Extreme Throughput Data Taking, Management & Analysis

Real Life Experiences of Generating/Evaluating Extreme Data Sets Around the World

G. Juckeland

Helmholtz-Zentrum Dresden-Rossendorf

https://www.fairfaxcounty.gov/news2/wp-content/uploads/2016/03/Fire-hydrant-with-water-blur.jpg

https://pluspng.com/img-png/bathtub-png-bathtub-png-1662.png

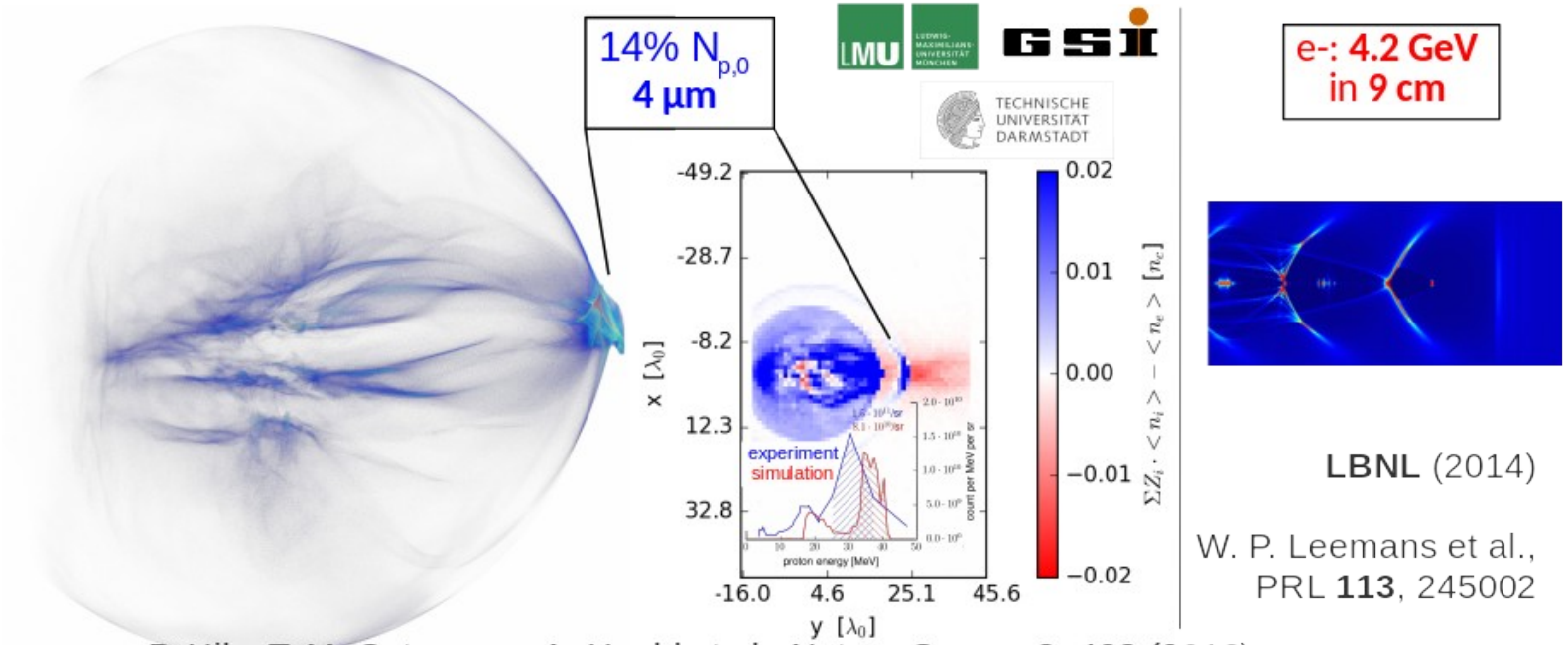
Matter meets Information, DESY, Jan 14-15, 2019

Our Background

well, one of many...

Dr. Guido Juckeland I g.juckeland@hzdr.de I Computational Science Group I www.hzdr.de/fwcc

Laser Plasma Acceleration

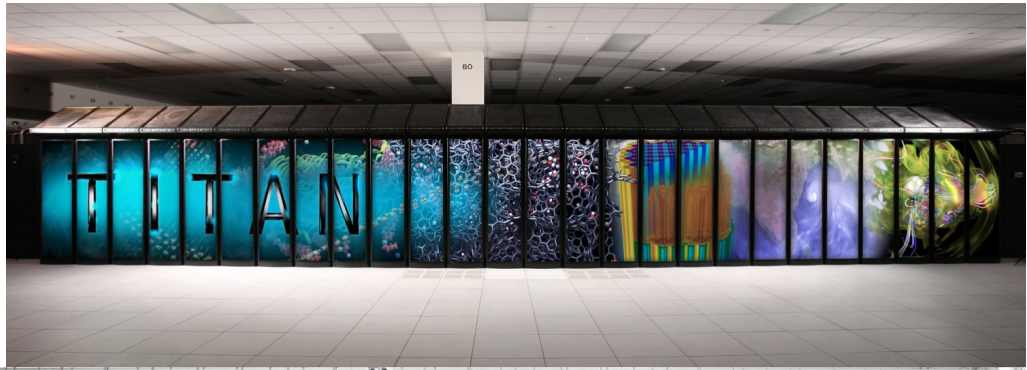


P. Hilz, T. M. Ostermayr, A. Huebl et al., Nature Comm. **9**, 423 (2018)

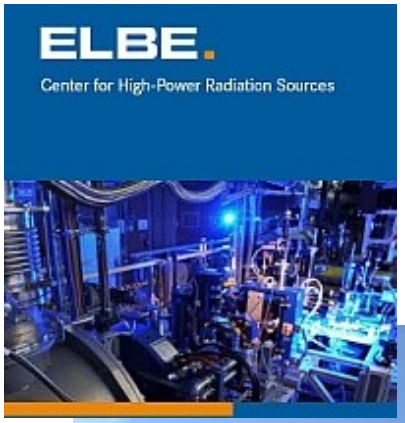
Particle-in-Cell Simulations a.k.a. digital twins

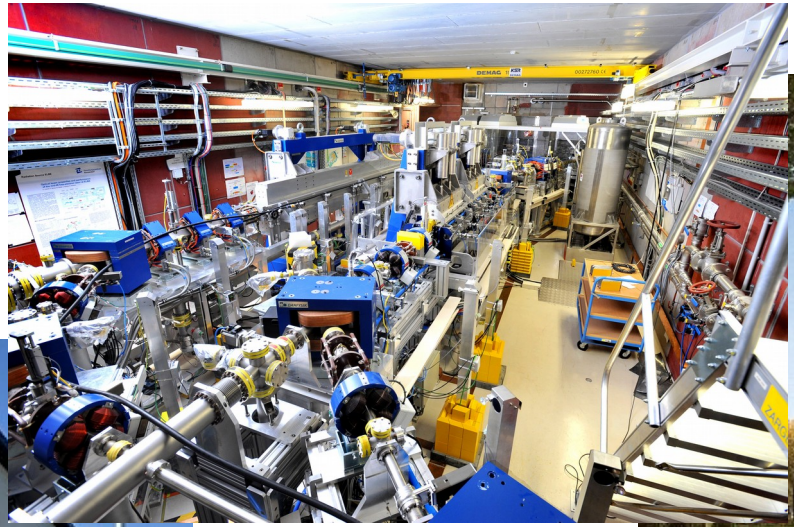
PIConGPU

- Ab initio, electro-magnetic plasmas
- Scaling to the full-size of Titan & Piz Daint
- Gordon Bell finalist 2013



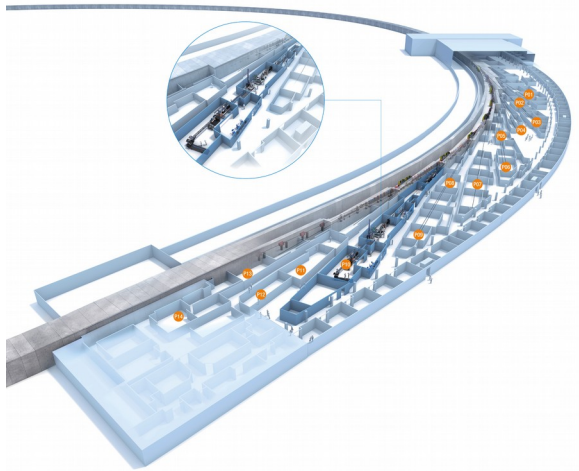
LK-II Facilities for Plasma Physics, Material Science,...

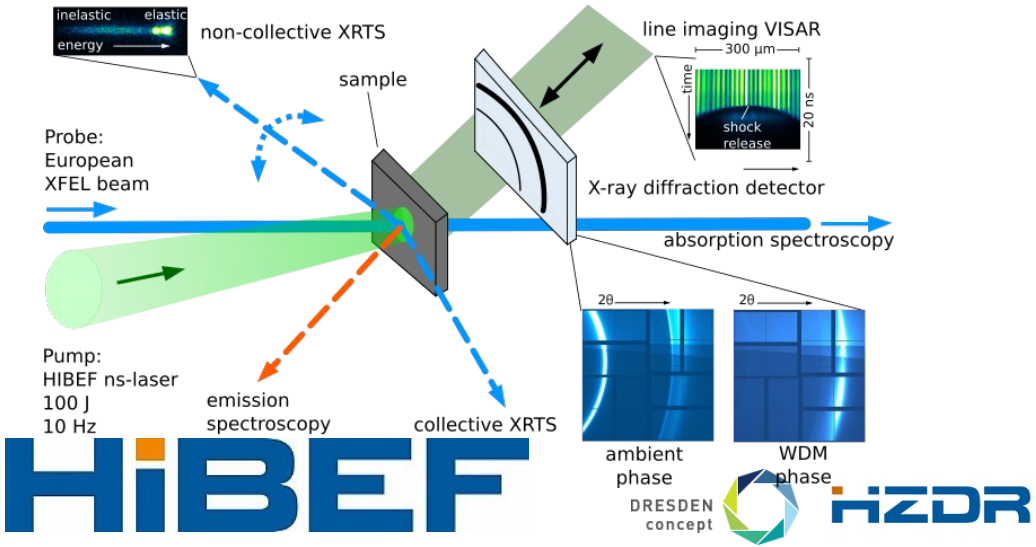




„USS Enterprise“ Experiments

- Novel methods, setups, (commodity) detectors
- Unprecedented bandwidth/latency requirements
- Photon Science different from HEP





Data Challenges at Extreme Scale

(also coming to your site! Soon.)

Typical Post-Processing

A. Huebl et al., DRBSD-1 - ISC'17 (2017),
DOI:10.1007/978-3-319-67630-2_2, arXiv:1706.00522

generate data
measure or sim.

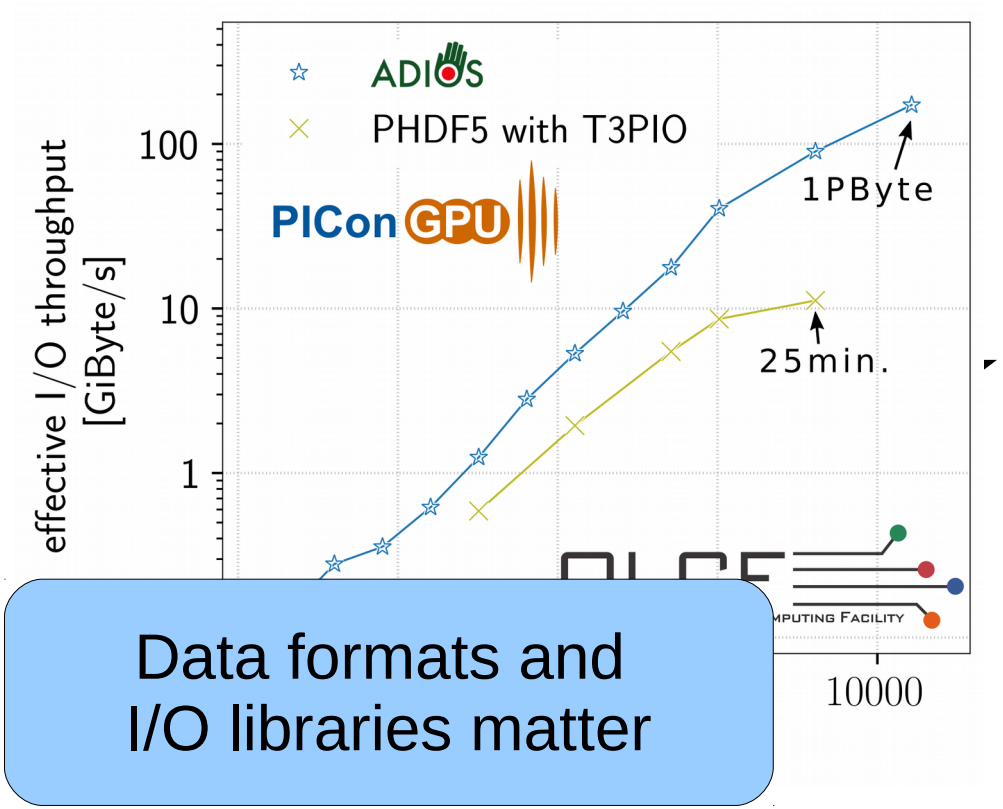
store

post-process

K20x GPU (2013), **bw :=**
6 GByte · 10 Hz steps

PCI-Express: **6 Gbyte/s**
1 / 10 th bw

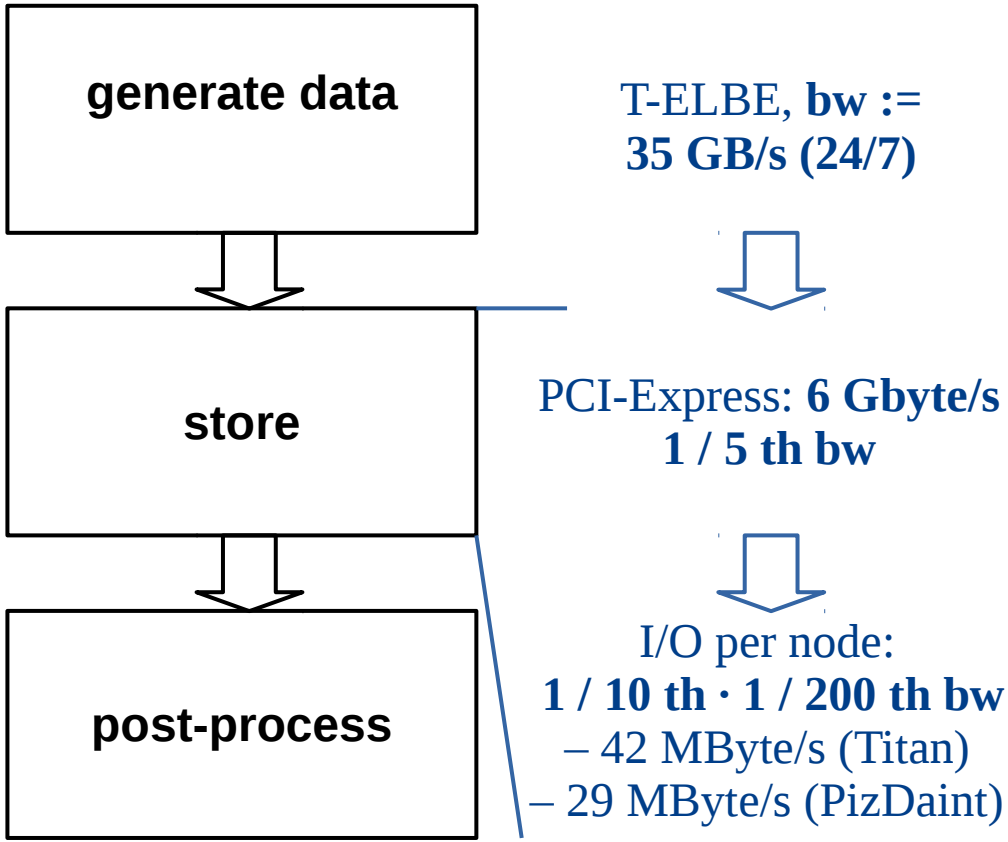
I/O per node:
1 / 10 th · 1 / 200 th bw
– 42 MByte/s (Titan)
– 29 MByte/s (PizDaint)



Data formats and
I/O libraries matter

Summit (ORNL, 2018): ratio **4x "worse" - gap of 10^4**

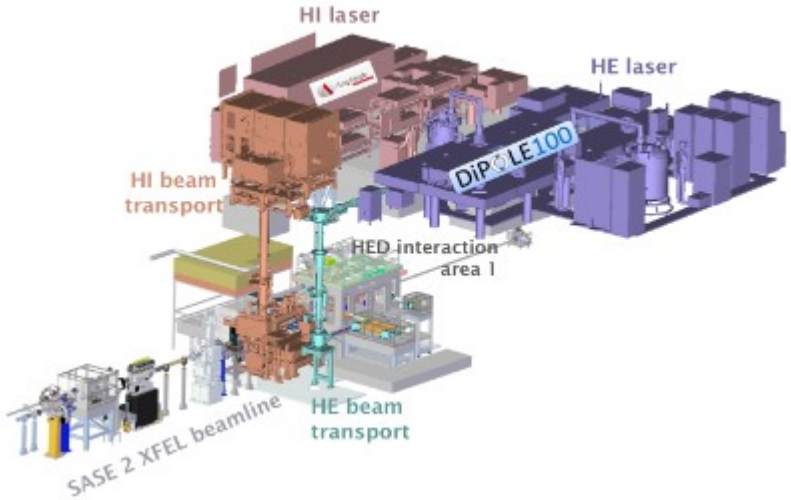
Next-Generation Experiments



T-ELBE, **bw :=**
35 GB/s (24/7)

PCI-Express: **6 Gbyte/s**
1 / 5 th bw

I/O per node:
1 / 10 th · 1 / 200 th bw
– 42 MByte/s (Titan)
– 29 MByte/s (PizDaint)



When the Bath-Tub is Full… - Site Storage Policies for >1 PB data sets

	Site A	Site B	Site C
Capacity	250 PByte	6 PByte	3 PByte
Capacity/FLOP	2 Byte/FLOP/s	0.3 Byte/FLOP/s	10 Byte/FLOP/s
Bandwidth	2.5 TB/s	100 GB/s (estd.)	40 GB/s
Bandwidth/FLOP	20 µByte/FLOP	5 µByte/FLOP	133 µByte/FLOP
Retention Time	90 days + archive	30 days	∞

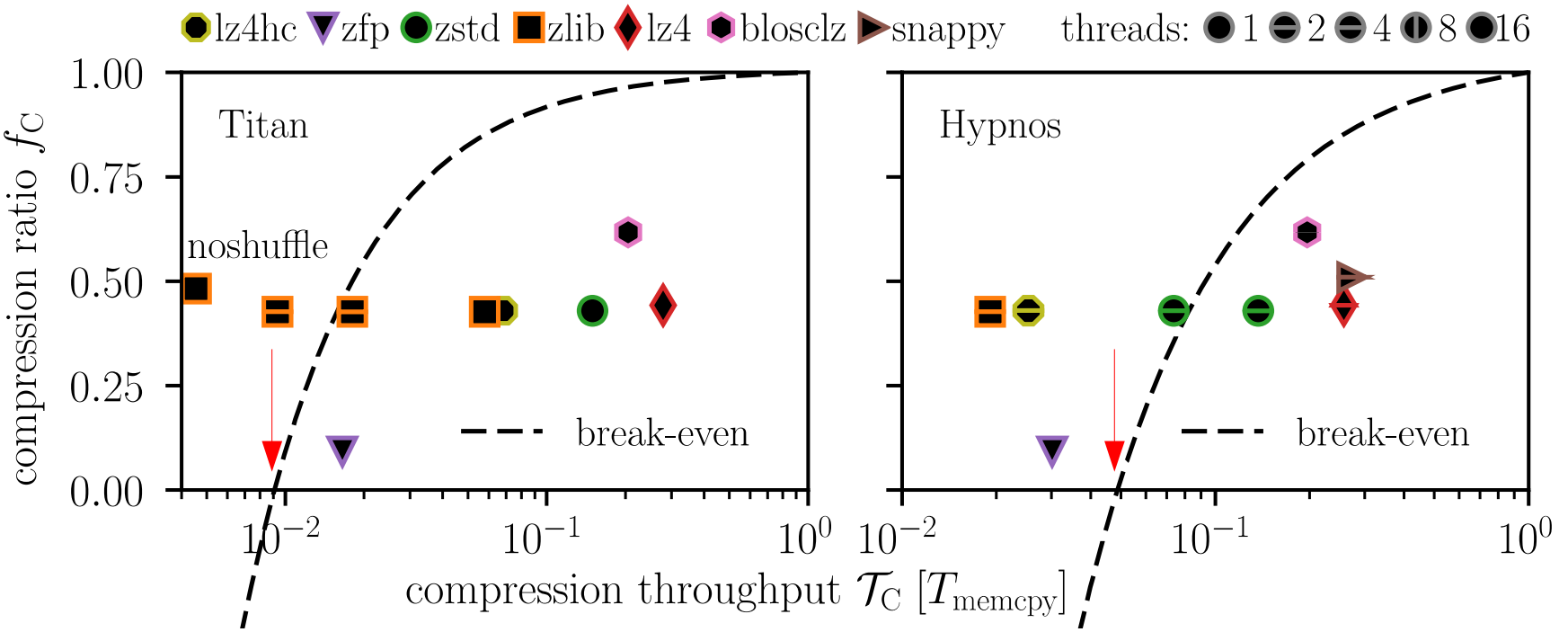
Which to choose??

Potential Solutions

and first steps taken

On-the-fly (Lossless) Compression

$$\mathcal{T}_C > \frac{\mathcal{T}_{FS}}{\mathcal{T}_{FS}+1}$$

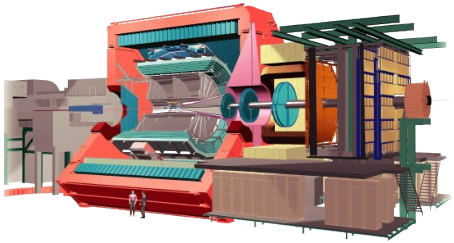


Zfp 0.5.1: three uncompressed bits / scalar; on *particle data*

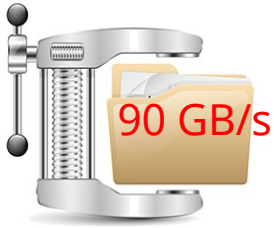
Blosc 1.11.4-dev: add bitshuffle pre-conditioner

In Transit Data Reduction

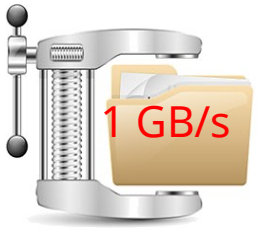
Blurring Online and Offline by ALICE and FAIR Experiments

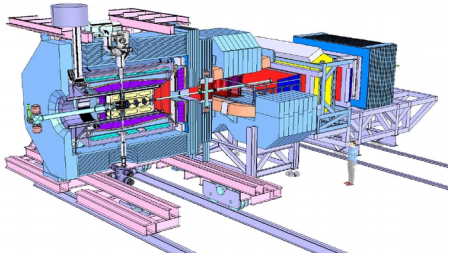


3.4 TB/s

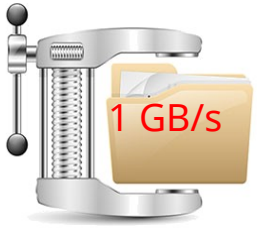


1 TB/s





0.3 TB/s



The unifying element: Detector Readout

- Continuous readout with self-triggered front-end electronics
- No or limited hardware triggers
- Event definition & selection requires full reconstruction in online compute farms

FAIR

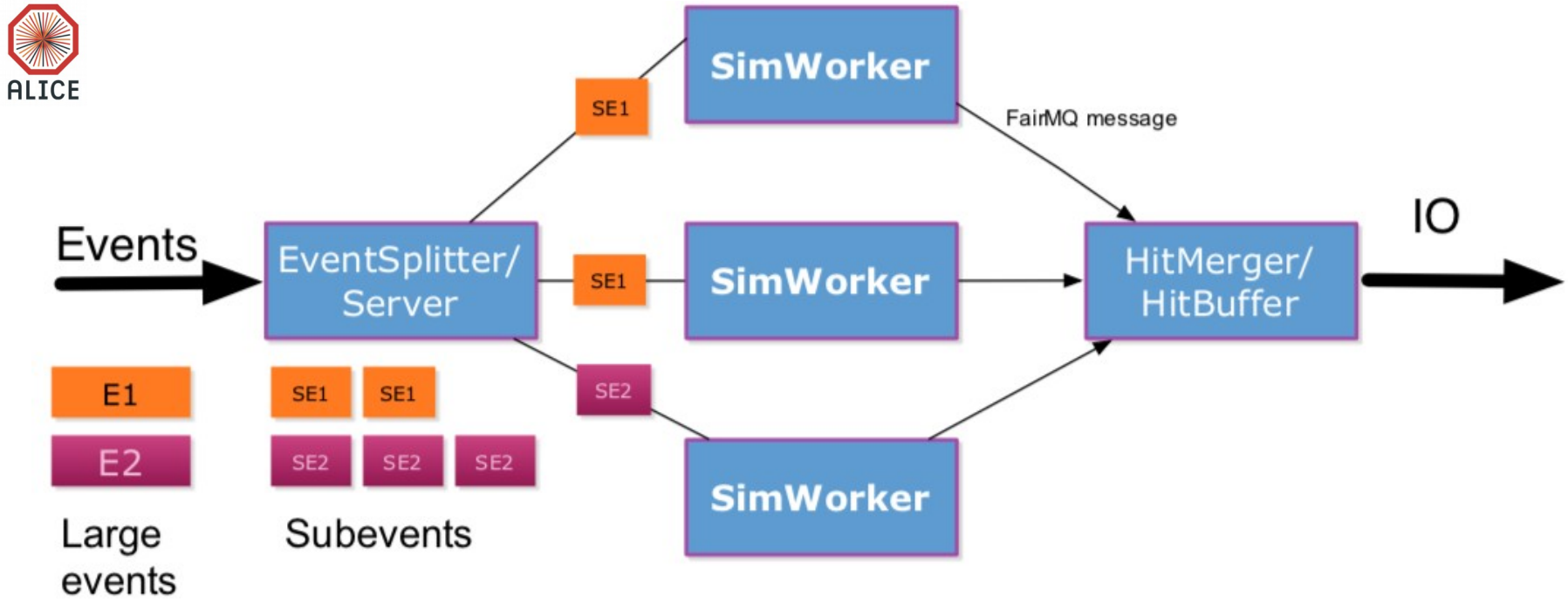
ALICE

DRESDEN concept

HZDR

In Transit Data Processing

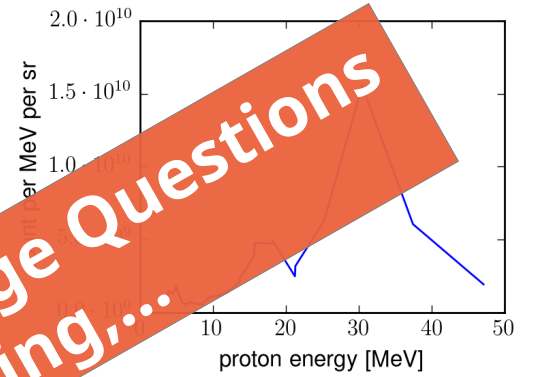
E.g. Using FairMQ



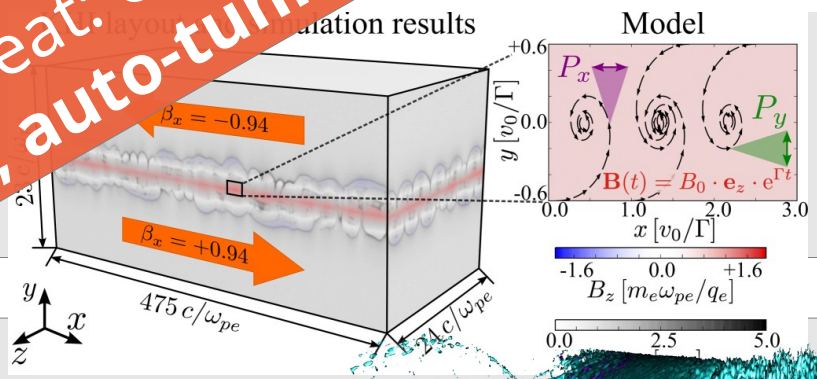
Human In-the-Loop In-Situ Data Reduction

A. Huebl et al. (2014), DOI:10.1109/TPS.2014.2327392
R. Pausch et al. (2017), DOI:10.1103/PhysRevE.96.013316
A. Matthes, A. Huebl et al., ISC'16 (2016), DOI:10.14529/jsfi160403
A. Huebl et al., ISC'17 (2017), DOI:10.1007/978-3-319-67630-2_2

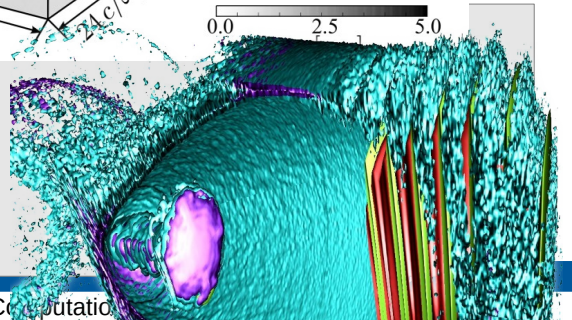
Binning of a **spectrogram**
Creation of a **phase space image**



In situ **radiation diagnosti**



Ray-cas photo-realistic ray-trace
Lossy da **pression**

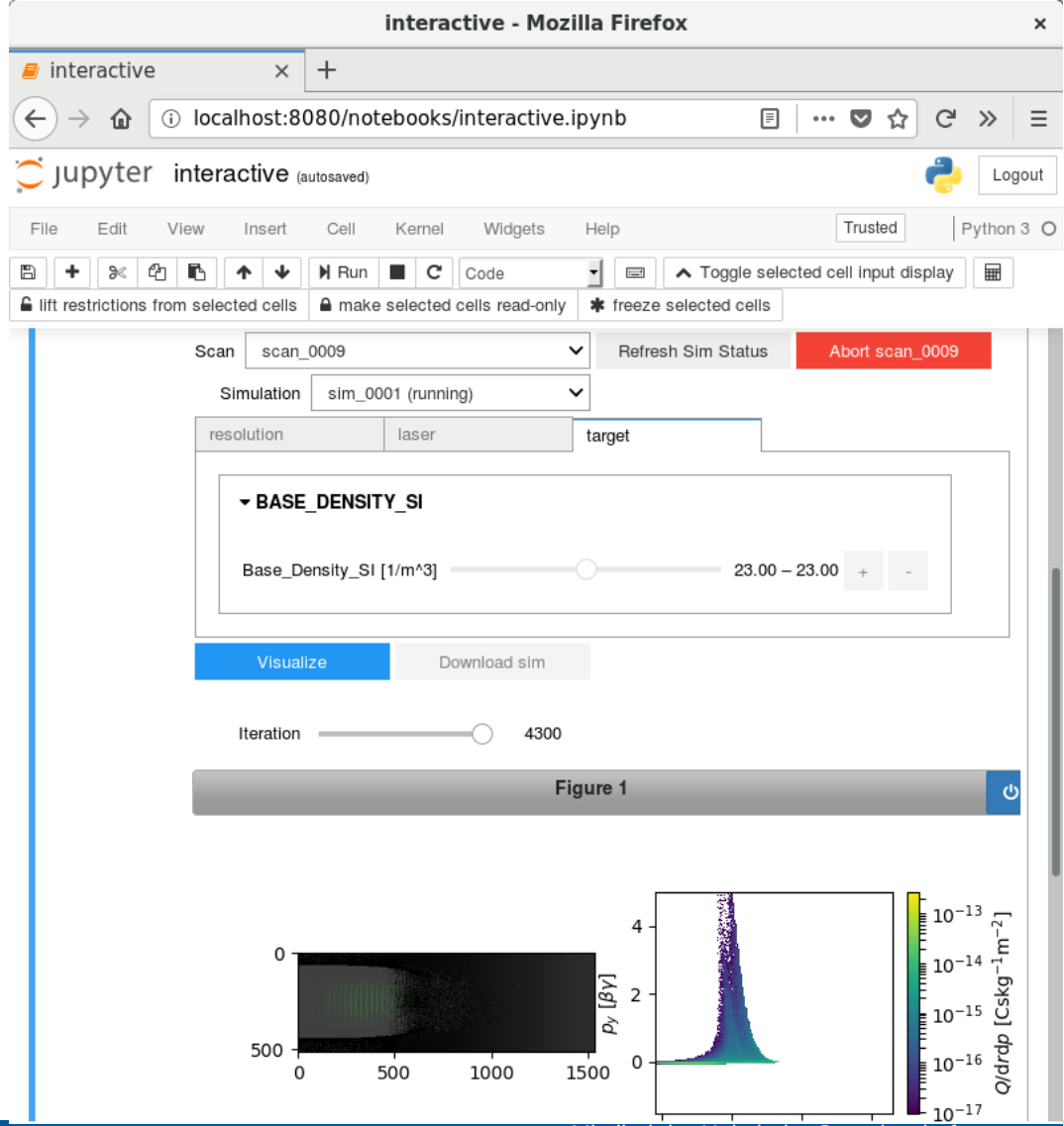


Observe, Correlate & Repeat: Change Questions
Later: Add ML, auto-tuning,...

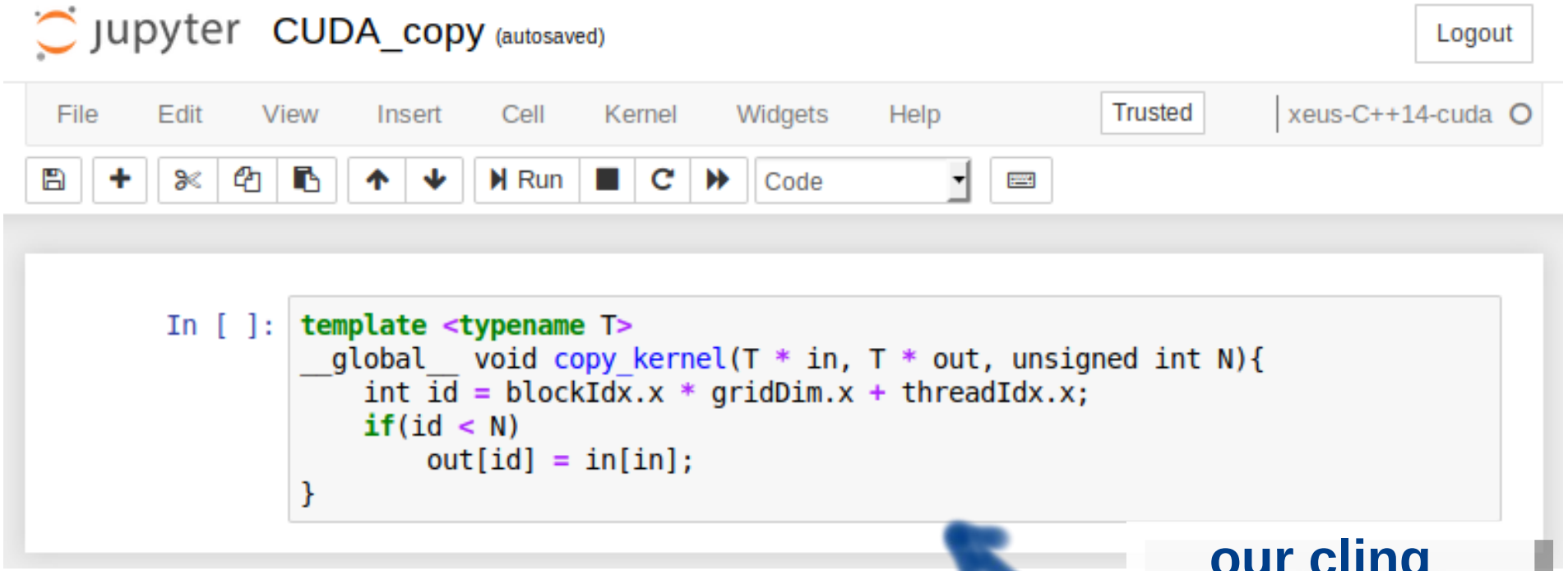
Dr. Guido Juckeland I g.juckeland@hzdr.de I C

Data-Driven, Interactive Exploration

- **data** (reduced or not)
  open meta standard
- **scientific questions**
  python & GUI scripts
  transparent (hackable)
  Jupyter integration
- **installability**
  containerization
  package manager

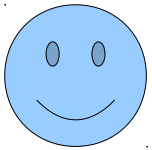


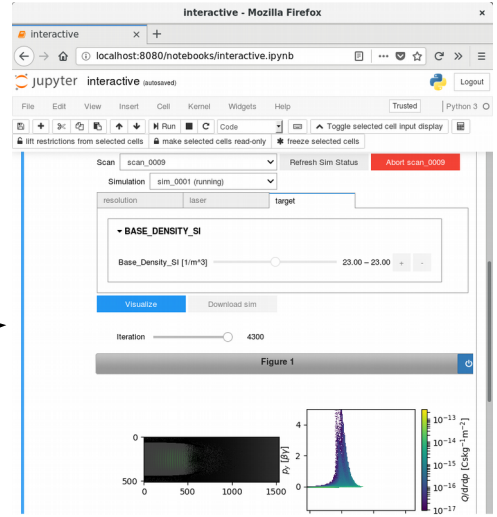
Interactive JIT CUDA/C++/...

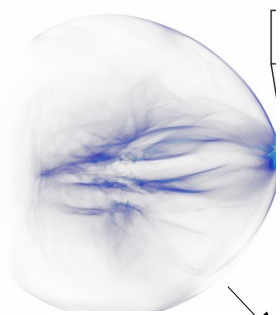


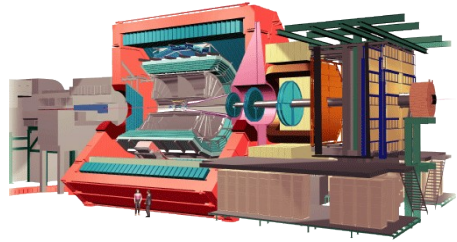
Cling CUDA: S. Ehrig (HZDR, TU Dresden), Diploma Thesis (2018)

Bringing it all together



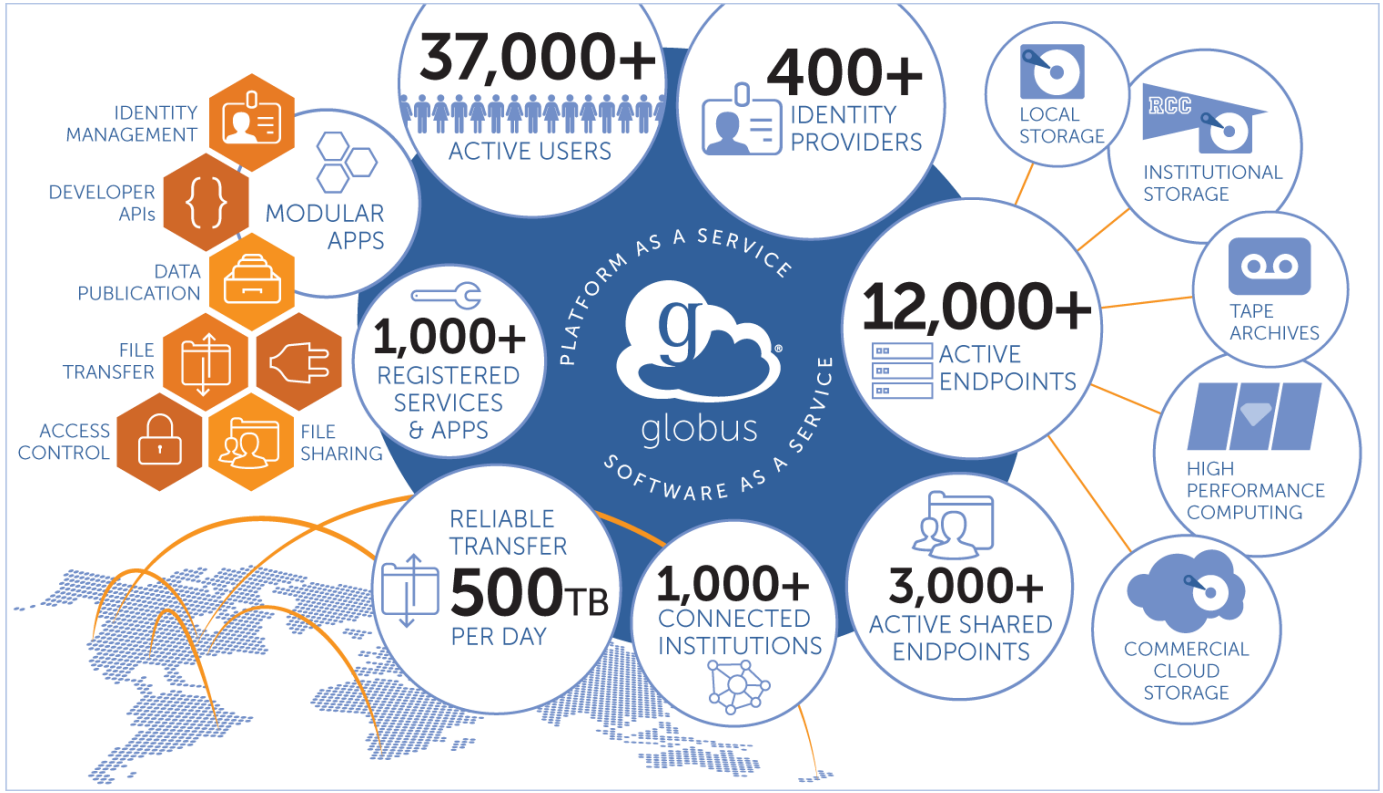






Surprise: Others call it Cloud Computing

Data Transfers between Data Centers



Summary + Outlook

- You want to reduce data (not blindly compress)
- It‘s hard to know the right <0.01% beforehand in explorative science (bleeding-edge experiments and simulations)
- The data will come (sooner than later)
- Keep data „in-flight“
- In-situ techniques (with or without a human) help producing „good“ data sets
- Use same software stack for simulation and experiment evaluation

- **Moving data is expensive** – do it at least asynchronously