HELMHOLTZ

MATTER AND TECHNOLOGIES

RESEARCH FOR GRAND CHALLENGES

DATA MANAGEMENT & ANALYSIS (DMA)

Michael Bussmann
Spokesperson Topic „Data Management & Analysis“
Research Area „MATTER“ in Helmholtz

DESY JÜLICH HI Jena Helmholtz-Institut Jena HIM Helmholtz-Institut Mainz GSI HZB Helmholtz Zentrum Berlin Helmholtz-Zentrum Geesthacht Centre for Materials and Coastal Research HZDR

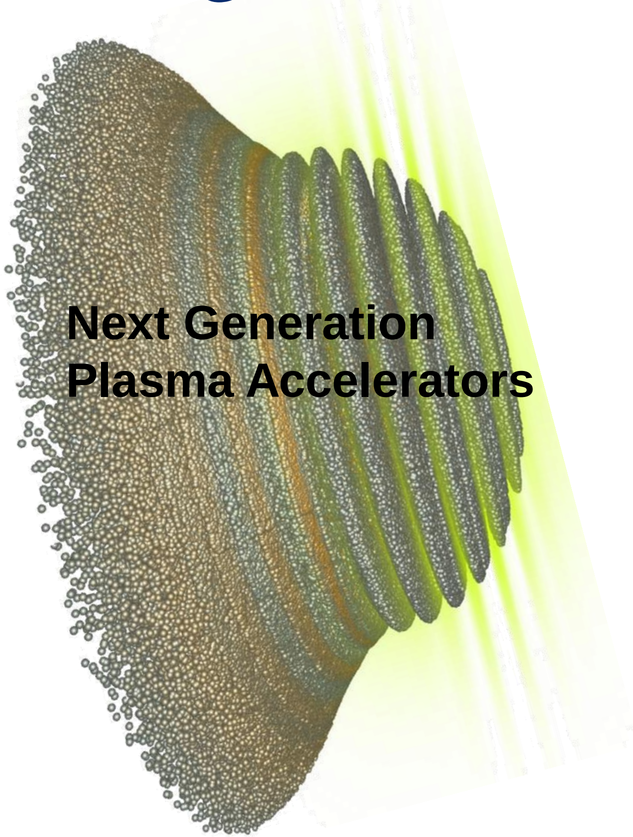
DIGITALIZATION @ HELMHOLTZ

- Helmholtz Incubator Platforms
  - Helmholtz Analytics Framework (HAF)
  - Helmholtz Infrastrucutre for Federated ICT Services (HIFIS)
  - Helmholtz Artifical Intelligence Cooperation Unit (HAICU)
  - Helmholtz Imaging Platform (HIP)
- Helmholtz Digitalisation Strategy
- Helmholtz Innovation Pool

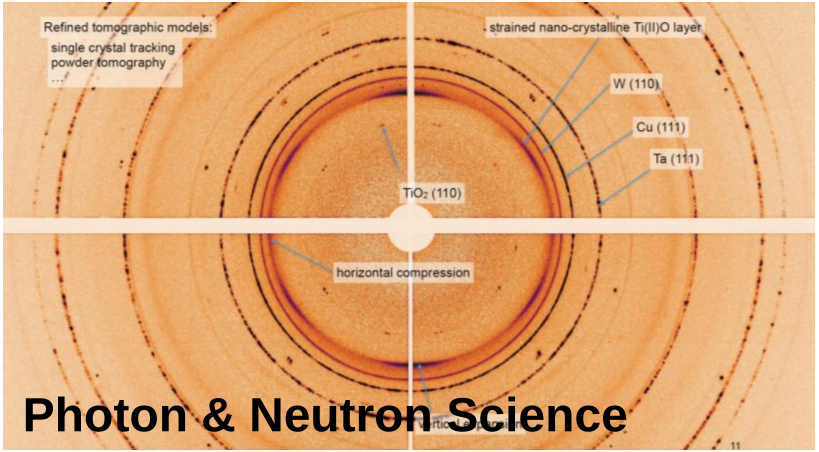
HELMHOLTZ
RESEARCH FOR GRAND CHALLENGES

- Germany’s largest research organization
- Annual budget of ~ €4,7 billion
- ~ 40,000 employees
- World-Class science infrastructure
- 19 independent research centers

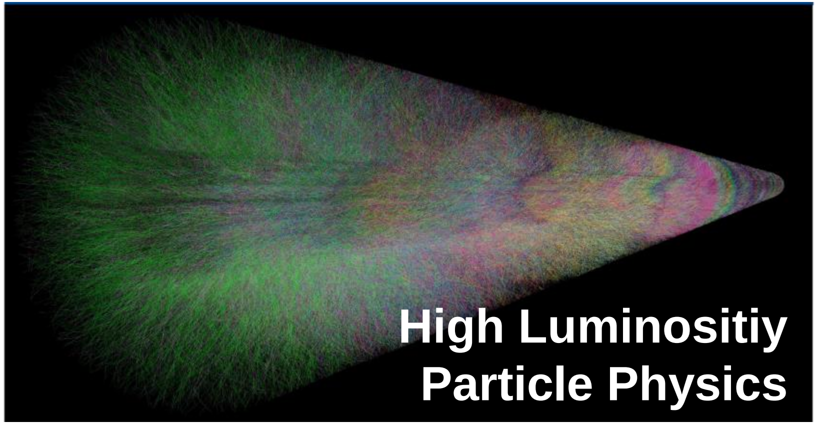
DMA @ HELMHOLTZ RESEARCH AREA MATTER



Next Generation
Plasma Accelerators



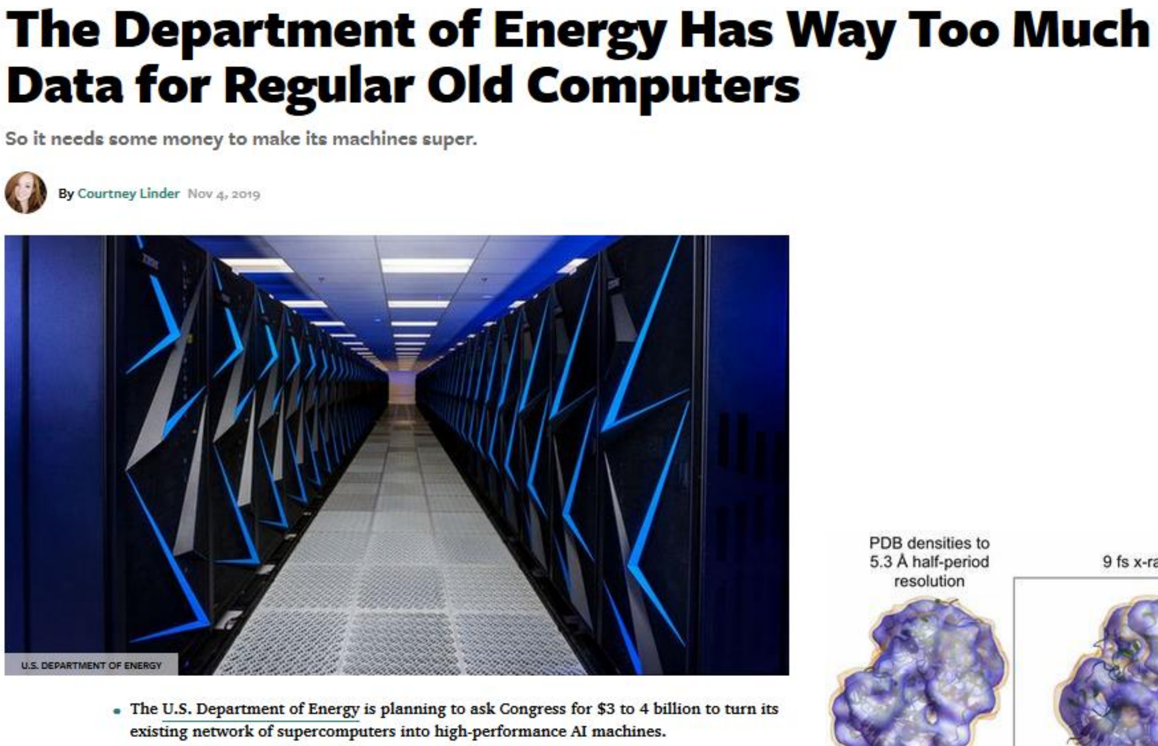
Photon & Neutron Science



High Luminositiy
Particle Physics

And more!
Applications!
Users!

COMPLEXITY (AND SYSTEMS) IS THE CHALLENGE



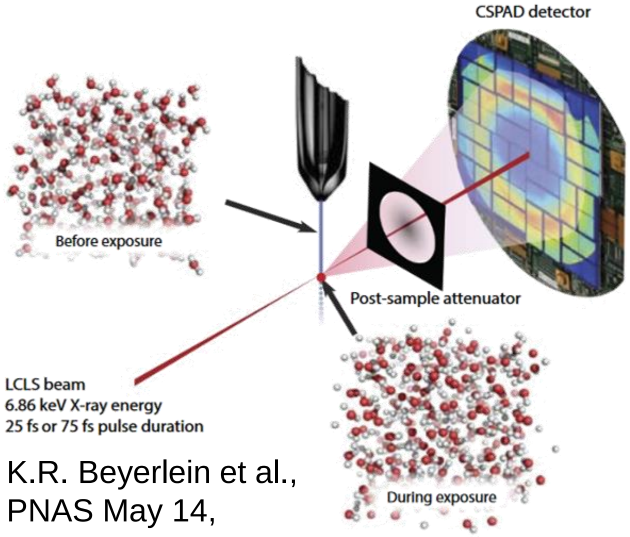
The Department of Energy Has Way Too Much Data for Regular Old Computers

So it needs some money to make its machines super.

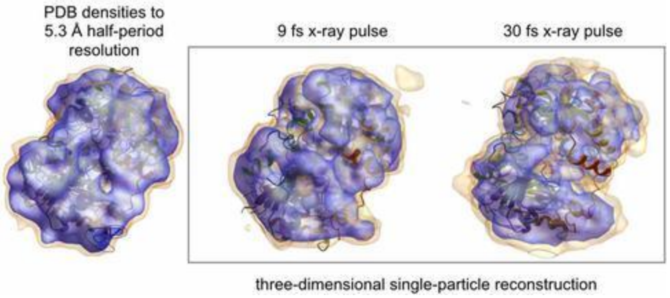
By Courtney Linder Nov 4, 2019

U.S. DEPARTMENT OF ENERGY

- The U.S. Department of Energy is planning to ask Congress for $3 to 4 billion to turn its existing network of supercomputers into high-performance AI machines.

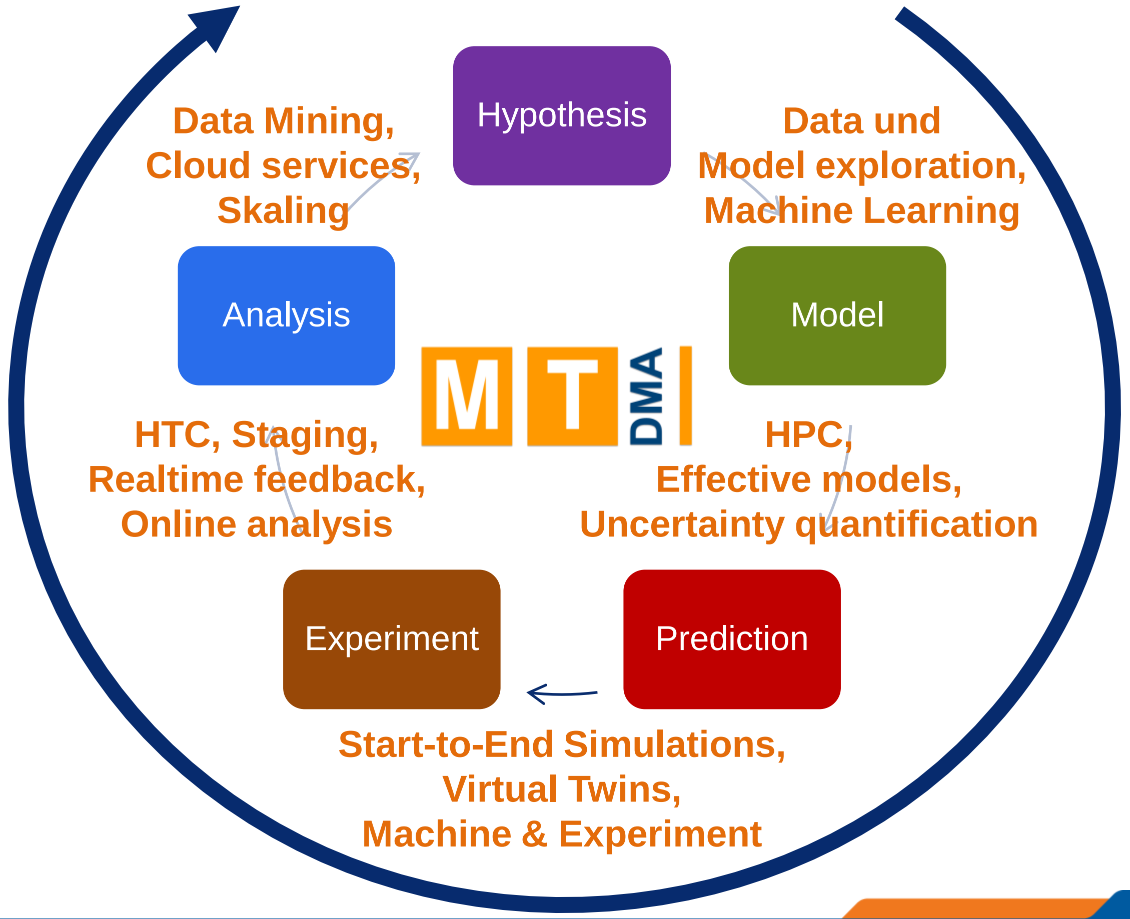


K.R. Beyerlein et al., PNAS May 14, 2018. 201711220


PDB densities to 5.3 Å half-period resolution
9 fs x-ray pulse
30 fs x-ray pulse
three-dimensional single-particle reconstruction

Chun Hong Yoon et al., Scientific Reports volume 6, Article number: 24791 (2016)

VISION – THE SCIENTIFIC METHOD MADE DIGITAL



Hypothesis

Data und
Model exploration,
Machine Learning

Model

HPC,
Effective models,
Uncertainty quantification

Prediction

Start-to-End Simulations,
Virtual Twins,
Machine & Experiment

Experiment

HTC, Staging,
Realtime feedback,
Online analysis

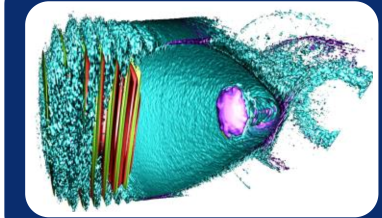
Analysis

Data Mining,
Cloud services,
Skaling

DMA DIVIDE & CONQUER (SUBTOPIC STRUCTURE)

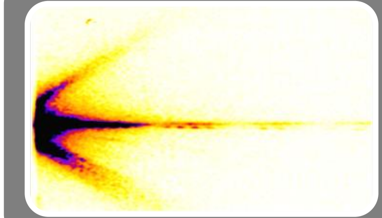
ST1: The Matter Information Fabric

- IT infrastructure (Hard- & Software) for facilities
- Automization of Data Lifecycle Management (LK II)
- Solutions für Communities



ST2: The Digital Scientific Method

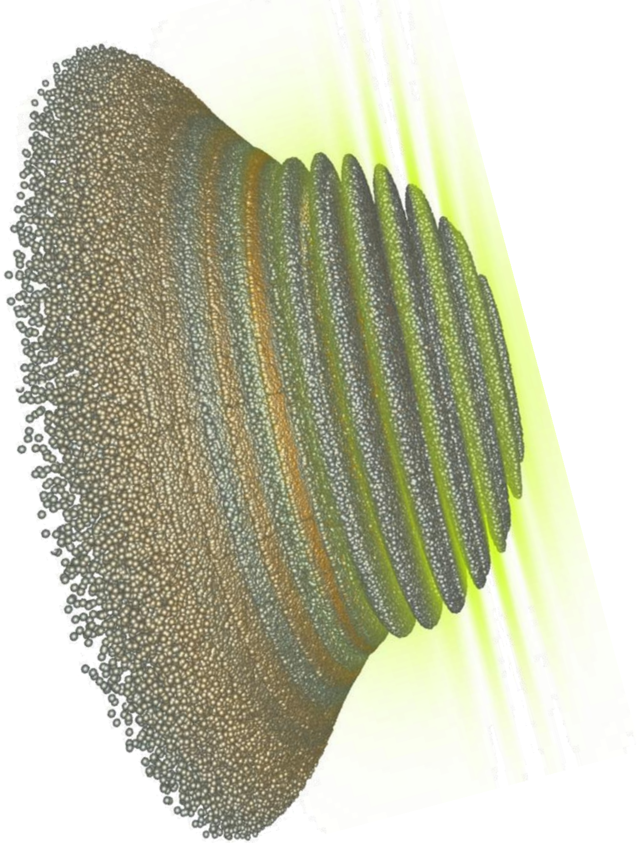
- Matter-specific research in Data Analysis & Simulation methods
- e.g. Machine Learning, Simulation, Visual Analytics, Scientific Workflow
- Developing methods für heterogeneous HPC, HTC, I/O for Matter applications



ST3: The Digital Experiment and Machine

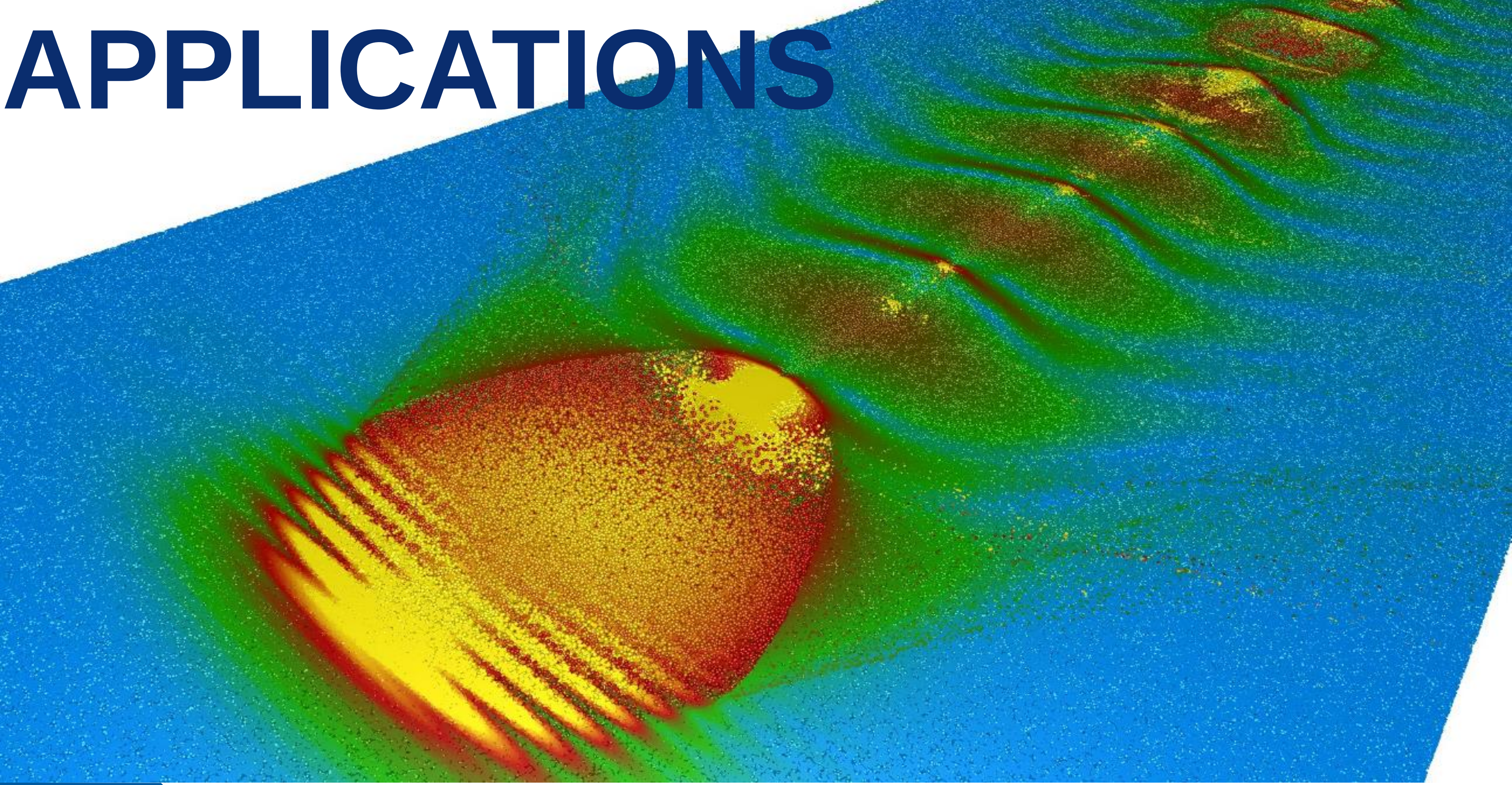
- Start-to-End Simulations (Machine/Interaction/Detectors)
- Fast feedback & machine control („Human in the Loop“)
- Quantifying data quality, meta data acquisition & analysis

DMA COMPETENCES



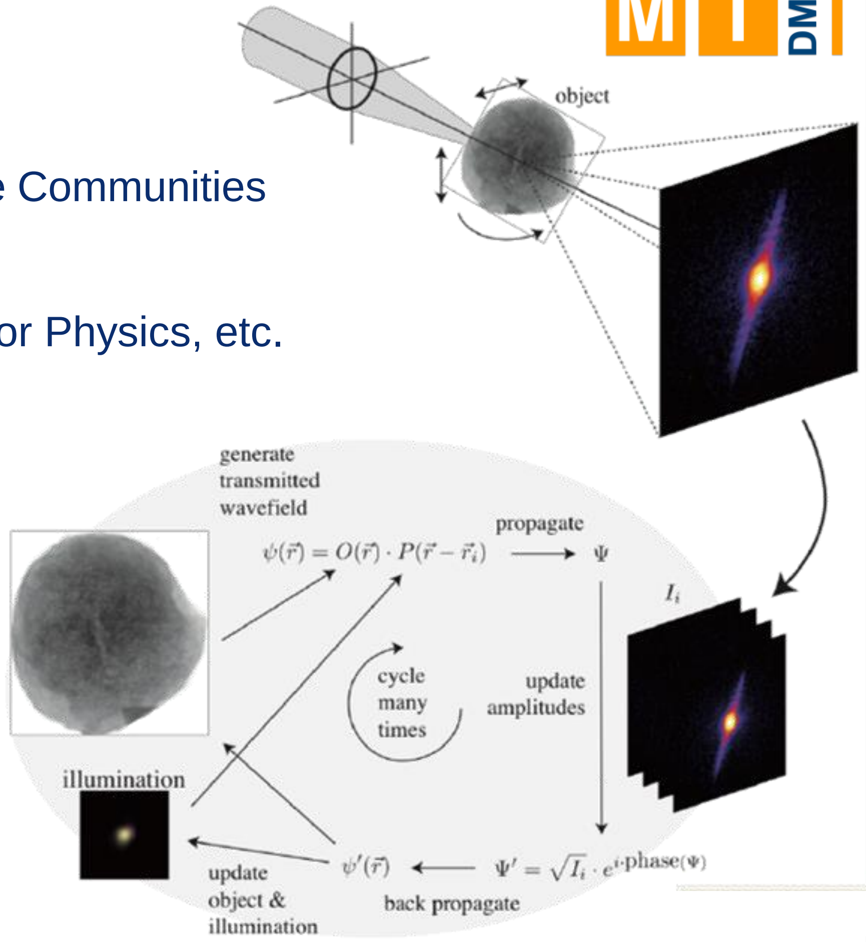
- Large-scale Data Management
- Applications
- Scalability
- Intelligence
- Bringing Facilities + Users together

APPLICATIONS

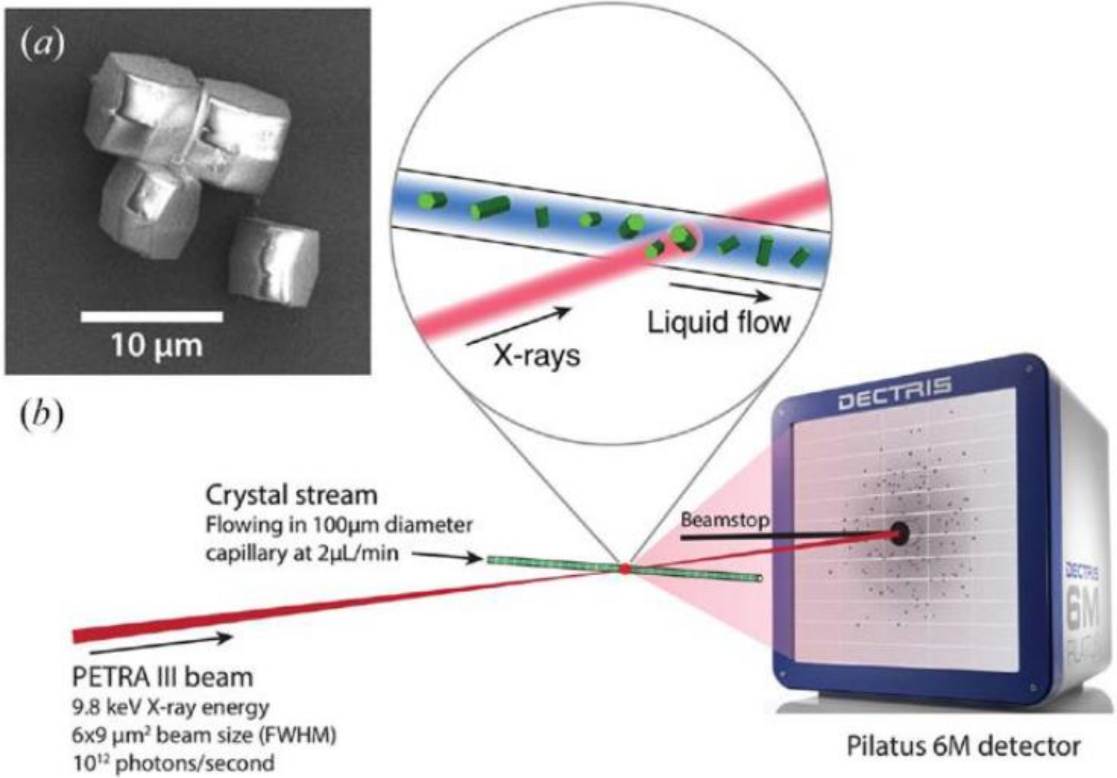


EXAMPLE: PTYCHOGRAPHY

- Strong involvement of Photon & Neutron Science Communities
- Particle Physics, Astroparticle Physics, Accelerator Physics, etc.
- Example: Ptychography
- CPUs? GPUs? FPGAs?


object
generate transmitted wavefield
$\psi(\vec{r}) = O(\vec{r}) \cdot P(\vec{r} - \vec{r}_i)$
propagate
Ψ
I_i
cycle many times
update amplitudes
illumination
update object & illumination
$\psi'(\vec{r})$
$\Psi' = \sqrt{I_i} \cdot e^{i \cdot \mathrm{phase}(\Psi)}$
back propagate

SCALABILITY

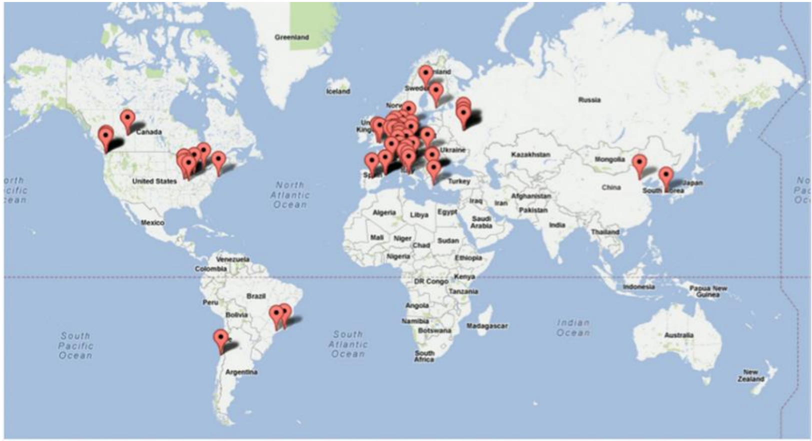


F. Stellato, et al., IUCrJ **1**, 204 (2014).

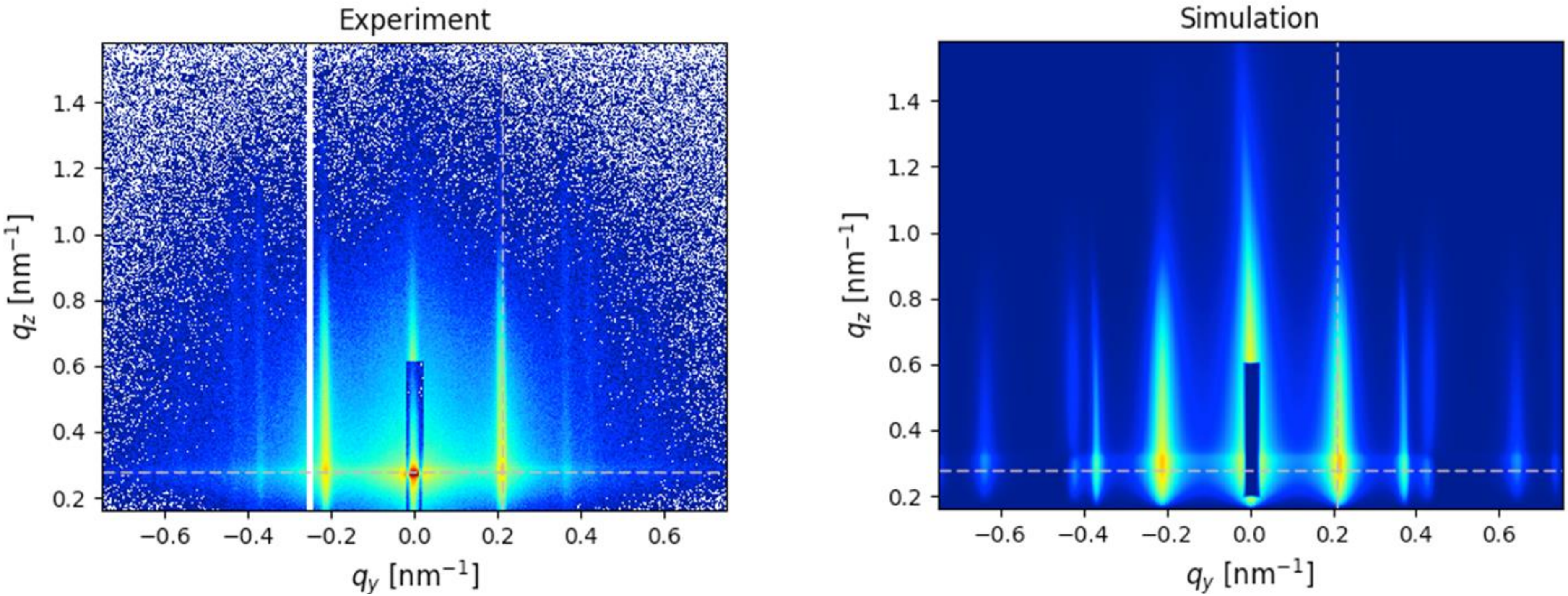
EXAMPLE: I/O AND STORAGE

“Overall , this is an outstanding proposal. […]
The PIs should try to reduce the data requirements
and try to find a solution that is technically possible for CSCS.”

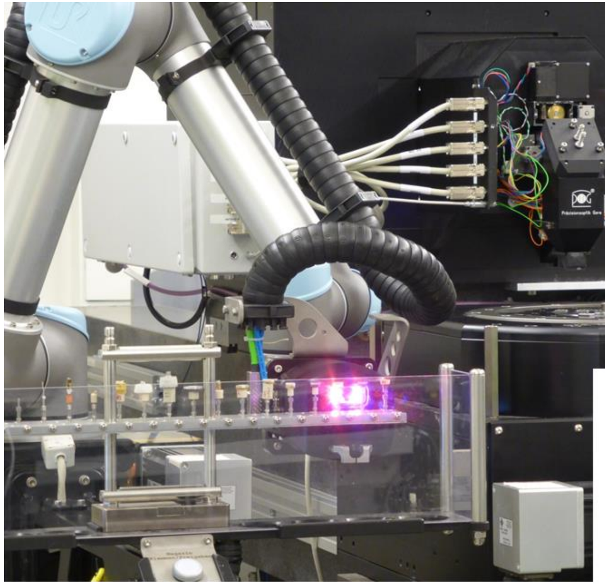
1 Pb produced in less than a week by a single PhD student

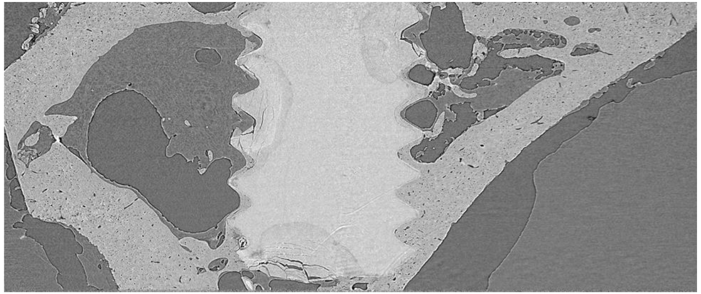


INTELLIGENCE

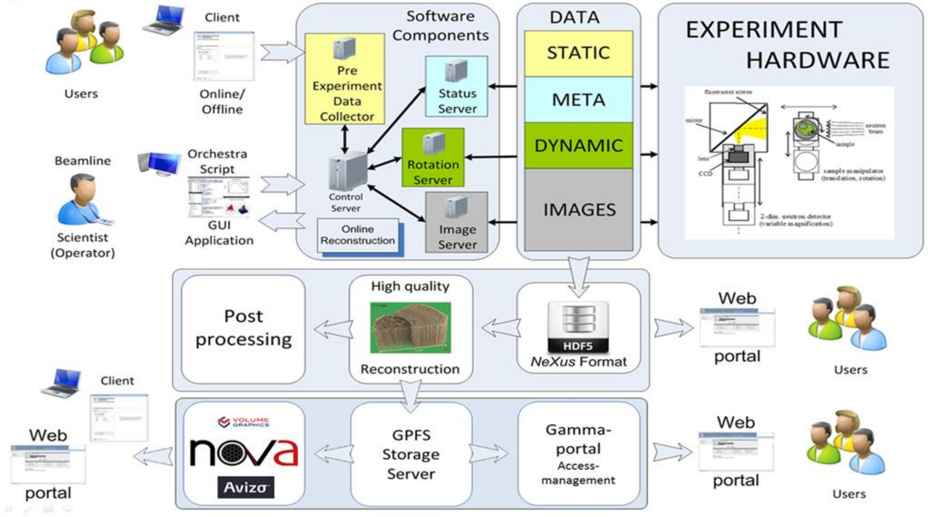


CONTROL + DATA + META DATA + …





Push-out: Sequence of vertical slicing through reconstructed volume



Users

Client

Online/ Offline

Beamline

Scientist (Operator)

Orchestra Script

GUI Application

Software Components

Pre Experiment Data Collector

Status Server

Control Server

Rotation Server

Online Reconstruction

Image Server

DATA

STATIC

META

DYNAMIC

IMAGES

EXPERIMENT HARDWARE

fluorescent screen
mirror
lens
CCD
neutron beam
sample
sample manipulator (translation, rotation)
2-dim. neutron detector (variable magnification)

Post processing

High quality Reconstruction

HDF5 *NeXus* Format

Web portal

Users

Web portal

Users

Client

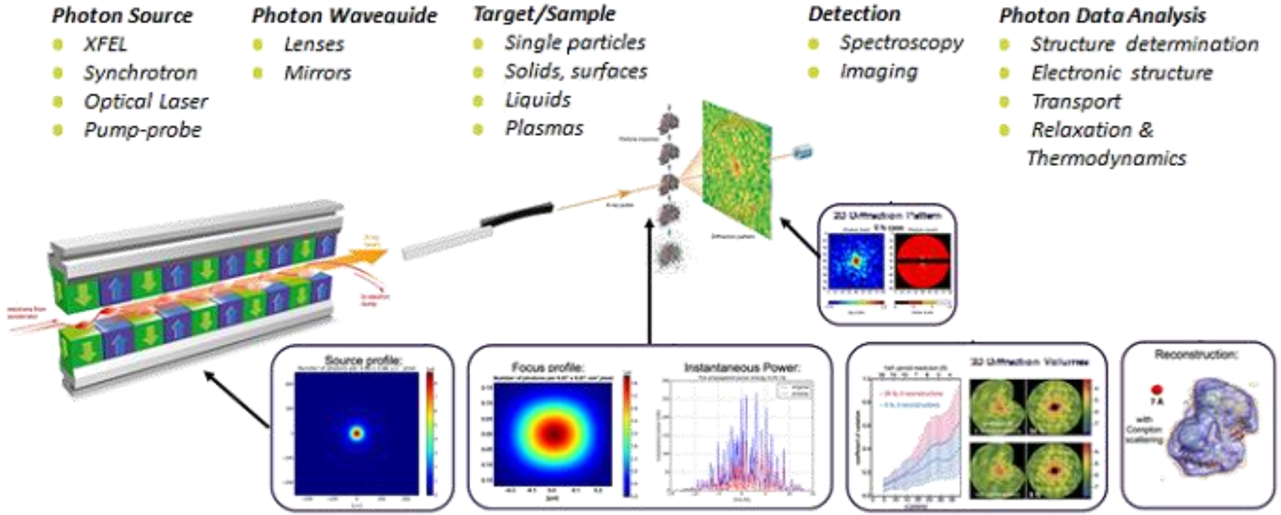
Web portal

GPFS Storage Server

Gamma-portal Access-management

- In-situ feedback & control
- Scalable Meta Data Collection & Understanding
- Transient data workflows

COMPLEXITY NEEDS INTELLIGENCE



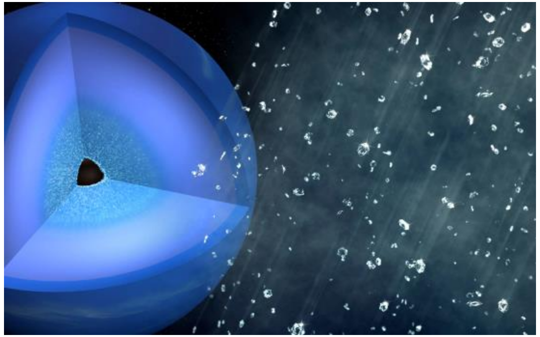
Photon Source
- *XFEL*
- *Synchrotron*
- *Optical Laser*
- *Pump-probe*

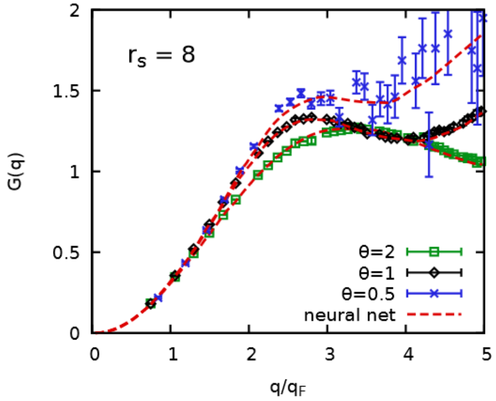
Photon Waveguide
- *Lenses*
- *Mirrors*

Target/Sample
- *Single particles*
- *Solids, surfaces*
- *Liquids*
- *Plasmas*

Detection
- *Spectroscopy*
- *Imaging*

Photon Data Analysis
- *Structure determination*
- *Electronic structure*
- *Transport*
- *Relaxation & Thermodynamics*





- Data fusion of experimental and simulation data will become the norm
- Data won’t be final! Data is at best alive and changes meaning
- Data analysis will stay transient much longer in the future (HPC+HTC)

The challenge of complexity and how to attack it

Michael Bussmann, CASUS

05/11/2019 | 1st annual PANOSC meeting

CASUS
CENTER FOR ADVANCED
SYSTEMS UNDERSTANDING

www.casus.science

```
mirror object to mirror
mirror_mod.mirror_object
operation == "MIRROR_X":
mirror_mod.use_x = True
mirror_mod.use_y = False
mirror_mod.use_z = False
operation == "MIRROR_Y":
mirror_mod.use_x = False
mirror_mod.use_y = True
mirror_mod.use_z = False
operation == "MIRROR_Z":
mirror_mod.use_x = Fal
```

HZDR HELMHOLTZ ZENTRUM DRESDEN ROSSENDORF

TECHNISCHE UNIVERSITÄT DRESDEN

CBG Max Planck Institute of Molecular Cell Biology and Genetics

HELMHOLTZ CENTRE FOR ENVIRONMENTAL RESEARCH – UFZ

Uniwersytet Wrocławski

SPONSORED BY THE

Federal Ministry of Education and Research

STAATSMINISTERIUM FÜR WISSENSCHAFT UND KUNST

Freistaat SACHSEN

The four horsemen of the datacalypse

How do we fill the data catalogues wisely?

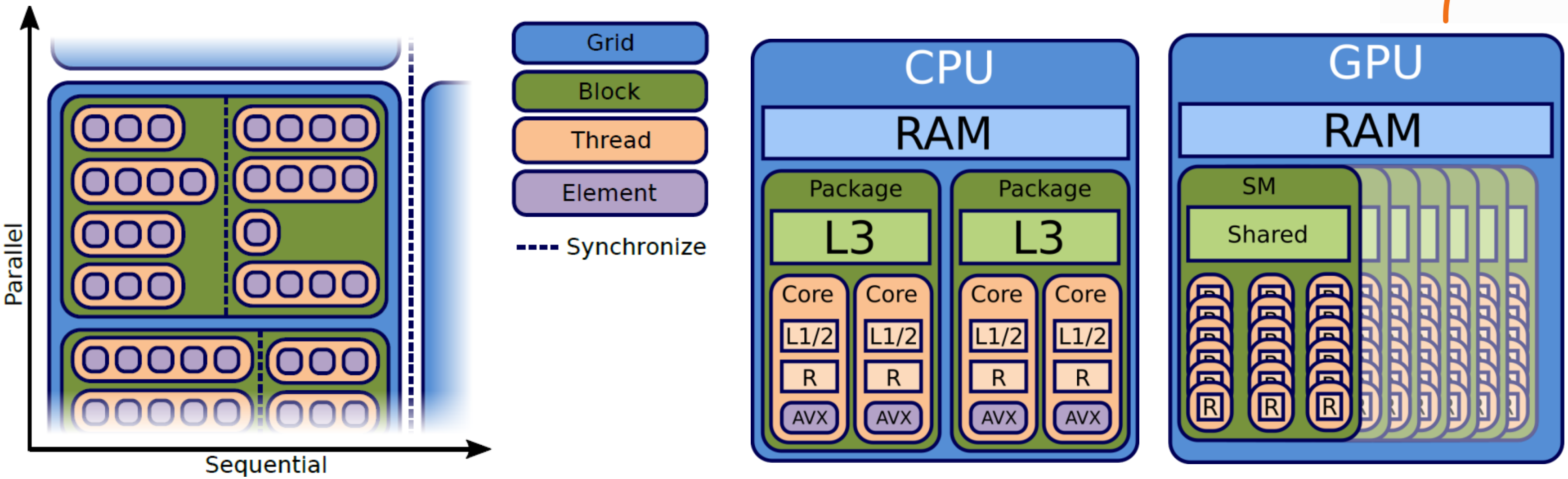
We face **four challenges**

- **High data quality** prevents us from serious initial automatic reduction
- **High data rates** give us Pbytes of data in a few hours
- **Short data lifetime** gives PhD students stress
- **Poor understanding** of the system investigated requires in-depth expert intervention

Actually, using computers efficiently is still pretty hard

Unified programming interfaces to CPUs, GPUs, FPGAs

alpaka



NVIDIA CUDA

GPUOpen

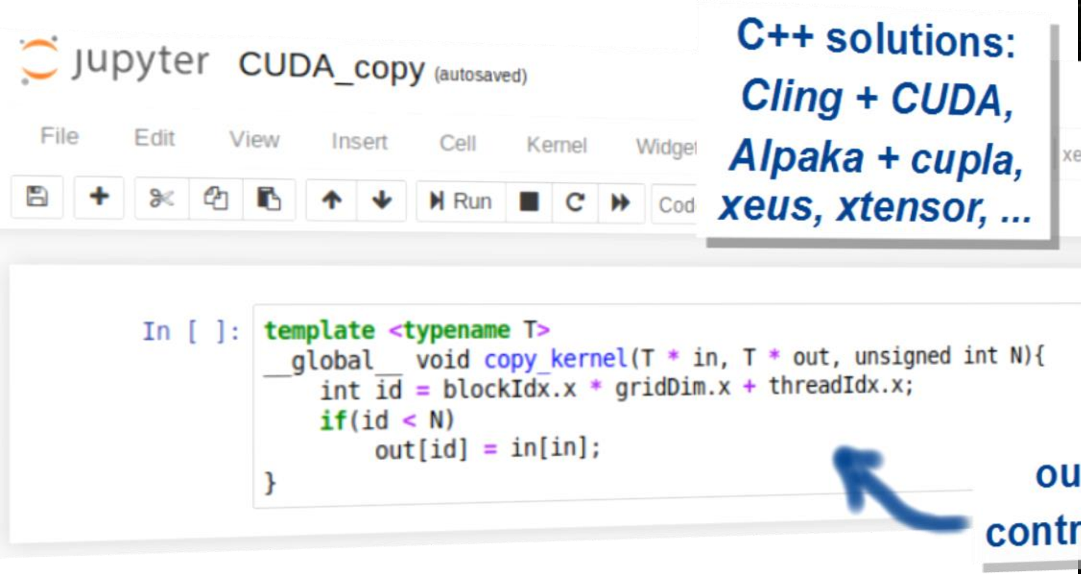
OpenMP

SYCL™

TBB

HTC / HPC interactivity @ Exascale

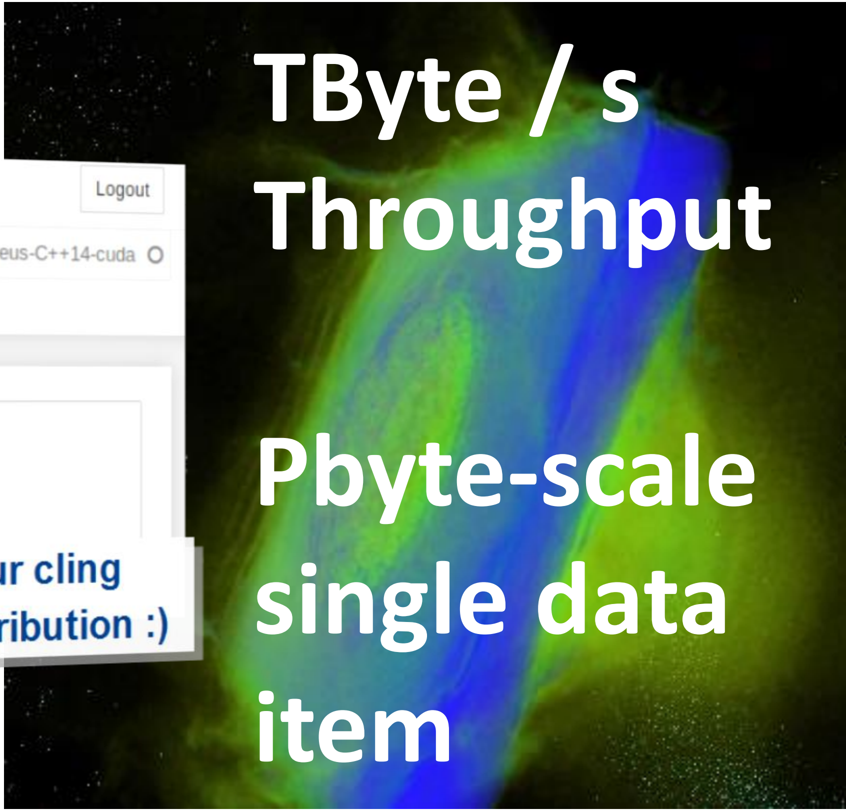
Scalable across full systems, full JIT capability, visual analytics,…



```
In [ ]: template <typename T>
        __global__ void copy_kernel(T * in, T * out, unsigned int N){
            int id = blockIdx.x * gridDim.x + threadIdx.x;
            if(id < N)
                out[id] = in[in];
        }
```

C++ solutions:
Cling + CUDA,
Alpaka + cupla,
xeus, xtensor, ...

our cling contribution :)

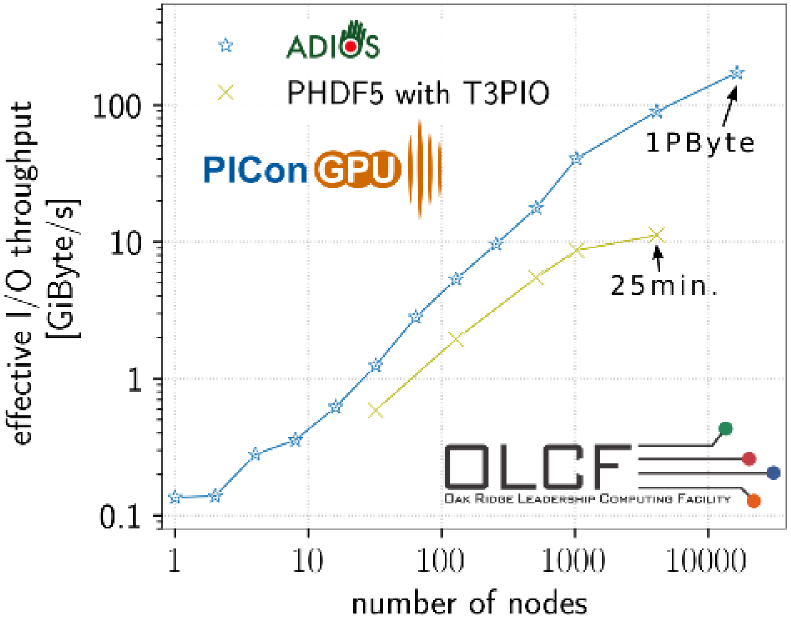


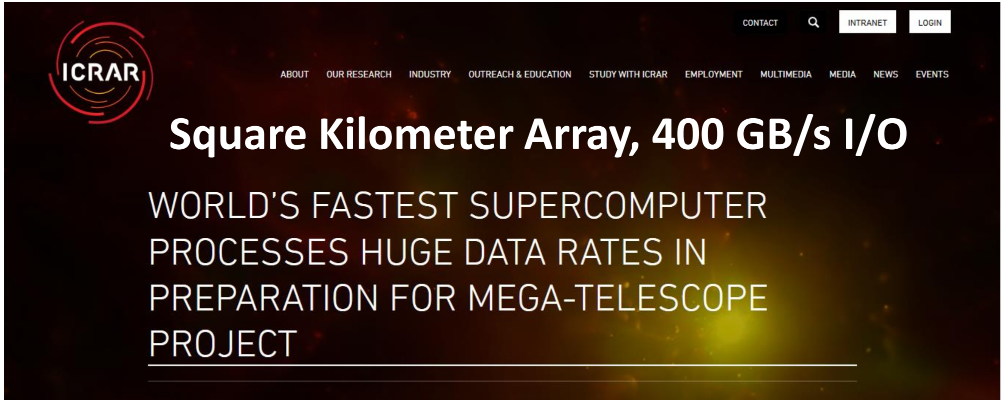
TByte / s Throughput

Pbyte-scale single data item

Transient data analysis or data storage or both or …?

We have a troublesome throughput hierarchy



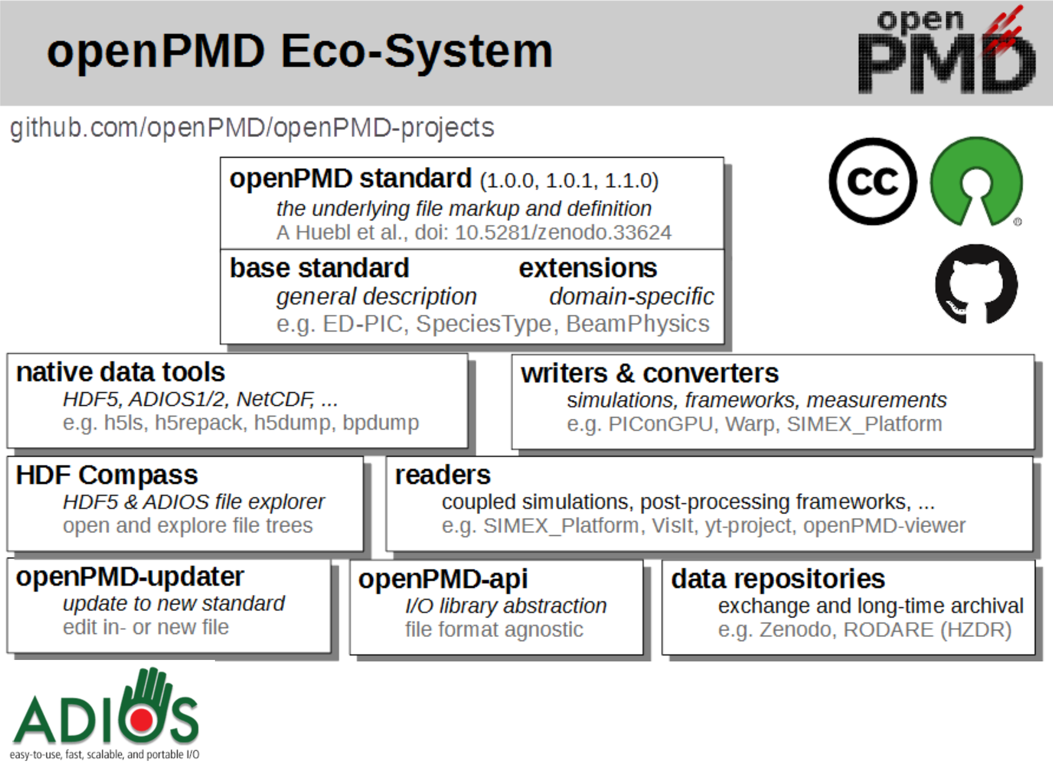


Square Kilometer Array, 400 GB/s I/O

WORLD'S FASTEST SUPERCOMPUTER PROCESSES HUGE DATA RATES IN PREPARATION FOR MEGA-TELESCOPE PROJECT

Open, self-explaining meta data formats & ecosystems

In-memory workflow coupling becomes standard



openPMD Eco-System

github.com/openPMD/openPMD-projects

openPMD standard (1.0.0, 1.0.1, 1.1.0)
the underlying file markup and definition
A Huebl et al., doi: 10.5281/zenodo.33624

base standard
general description

extensions
domain-specific

e.g. ED-PIC, SpeciesType, BeamPhysics

native data tools
HDF5, ADIOS1/2, NetCDF, ...
e.g. h5ls, h5repack, h5dump, bpdump

writers & converters
simulations, frameworks, measurements
e.g. PIConGPU, Warp, SIMEX_Platform

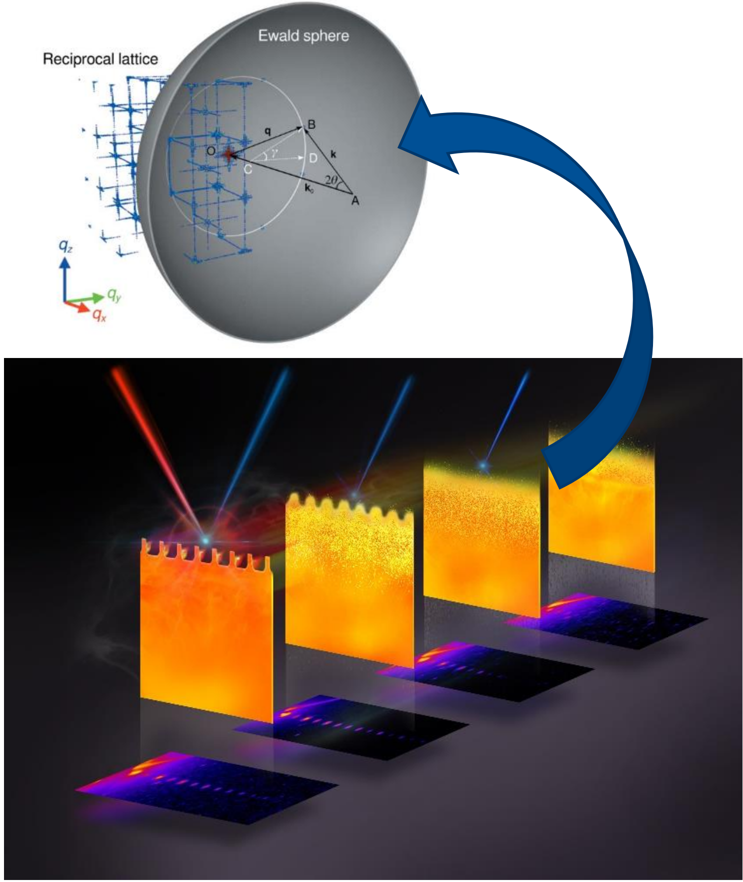
HDF Compass
HDF5 & ADIOS file explorer
open and explore file trees

readers
coupled simulations, post-processing frameworks, ...
e.g. SIMEX_Platform, VisIt, yt-project, openPMD-viewer

openPMD-updater
update to new standard
edit in- or new file

openPMD-api
I/O library abstraction
file format agnostic

data repositories
exchange and long-time archival
e.g. Zenodo, RODARE (HZDR)



- **Complexity** is the central problem if facilities produce **high quality** data and **share** it

- Data reduction will become synonymous with **knowledge extraction** (+ meta data)

- ?aaS will require **expert domain knowledge**, interactivity and **transient** data analysis capabilities (lifetime!)