Data Taking & NFS Services

Two (independent) use cases in physics data processing

Martin Gasthuber
GPFS UG @SC17

HELMHOLTZ | ASSOCIATION

DESY

outline

- EuXFEL
  - data taking
  - data analysis

- Particle Physics Analysis Facility (for LHC data)
  - file services for batch and interactive computing

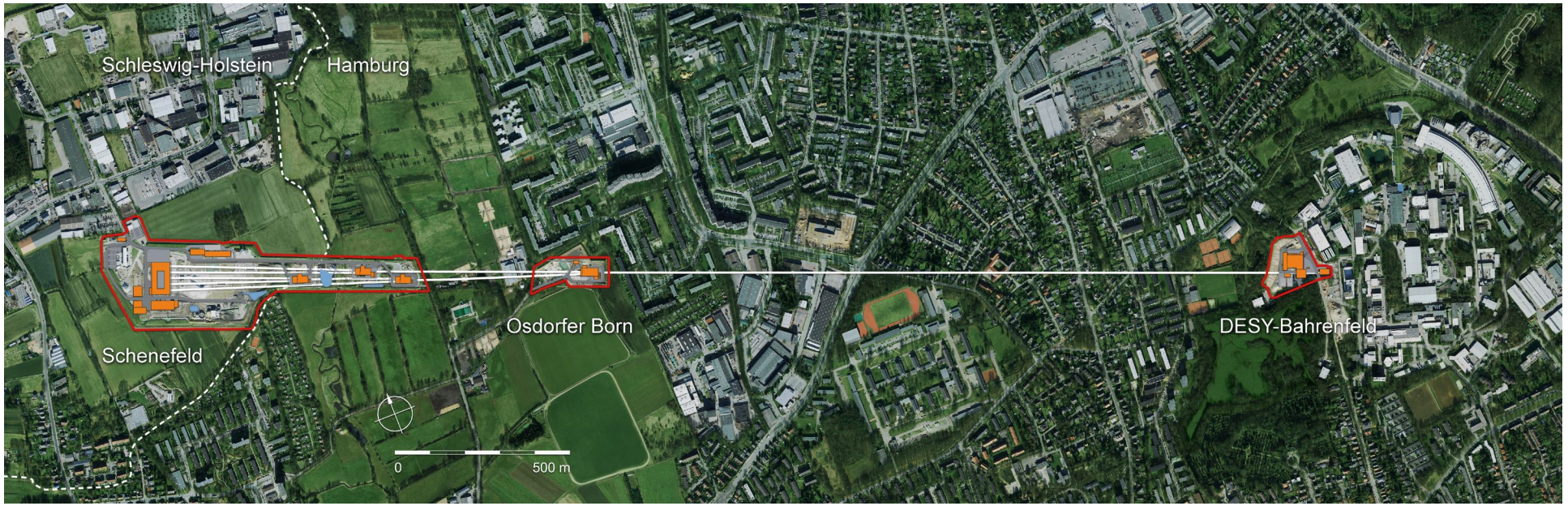
EuXFEL - European XFEL (X-Ray Free-Electron Laser)

- Organized as a non-profit corporation in 2009 with the mission of design, construction, operation, and development of the free-electron laser
- Supported by 11 partner countries
- Germany (federal government, city-state of Hamburg, and state of Schleswig-Holstein) covers 58% of the costs; Russia contributes 27%; each of the other international shareholders 1–3%
- Total budget for construction (including commissioning)
  - 1.22 billion € at 2005 prices
- User facility with ~300 staff members (+250 from DESY)
- user operations starts on September 1, 2017

European XFEL

DESY

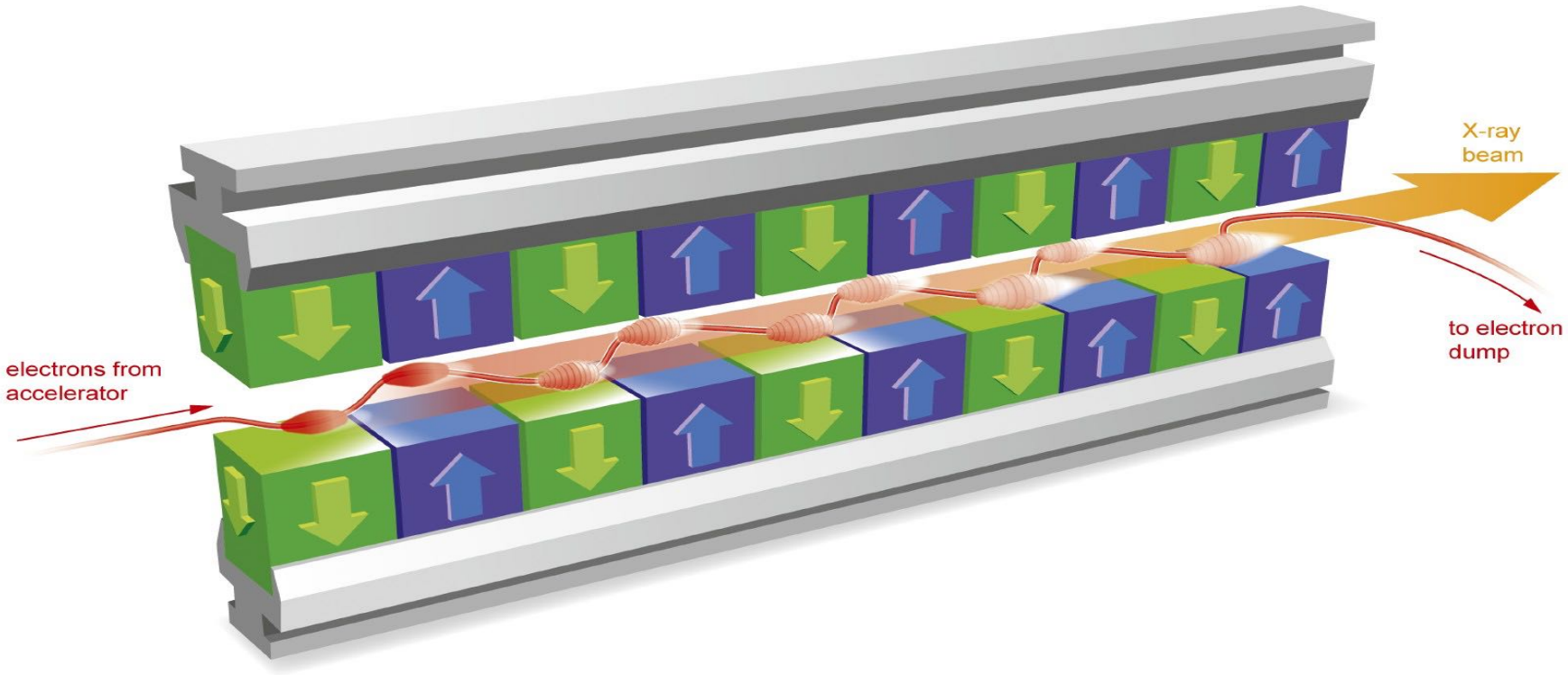
EuXFEL – a new research facility/instrument



lets go flying

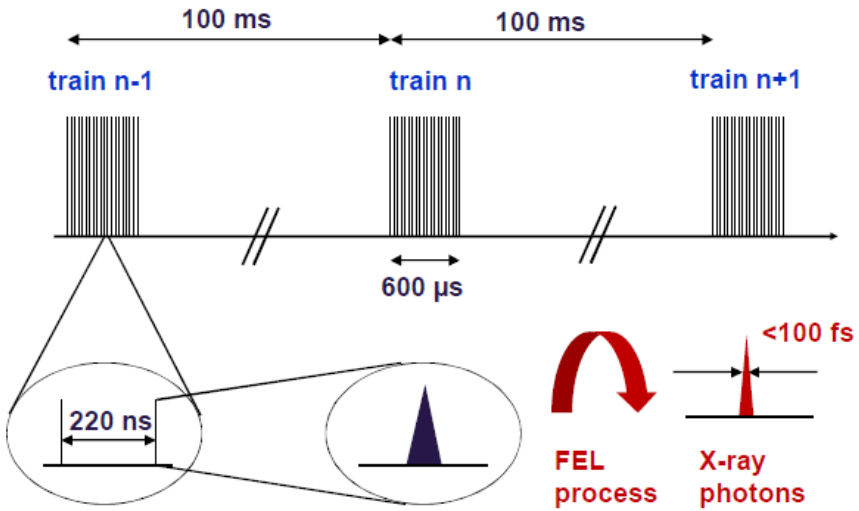
https://media.xfel.eu/XFELmediabank/ConvertAssets/Tunnelflug_2017_1920.mp4

from electrons (bunch) to coherent photons (laser)



photon beam – energy (wavelength) & coherence

- Readout rate driven by bunch structure
  - 10 Hz train of pulses
  - 4.5 MHz pulses in train (1-2700 pulses)
- Data volume driven by detector type

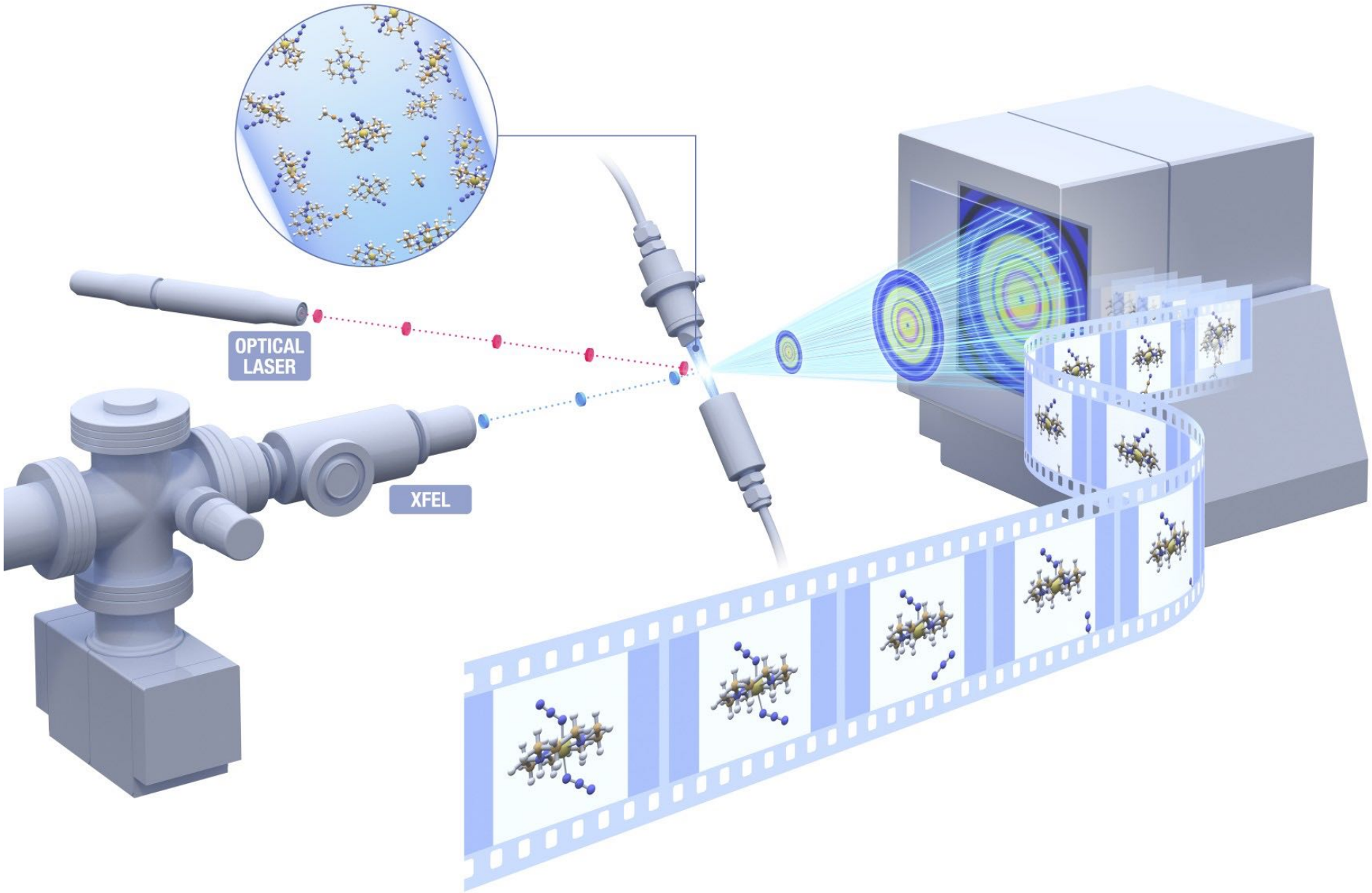


Detector type	Sampling	Data/pulse	Data/train	Data/sec
1 channel digitizer	5 GS/s	~2 kB	~6 MB	~60 MB
1 Mpxl 2D camera	4.5 MHz	~2 MB	~1 GB	~10 GB
4 Mpxl 2D camera	4.5 MHz	~8 MB	~3 GB	~30 GB*

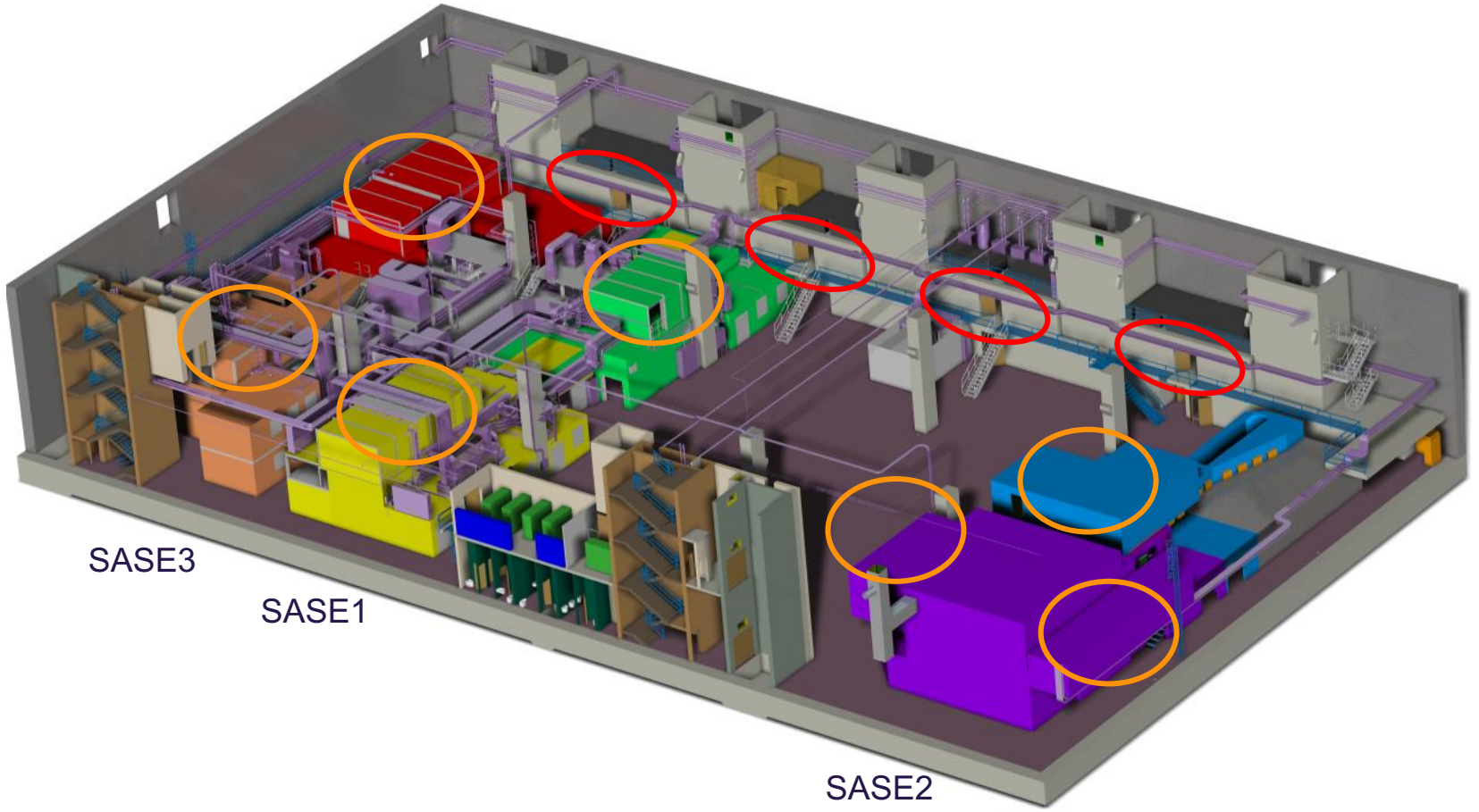
- volume depends on detector type and pulses per train
- 1-N trains per file -> 1GB file or larger

* Limited by AGIPD detector internal pipeline depth (352 img/sec), hence factor 3 compare to LPD 1MPx

at the end…

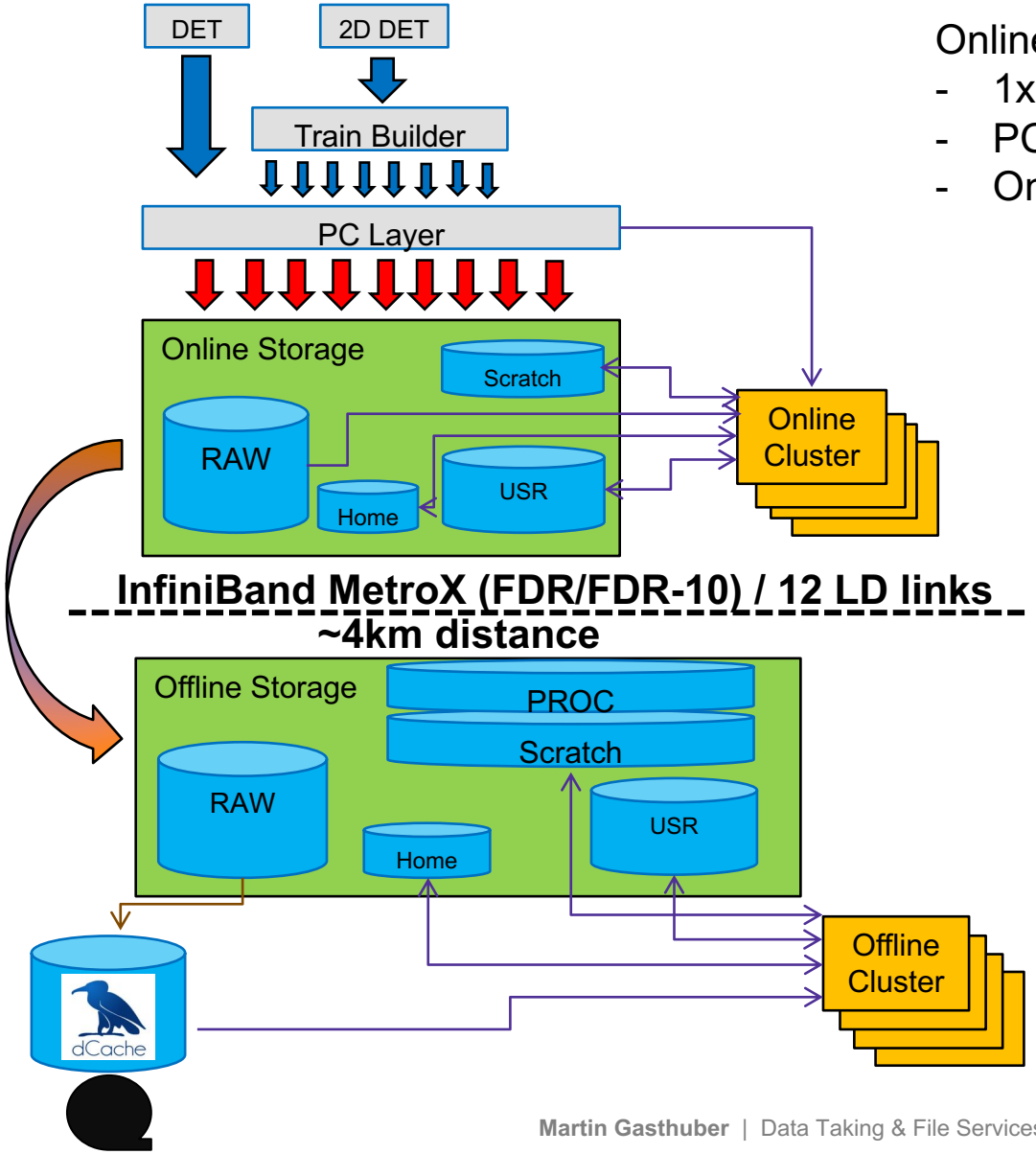


infrastructure locations

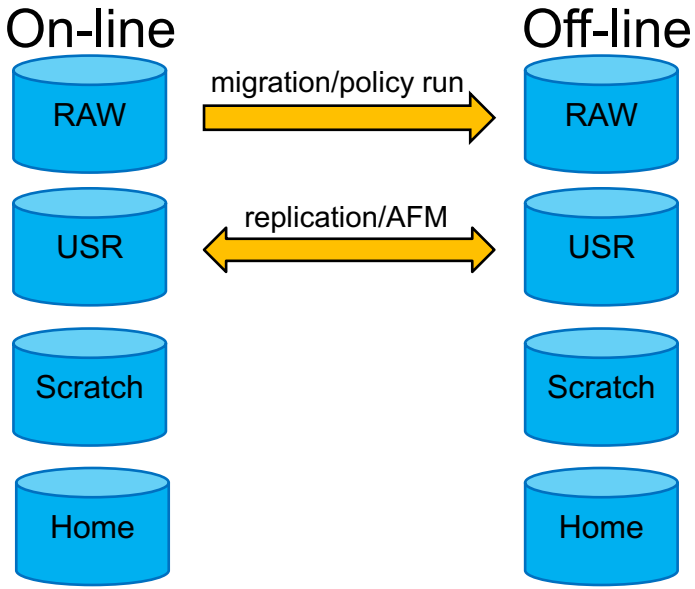


- 4 computer rooms in the experiment hall (red, a.k.a. balcony rooms)
- Dedicated rack rooms for the instruments (orange)

the simpler part

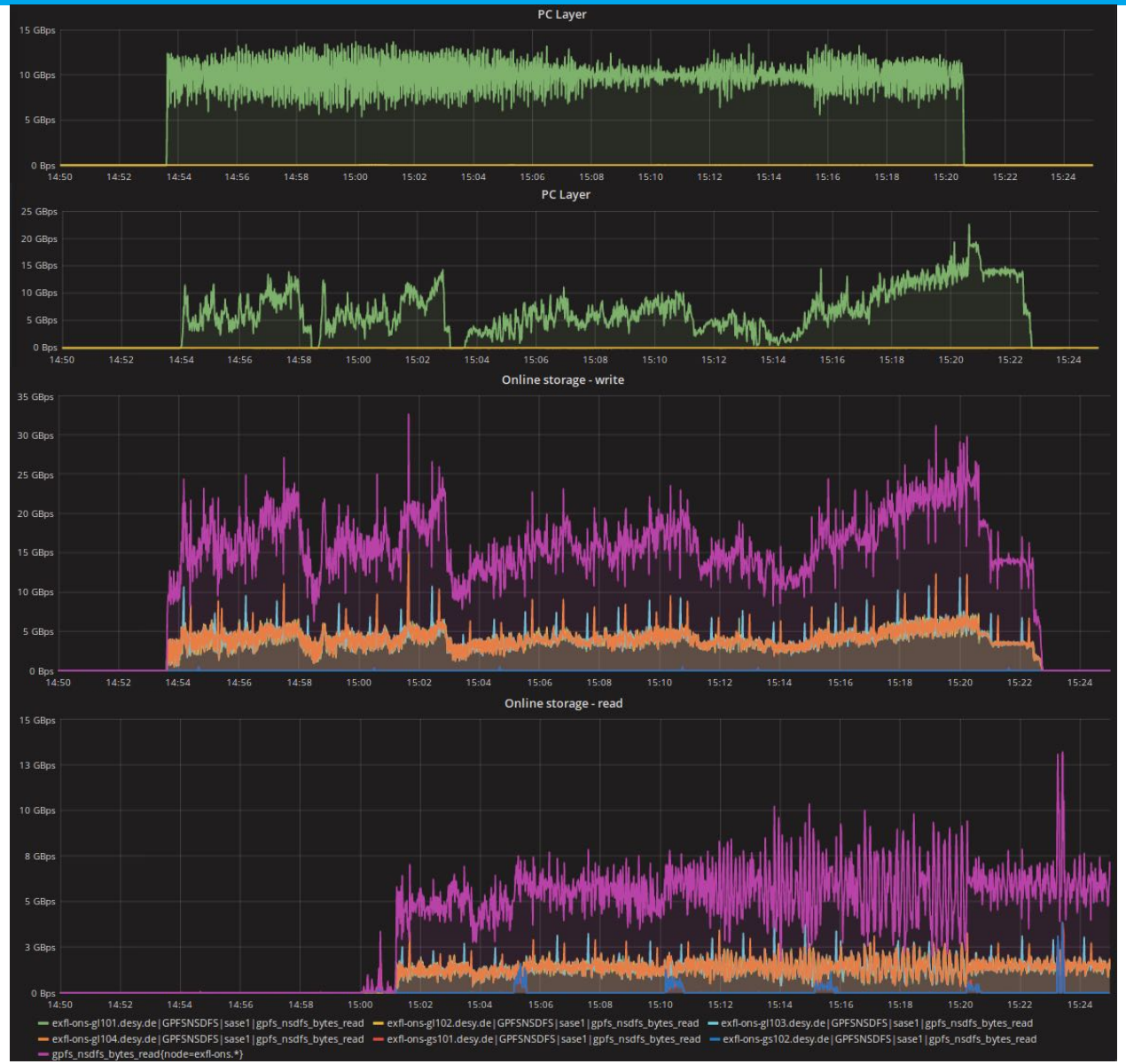


Online (@experimental hall)
- 1xGS1, 2xGL4, 2xCES, 4xProxy(CR, AFM)
- PC Layer – 28 nodes
- OnlineCluster – 20 nodes (8GPU)

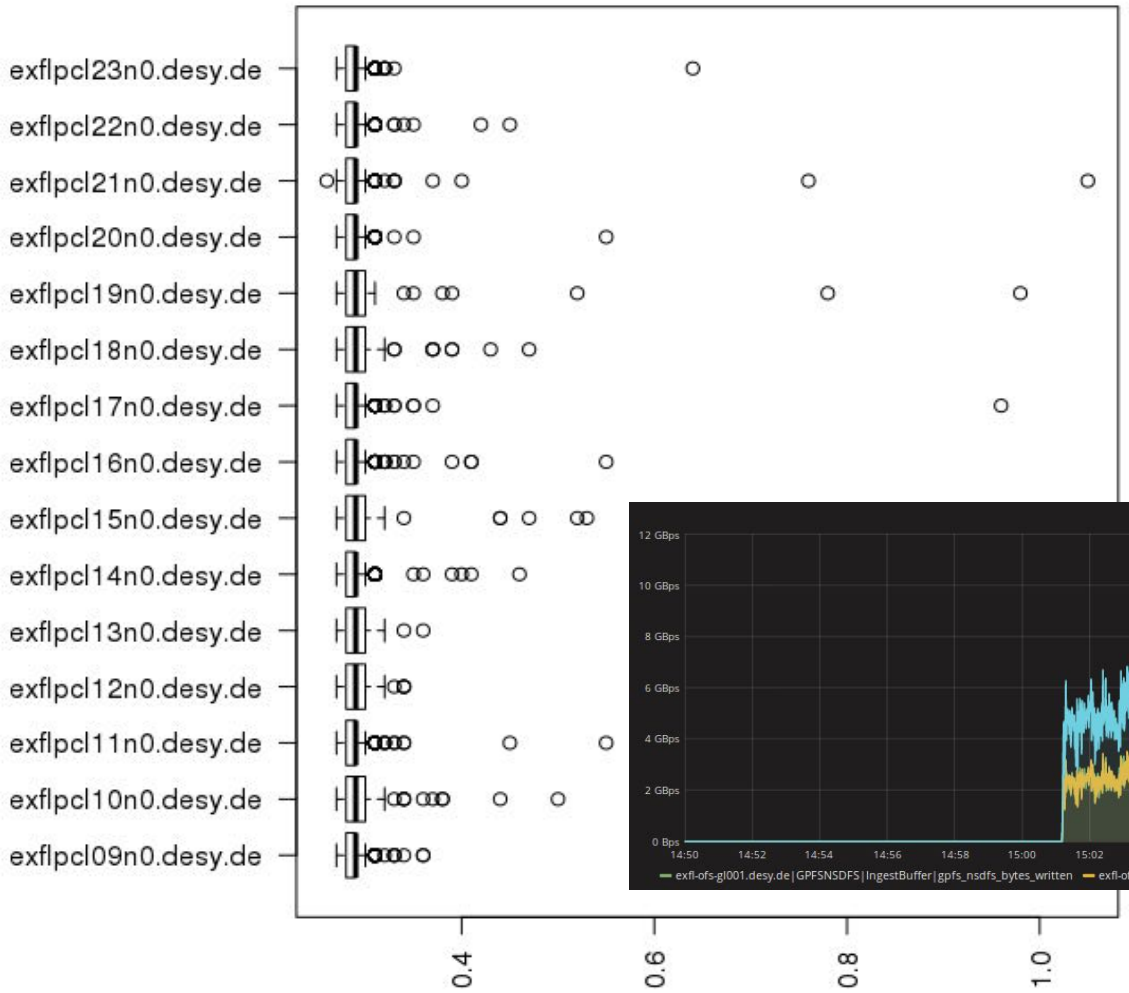


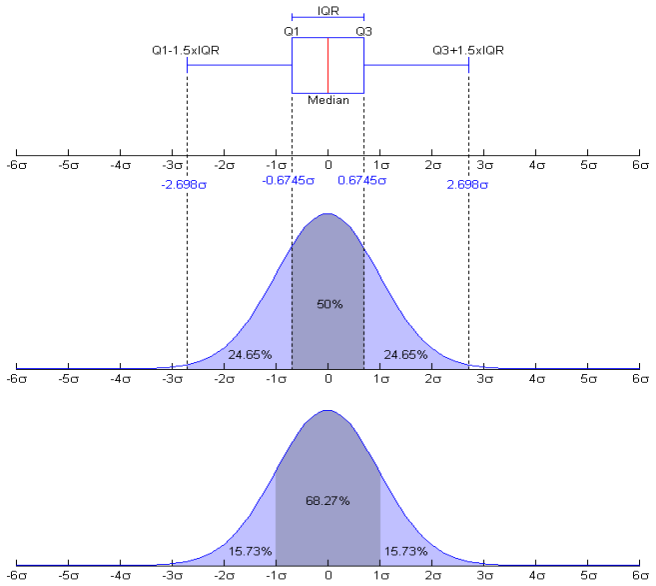
Offline (@DESY data center)
- 1xGS1, 3xGL4, 2xCES, 4xdCache
- 4xFTP, 1xUtility
- OfflineCluster – 221 nodes (shared)

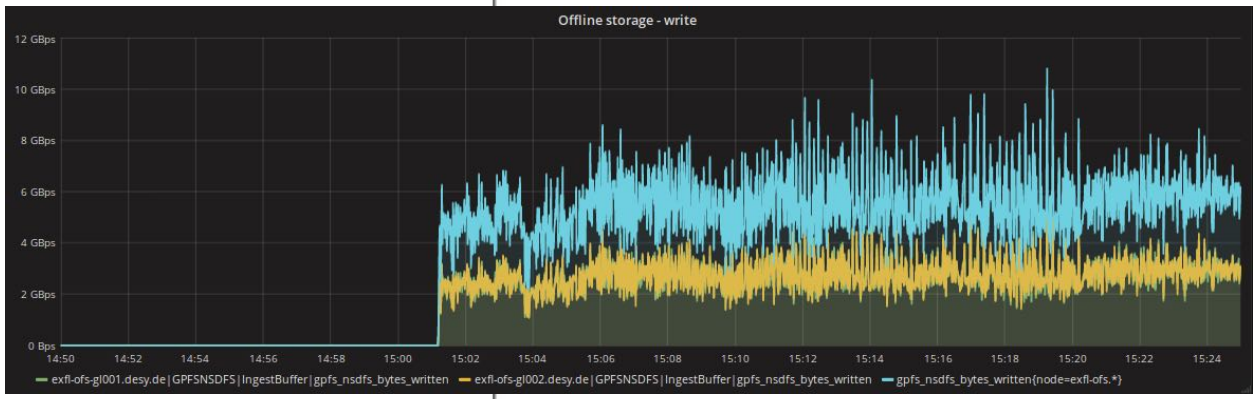
DAQ writes + copy2offl + user



continued


exflpcl23n0.desy.de
exflpcl22n0.desy.de
exflpcl21n0.desy.de
exflpcl20n0.desy.de
exflpcl19n0.desy.de
exflpcl18n0.desy.de
exflpcl17n0.desy.de
exflpcl16n0.desy.de
exflpcl15n0.desy.de
exflpcl14n0.desy.de
exflpcl13n0.desy.de
exflpcl12n0.desy.de
exflpcl11n0.desy.de
exflpcl10n0.desy.de
exflpcl09n0.desy.de
0.4
0.6
0.8
1.0


IQR
Q1
Q3
Q1-1.5xIQR
Q3+1.5xIQR
Median
-6σ -5σ -4σ -3σ -2σ -1σ 0 1σ 2σ 3σ 4σ 5σ 6σ
-2.698σ
-0.6745σ
0.6745σ
2.698σ
50%
24.65%
24.65%
68.27%
15.73%
15.73%


Offline storage - write
12 GBps
10 GBps
8 GBps
6 GBps
4 GBps
2 GBps
0 Bps
14:50 14:52 14:54 14:56 14:58 15:00 15:02 15:04 15:06 15:08 15:10 15:12 15:14 15:16 15:18 15:20 15:22 15:24
exfl-ofs-gl001.desy.de|GPFSNSDFS|IngestBuffer|gpfs_nsdfs_bytes_written
exfl-ofs-gl002.desy.de|GPFSNSDFS|IngestBuffer|gpfs_nsdfs_bytes_written
gpfs_nsdfs_bytes_written{node=exfl-ofs.*}

observations…

- predictable IO not always easy (DAQ writes) – still investigating
- tiny (appending) HDF5 IO (slow control data) – to be tuned ;-)
- fileset (quantity) limits too low !
- filesets with more then one mountpoint (symlink inflation)
- remote cluster – rootsquash AFM, QOS
- Grafana bridge is essential monitoring component
- interfacing/glueing with XFEL services (create FSet, ACLs, …)
  - RestAPI not (yet) in use
- usage: ~300TB of raw data generated (since Sept. 1) - ~600TB total used

next steps

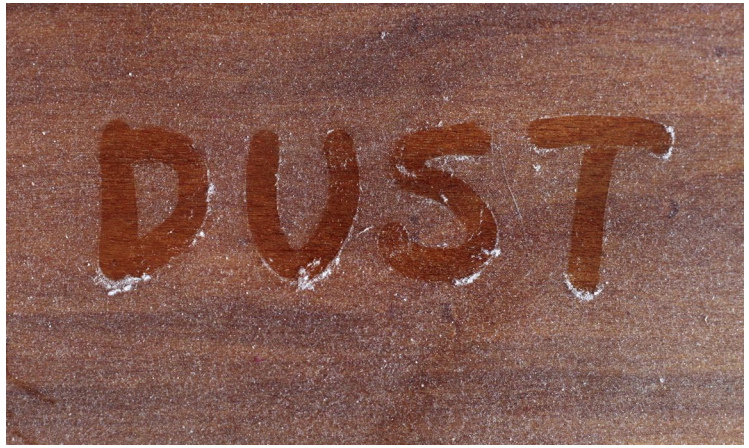
- replace online GL4s with GS4Ss – flash-only for online storage (DAQ)
  - rotating spindles grouped in offline cluster
  - tuning (GS4S) expected and scheduled – expect largely independent performance of randomness and blocksizes
- further extending offline storage (MD & Data) – new GL4S by next week
- PoC on ‘inotify’ for GPFS – cluster wide inotify (known as ‘file audit logging’ ?)
  - replace ‘policy run’ based copy of RAW data by ‘event’ based copy

    important for other local systems !

NFS file services for LHC data analysis / interactive and batch

complementary to Grid (WLCG) LHC data analysis

larger scale Ganesha use case – particle physics



2 x GS1, 2 x GL6(6TB)
2 FDR fabrics
6 x CES (each 4 x 10GE) – 256G, NVMe SSD (LROC)

bonded (LACP, Layer 3+4) all 10GE ports active

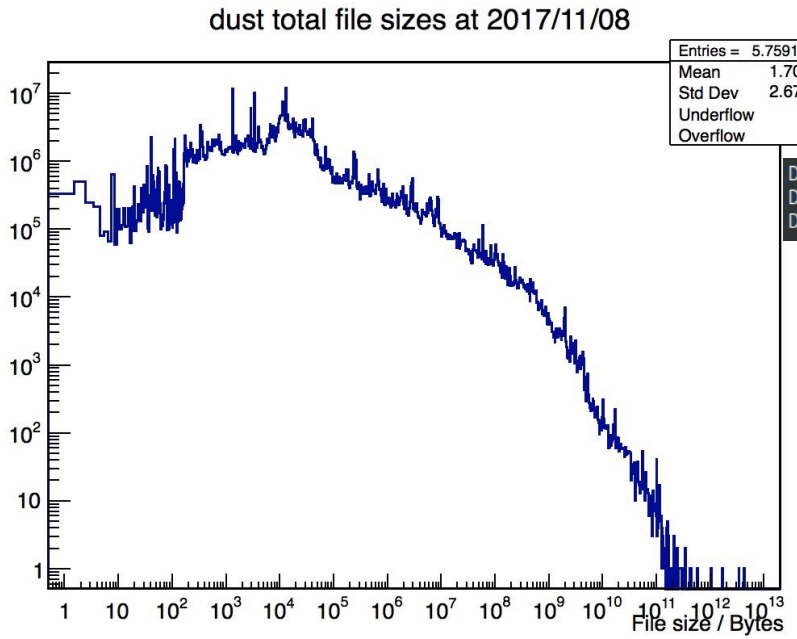
~500 batch (4CES) & 20 interactive (2CES) clients

more details

- Integration
  - LDAP
  - Amfora (priv. development) FileSet generation, Quota mgmt by group admins
  - monitoring – ZiMon and Grafana bridge
  - NFSv4 ACL
- initial setup (10m ago) – 1xGS1, 1xGL6, 4xCES nodes with 2x10GE
  - GS1 filled up too rapidly ;-) – CES nodes filled all 8 10GE links for days

findings

- running 4.2.3.X, 8M/1M blocksize
- ~1PiB used, >500M files
- low GPFS load (although snapshots got “quiesce of SG timed out …)
- very rare IO errors on client (hard to debug/trace – more tools ?)
- need to extend GS1 (GS2) – too many too small files ;-)



LROC – worth the buck

```
Data objects stored 1887356 (1863688 MB) recalled 1854172 (9321497 MB) = 98,24 %
Data objects queried 0 (0 MB) = 0,00 % invalidated 2860549 (14672232 MB)
Data objects failed to store 959624 failed to recall 2223101 failed to query 0 failed to inval 0
```

FileSets (6m ago)