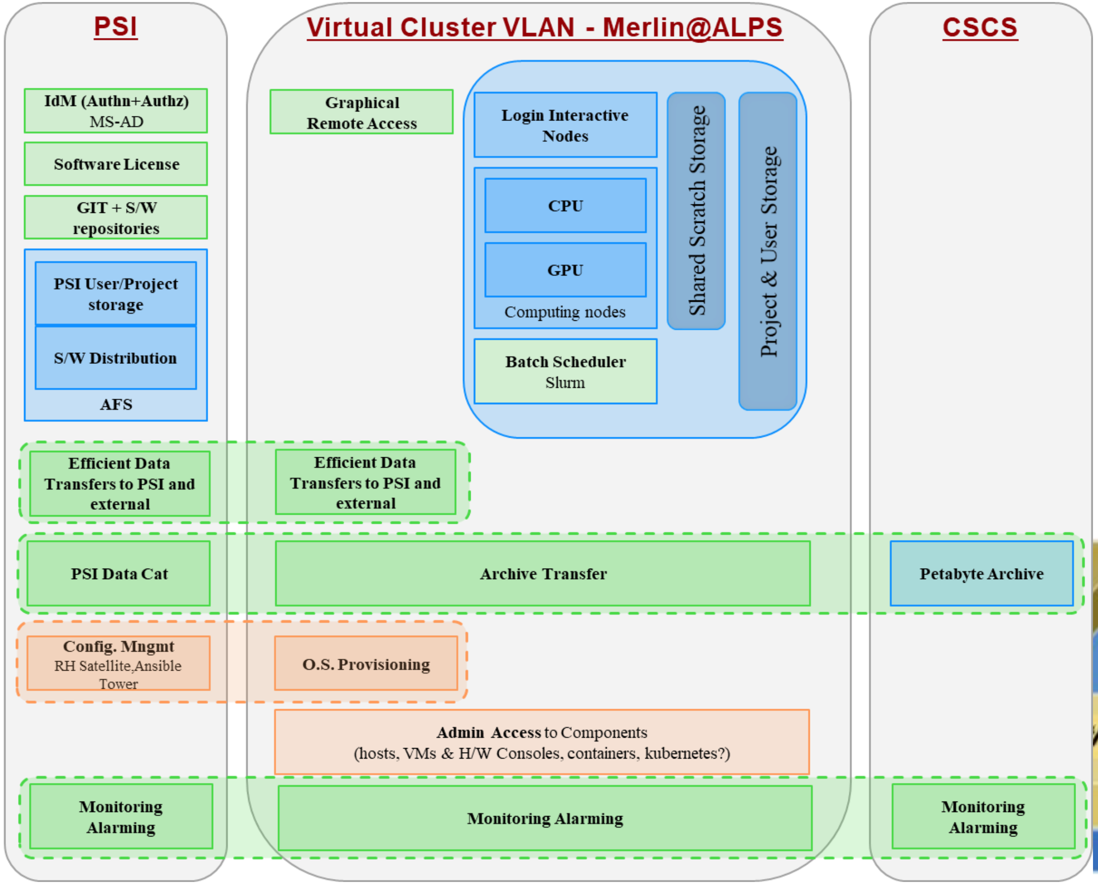
Challenging due to cost of integrating nonstandard platforms.

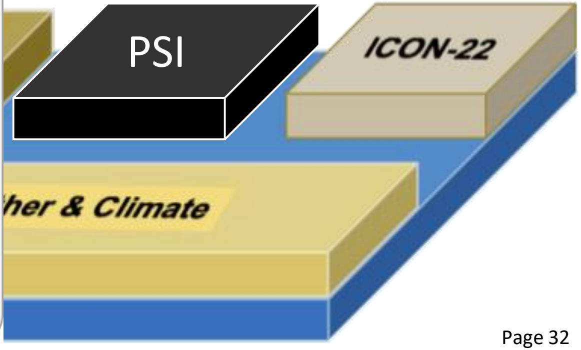
- Security, network segregation of computation and storage
  – Continue with *black box* service for PSI researchers.

Additional costs due to e.g. dedicated storage



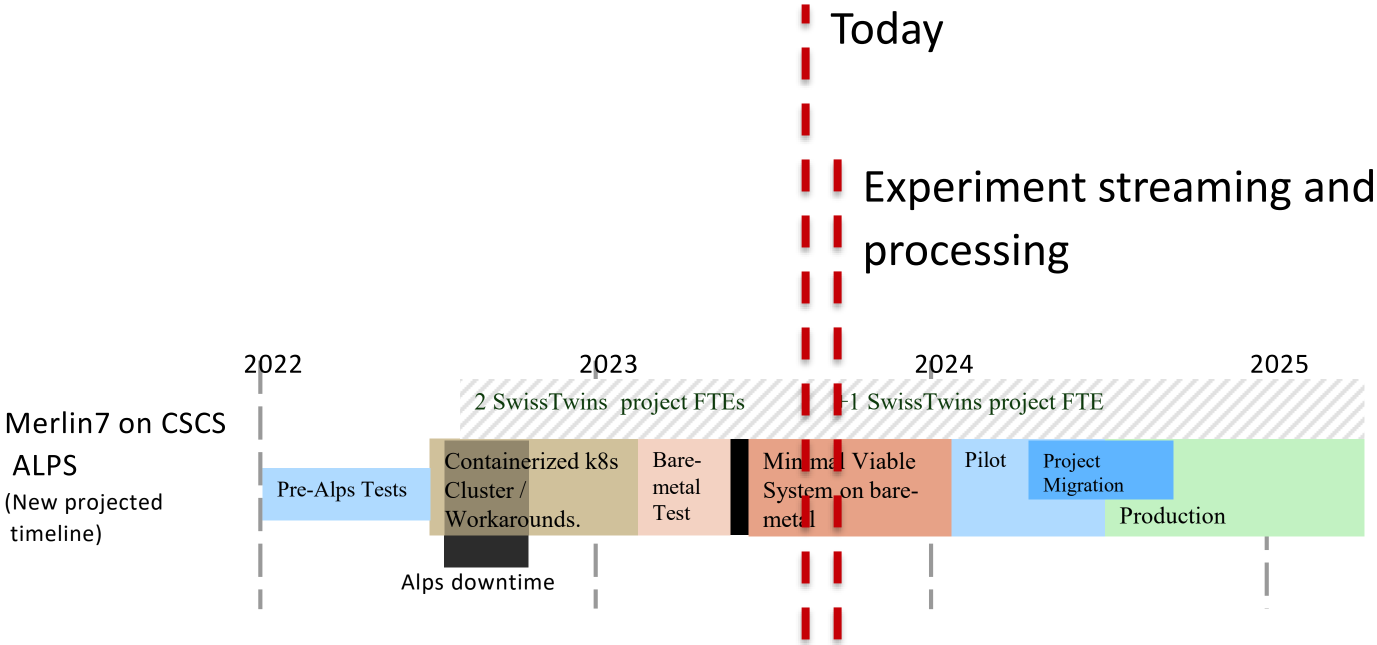
Architecture/work





PSI Services that System should get interfaced with

TranAlps project progress

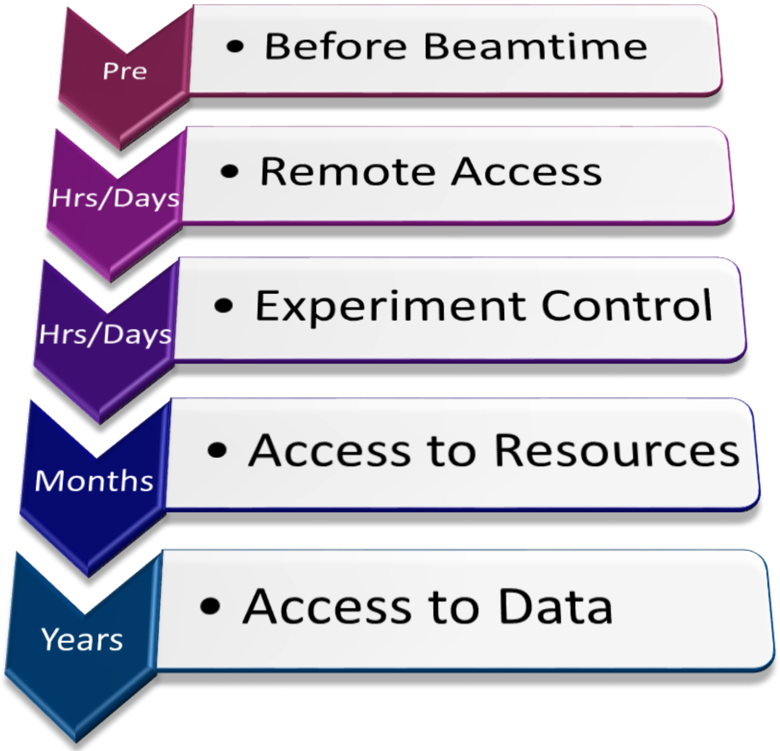


PAUL SCHERRER INSTITUT

Miscellaneous activities

User Accounts Post SLS 2.0

- User accounts
  - Discussions and planning underway to rationalise user accounts for:
    - Pre experiment (DUO)
    - Remote access
    - Experiment control
    - Access to resources
    - Access to data
  - Currently anything up to 4 accounts involved with high security risk for Data and access + user and staff frustration!



PAUL SCHERRER INSTITUT

Test Setups

- Thanks to the support from Kurt Bitterli, Daniel Grolimund, and Joerg Raabe, we would like to propose the following **testing infrastructure base, extension locations and Vitual:**
  - **Base location**: one corner in WBGA B18 (where optical table is; this area is outside the SLS building and has office network, and will not be affected by power outage and construction noise/dust/etc. due to SLS 2.0 construction.)
  - **Extension location**: MicroXAS beamline (MicroXAS will move to a new location, and it is currently scheduled as a Phase 2 beamline. Existing IT infrastructure and most beamline components are available during the dark time as long as electricity is available.)
  - **Virtual**: for testing and integration developments for beamlines moving to BEC

First hardware installation at the CaSIT test infrastructure / base location, courtesy of Kurt Bitterli