Leonardo Sala :: AWI :: Paul Scherrer Institute

Challenges for the next generation of experiments at large scale research facilities

SOS 26, Cocoa Beach, Florida - 2024-03-11

Nothing new under the sun… but first: what / where is PSI?

PSI and the ETH-Domain

Facilities at the PSI Campus

Current status

- Data acquisition is mostly run on **CPUs**
- Data analysis is mostly run on **CPUs, and files**
- Storage
  - mostly focused on **capacity** rather than speed (HDDs)
  - based on IBM Storage Scale
- Compression and reduction are mostly done by beamlines, and after data is written
- **Keep** most of data (RAW data == uncorrected data)

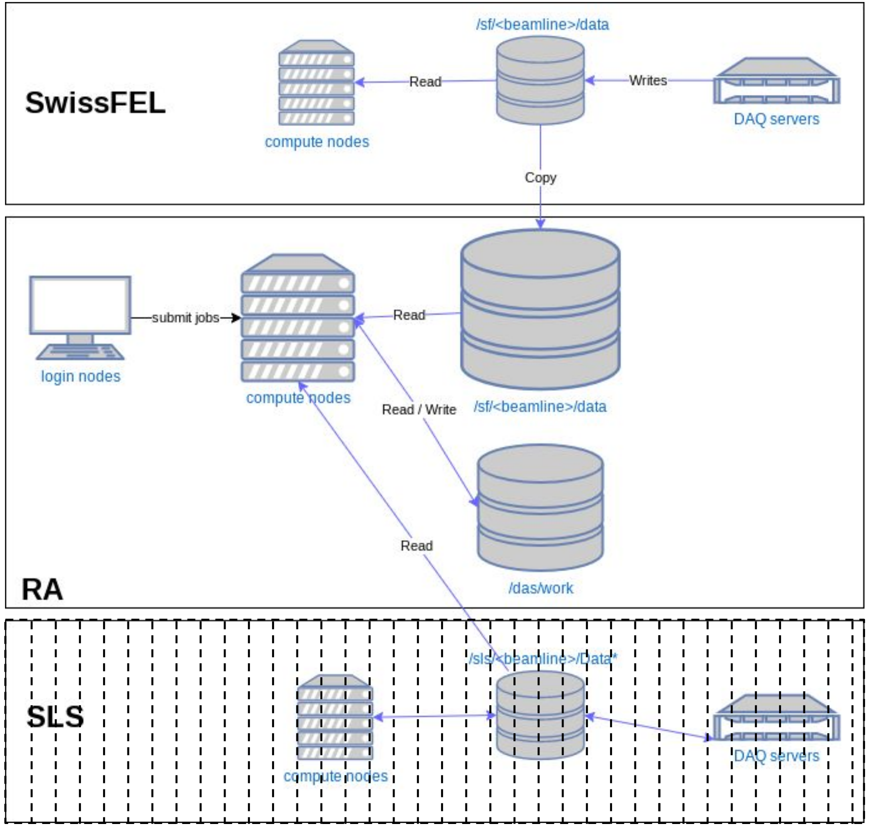
PAUL SCHERRER INSTITUT PSI

Overview – Photon Science Resources

The Photon Science (Ra) data analysis cluster has (11 PB, ~3600 cores, 16 GPUs

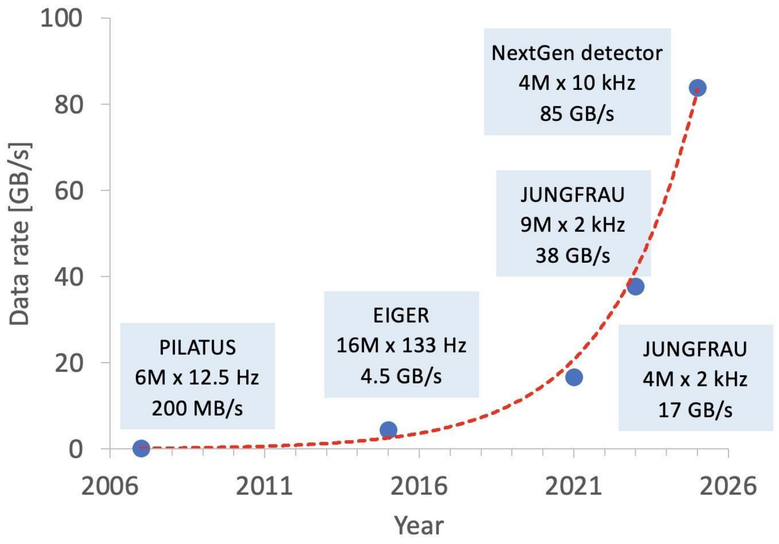
SwissFEL (and **SLS** before shutdown) have dedicated online compute nodes.

- **SwissFEL** only have a dedicated short-term buffer storage
- **SLS** had fully dedicated storage infrastructure

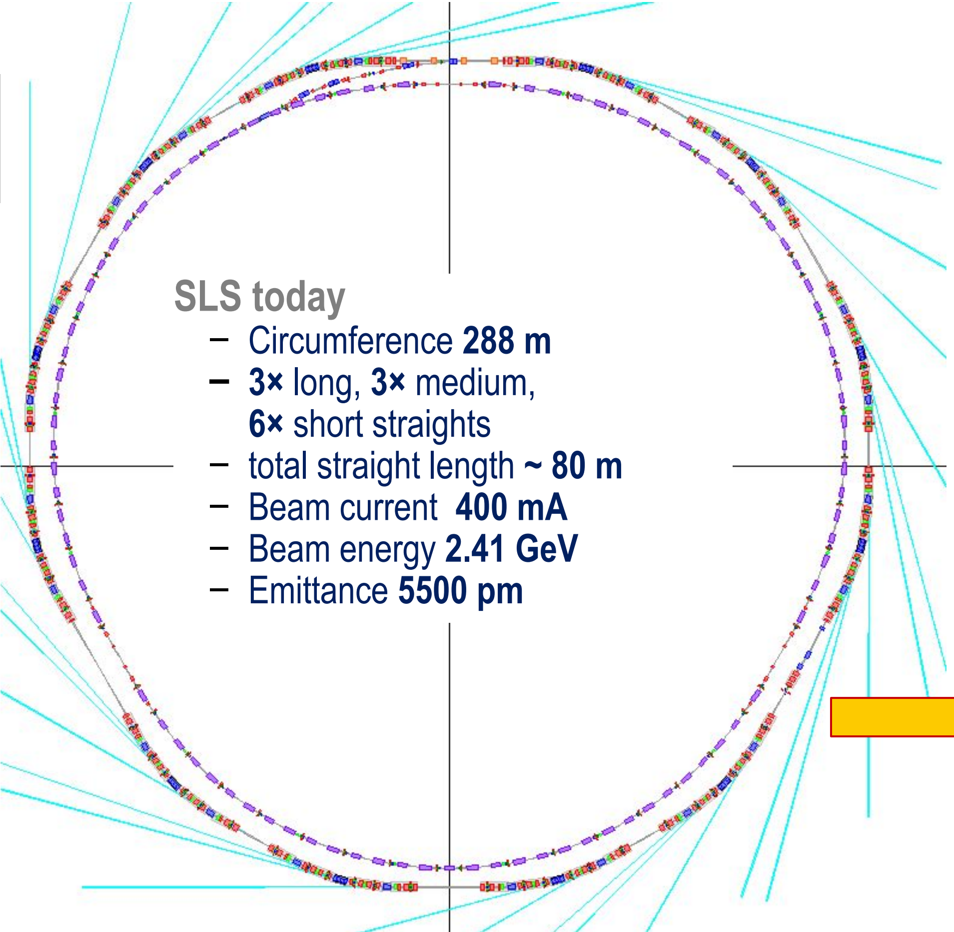


What are the future challenges?

- **SLS upgrade**
  - increased brilliance -> increased photons -> larger data
- **New detectors**
  - faster, smaller pixel size
  - does not scale with Moore
  - 85 GB/s
- **CryoEM**
  - long running experiments
  - few PB / year, 1000s GPUs

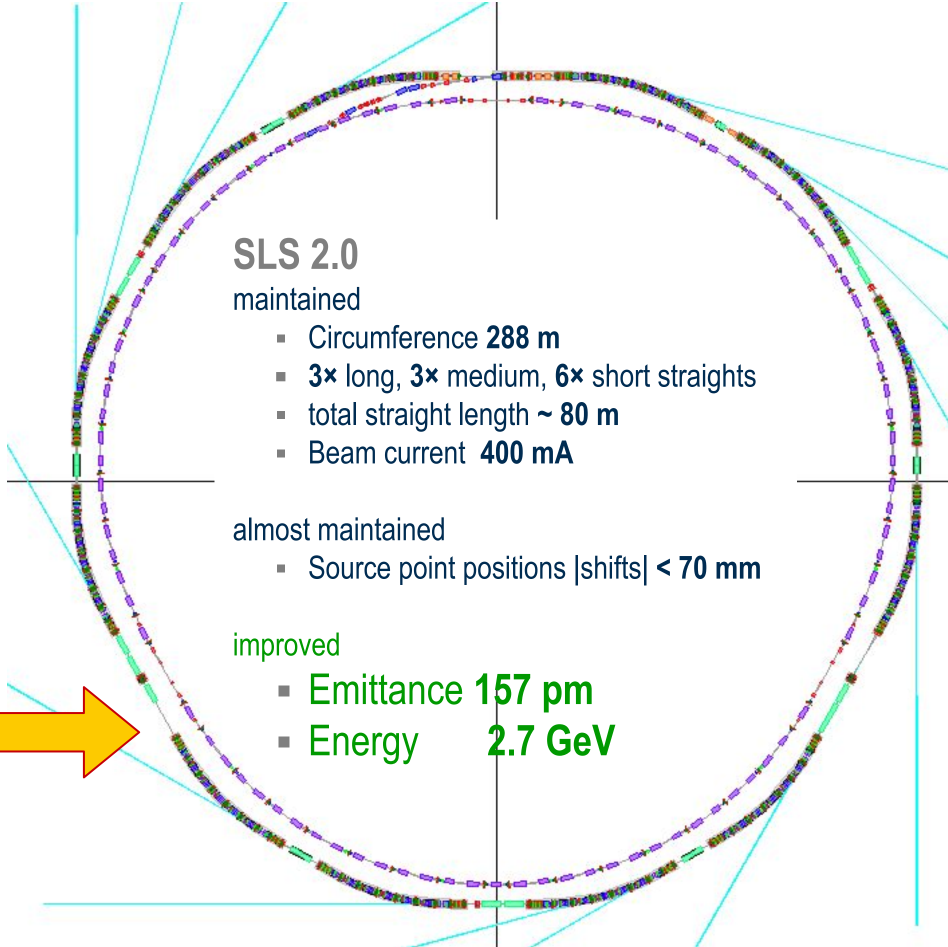


SLS → SLS 2.0



SLS today

- Circumference **288 m**
- **3×** long, **3×** medium, **6×** short straights
- total straight length **~ 80 m**
- Beam current **400 mA**
- Beam energy **2.41 GeV**
- Emittance **5500 pm**



SLS 2.0

maintained

- Circumference **288 m**
- **3×** long, **3×** medium, **6×** short straights
- total straight length **~ 80 m**
- Beam current **400 mA**

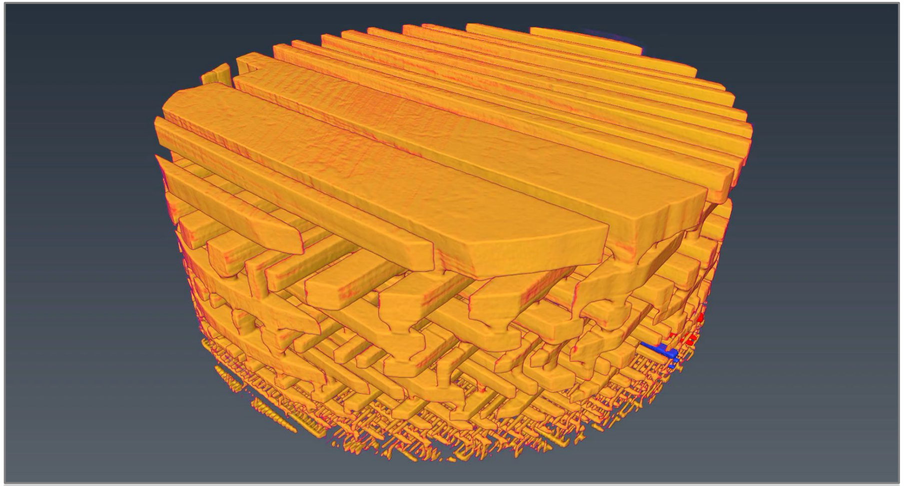
almost maintained

- Source point positions |shifts| **< 70 mm**

improved

- Emittance **157 pm**
- Energy **2.7 GeV**

SLS 2.0 Example: Ptychography



M. Holler *et al.*,
Nature **543**, 402 (2017)

DEVELOPMENT	RESOLUTION (nm)	VOLUME (μm^3)	TIME	computing power (a.u.)
State of the art	**14.6**	**15x15x8**	**22 h**	**1**
SLS 2.0	6.2	85x85x8	41 min	32
+ new undulator	4.6	150x150x8	13 min	100
+ broadband	2.6	475x475x8	1.3 min	1000
+ efficient optics	1.5	1500x1500x8	8 s	10000

PAUL SCHERRER INSTITUT

Lessons from the past

- **Control** of data flow is essential
  - pre-defined data structures
  - allows for automatic procedures, e.g. archiving
  - this improves also metadata treatment
- **early reduction** is mandatory, speeds up further data treatment
  - sometimes difficult to get accepted
  - old way of "saving all to tape" do not scale with costs
- **scheduling** can be hard
  - changes in e.g. sample preparation can lead to 10x more data
  - some experimental techniques are well defined and predictable
  - this can lead to waste of dedicated static resources

PAUL SCHERRER INSTITUT PSI

Way forward for SLS 2

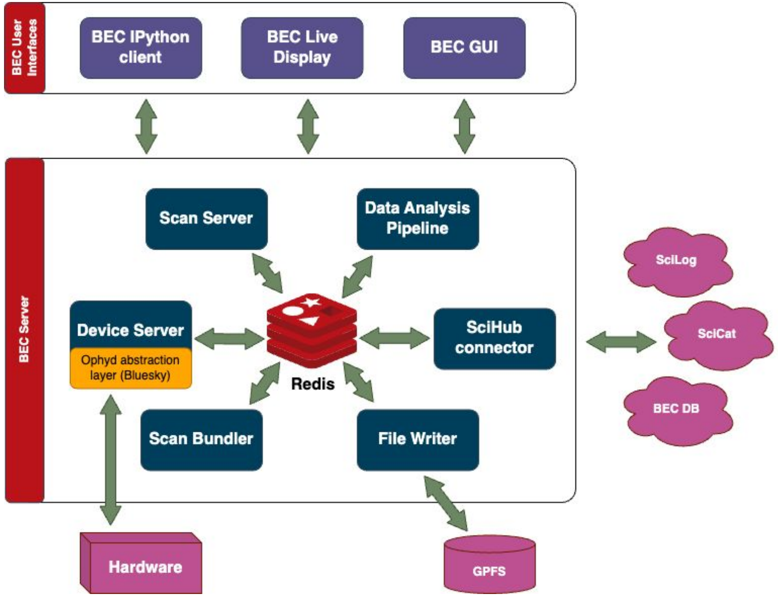
- Get **better control** of the data flow
  - Beamline Experiment Control (BEC) based on BlueSky / Ophyd
  - Strong integration with e-log (SciLog) and data catalog (SciCat)
- **Tiered acquisition**
  - low- to mid-range: on CPU, with centrally supported tools (std_daq)
  - high-range: FPGAs + GPUs
  - Fast NVMe storage to help reduction
- **Burst** scale-out for analysis
  - cater to "20%" challenging cases
  - Collaboration with CSCS (similar to Superfacility API)

BEC: Beamline Experimental Control

The BEC is a cross divisional initiative, **extending the bluesky** international initiative led from NSLSII, USA.

The BEC deployed on SLS2.0 beamlines will help enable:

- Interoperability
- Open research data (FAIR)
- Exploiting emerging technologies (AI/ML)



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

PAUL SCHERRER INSTITUT PSI

ALPS@CSCS

Ongoing project to migrate local HPC resources to CSCS Alps Infrastructure:
- Based on HPE Cray EX
- Build up by the Swiss National Supercomputing center (CSCS)

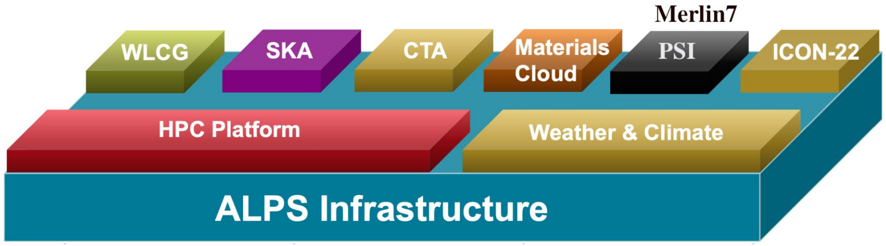
Quantity	What	Notes
1024	AMD Rome 7742 Nodes	128 cores, 256/512G RAM / node
144	Nvidia A100 GPU nodes	4 GPUs / node
X * 1000	Nvidia Grace Hopper modules	4 Grace Hopper / node
	HPE Cray Slingshot network	
100 PB	Lustre Storage (HDD)	
5 PB	Lustre Storage (SSD)	

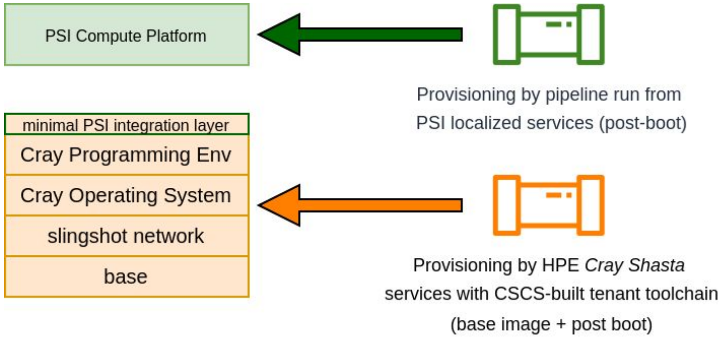
PAUL SCHERRER INSTITUT

ALPS@CSCS

PSI's next HPC simulation/modelling cluster is being implemented as a **vCluster** on top of CSCS Alps

vCluster: Versatile software-defined cluster
→ Convergence of HPC and Cloud technologies

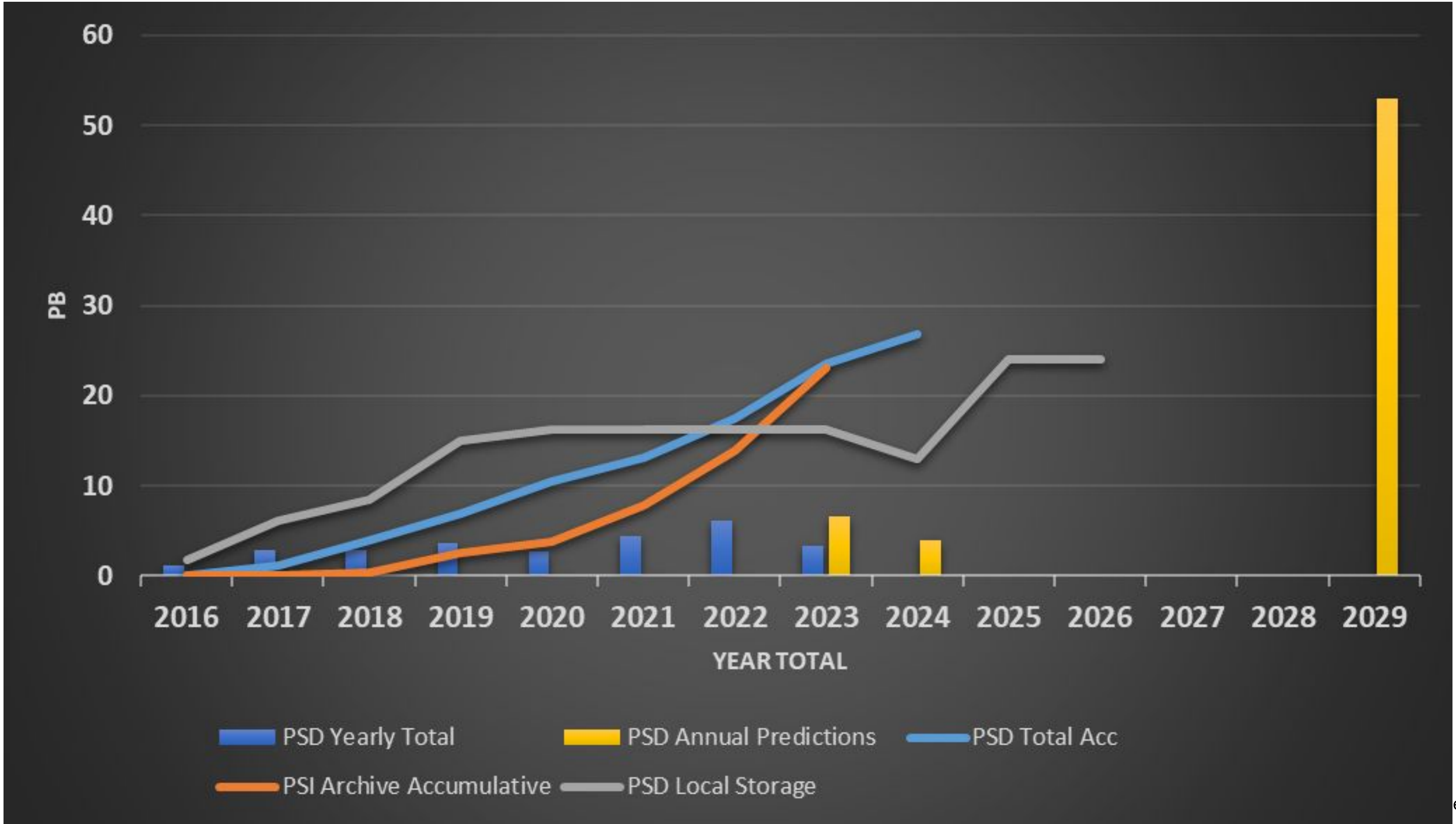




Challenges

- FPGAs are hard(-coded)
  - how many cases can we fit?
- HPC centers are historically not very flexible
  - very homogeneous
  - ARM means code recompilation
  - Need tools to transparently integrate, focus on advanced use cases
- Data volumes
  - Mandatory to reduce as early as possible
  - Even tape has a cost!
- Automation
  - high-throughput experiments
  - what is AI/ML role in?

Forecast: data volumes



My thanks go to

- People I stole slides:
  - Alun Ashton
  - Derek Feichtinger
  - Filip Leonarski
  - Klaus Wakonig
  - Hans Braun

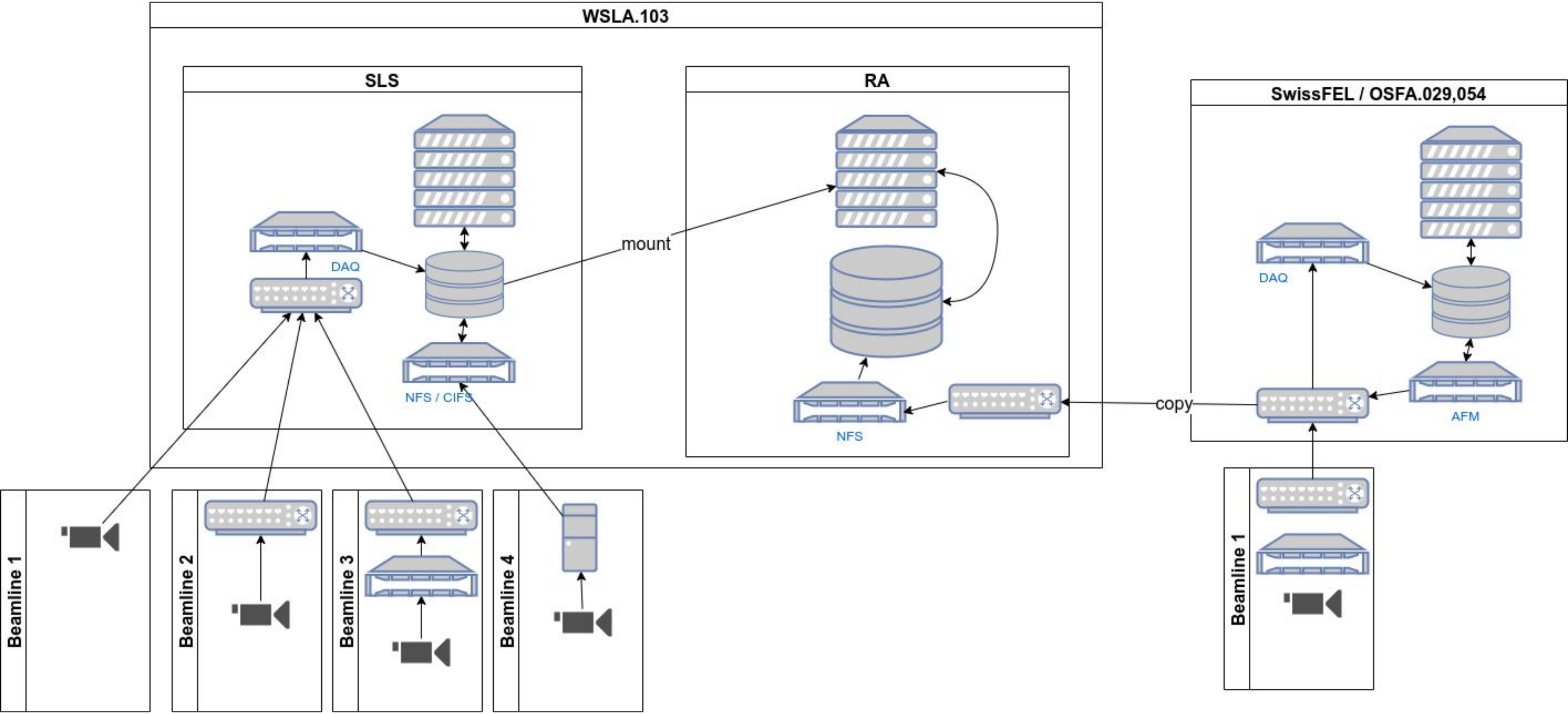
Backup slides

PAUL SCHERRER INSTITUT PSI

Bibliography / material

- Us: Science IT Infrastructure and Services (AWI)
- SLS 2 Upgrade project: https://www.psi.ch/en/sls2-0
- BEC: A BEAMLINE AND EXPERIMENT CONTROL SYSTEM FOR SLS 2.0
- JFJoch: Jungfraujoch: hardware-accelerated data-acquisition system for kilohertz pixel-array X-ray detectors
- High-resolution non-destructive three-dimensional imaging of integrated circuits
- ALPS: CSCS Alps Infrastructure

More detailed IT Architecture



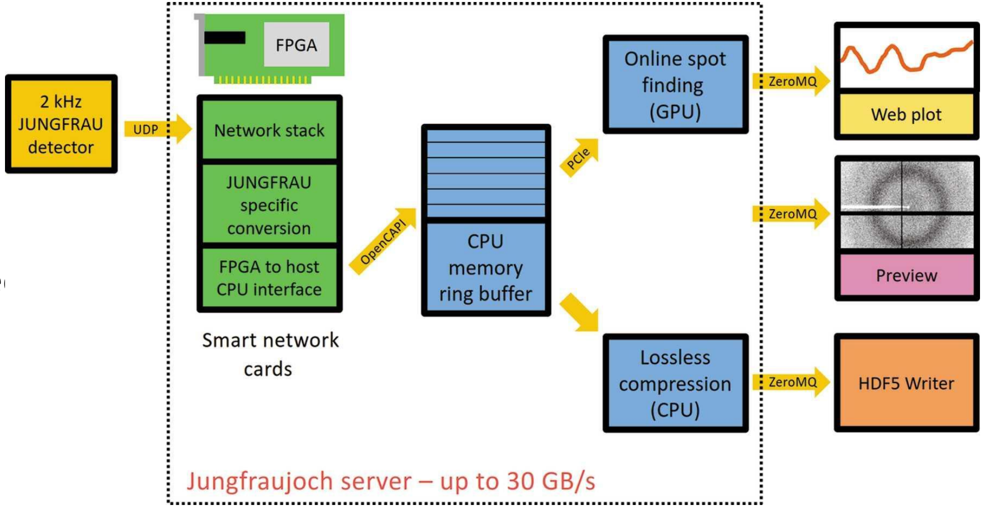
PAUL SCHERRER INSTITUT

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



JOURNAL OF SYNCHROTRON RADIATION

ISSN 1600-5775

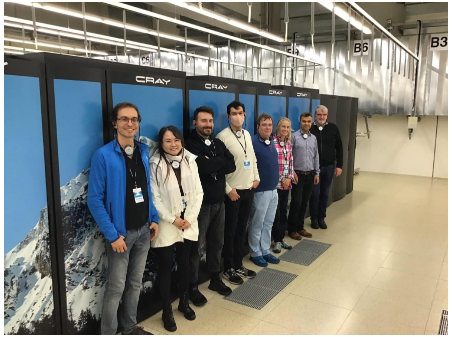
Jungfraujoch: hardware-accelerated data-acquisition system for kilohertz pixel-array X-ray detectors

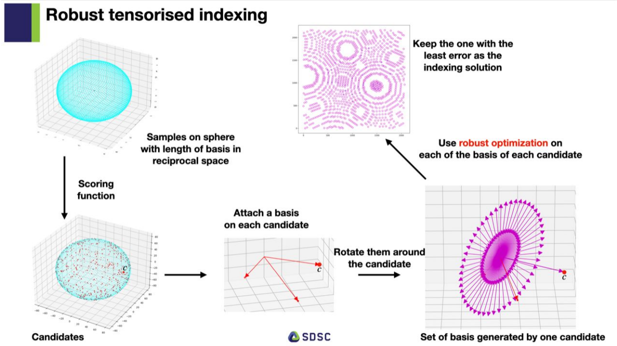
Filip Leonarski,[a]* Martin Brückner,[a] Carlos Lopez-Cuenca,[a] Aldo Mozzanica,[a] Hans-Christian Stadler,[b] Zdeněk Matěj,[c] Alexandre Castellane,[d] Bruno Mesnet,[d] Justyna Aleksandra Wojdyla,[a] Bernd Schmitt[a] and Meitian Wang[a]

Received 23 June 2022
Accepted 24 October 2022

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



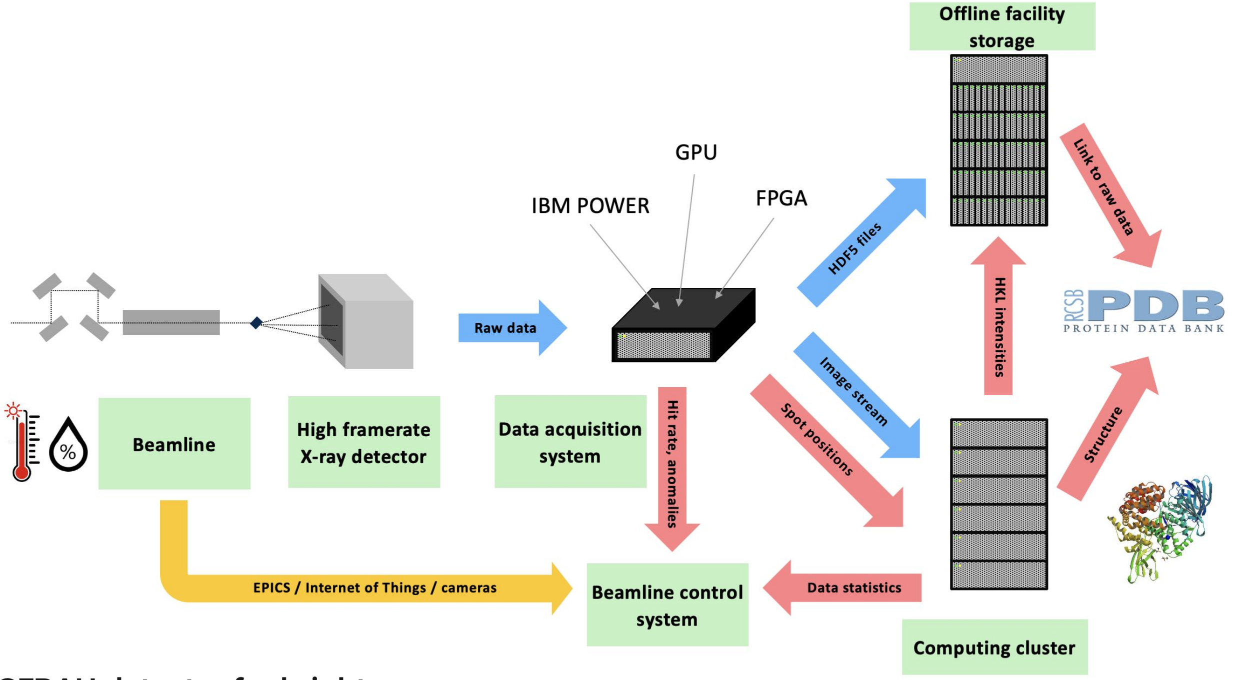


Acknowledgements:
A. Ashton, G. Assmann, L. Barba, B. Béjar,
P. Gasparotto, M. Janousch, T. Koka,
H. Mendoça, H.-C. Stadler

SDSC

CSCS
Centro Svizzero di Calcolo Scientifico
Swiss National Supercomputing Centre

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

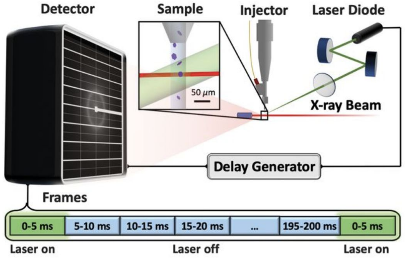
Filip Leonarski,[a] Aldo Mozzanica, Martin Brückner, Carlos Lopez-Cuenca, Sophie Redford, Leonardo Sala, Andrej Babic, Heinrich Billich,[b] Oliver Bunk, Bernd Schmitt, and Meitian Wang

Time-resolved serial synchrotron crystallography at SLS 2.0

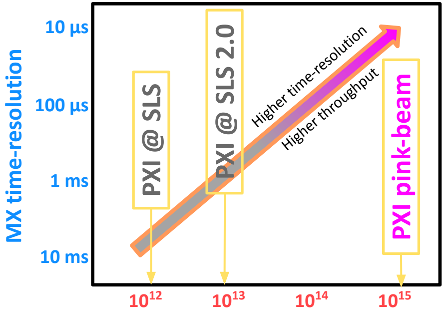
- **Serial crystallography** solves protein structures with diffraction images from thousands of crystals
- **PXI-VESPA**: A Versatile End-station for Scattering Pink-beam Applications

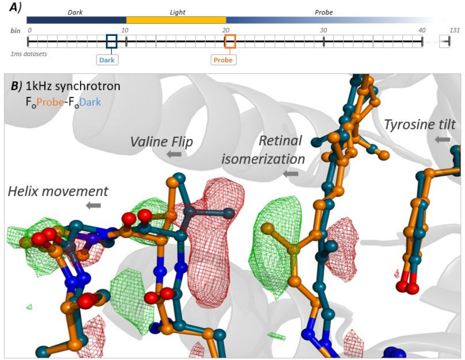
- **Pump-probe** with nanosecond laser
- Thanks to increased brilliance of SLS 2.0 and multilayer optics **microsecond** time resolution is possible

- **The most data intensive technique** for MX@SLS 2.0
  - Needs surplus of images (millions)
  - Long, continuous measurements
  - kHz frame rates
- **Online reduction** possible – only fraction of images useful



T. Weinert et al., *Science* (2019)
https://doi.org/10.1126/science.aaw8634





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

Technical ALPS Challanges

- Multi tenancy - control of the bare-metal compute nodes by PSI admins
  - PSI is the first CSCS customer institution to run a tenant managed vCluster
  - PSI vCluster is in a VLAN of the PSI DNS space. Develop security policies.
  - CSCS developing Manta to provide tenant admins possibility to manage the vCluster autonomously (WIP)
  - Integration of low level node information in PSI alarming / monitoring
- Architecture and workflows
  - Compute nodes have no local fast storage, so all I/O needs to be handled by the shared FS
  - Cluster used by ~90 research groups of all PSI divisions - must support multitude of applications and workflows, from HEP style high throughput to MPI, GPU and single core.
  - Grace Hopper ARM-based architecture will require adaptation / porting of many user applications (many commercial or non-OSS) over this year.
- Future
  - Achieve vCluster resource elasticity to support PSI facilities use cases