Planning ahead data management at the Extreme-Light-Infrastructure

State-of-play and open questions

F. Gliksohn, ELI Delivery Consortium, Associate Director

Laserlab-Europe workshop, Data Handling and Open Data

7 December 2018, Lisbon, Portugal

ELITRANS has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 676627

What is ELI?

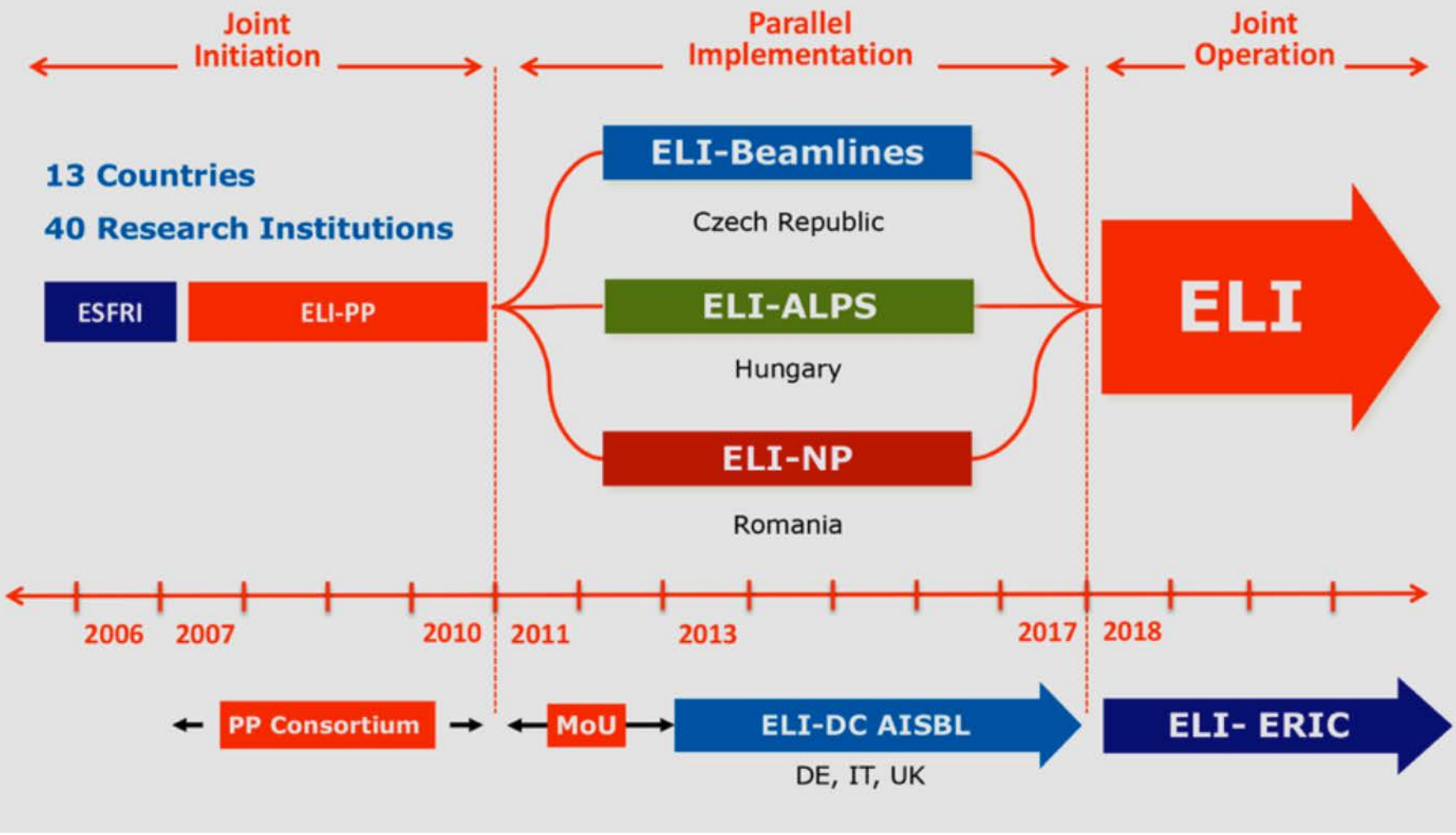
- ELI ERIC will be the world's most advanced **laser research infrastructure**
- First **international research infrastructure** built completely in **Central Europe**
- Funded in synergy between **ESIF, National and Framework** funds, after **EU approval** and delivered through a community effort

Current Status

- Host Countries are delivering ELI facilities **on time and budget**
- **Transitioning** from Implementation (Construction) to **Operations**
- The **three facilities are coming together as a single ELI ERIC**
- Continued **support of the ERIC from the EU during the start-up period is critical** to manage the complex organisational merger.



Timeline

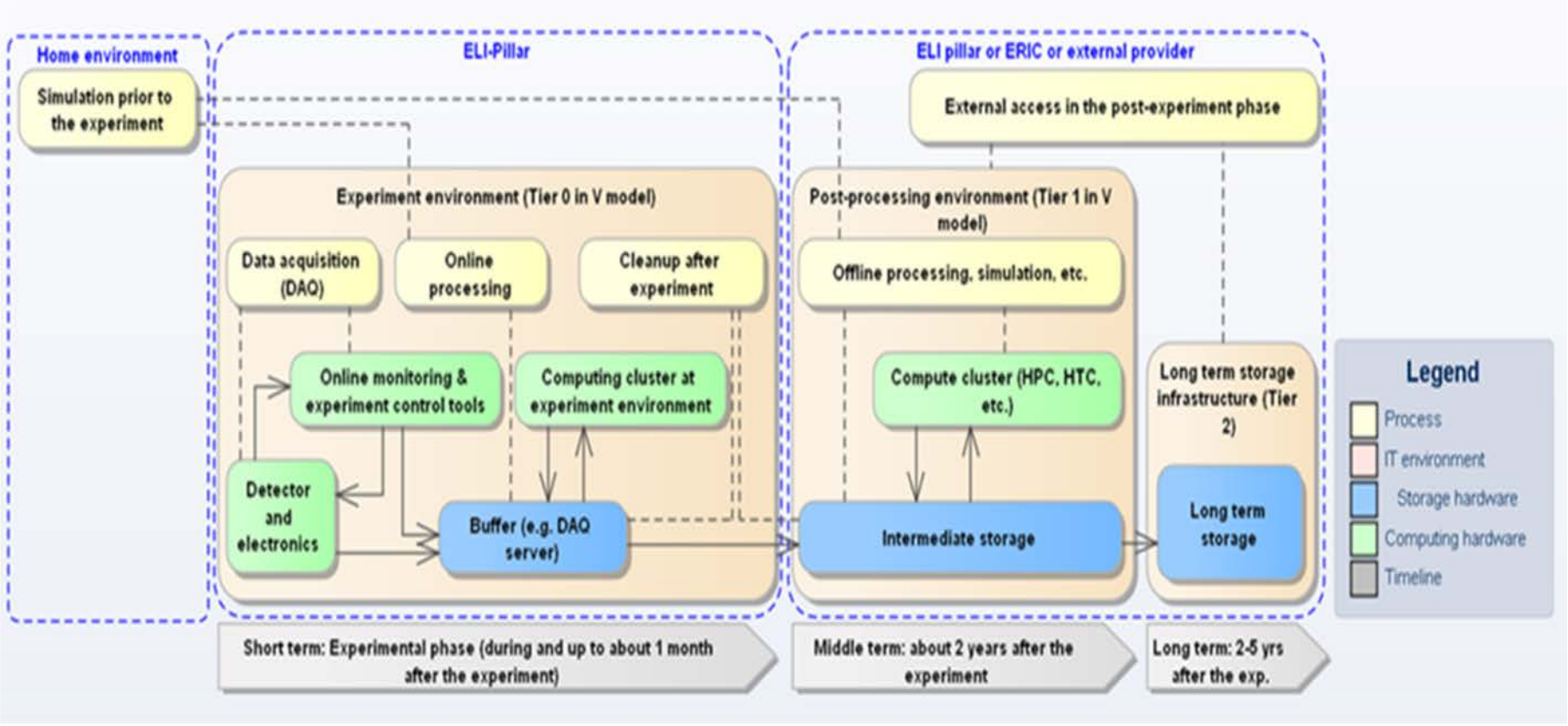

Joint Initiation
Parallel Implementation
Joint Operation
13 Countries
40 Research Institutions
ESFRI
ELI-PP
ELI-Beamlines
Czech Republic
ELI-ALPS
Hungary
ELI-NP
Romania
ELI
2006
2007
2010
2011
2013
2017
2018
PP Consortium
MoU
ELI-DC AISBL
DE, IT, UK
ELI- ERIC

Relevant ELITRANS deliverables

Deliverable	
DL03.01	**Report on ELI-wide Data Management concept**
DL03.02	**Evaluation report on e-Infrastructures**
DL03.03	**Conceptual design report and high-level implementation plan for the common data management service**
DL03.04	**Model for a Virtual Beamline in 3D**
DL03.05	High-level conceptual design and an implementation roadmap for ELI-ERIC's network infrastructure and cyber services
DL03.06	High-level conceptual design, implementation roadmap and integration plan for ELI-ERIC's ERP and User Management systems
DL03.07	High-level conceptual design and an implementation roadmap for the IT infrastructure and IT services of the future ELI-ERIC organisation

Activities and Achievements

- Creation of ELI research workflow
- General (technology-independent) data processing scheme was defined
- Identification of requirements for experiment data management at ELI Facilities



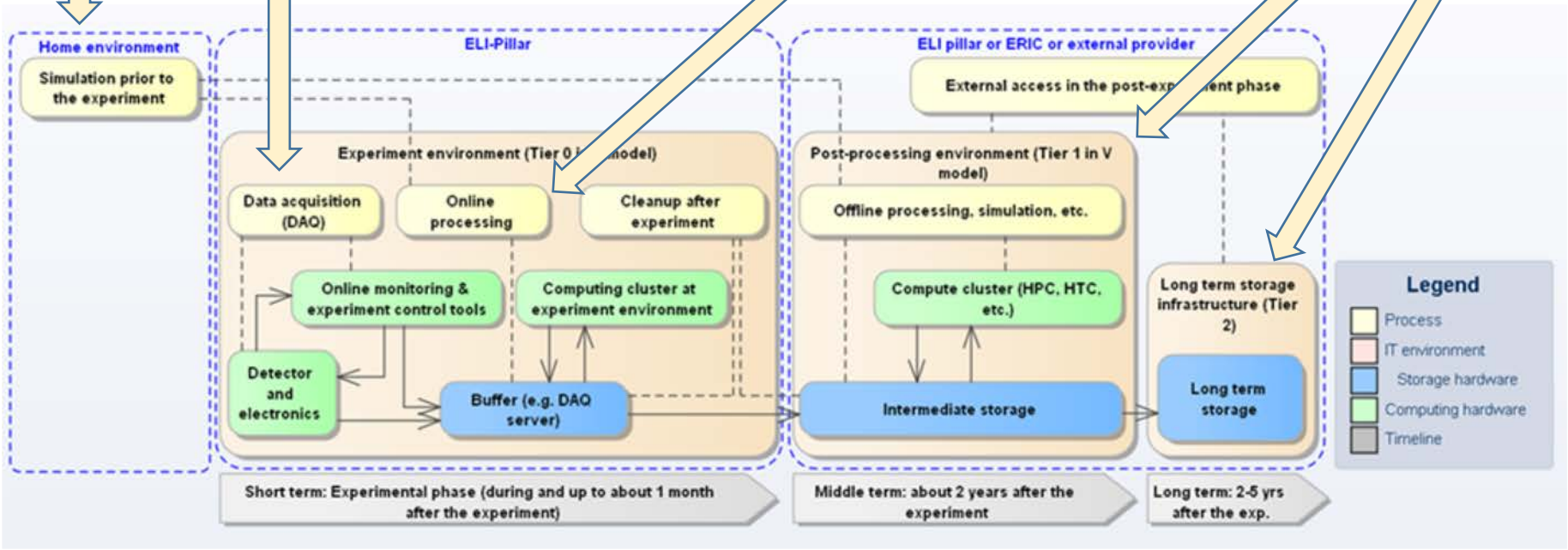
Data and Computing – General Workflow

1. Simulation
Some researchers will run simulation and compares them to experiment results

2. Data Acquisition (DAQ)
Detector data are read and stored for further processing

3. On-Site Evaluation
On-site evaluation can affect the experiment

4. Mid- and long term operations
Researcher must be able to access data after leaving the site. Data should also be stored for longer term



Use Cases

Experiment preparation: Facility maintenance staff member performs the necessary technical preparations on the experiment environment (including configuration of detectors, loading the experiment into the controller, providing access for the experiment team, etc.)

Online processing: During the experiment the experimenter will have some access (depending on the beamline) to data and software tools and perform certain processing on the data gathered during the experiment.

Clean-up after the experiment: This use case will be applied by the facility maintenance staff member. Data are moved to the intermediate storage and the buffer is cleared for the next experiment.

Post processing: This use case will be applied by the experimenter, either at the facility or via remote access. Software provided by ELI to run computations on data stored in the intermediate storage.

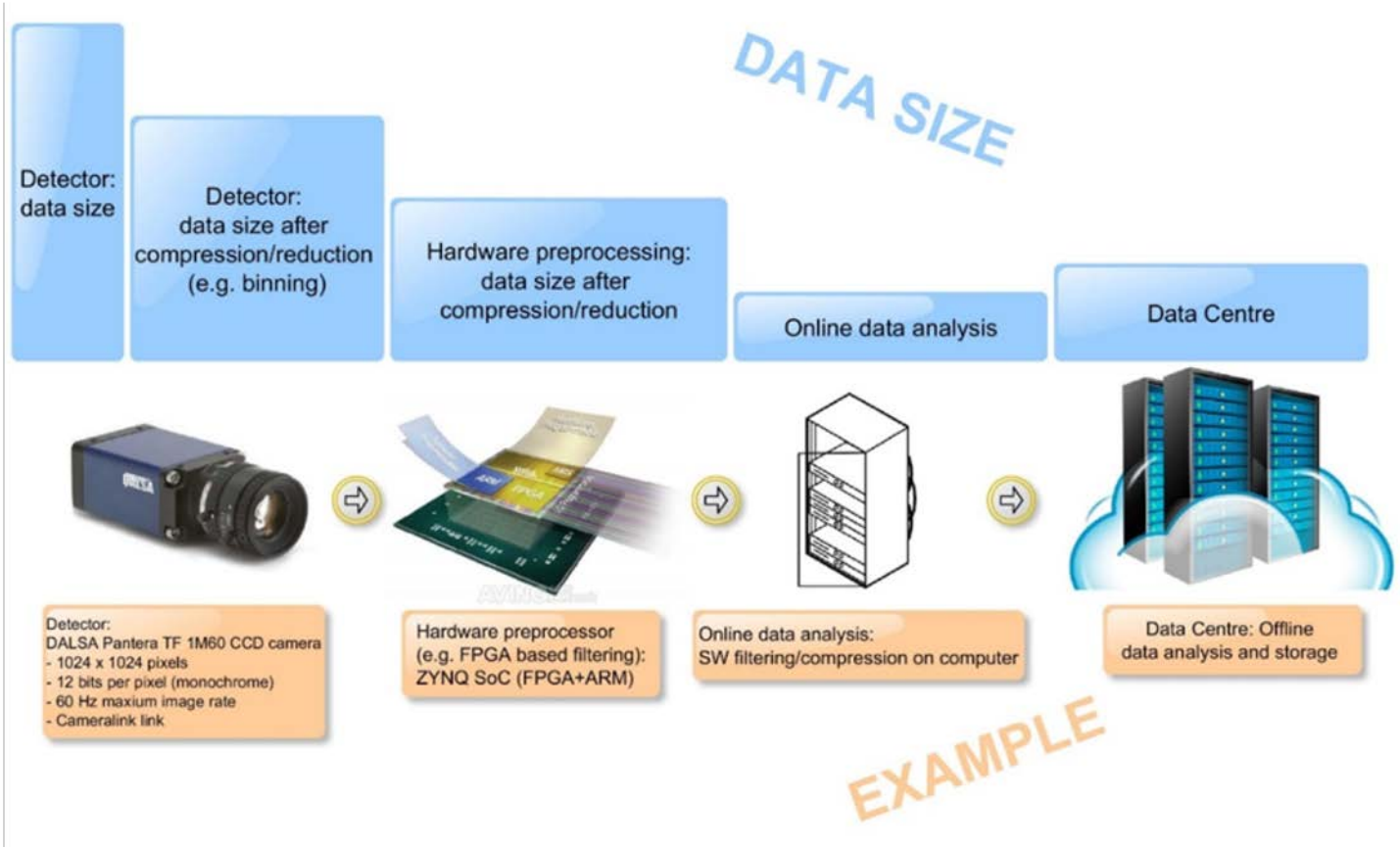
Use Cases

***Transfer of data between intermediate and long term storage*:** Facility maintenance staff (including the automated processes they operate) move data between the two storage type. Transfer from the intermediate to long term storage will happen mostly automatically, a certain period after the experiment. Transfer of other direction might happen upon the experimenter's (or other entitled person) request if it turns out that already archived data must be looked up for further processing.

Lookup and download of data: under conditions controlled by the data policy, experimenter or general public can perform queries on the metadata and may (or may not) download experimental data. We have to note that the facility can limit the access to some preliminary control, e.g. user registration.

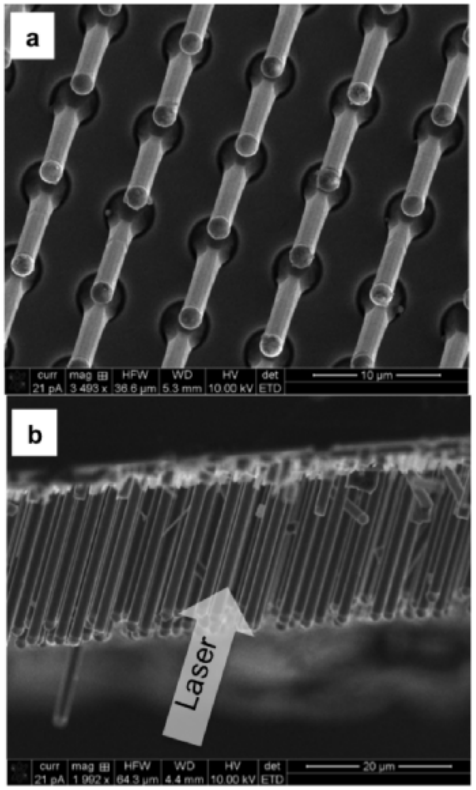
Challenge #1: Reduce Data to a Manageable Size

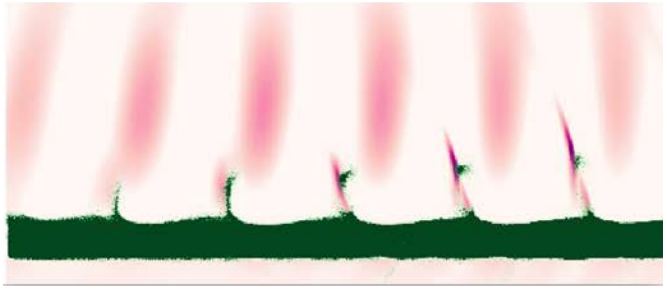
Detectors at certain beamlines will generate huge amount of data. During acquisition and on-site analysis, filtering must be made. After reduction, pillars are expected to archive 2-5 PBytes per year.

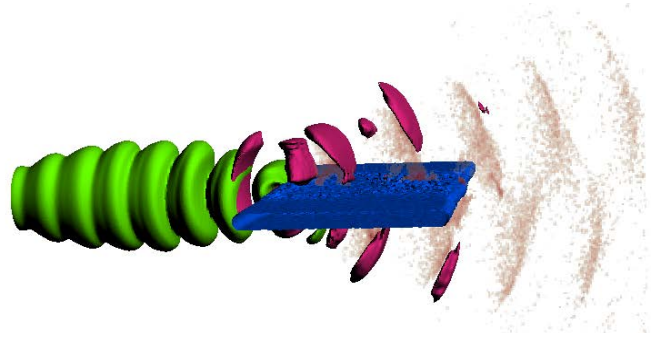


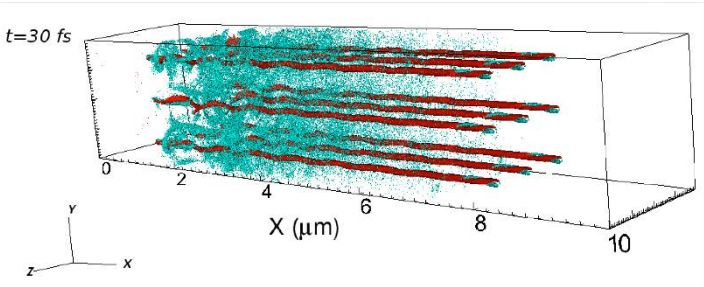
Challenge #2: Need for HPC

Simulation and evaluation of experiment data require thousands of processor cores.





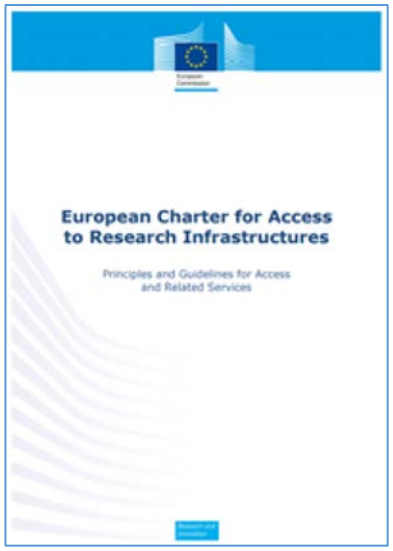




One important application is particle in cell (PIC) method to simulate solid high harmonic generation.

A “routine” task includes the computation on a mesh of ≈100 million cells in 1000 timesteps

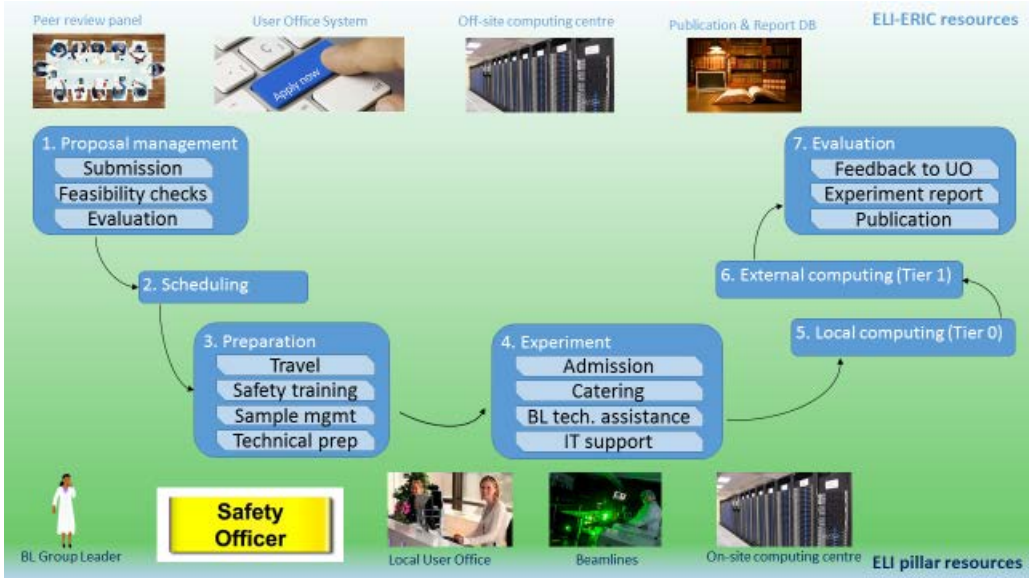
Challenge #3: Whose Data Is It, Anyway?

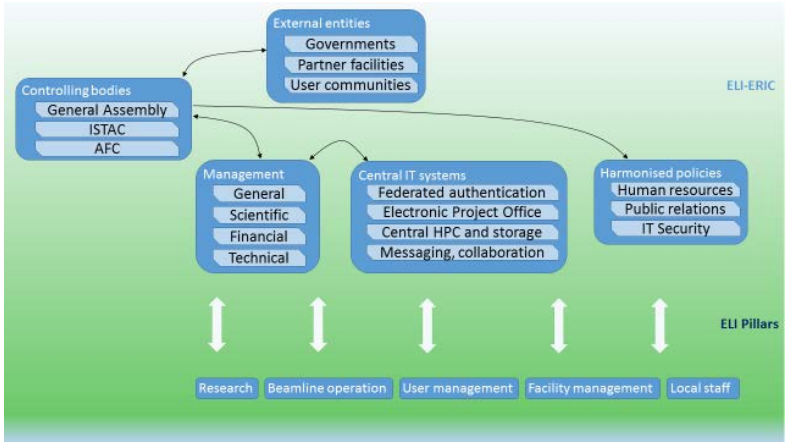


- There is a growing demand for making research data available for the broader scientific community and the general public.
- H2020 funded projects are supposed to contain Data Management Plan according to FAIR principles (**f**indable, **a**ccessible, **i**nteroperable and **r**e-usable research data)
- Research institutions must maintain transparent data management policy (clarify ownership, archiving times, embargo period, etc.)
- Fulfillment of these demands poses legal, organisational and technical challenges

Challenge #4: Two Layers of ELI ERIC

ELI ERIC will be a two-layered organisation. Some services will be managed centrally, others locally at the pillars. This organisational constraint also affects data management plans.





Elements of a Data Policy

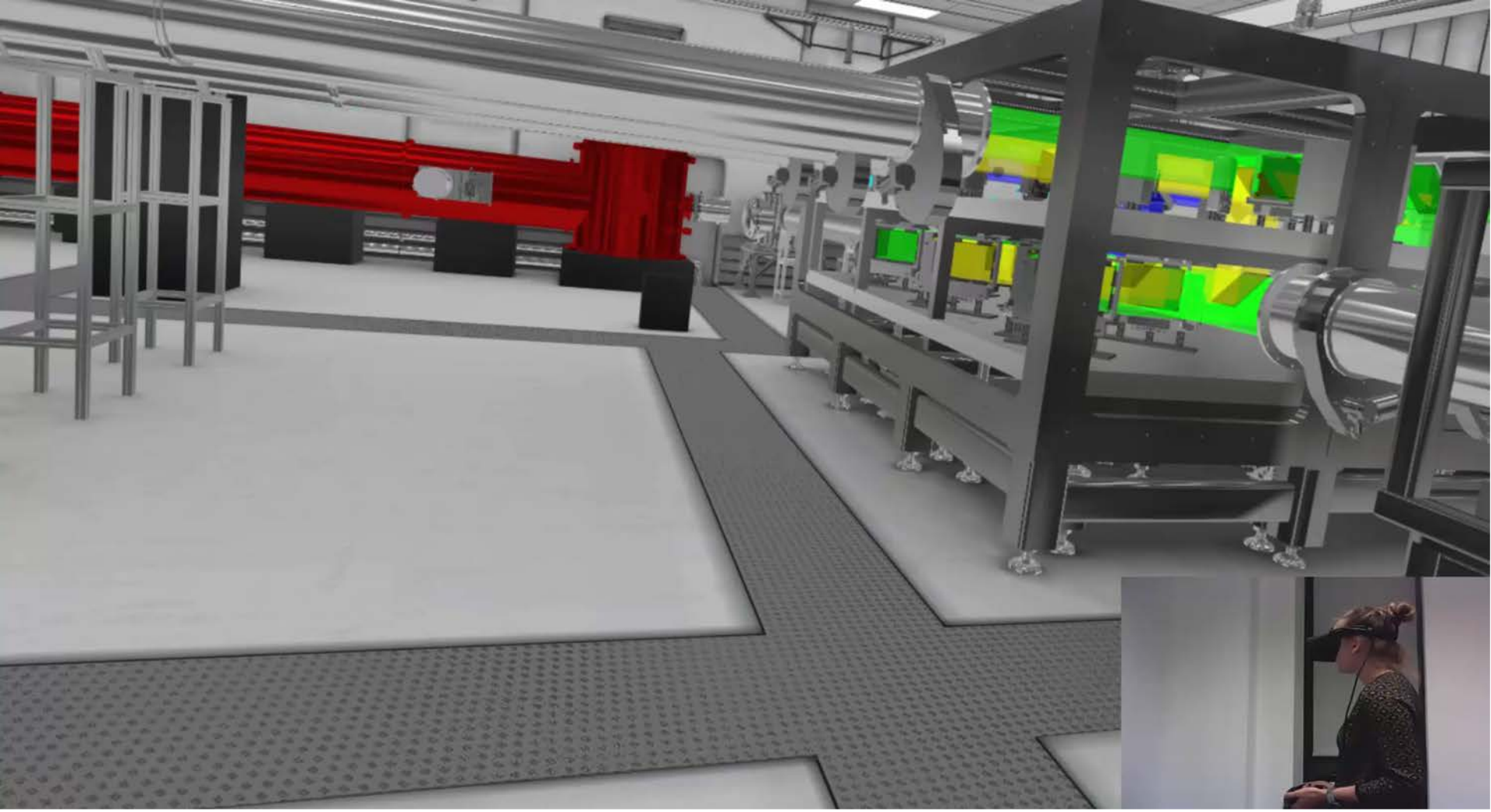
According to Charter, the research institution is recommended (but not obliged) to

- Define a transparent data policy, which is accepted by the users upon submitting the proposal
- Maintain a data management plan
- Encourage the open access of the data after a certain embargo period (during which the experimenter has exclusive access to the data and can publish results based on them)

Elements of a Data Policy

Regulations regarding data stewardship vary across RIs, however, a growing number of institutions begin to apply the following practice:

- After the experiment, the research institution is the custodian (but not the owner) of the data and as such, it is obliged to keep an archive of the data.
- The research institution also keeps a catalogue of metadata and makes this catalogue searchable (for a limited set of users or the general public, depending on the policy).
- Data ownership remains at the experimenter who also has the publication rights for the data. It must be noted, however, that the ownership has some limitations. For example, it does not extend to a right to delete data.
- After a certain embargo period (typically 2--5 years, depending on the field of science) the research institution has to make the data public. From then on, data can be publicly used, however, the experimenter still retains the copyright of data under (for example) CC-BY-x license).

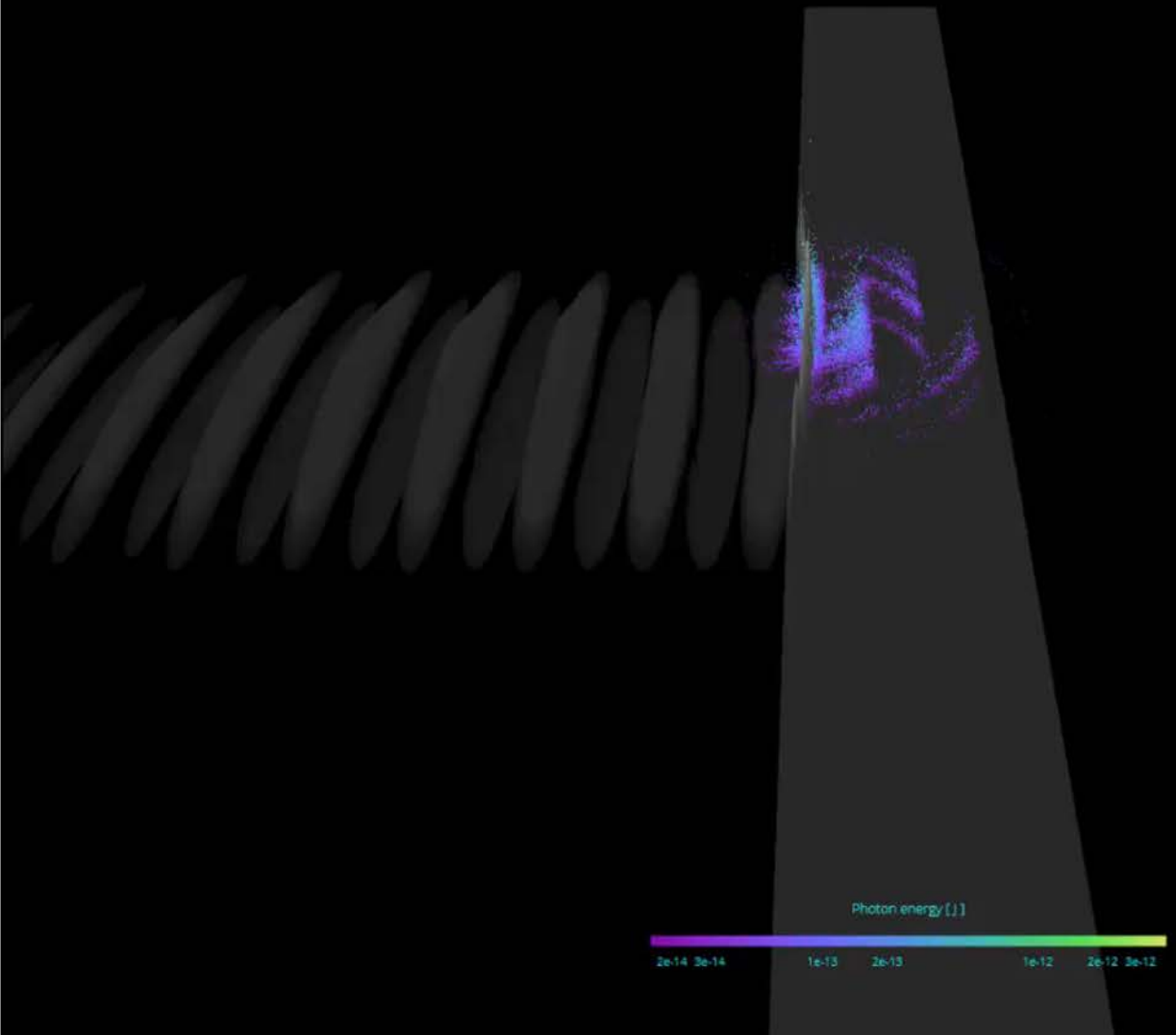


- Virtual and augmented reality applications
- Experimental halls with fully detailed models of scientific equipment and setups of experiments
- Data optimization of large CAD datasets
- Prototypes in works: interactive experimental layout configurator, in-house developed Web version with advanced data and performance optimization, integration with simulations

Laser produced gamma rays

Electrons accelerated from a thin foil irradiated by an ultra-intense laser pulse emit energetic photons in the gamma region. The pulse rips electrons out of the initially solid target into the free space in front of it. While the laser pulse still burns through the target, the fast electrons interact with its strong electromagnetic field, producing energetic photons through the process of quantum radiation reaction - a peculiar kind of self-force stemming from higher-order corrections to the Lorentz force only observable in the presence of extremely strong electromagnetic fields, such as those of produced by the ELI lasers.

VR mode is OFF



Key open questions

- Cohesive vision and strategy starting from what level of service is needed
- Who is the user? What data services are needed?
- Fragmentation of staff involved in data-related activities within organization
- Cultural issue among staff and within community
- Standardisation challenge (approach to metadata, diagnostics, etc.)
- Leveraging ELI's and Laserlab's critical masses to have community and "trickle-down" effects
- Readiness of ELI to engage with Laserlab within dedicated working group and also to ensure connection with PaNOSC and dissemination of results (observer status?)

THANK YOU
FOR YOUR ATTENTION!