LUNARC Townhall meeting

2021-02-17

MAX IV and LUNARC

- Research infrastructure
- Large data volumes are having high resource requirements
- Concern: Resource access can often be related to losing a momentum in data analysis
- Data analysis tools need to have user-friendly and interactive interface easy to understand for non-expert/novice users
- Example applications:
  - x-ray imaging
  - macromolecular crystallography

Photon and Neutron (PaN) facilities

EUROPEAN SPALLATION SOURCE

MAX IV LABORATORY

Swedish national
+ 2 Danish beamlines
+ 1 Finish-Estonian beamline

situated in LUND

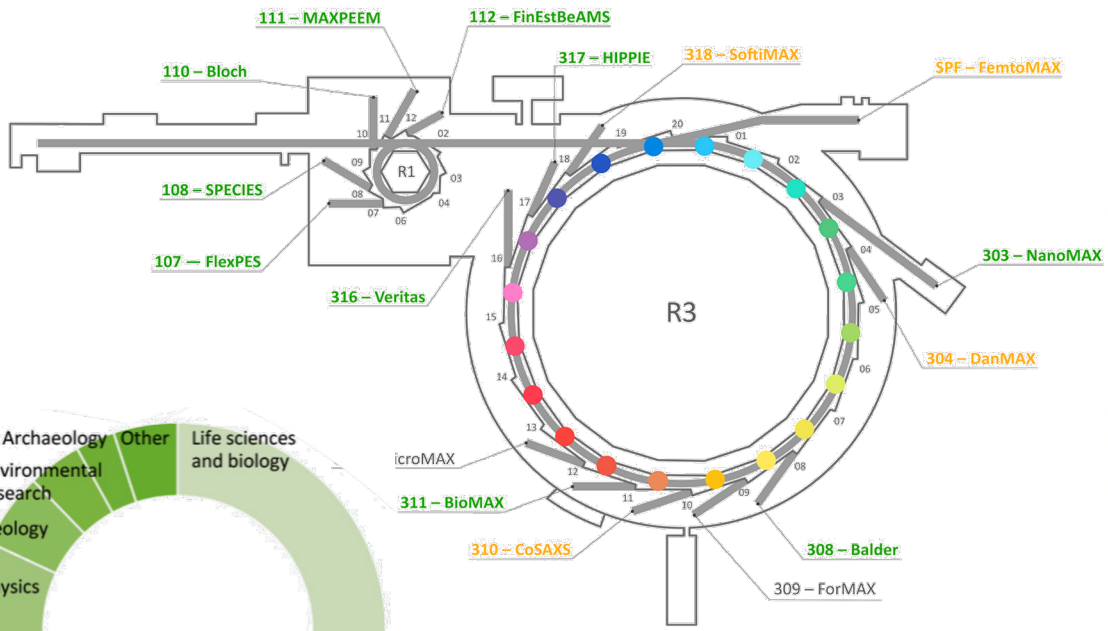
European (ESFRI)
Datacentre in Copenhagen

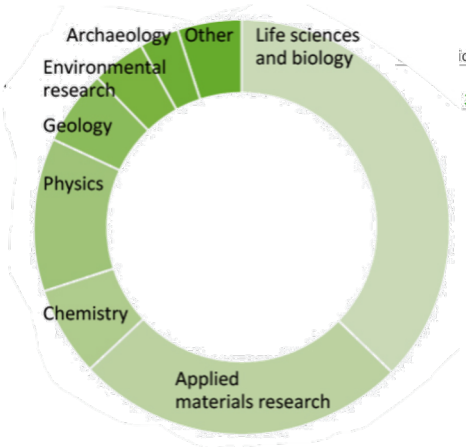
ExPaNDS
European Open Science Cloud Photon and Neutron Data Services

panosc
photon and neutron open science cloud

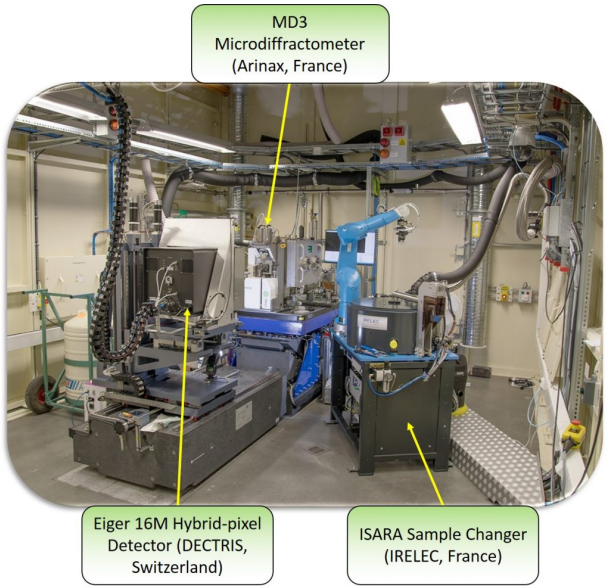
MAX IV

Beamlines – variety of sciences





BioMAX endstation



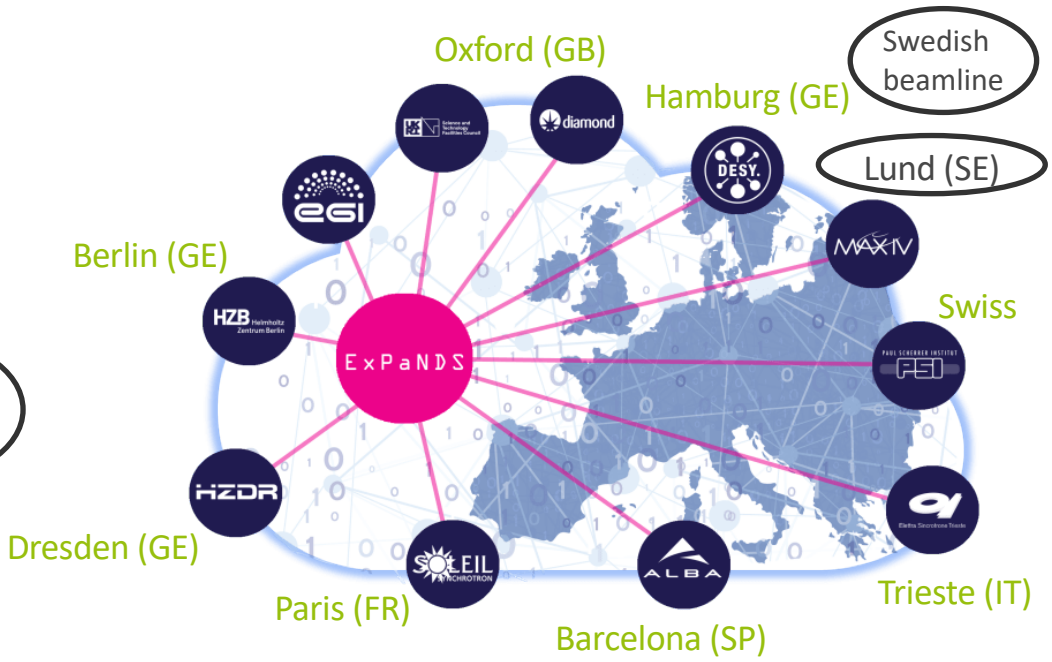
electron -> photon -> data -> scientific result

all beamlines running simultaneously

PaN facilities around Europe

PaNOSC: ESFRI project

LEAPS-initiative.eu: 16 members
LENS-initiative.org: 9 members



panosc
photon and neutron
open science cloud

ExPaNDS: EOSC project

ExPaNDS
European Open Science Cloud Photon
and Neutron Data Services

How many scientific visitors do you receive per year?

Facility	Current situation (2018/2019)	Forecast 2023
DESY	3667 (2019)	> 3667 (we'll have more beamlines by then)
PSI	-	-
Diamond	~6000	Unknown
ISIS	In 2018 ISIS had 2512 scientific visitors.	>2500
SOLEIL	In 2019: 4986 users visits	Forecast for 2023: 4000 on site and 1500 in remote or mail in
ESRF	2018 (the last year of operation): 6548 user visits. Some users are coming several times. The estimation of unique individuals visiting ESRF in 2018 is about 3 000.	~ 10 000
ALBA	2200	3100
HZB	ca. 3000 user visits at BESSY II, these are about 1600 individuals (most people come twice a year since we work in semesters)	maybe a few more, since beamtimes become shorter due to better infrastructure, detectors, sample environment, IT... we also expect "remote users", but it is not clear, how to count them yet. LEAPS etc. are working on definitions and a common metrics
HZDR	100	120
MAXIV	700 (312 experiments)	1500

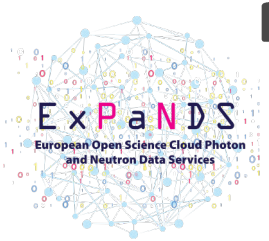
Sophie Servan: PaNOSC/ExPaNDS Technical-Workshop-Survey (2020)
https://github.com/ExPaNDS-eu/ExPaNDS/blob/master/WP4/20201009-Technical-Workshop-Survey.pdf

Note: only subsets of users & experiments have high computing needs

How much data is produced per year (in TB, PB,...)?

Facility	Current situation (2018/2019)	Forecast 2023
DESY	around 2 PB / year (01.2019-12.2019)	
PSI	3PB	
Diamond	2019: 5.35 PB	
ISIS	About ~15 TB of raw data from April 2017 to April 2018.	
SOLEIL	700 TB in 2019	
ESRF	2018: 9 PB	2023: > 50 PB. The error bar on this number is large. It could be much more if our IT infrastructure is capable of dealing with it.
ALBA	We expect a big increase in the quantity of data produced in following years: 2020 < 200TB/year, 2021-950 TB/year, 2022 - 1.7PB/year, 2023-4PB/year	
HZB	The HZB Data Policy is not yet fully implemented and thus we do not have the full picture yet. The planning for the storage capacity were based on a rough estimate of 2 PB data production per year for all HZB facilities.	
HZDR	about 1 PB (2019)	
MAXIV	1 PB	

Sophie Servan: PaNOSC/ExPaNDS Technical-Workshop-Survey (2020)
https://github.com/ExPaNDS-eu/ExPaNDS/blob/master/WP4/20201009-Technical-Workshop-Survey.pdf

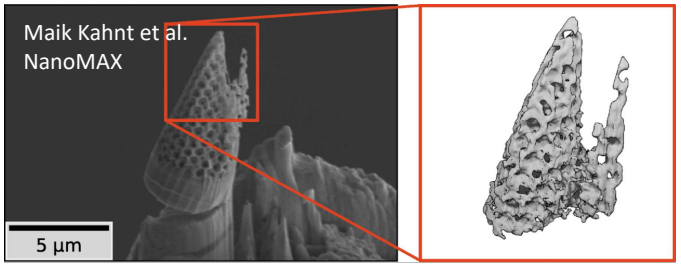


MAXIV

Data analyses services

ExPaNDS applications cases – datasets

Ptychographic X-ray computed tomography



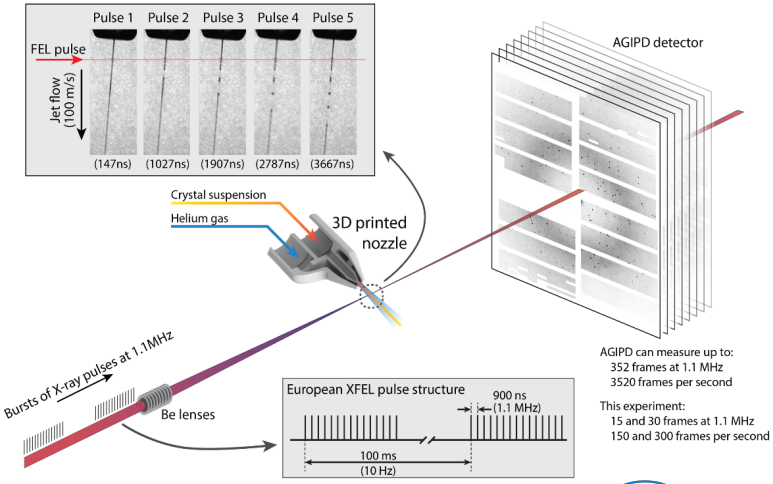
PtyPy

https://doi.org/10.5281/zenodo.3702582
https://doi.org/10.1107/S160057672001211X

MAX IV

1.5 TB / 16 GB

Serial crystallography



http://cxidb.org/id-32.html
https://doi.org/10.1038/sdata.2016.21

10 TB / 79 GB

Cheetah & CrystFEL

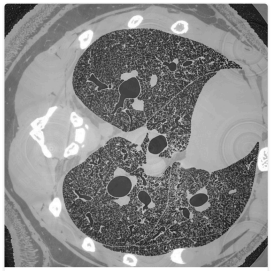
Micrometer-resolution X-ray tomographic imaging of a complete intact post mortem juvenile rat lung

Elena Borisova,Goran Lovric,Arttu Mietinen,Luca Fardin,Sam Bayat,Anders Larsson,Marco Stampanoni,Johannes C. Schittny,Christian M. Schlepütz; PSI (2020)

PAUL SCHERRER INSTITUT PSI

Abstract

In the associate article to these data sets, we present an X-ray tomographic imaging method that is well suited for pulmonary disease studies in animal models, to resolve the full pathway from gas intake to gas exchange. Current state-of-the-art synchrotron-based tomographic phase-contrast imaging methods allow for three-dimensional microscopic imaging data to be acquired non-destructively in scan times of the order of seconds with good soft tissue contrast. However, when studying multi-scale hierarchically structured objects, such as the mammalian lung, the overall sample size typically exceeds the field of view illuminated by the X-rays in a single scan, and the necessity for achieving a high spatial resolution conflicts with the need to image the whole sample. Several image-stitching and calibration techniques to achieve extended high-resolution fields of view have been reported, but those approaches tend to fail when imaging non-stable samples, thus precluding tomographic measurements of large biological samples, which are prone to degradation and motion during extended scan times. In this work, we demonstrate a full-volume three-dimensional reconstruction of an intact rat lung under immediate post mortem conditions and at an isotropic voxel size of (2.75 µm)^3. We present the methodology for collecting multiple local tomographies with 360 degree extended field of view scans followed by locally non-rigid volumetric stitching. Applied to the lung, it allows to resolve the entire pulmonary structure from the trachea down to the parenchyma in a single dataset. For related publication see https://link.springer.com /article/10.1007/s00418-020-01868-8



https://doi.psi.ch/detail/10.16907/7eb141d3-11f1-47a6-9d0e-76f8832ed1b2

Full field tomography

~ 1.2 TB*

- Kinetic SAXS, 2D Scanning SAXS
- Neutron-imaging/tomo
- Neutron-reflectometry
- CryoEM
- Terahertz-spectroscopy
- PXRD
- 4DSTEM

MAX IV

Examples from other facility 1

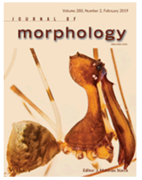
Tomography computing at ALS

ALS microCT beamline

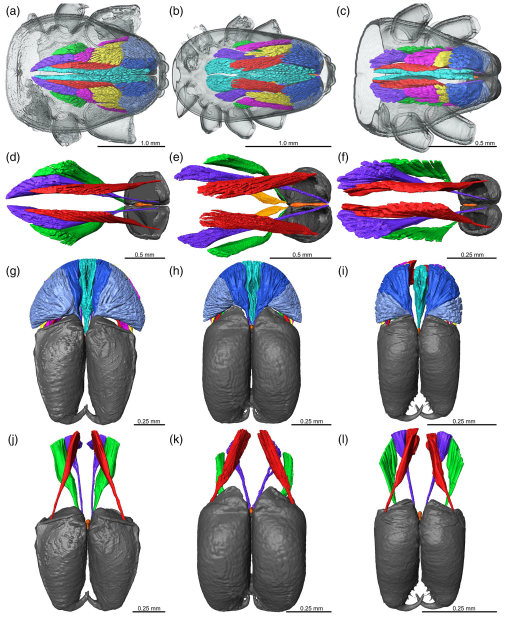
- data can be handled on a strong desktop-gpu workstation
- leading collaboration with **NERSC**

Comparative morphology of cheliceral muscles using high-resolution X-ray microcomputed-tomography in palpimanoid spiders

JANUARY 17, 2019



Spiders are important predators in terrestrial ecosystems yet we know very little about their principal feeding structures—the chelicerae—an extremely important aspect of spider biology. Here, using micro-Computed-Tomography scanning techniques, researchers perform a comparative study to examine cheliceral muscle morphology in six different spider specimens. Read more ›



Journal of Morphology, Volume: 280, Issue: 2, p. 232-243,
First published: 17 January 2019, DOI: (10.1002/jmor.20939)

Examples from other facility 2

Tomography computing at APS

globus

Argonne-led team wins technology challenge at SC19

December 17, 2019

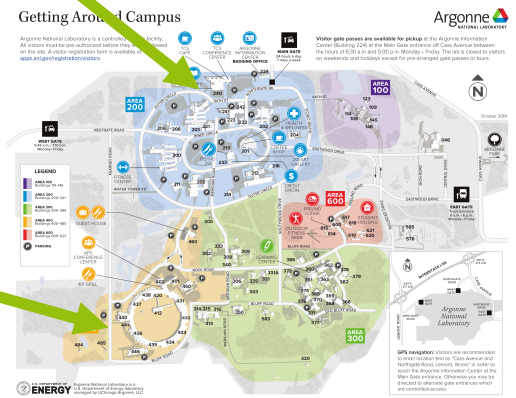
Argonne NATIONAL LABORATORY

… the team demonstrated real-time analysis of light source data from Argonne's Advanced Photon Source (APS) streamed at close to 100 gigabits per second to the Argonne Leadership Computing Facility (ALCF), …

… The team analyzed a portion of the data using 16,384 cores on Argonne's Theta supercomputer. …

... The supercomputer processed the data from one of the beamlines in real time, storing the remainder for later analysis. This real-time processing step involves using the lab's Cooley visualization cluster for iterative reconstructions of a 3D volume from 2D images obtained at a microtomography imaging beamline.

https://www.anl.gov/article/argonneled-team-wins-technology-challenge-at-sc19



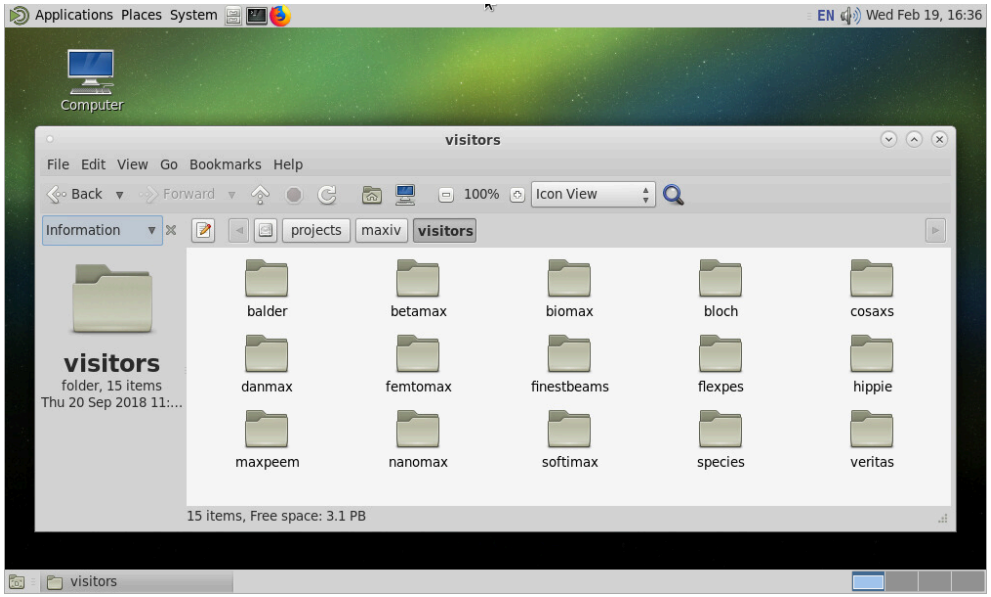
MAX IV

LUNARC for MAX IV staff and users

- MAX IV staff
  - accelerator physics group
  - radiation safety group
  - . . . *maybe somebody listening here* . . .
- MAX IV and LUNARC collaboration
  - Tomograms project together with LINXS
  - Infrastructure IT collaboration & service
    - MAX IV data acccess
    - software
    - support for SNIC infrastructure
    - knowledge and expertize sharing
- MAX IV users
  - Interactive visualisation for MAX IV users (Imaging)
  - PReSTO sw stack for macromollecular crystallography

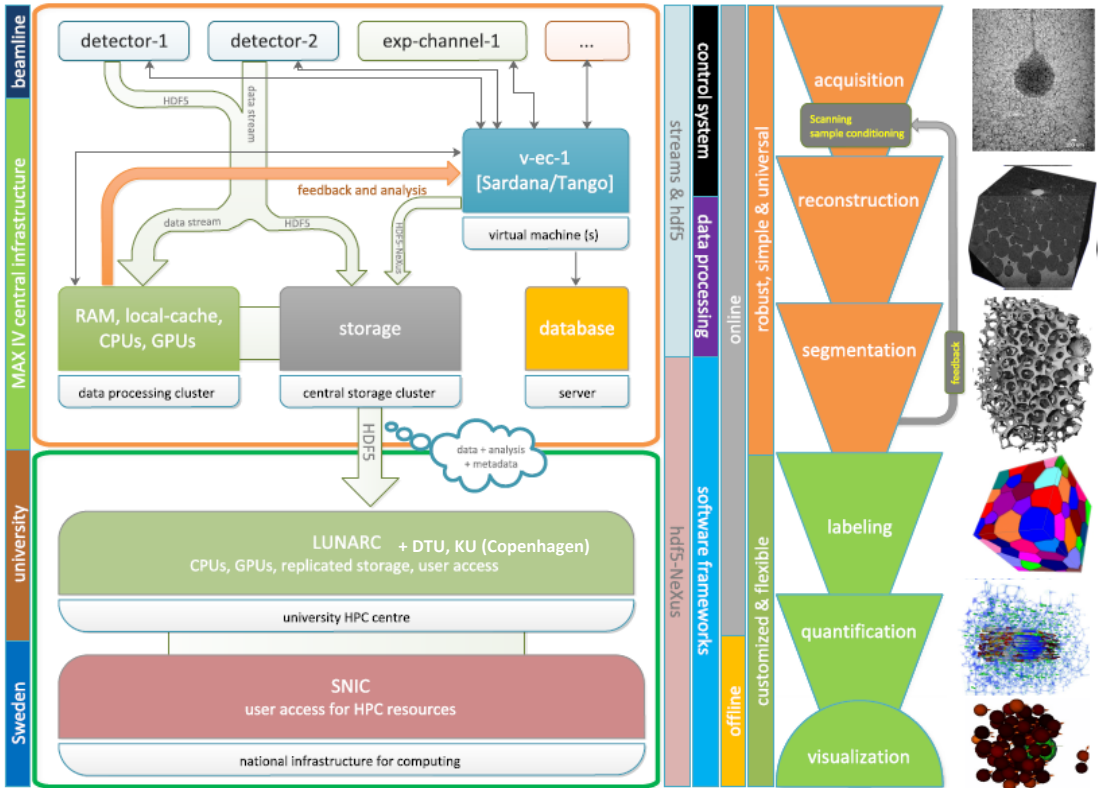
LUNARC and MAX IV data

- direct mount of MAX IV data
  - NFS (LUNARC was Lustre) ✓
  - *future*: GPFS possible (LUNARC is gpfs now) ?
- user ID mapping
  - without overlap (minor issues) ✓
  - *future*: matching proposals from user-office and data access ?



Data analysis: 3 stages

Tomography & imaging



Stage 1

robust, memory bandwidth limited calculations, heavy FFTs, heavy I/O, real-time computing during experiment, help of synchrotron staff

result: large reconstructed data

Stage 2

input: large data

imaging expertise needed

result: reduced data

Stage 3

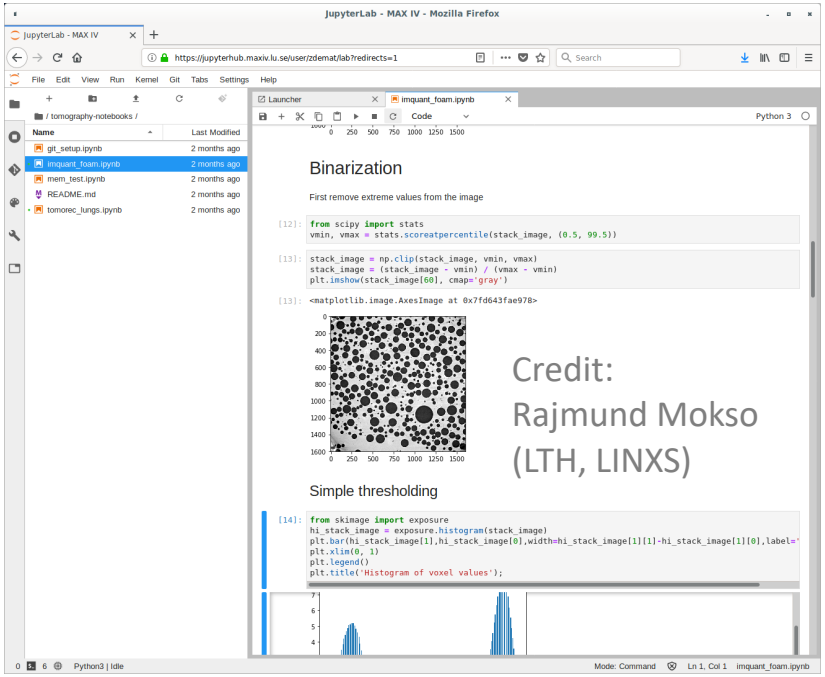
complex, imaging & scientific domain expertise required, specialised sw, collaborative environment

result: publication ready data

Stage 2: segmentation

- input: large data
- imaging expertise needed
  - imaging software aligned either to desktops but has to deal with large data
  - Python: scikit-image, tensorflow, pytorch, . . .
  - remote desktop: fiji & ImageJ, Matlab-QiP, Avizo, Paraview, Tomviz, medical imaging sw,. . . .
- result: reduced data

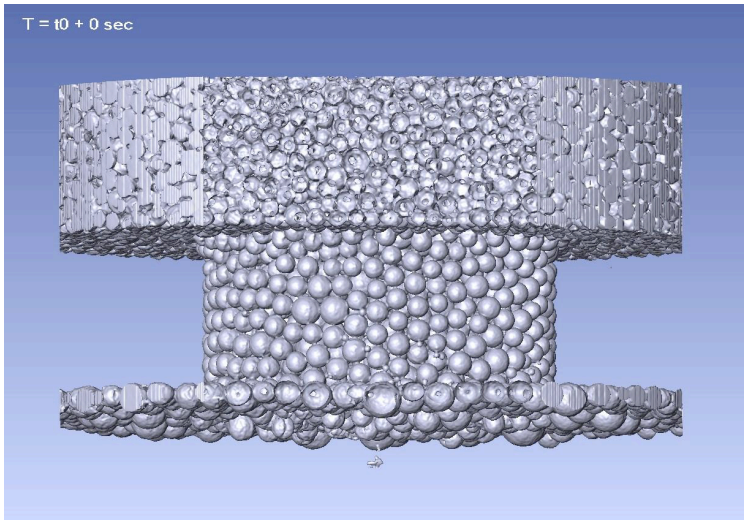
Example:
input: > 2 TB reconstructed data
2 GB data-slab selected
notebook runtime ~ 1.5 hour
99% of time single core used
memory required: 200 GB



some imaging sw will always run ineffectively
universality preferred before performance

Stage 3
labelling, quantification, visualisation

- most complex
- imaging & scientific domain expertise and collaboration
- input: segmented data
- high performance 3D desktop (or innovative) visualization environment needed



R. Mokso et al., https://doi.org/10.1107/S1600577517013522

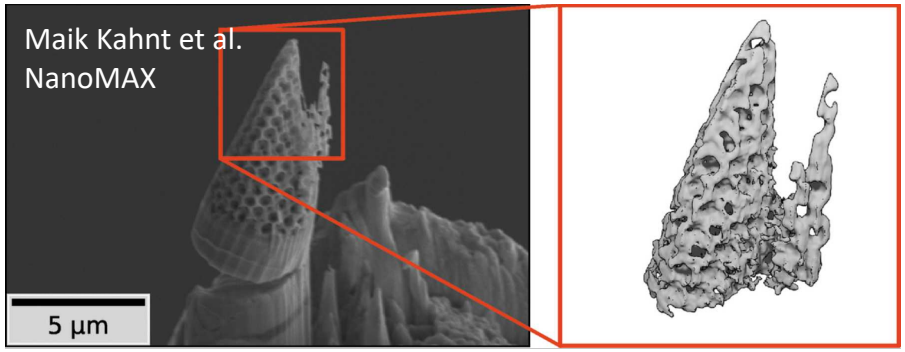
Johan Hektor – QiM application expert at LUNARC (2019), former postdoc experience at DESY

Ptychographic X-ray tomography imaging

The community in a high compute need

- Ptycho-tomo data
  - large volumes: 2D image x 2D scan x 1D rotation (x 1D time) … 5(-6) dims
  - **can be compressed** (4-10x)
- **better compute vs. storage ratio**
- other facilities:
  - ESRF: 16x Power9 multi-GPUs + -> 150 GPUs
  - Argonne: utilizing ThetaGPU – 24x NVIDIA DGX A100 nodes

MAX IV datasets available and ramping up



https://doi.org/10.5281/zenodo.3702582
https://doi.org/10.1107/S160057672001211X

1.5 TB / 16 GB

contrary to full-field tomo cannot easily parallelize over slices

Ptychographic phase retrieval
compute demanding, heavy FFTs, MPI & GPUs

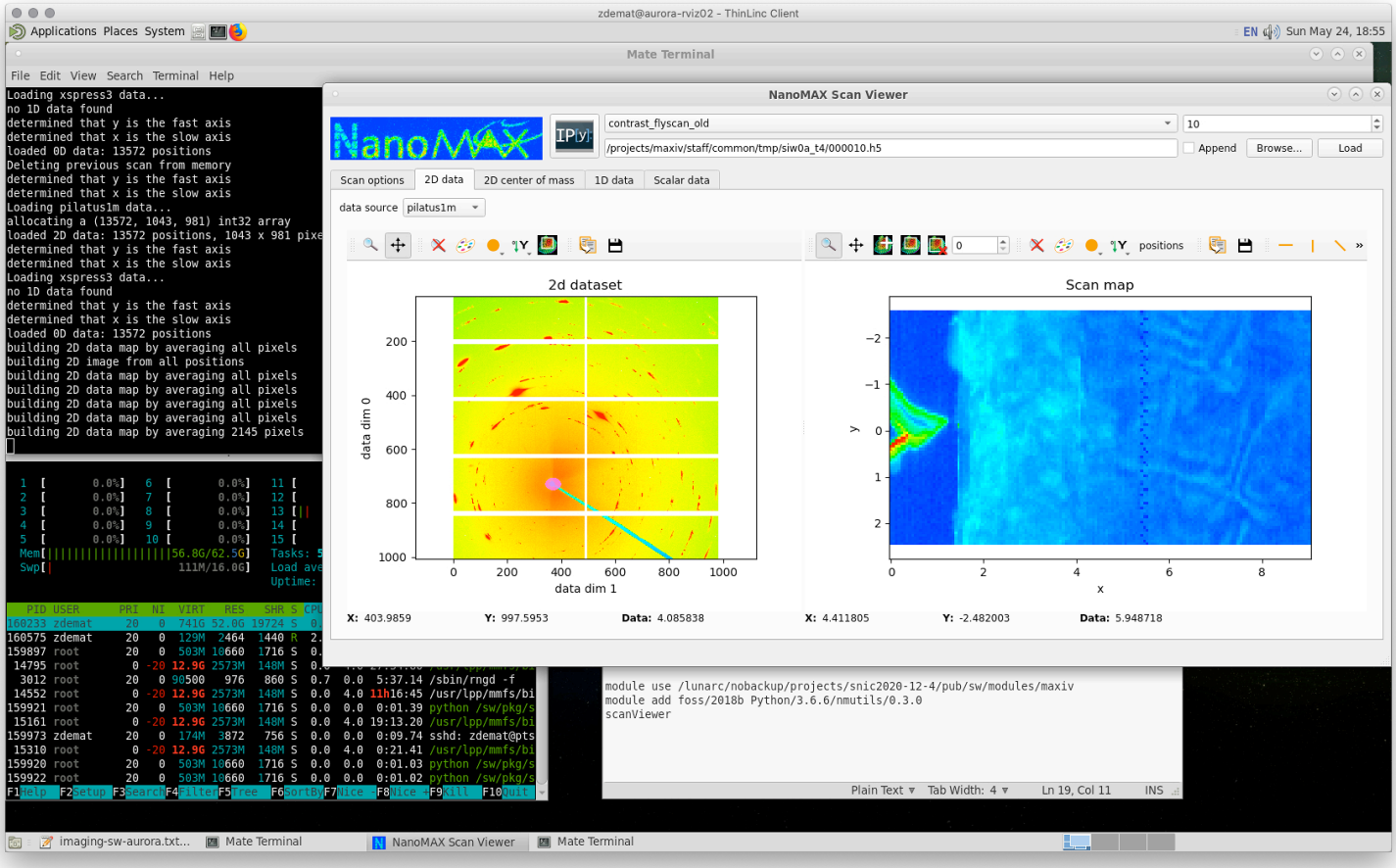
image analysis and realignment
complex (ML)

Tomographic reconstruction
compute demanding, FFTs, MPI & GPUs

iterative

NanoMAX scan viewer

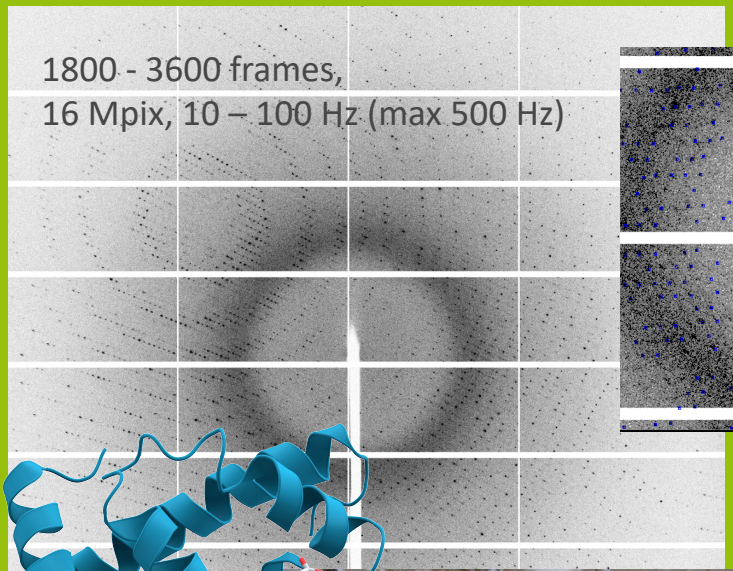
https://github.com/maxiv-science/nanomax-analysis-utils

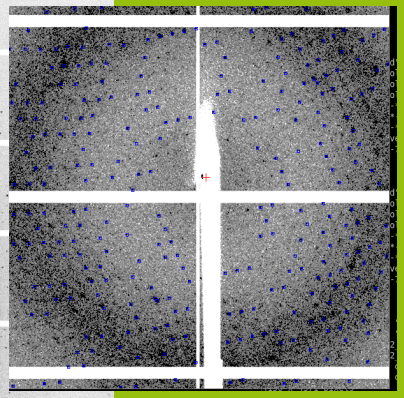


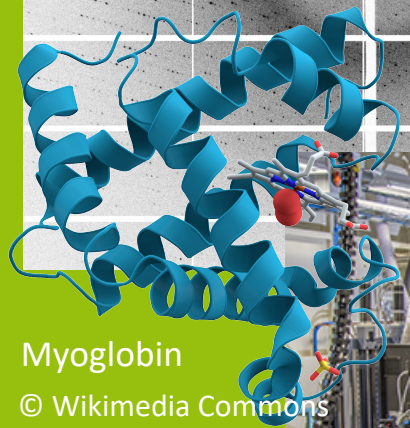
Data and science case: Gudrun Lotze

- started (5y ago) at laptop with Matlab
- Nowadays at MAX IV up to 360 GB RAM
- uses silx toolkit
- works at LVIS nodes

Macromollecular X-ray crystallography







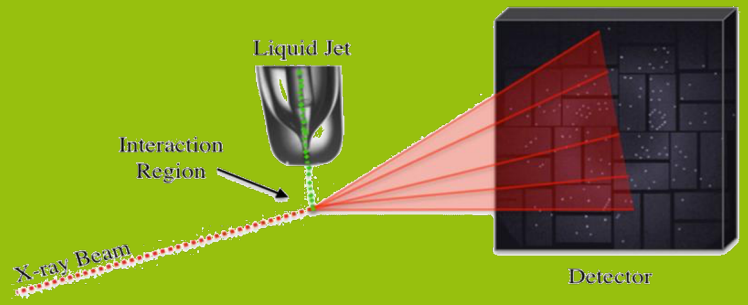
Myoglobin

© Wikimedia Commons

BioMAX endstation

Serial crystallography

MicroMAX & BioMAX

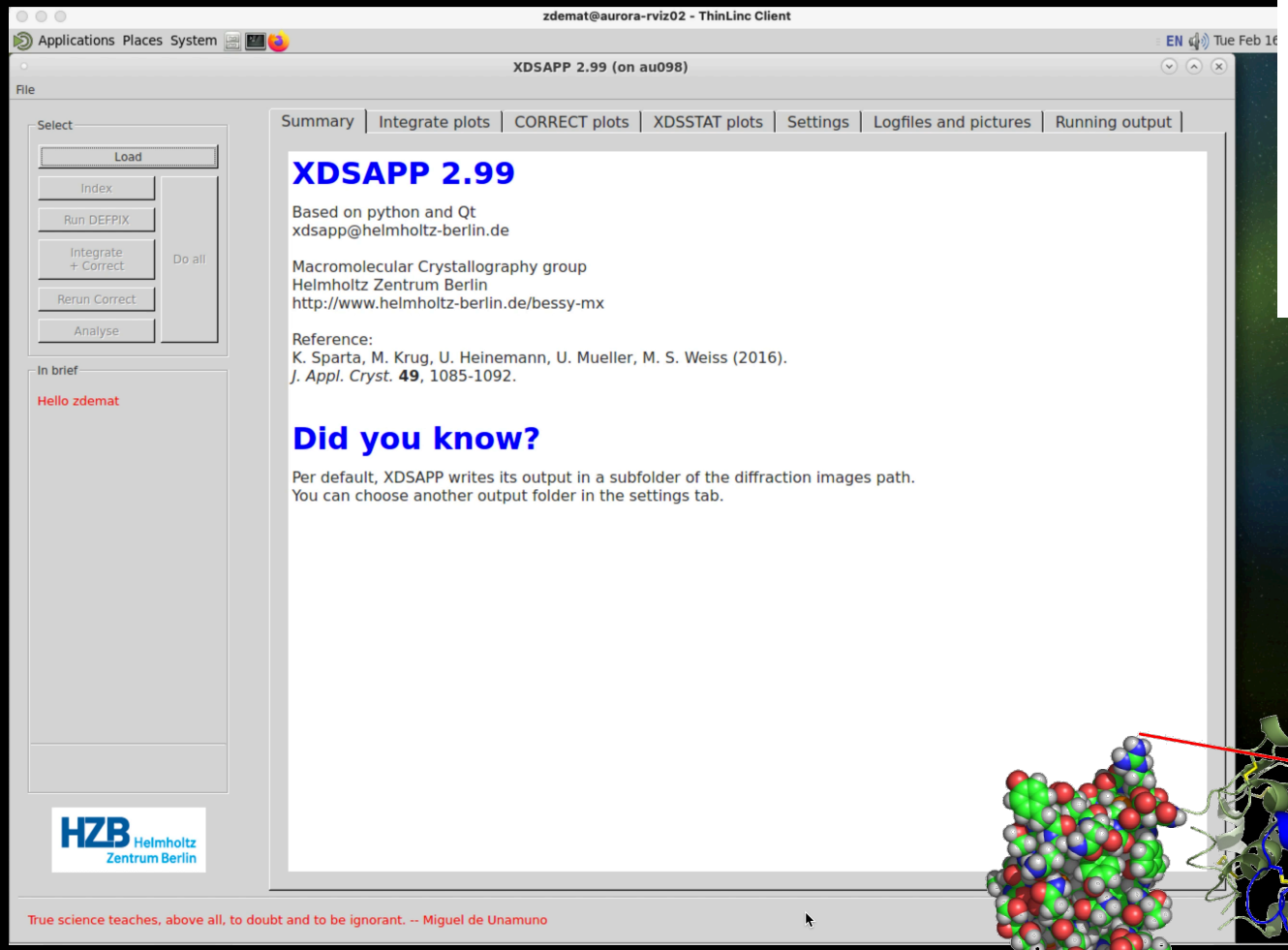


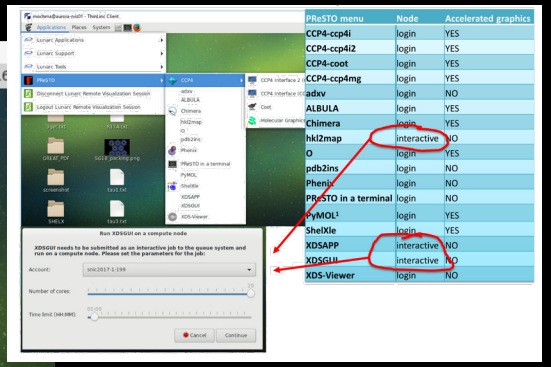
~ 100k crystals (frames),
4 Mpix, 2 kHz

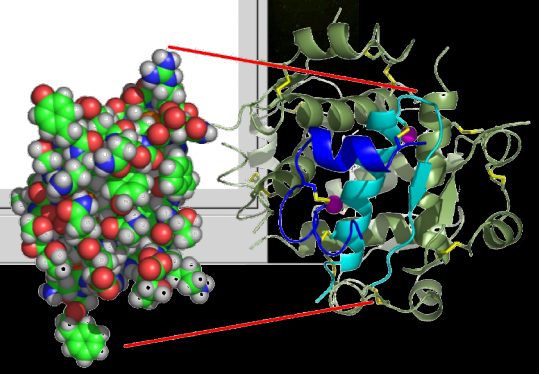
Credit: Jose M. Martin-Garcia et al., 10.1016/j.abb.2016.03.036

MAX IV

PReSTO – macromollecular crystallography







Insulin monomer
© Wikimedia Commons

PSF Protein Science Facility

KAROLINSKA INSTITUTET ANNO 1810

NSC

LUNARC

MAX IV