High performance and high capacity storage systems

Janusz Malka[1]
Co-auth:
Nasser Al-Qudami[1],Djelloul Boukhelef[1], Lennart Boege[1], Stefan Dietrich[2], Uwe Ensslin[2], Kimon Filippakopoulos[1], Martin Gasthuber[2], Juergen Hannappel[2], Ralf Lueken[2], Luis Maia[1], Maurizio Manetti[1], Birgit Lewendel[2], Carsten Patzke[2], Gianpietro Previtali[1], Janusz Szuba[1], Christian Voss[2], Krzysztof Wrona[1]
[1]EuXFEL, [2]DESY

Online, 27.10.2020

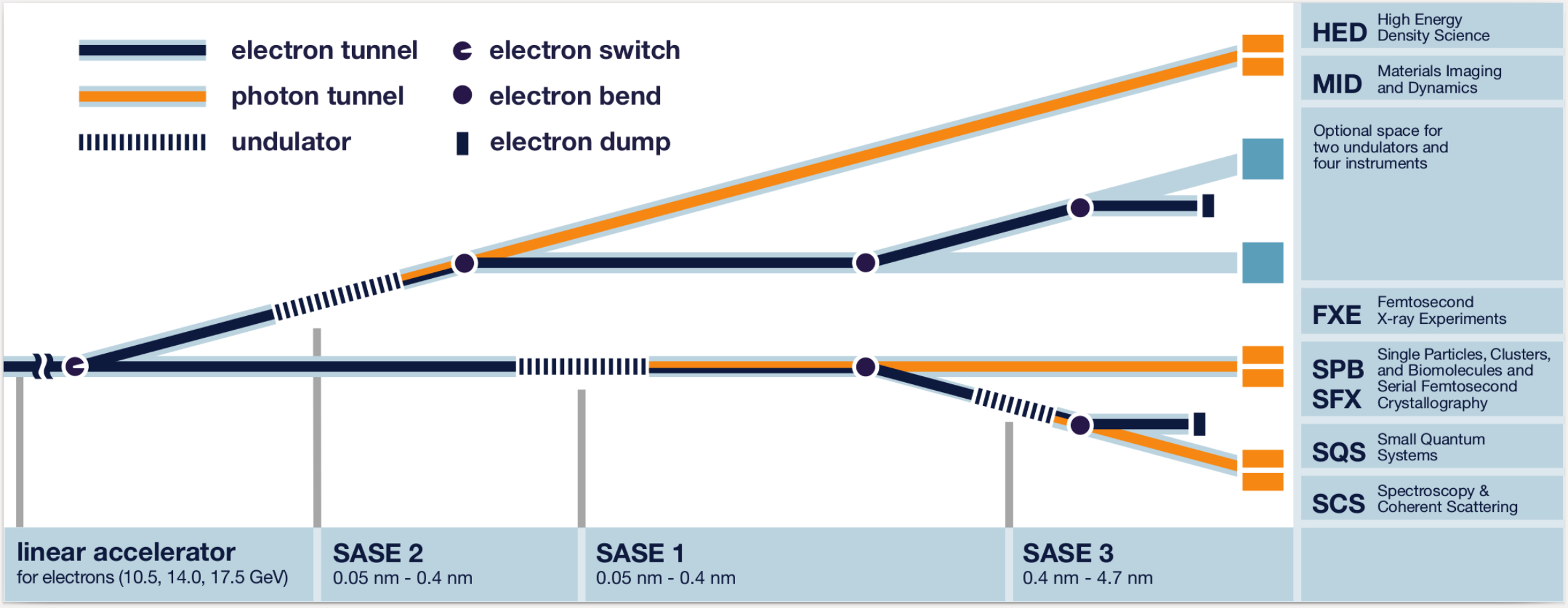
European XFEL – overview

The European XFEL is a research facility which will use high intensity X-ray light to help scientists better understand the nature of matter.

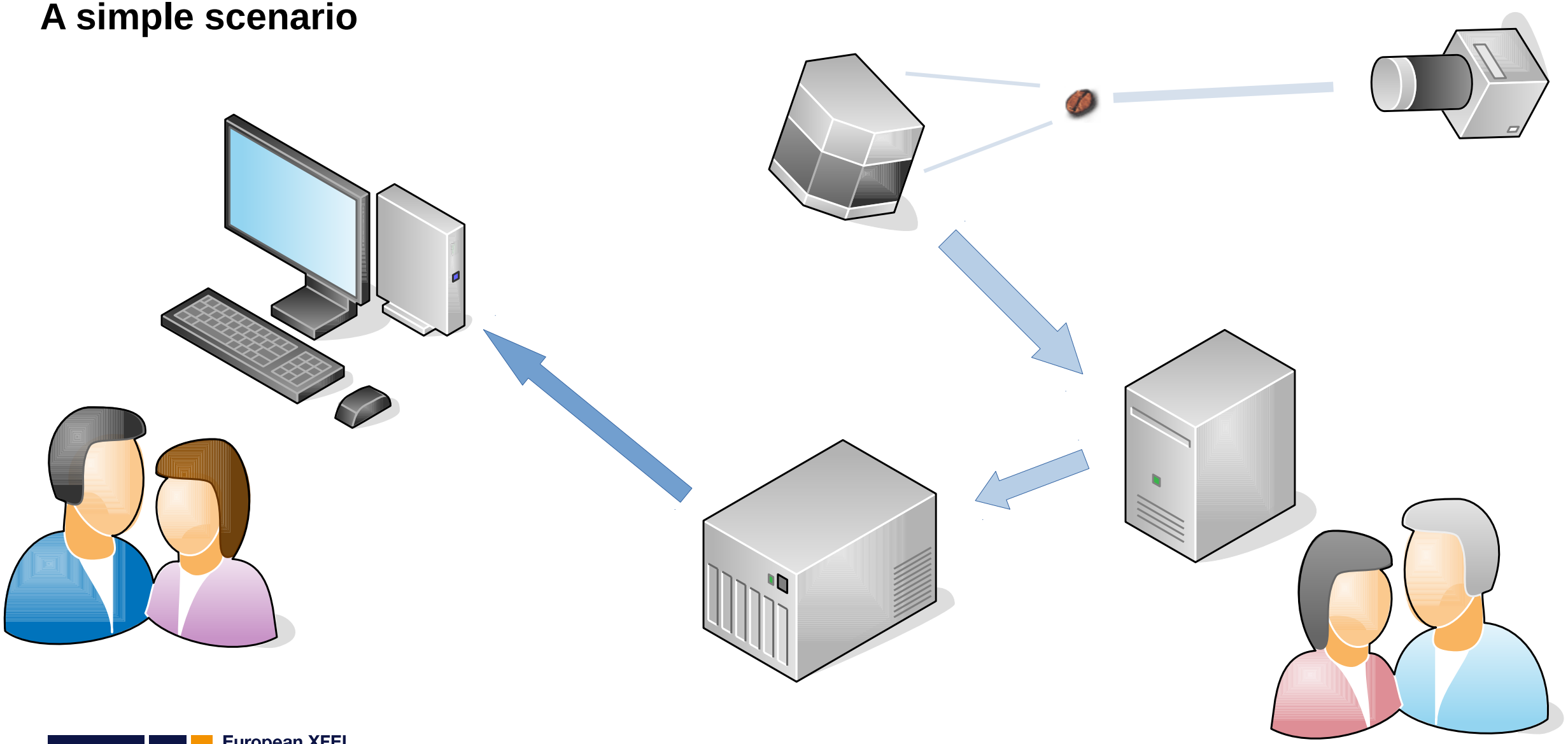
https://youtu.be/Tq5-inrws6k

https://www.xfel.eu/

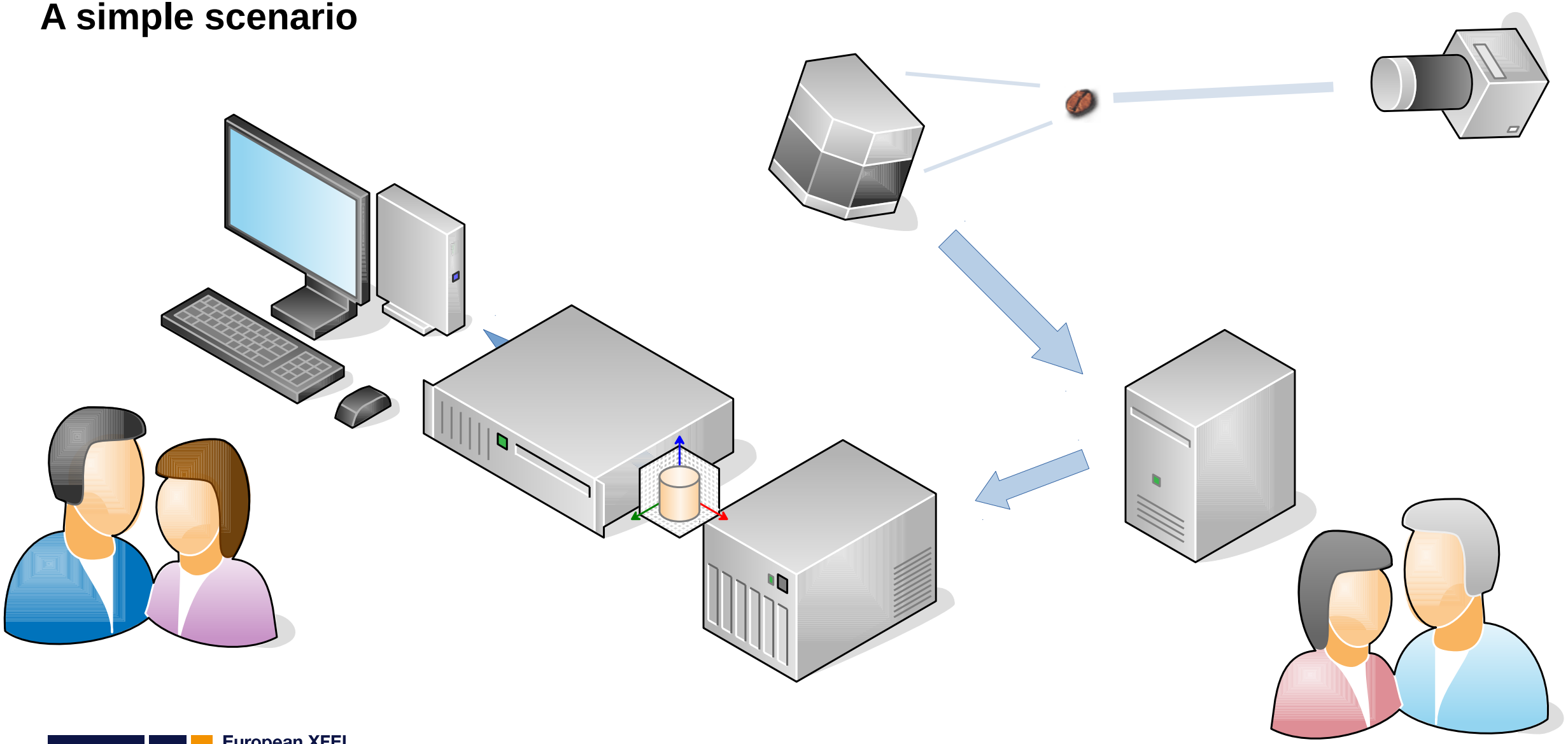
Beam-lines of the European XFEL - Self-amplified spontaneous emission (SASE)

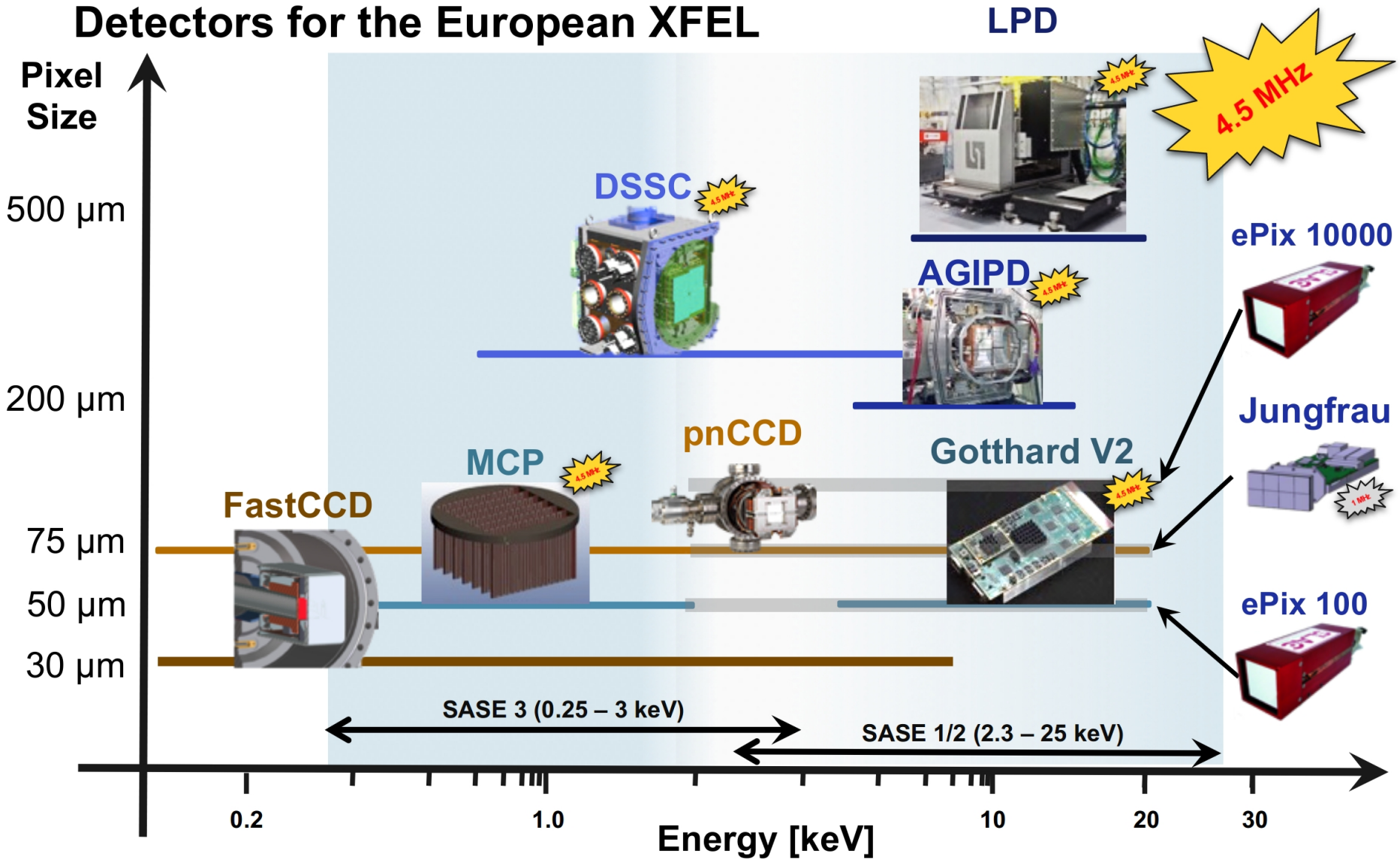


A simple scenario

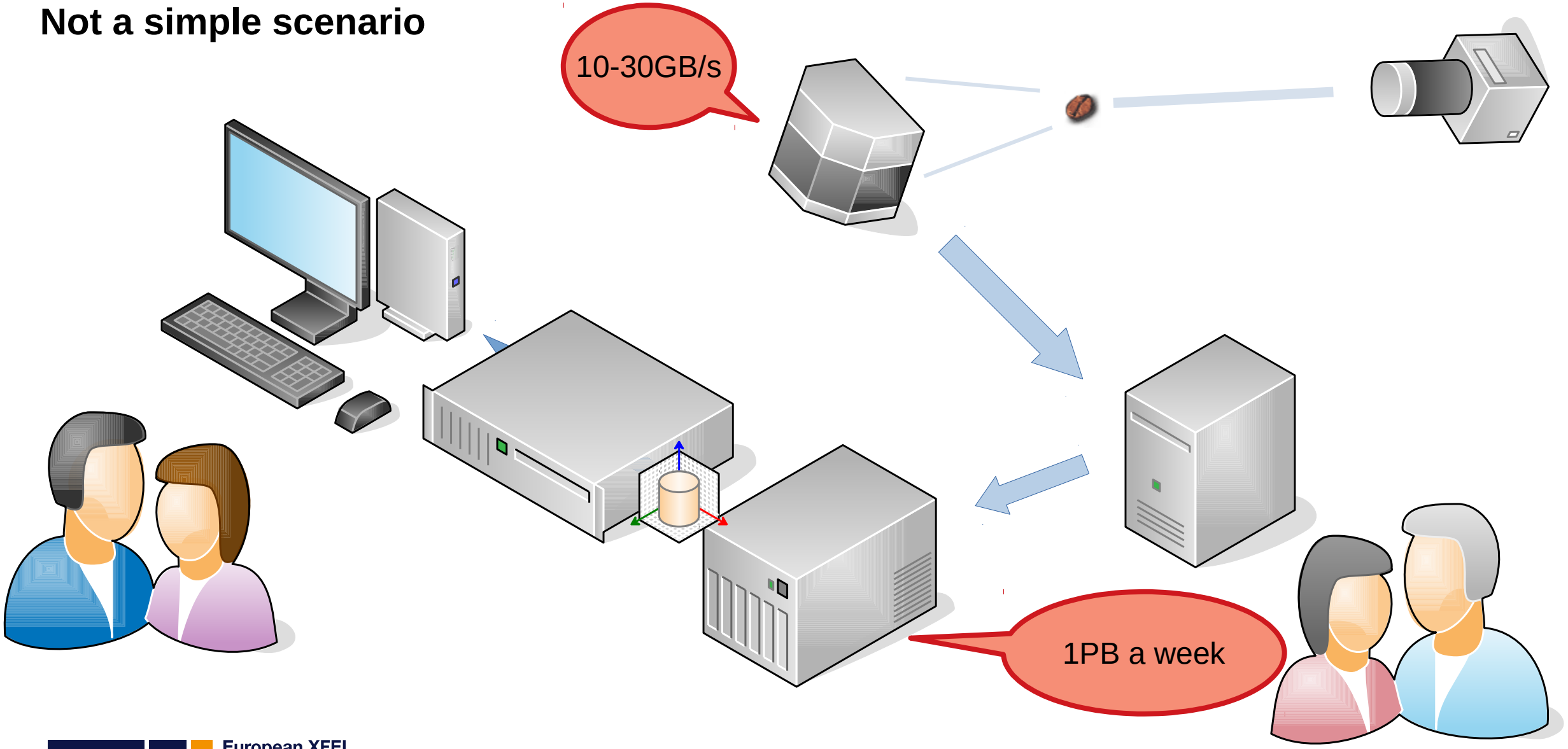


A simple scenario

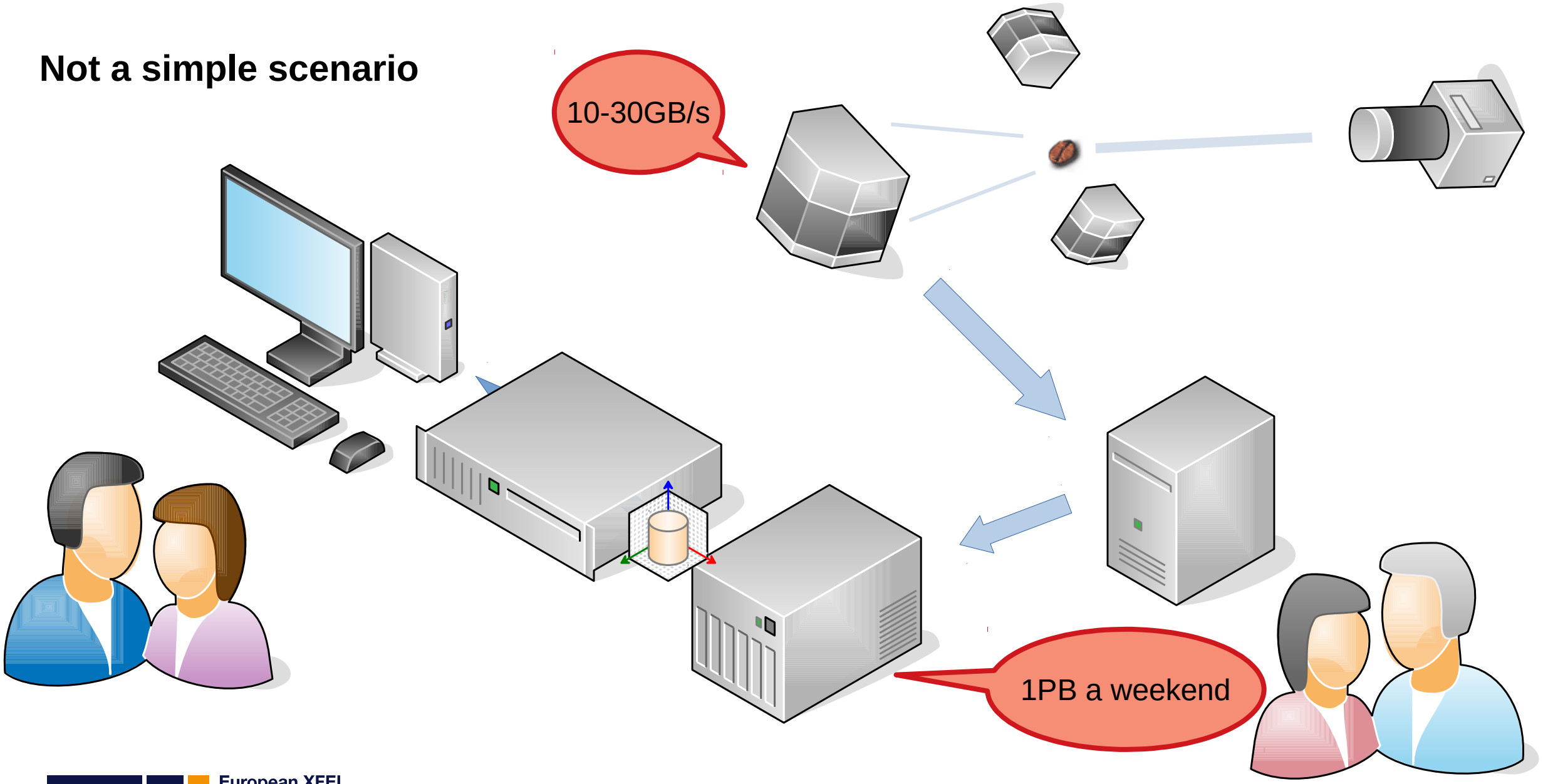



Detectors for the European XFEL
Pixel Size
500 µm
200 µm
75 µm
50 µm
30 µm
LPD
4.5 MHz
DSSC
AGIPD
ePix 10000
Jungfrau
pnCCD
MCP
Gotthard V2
FastCCD
ePix 100
SASE 3 (0.25 – 3 keV)
SASE 1/2 (2.3 – 25 keV)
0.2
1.0
10
20
30
Energy [keV]

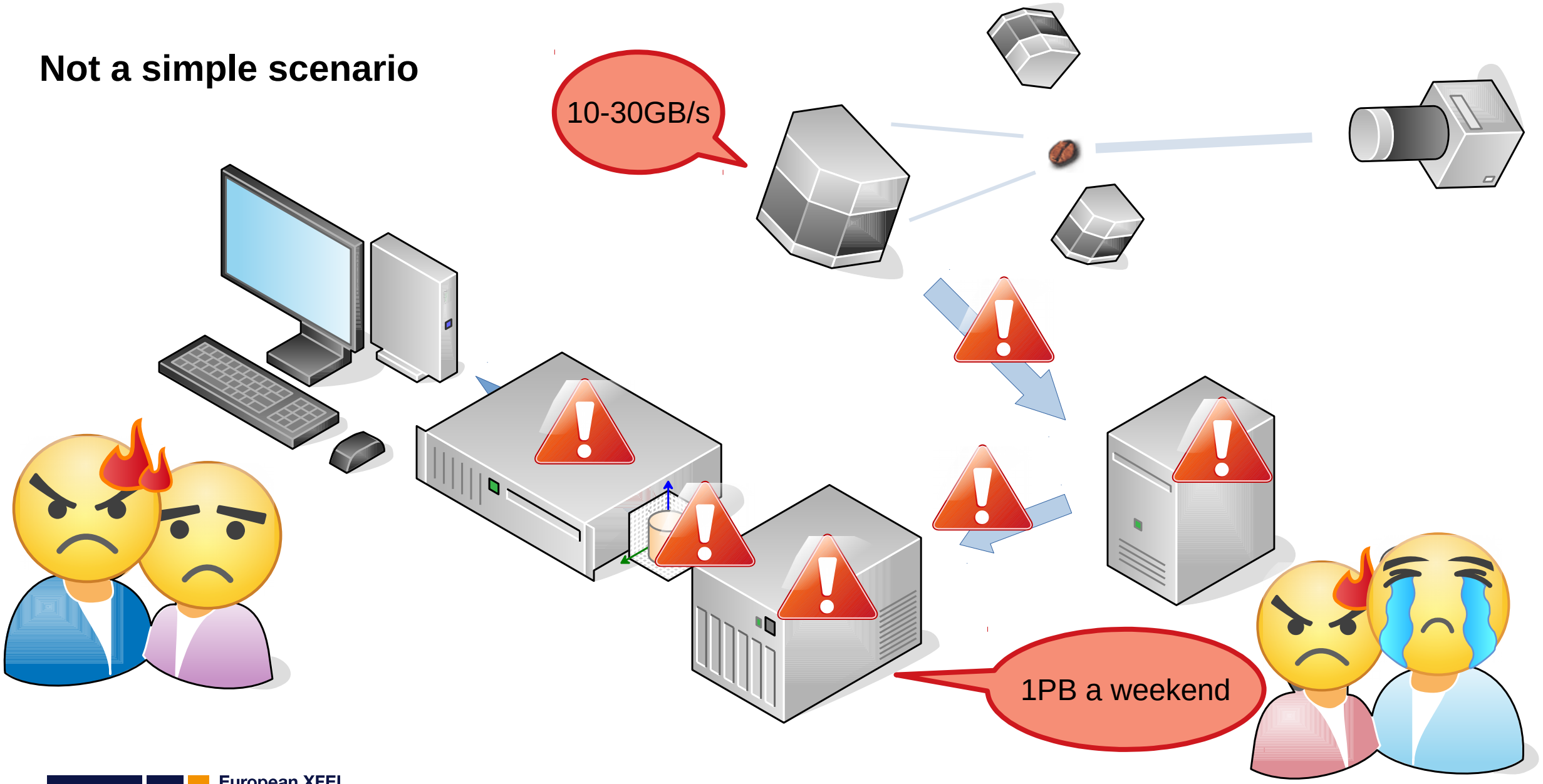
Not a simple scenario



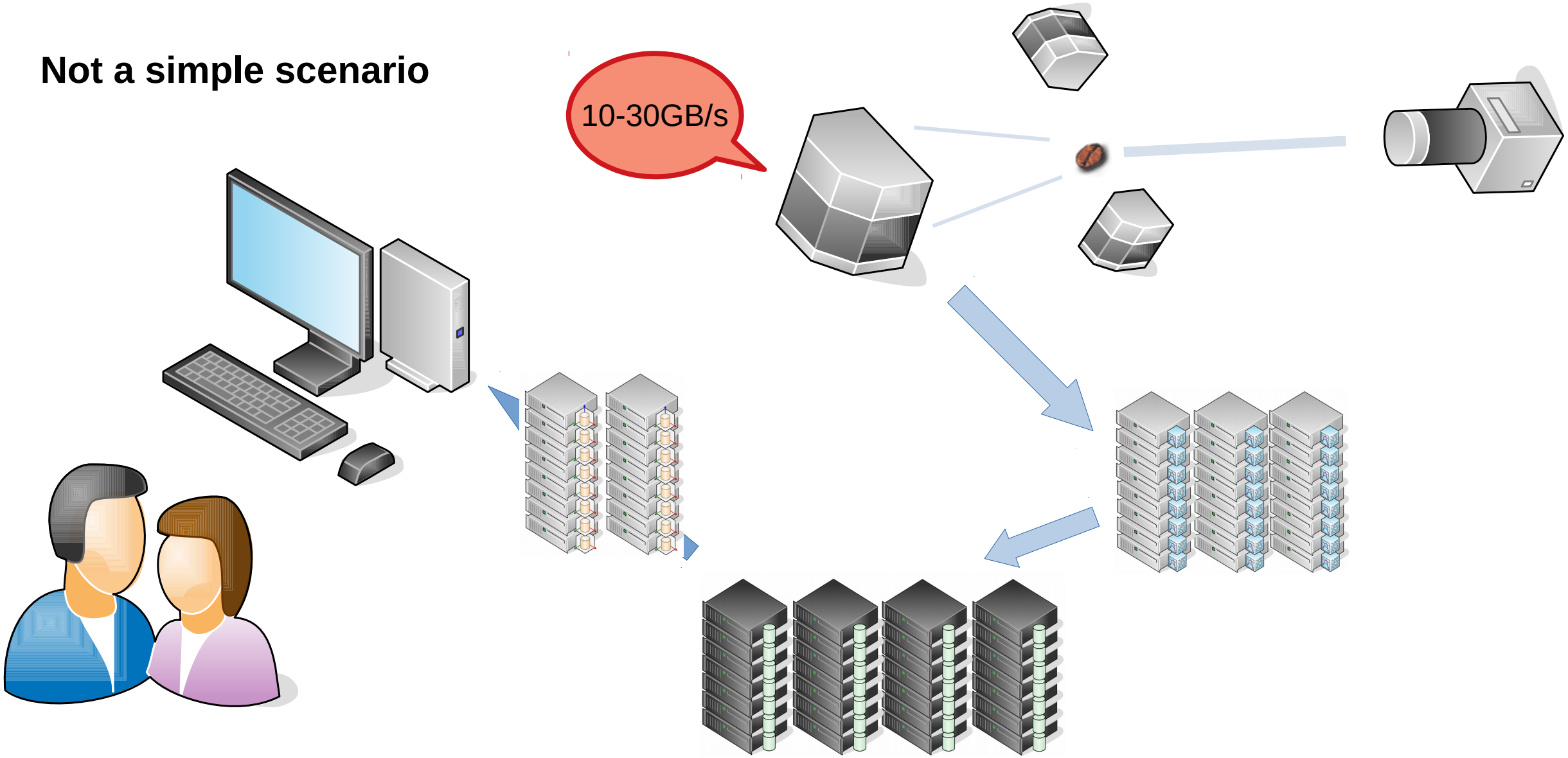
Not a simple scenario



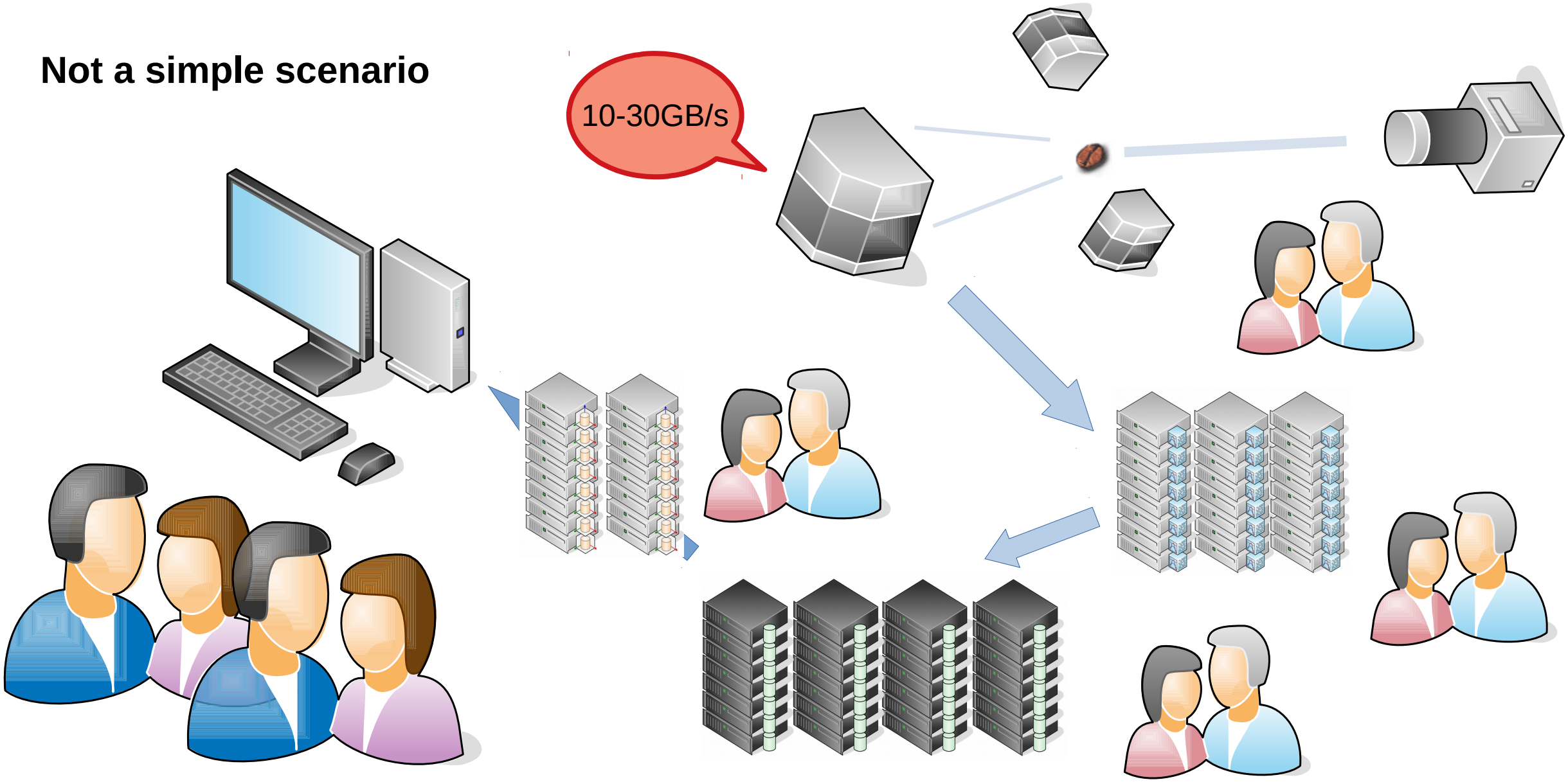
Not a simple scenario



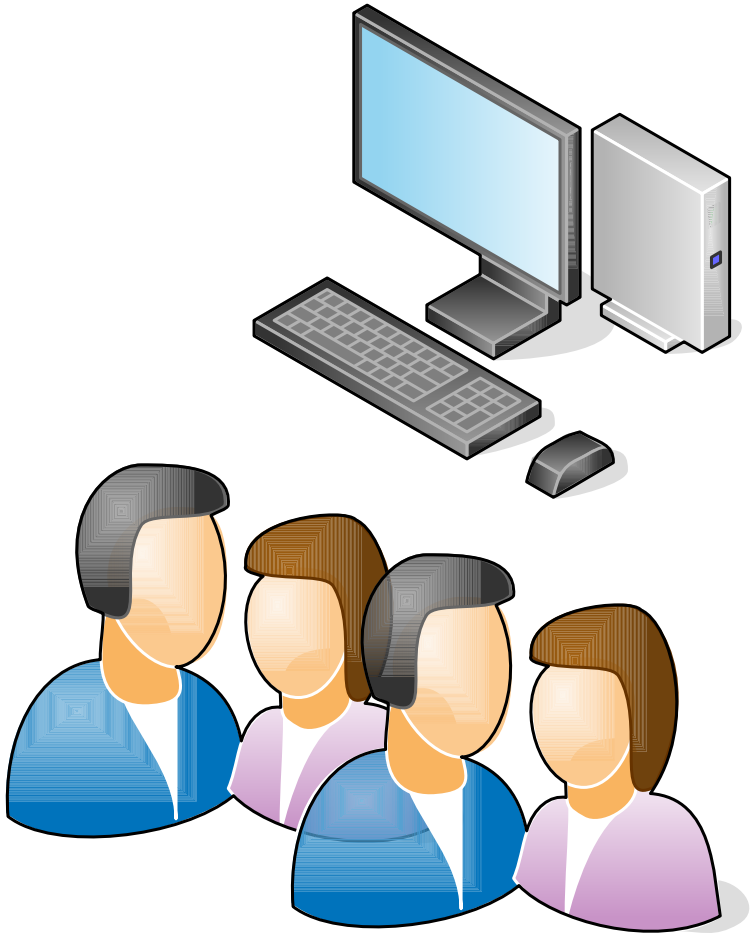
Not a simple scenario

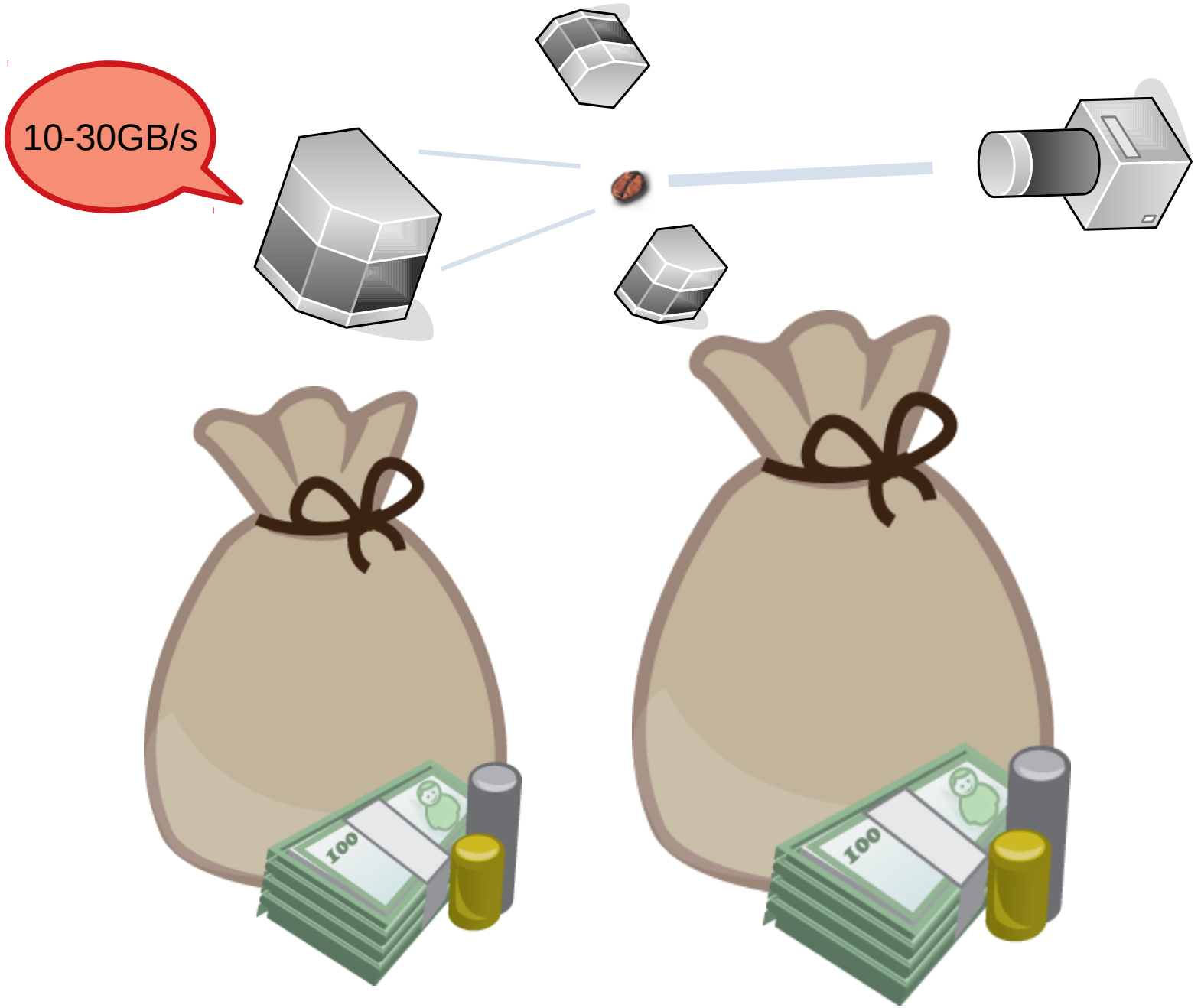


Not a simple scenario



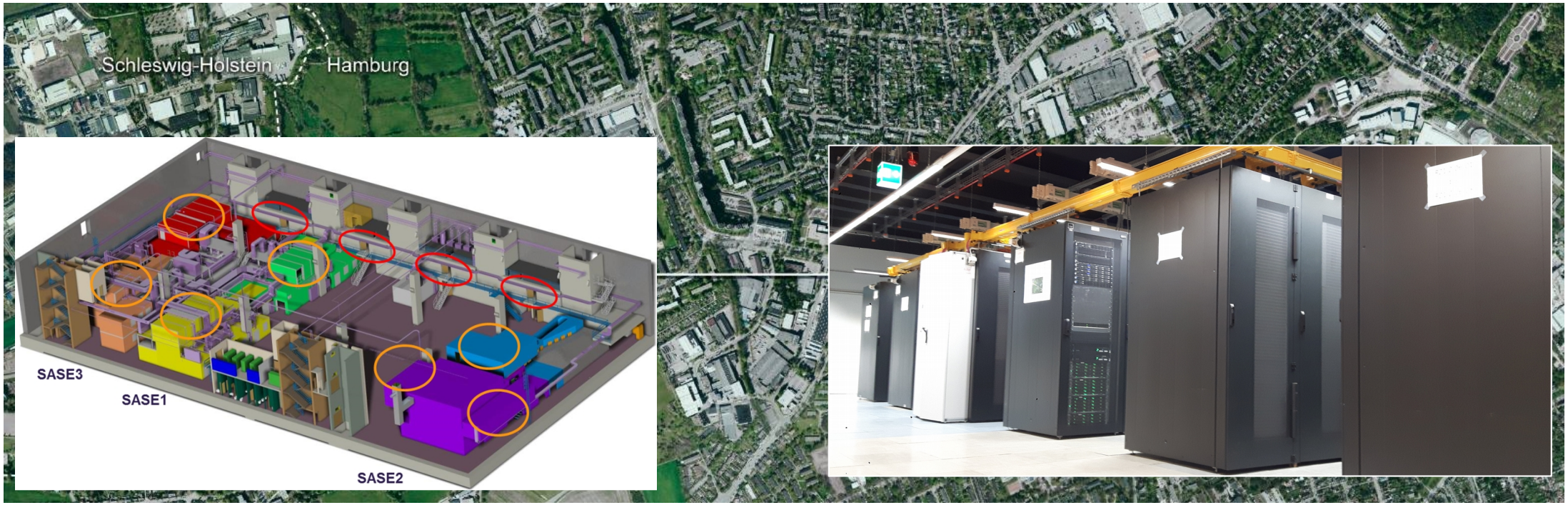
Not a simple scenario





European XFEL – overview

The European XFEL is a research facility which will use high intensity X-ray light to help scientists better understand the nature of matter.



3 computer (balcony) rooms in the experiment hall

EuXFEL hardware in DESY Computer Center

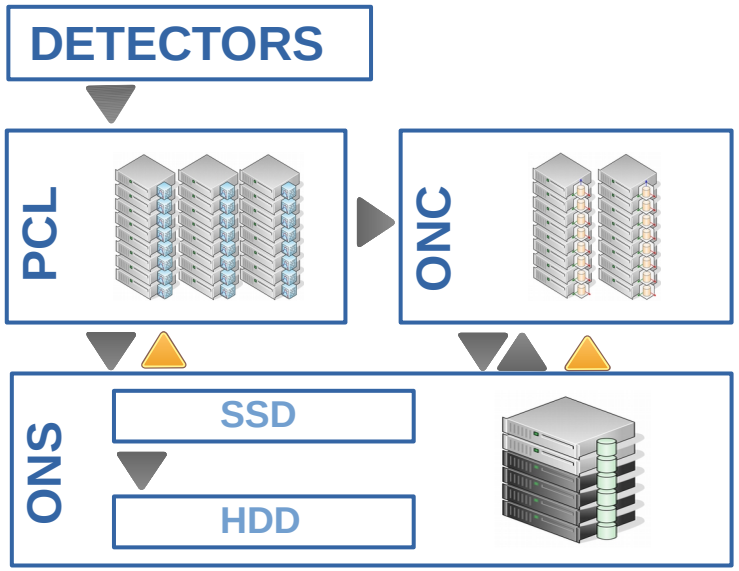
~4430m of optical fiber

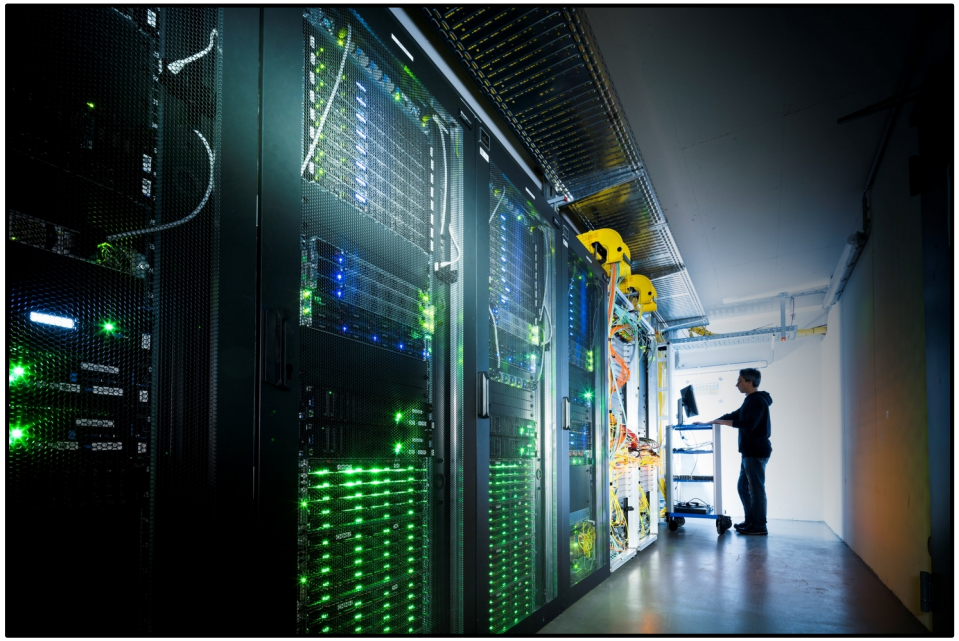
Different dimensions of scientific data

- On-line vs Off-line
  - Time – before, during, after experiment
  - Locality – Schenefeld, DESY
- Access authorization
  - User specific
  - Experiment specific
  - Common for instrument
- Data sources
  - detectors, cameras, sensors ...
  - calibration constants generation
  - data processing
  - user analysis data
- Data life time
  - hours
  - weeks
  - years
- Data safety
  - Backup, Snapshots, Archive
  - Scratch
- Data Size
  - Small files – scripts (kB)
  - Big HDF5 files (GB)
- Data throughput
  - slow data
  - fast data (~20GB/s)
- Data access
  - Globus, NFS, FTP
  - Central storage

On-line infrastructure (per SASE)

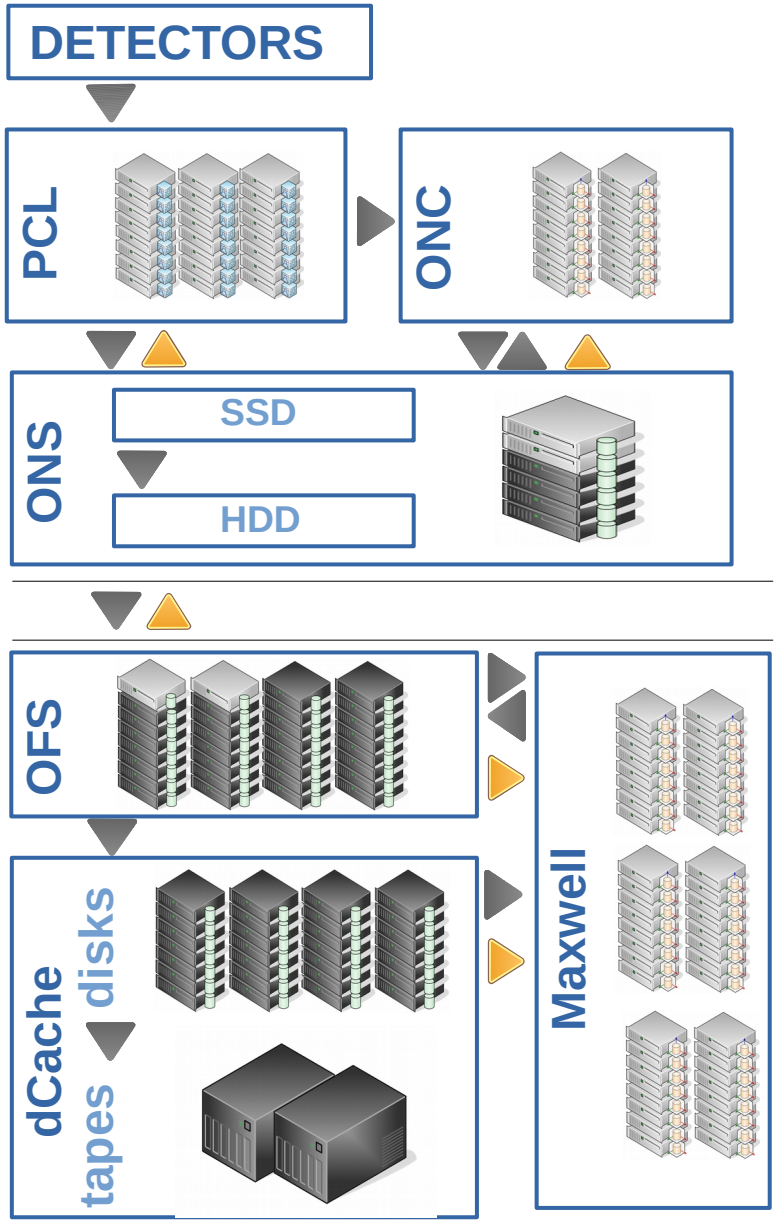
- **PC L**ayer
  - data aggregation from different sources
  - HDF5 file creation
  - data source for the on-line data analysis
- **ON**line **C**luster
  - captures raw data stream from PCL
  - initial on-line calibration of raw data
  - on-line data preview
  - on-line data analysis
- **ON**line **S**torage
  - storage for the raw data from PCL
  - generic storage resource for “on-line users”
  - monitoring and first level clean-up of bad data
  - source for raw data draining to the off-line storage





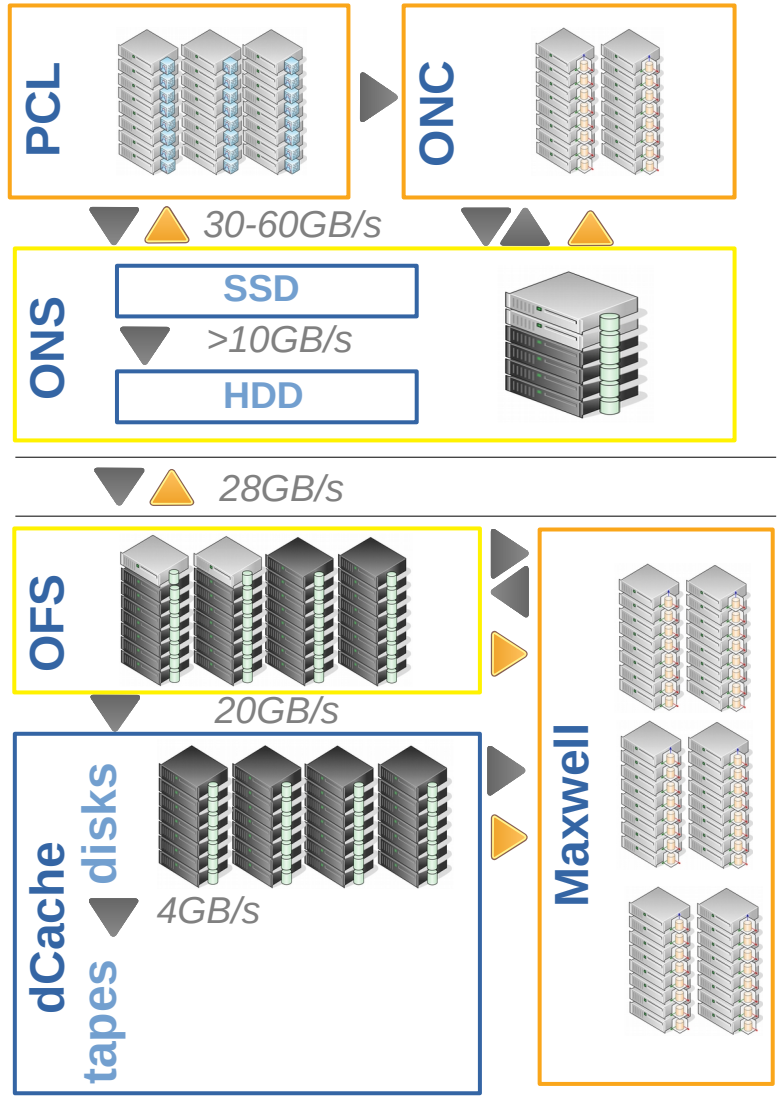
Off-line infrastructure – (common)

- **OF**line **S**torage
  - shared data storage for all SASEs
  - data source for the off-line data analysis
  - source for data transfer to the dCache system
- MAXWELL
  - shared off-line computing cluster
  - generation of calibrated constants
  - raw data reprocessing
- dCache – non GPFS resource
  - storage for the raw data and calibration constants
  - interface to the type archive



Technologies

- IBM Spectrum Scale
  - remote clusters
- IBM Elastic Storage Server
  - on-line and off-line storage
- dCache
  - off-line storage
- Interconnection
  - Mellanox InfiniBand Switches



IBM Elastic Storage Server and IBM Spectrum Scale (GPFS)

- IBM Elastic Storage Server
  - GPFS Native RAID (Erasure Coding)
  - Short rebuild times
  - End-to-end check-summing for data integrity
  - up to 30GB/s of throughput for single building block
- The following GPFS features are used to manage the current storage:
  - Spectrum Scale Policy
  - Callback events
  - Quality of Service
  - Independent, dependent file-sets, snapshots, quotas
  - NFSv4ACL
  - Independent writer (IW) Active File Management (AFM) mode
  - Cluster Export Services (CES) for NFS export
  - Grafana Bridge

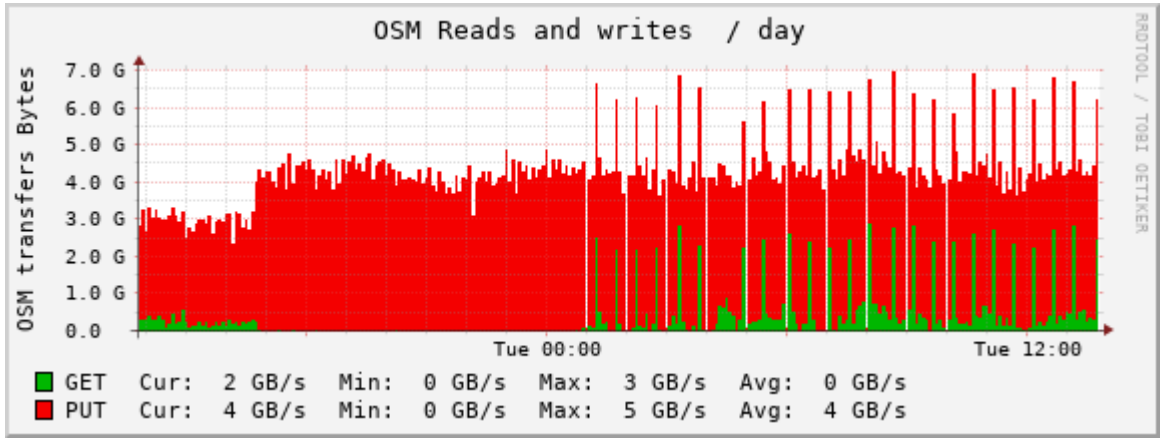
IBM Spectrum Scale numbers

- On-line
  - 12 GPFS clusters
  - 28 file systems
  - 18 storage servers
  - 175 clients
  - available space :
    - 3.6PB SASE1
    - 1.8PB SASE[2,3]
    - 0.2PB LAB
- Off-line
  - 2 GPFS clusters
  - 7 file systems
  - 22 storage servers
  - > 600 clients
  - 16.5PB (+11PB this year)

dCache

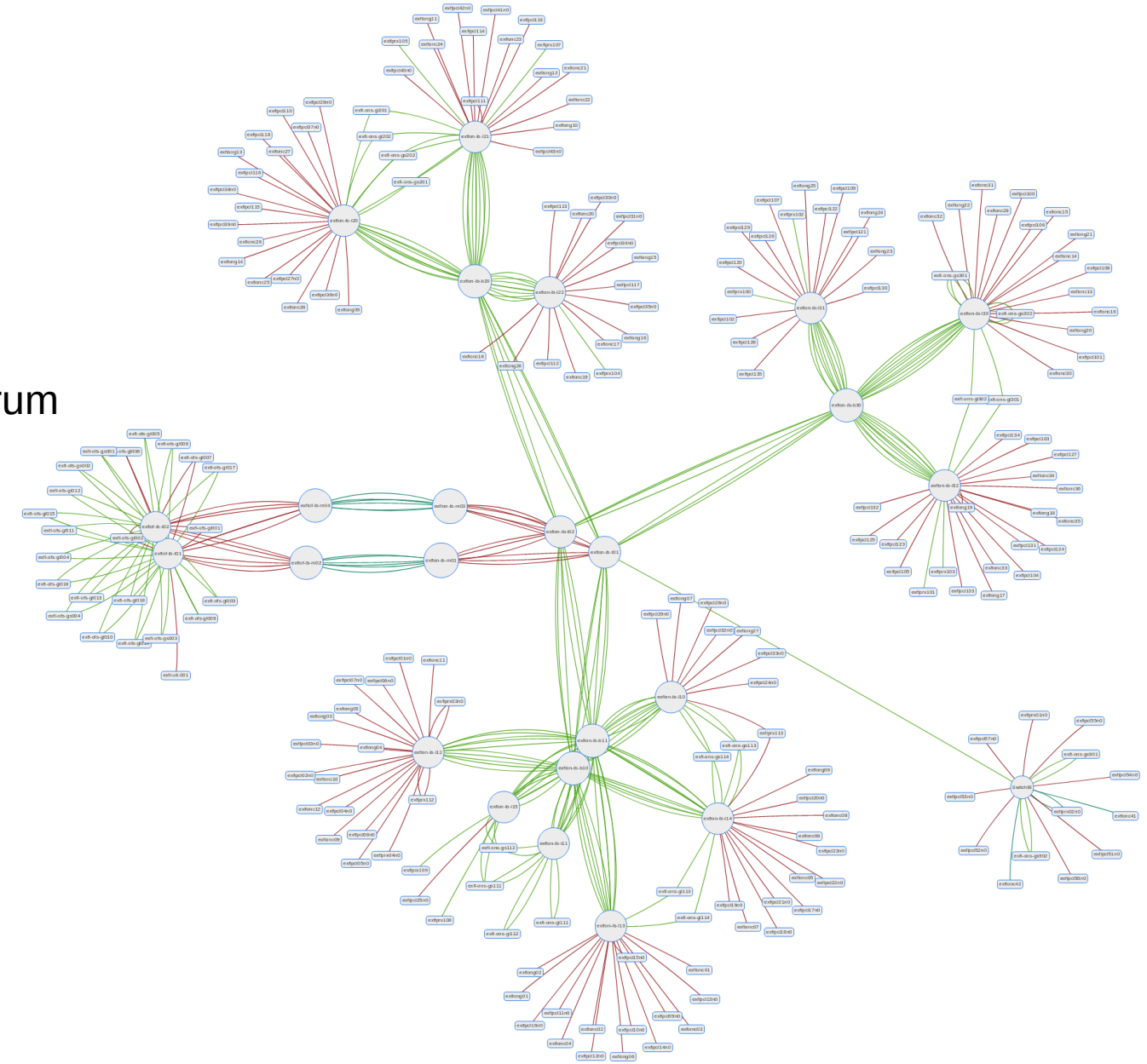
- Joined effort between Deutsches Elektronen-Synchrotron, Fermi National Accelerator Laboratory and Nordic e-Infrastructure Collaboration
- Optimizing the access to the mass storage consisting of a tape library and many disk servers
- Usage: High speed **data ingest, batch processing**, interactive analysis, wide area transfer
- Can be accessed using the pNFS, NFS4 mount or HTTP, XROOTD or GridFTP protocol
- Open Storage Manager (OSM) is performing the access to the files on tape transparently

- dCache is used to implement the raw data and calibration constants repository
- Acts as gateway and management system for cold data (tape archive)
- Scale out is achieved by using building blocks capable of a sustained throughput of 1 GB/s
  - offers cache retention management
  - current size 31.5 PB (+30PB to add this year)
  - 2530 tapes: LTO8, LTO6, LTO8M



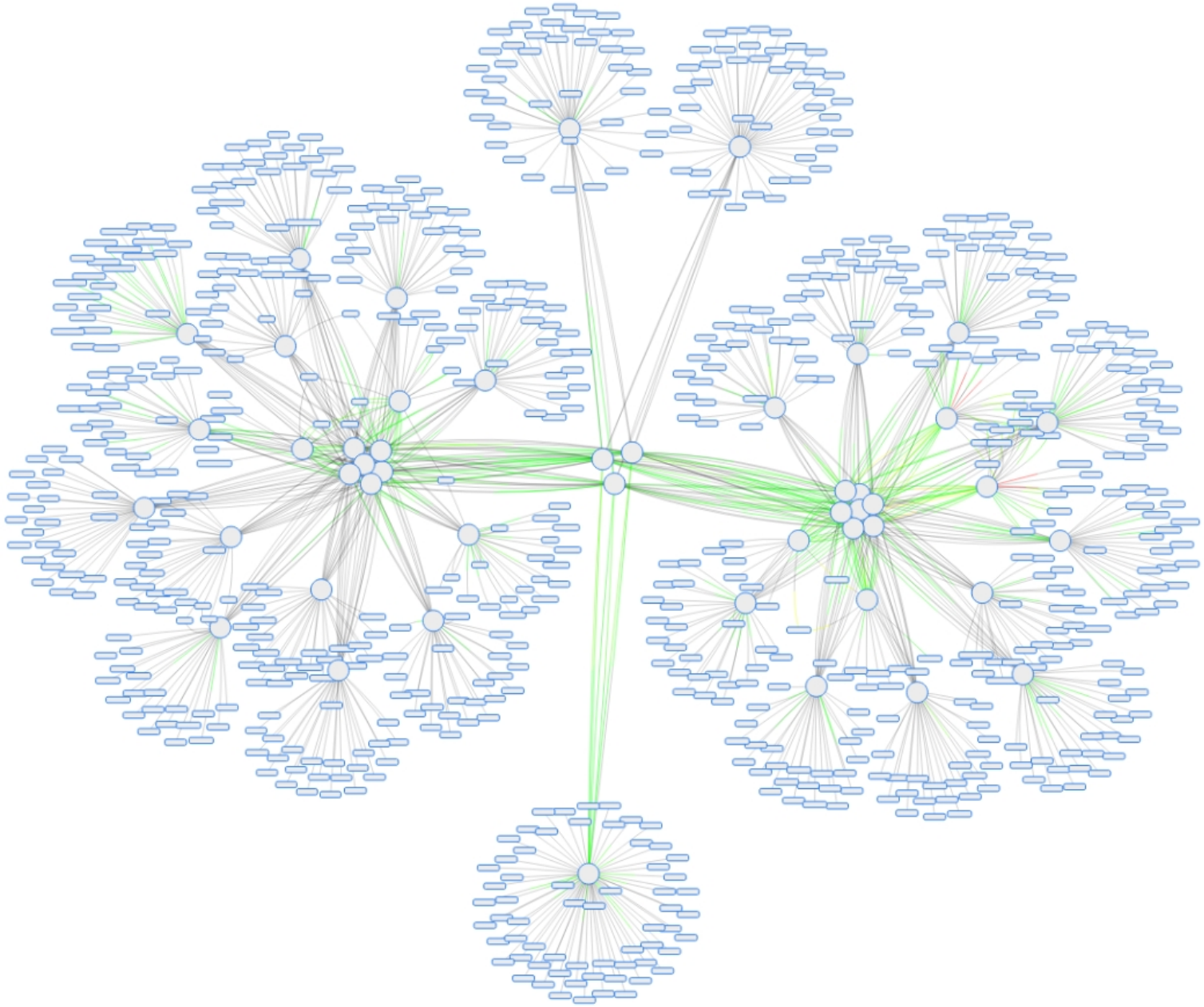
Infiniband network

- EuXFEL private fabric
  - Data taking
  - On-line analysis
  - Data transfer
  - Dedicated long-haul link connecting Spectrum Scale clusters over 4km resulting in single RDMA capable fabric connected via Mellanox MetroX 12xFDR10



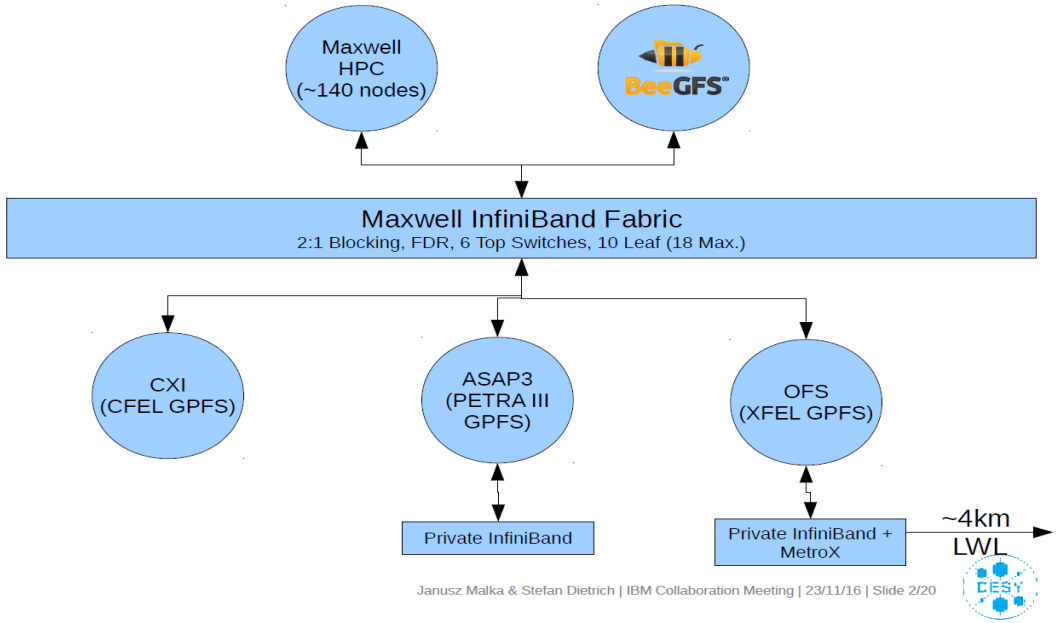
Infiniband network

- Maxwell computing fabric
  - Off-line data analysis
  - Data transfer to dCache
  - Data exports to outside



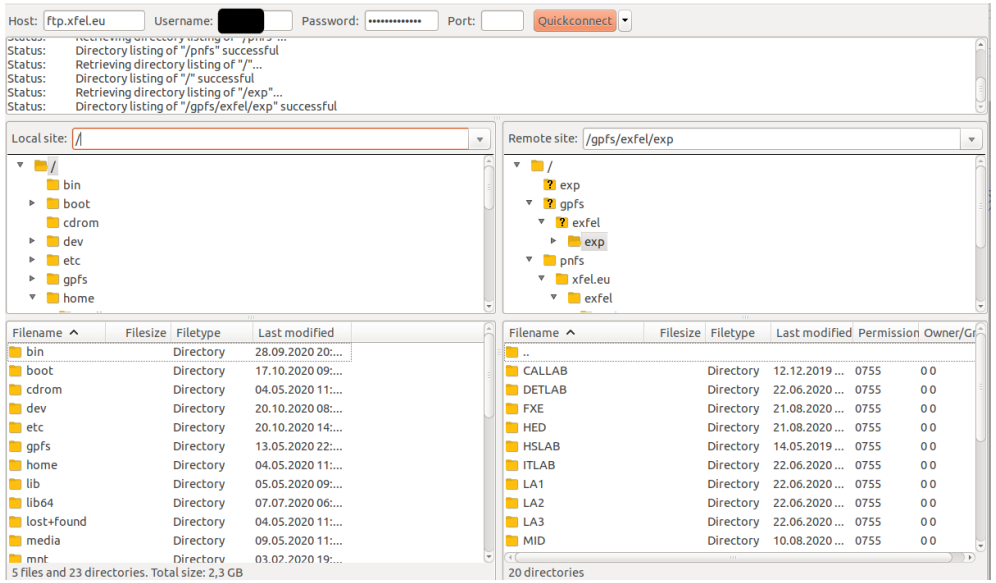
Maxwell cluster

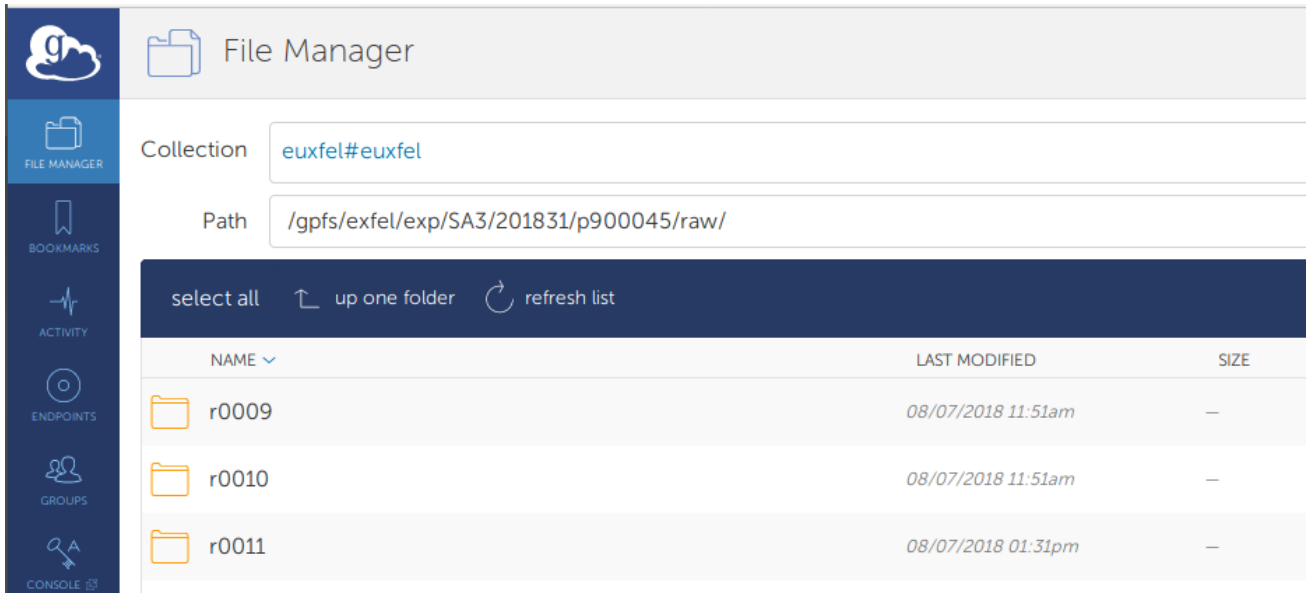
- The Maxwell-Cluster is a resource dedicated to parallel and multi-threaded application, which can use at least some of the specific characteristics
- In addition to serving as a medium scale High-Performance-Cluster
- Maxwell incorporates resources for Photon Science data analysis, resources of CFEL, CSSB, Petra4 and European XFEL.
- In total 615 hosts, 37160 CPUs
- European XFEL contribution:
  - 276 nodes with CPU and 7 nodes with GPU
  - upex partition for users
  - reservation for each beamtime

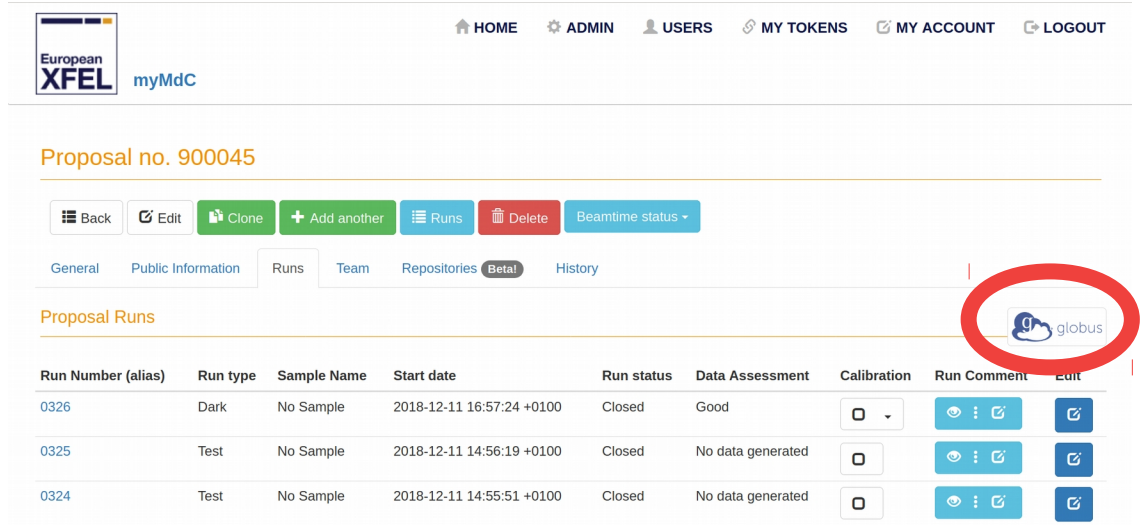


Data Exports

- XFEL is connected to the German research network
  - Inbound – 2 x 1Gb/s
  - outbound – 2 x 15Gb/s (shared with DESY)
- Services in production:
  - **FTP – ftp.xfel.eu**
  - **Globus - euxfel#euxfel**







Monitoring and maintenance

- Hardware and services monitoring
  - Icinga - off-line and GPFS
  - Nagios - on-line
- Log files monitoring
  - Elastic Stack
- Performance and Data Management monitoring
  - Influx
  - Grafana
  - GPFS internal monitoring tools
- Provisioning
  - Foreman
  - Puppet, xCat

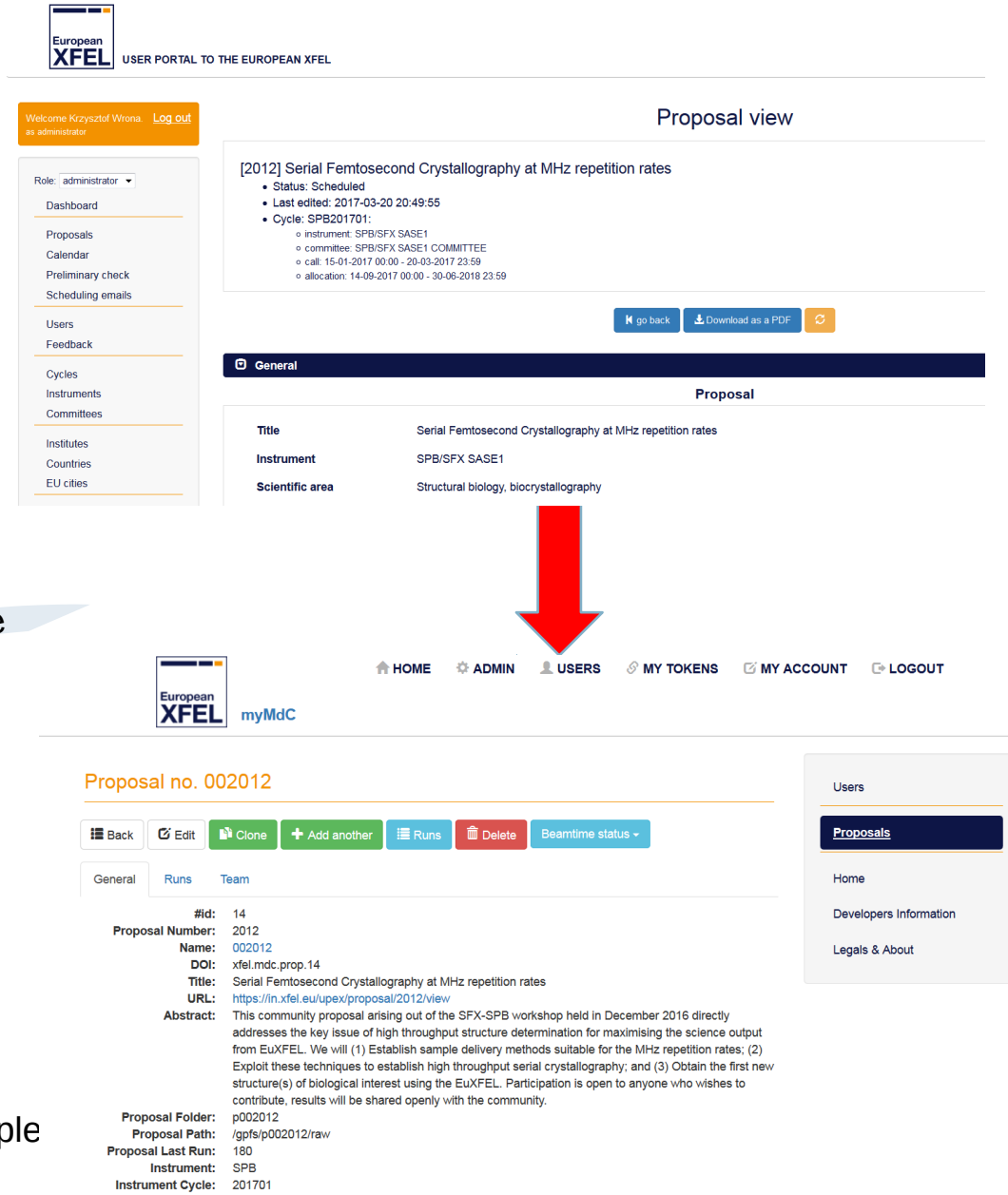
Metadata Catalog

- If proposal is scheduled
- … some information is automatically migrated from UPEX (user portal) to the Metadata Catalogue (myMdC)

- UPEX:
  - 1 Proposal
    - Team members
    - 1 x Instrument
    - N x Samples

Migration:
~4 weeks before beamtime

- myMdC:
  - 1 Proposal
    - Group (with team members)
    - 1 x Instrument
    - N x Experiments & N x Sample



Integration with Metadata Catalog

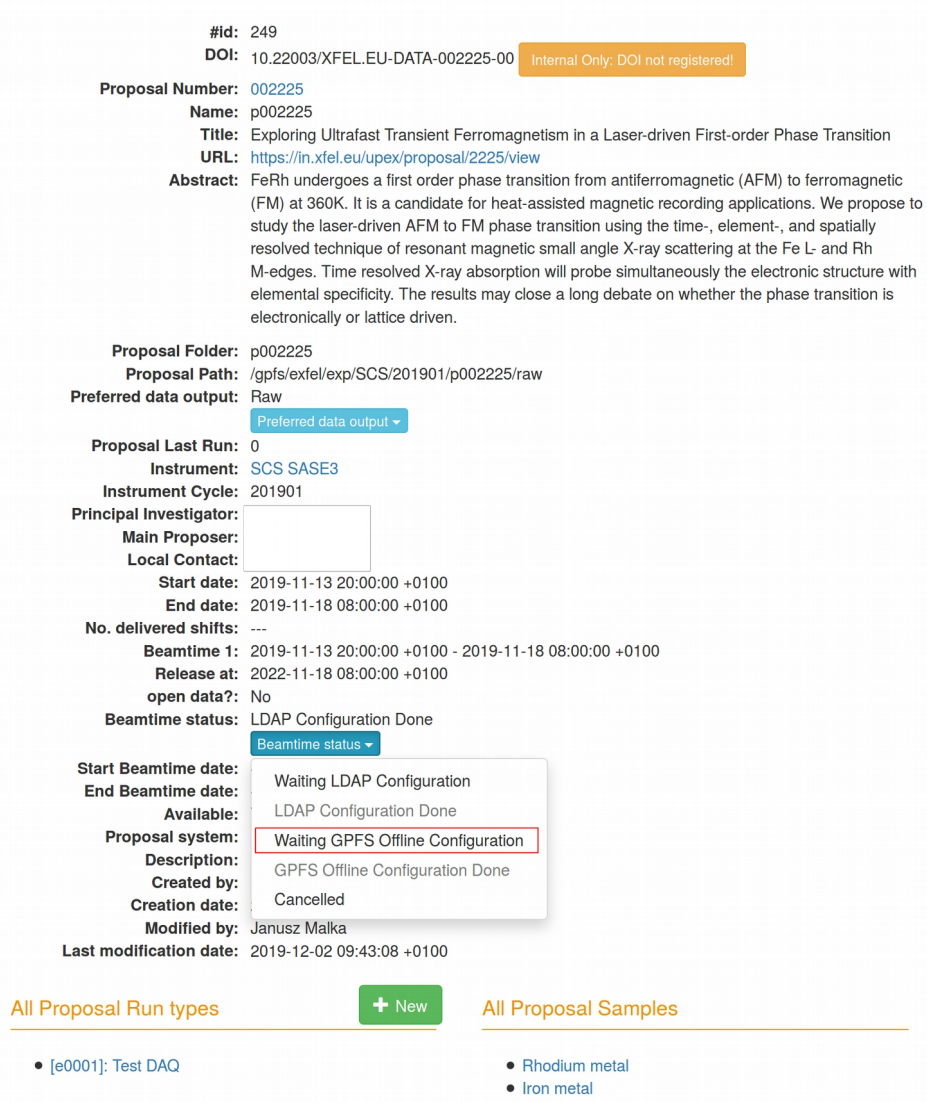
- Communication channel
  - Triggered by myMdC by sending ZMQ message
  - Execute the python code which interact with GPFS
  - Python code sends feedback to myMdC

```
/gpfs/exfel/exp/TST/202031/p900149/
├── proc -> /gpfs/exfel/d/proc/TST/202031/p900149
├── raw -> /gpfs/exfel/d/raw/TST/202031/p900149
├── scratch -> /gpfs/exfel/u/scratch/TST/202031/p900149
└── usr -> /gpfs/exfel/u/usr/TST/202031/p900149

/gpfs/exfel/u/usr/TST/202030/p900117/
├── CalOutput
├── Reports
├── Shared
├── Software
└── UserLab
```

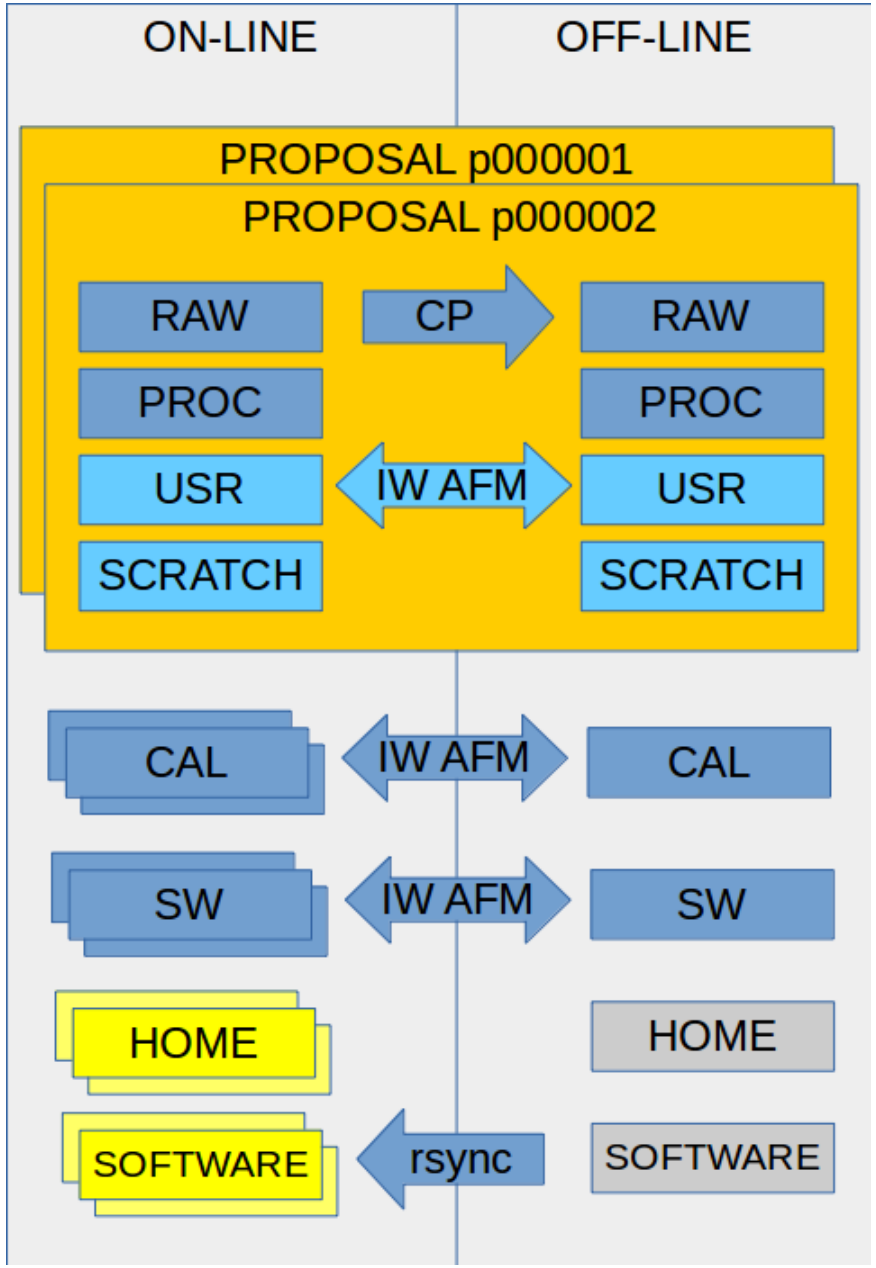
```
/gpfs/exfel/exp/SPB/202001/p002467/
├── proc
├── raw
├── scratch
└── usr
```

GPFS resources creation per proposal



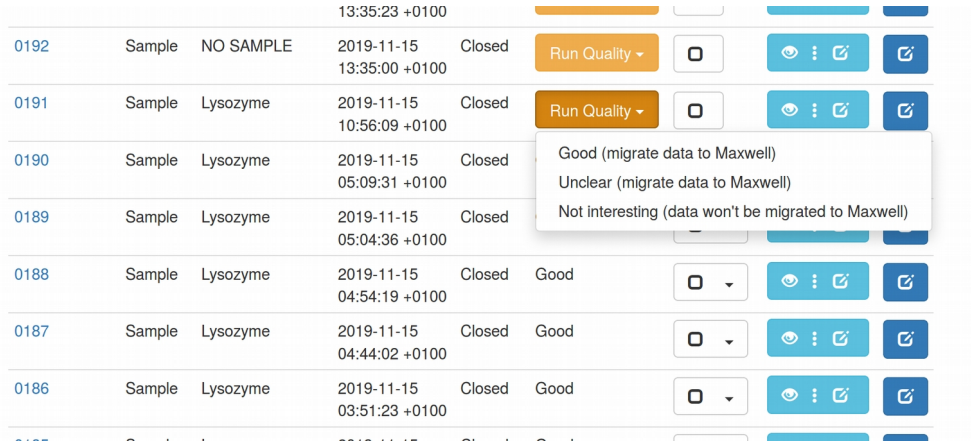
Data placement

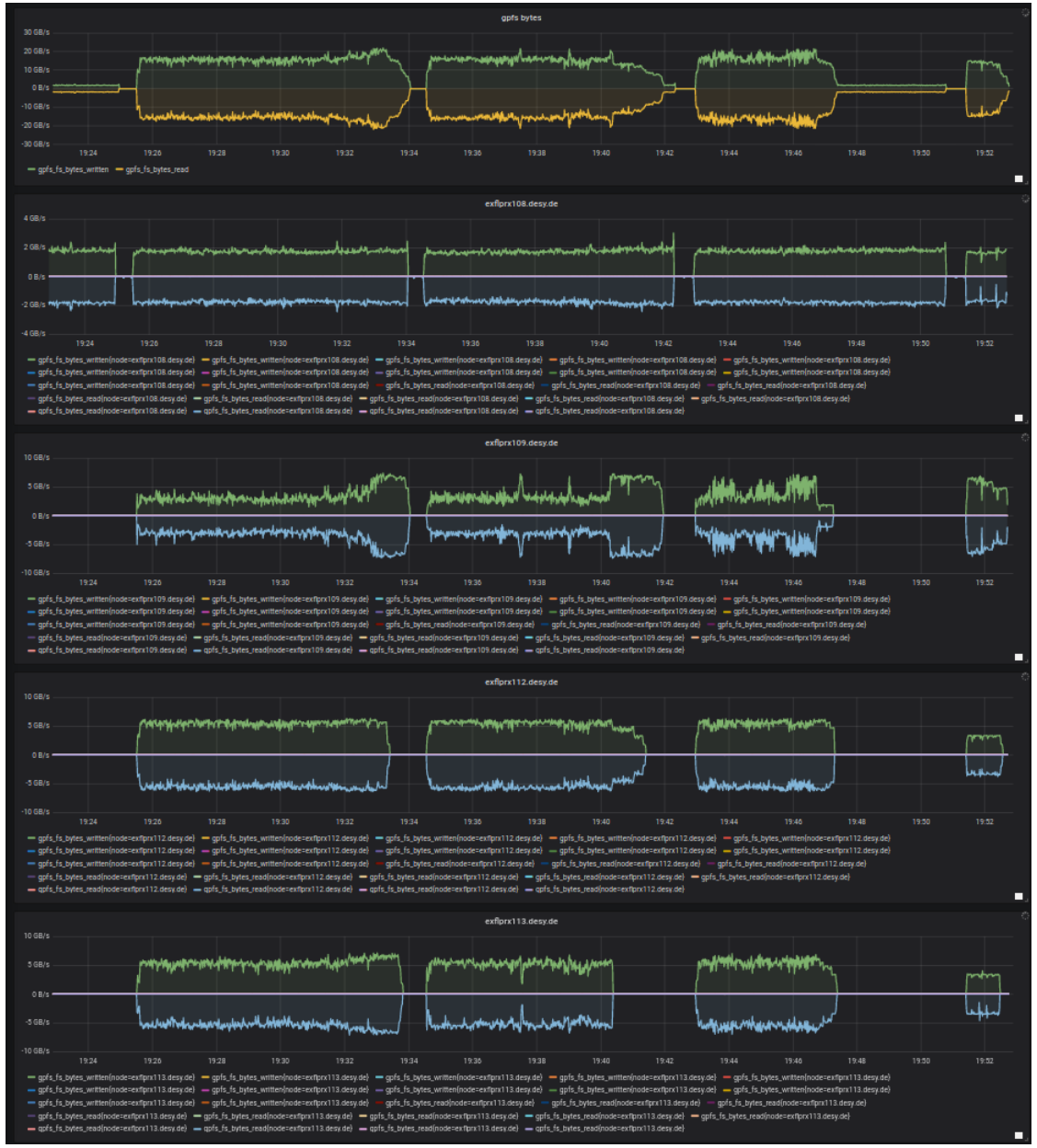
- Proposal specific (on request)
  - RAW – raw data from detector
  - PROC – processed data
  - USR – user data, scripts, programs
  - SCRATCH – intermediate results
- Proposal independent
  - CAL – calibration constants
  - SW – internal control software
  - HOME – $home folder of computing clusters
  - SOFTWARE – analysis software, python, matlab, …
  - ….



Grafana - Zimon via Grafana Bridge

Run Migration

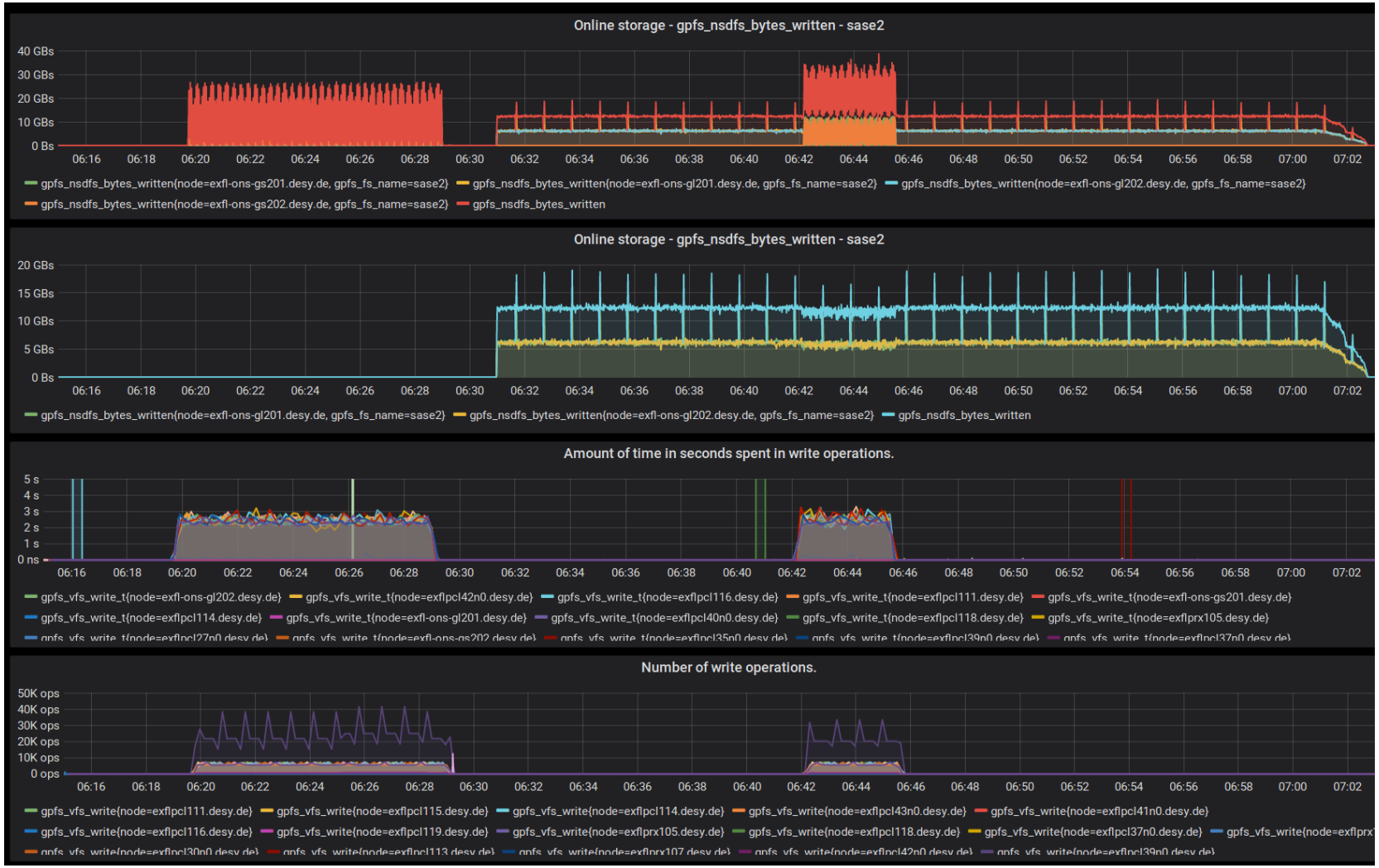




Performance data

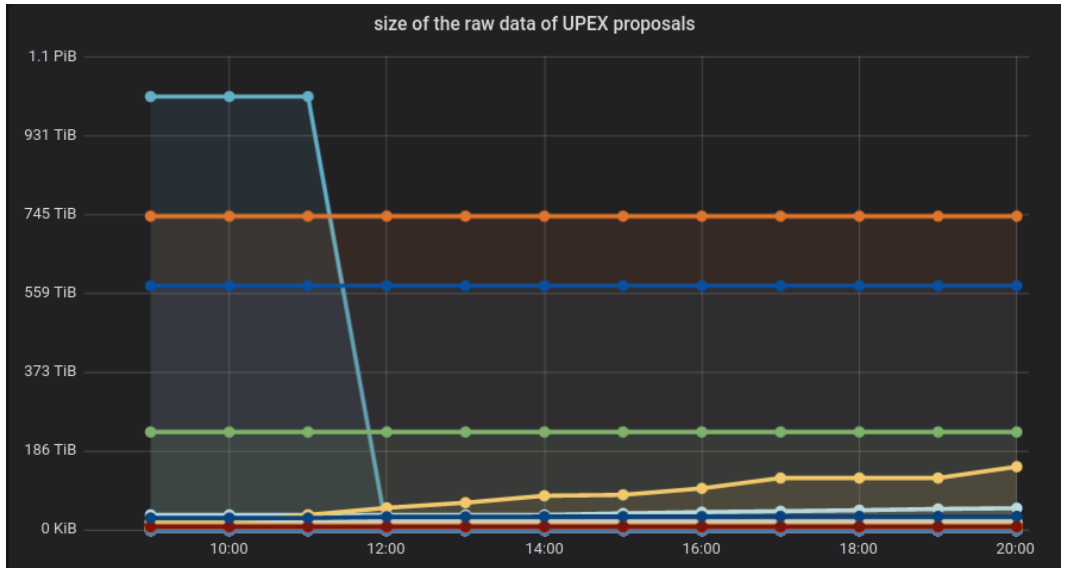
- AGIPD detector on SASE2
- Migration from cache (SSD)

  to data pool (HDD)



Grafana - Zimon via Grafana Bridge

- Space per proposal

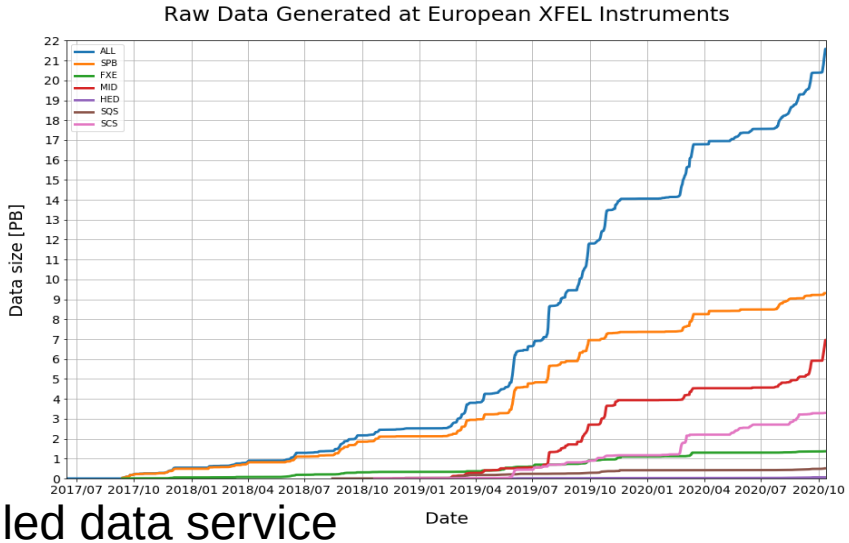


- Space per place

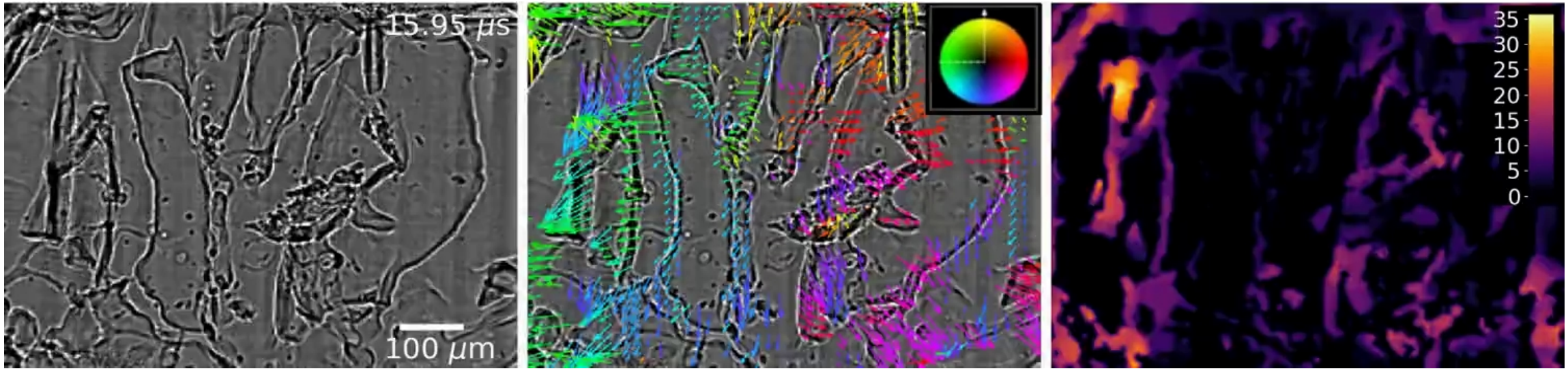
Summary

The number of data has increased significantly over last months:

- PROC: ~10 PB,
- RAW: ~22 PB,
- USERS + EuXFEL data: ~1.5PB



- The current system is still further develop in order to provide excelled data service



Vagovic and All. (2019). Megahertz x-ray microscopy at x-ray free-electron laser and synchrotron sources. Optica. 6. 1106-1109. 10.1364/OPTICA.6.001106.