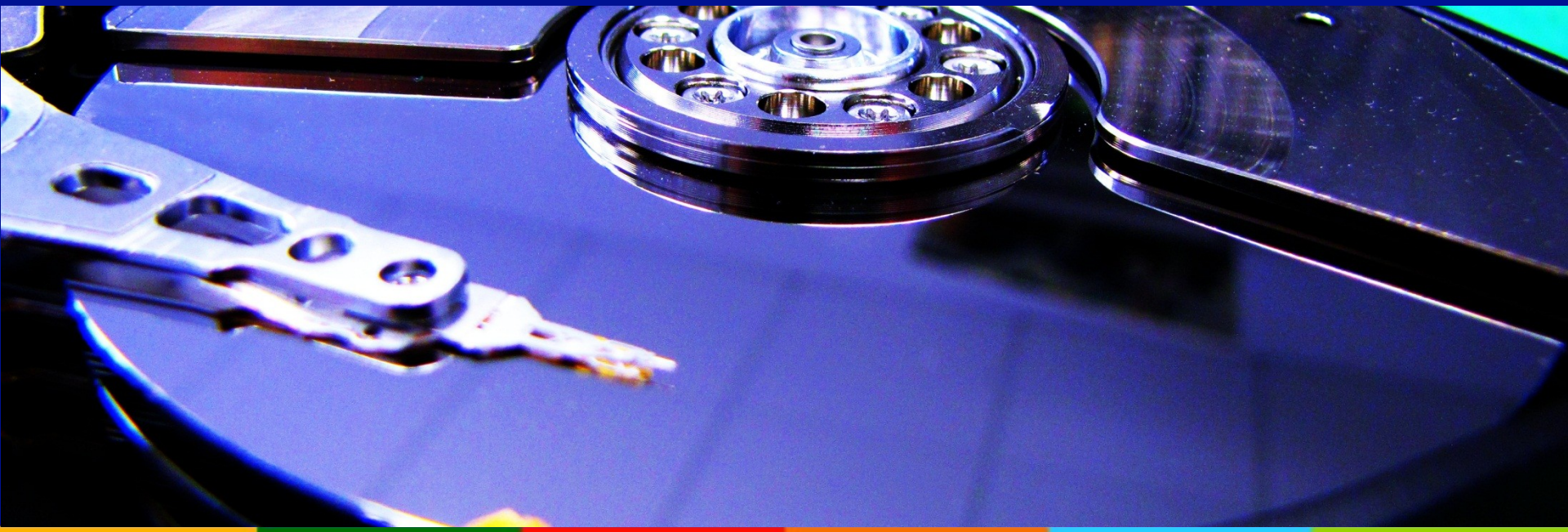
DLCL Matter

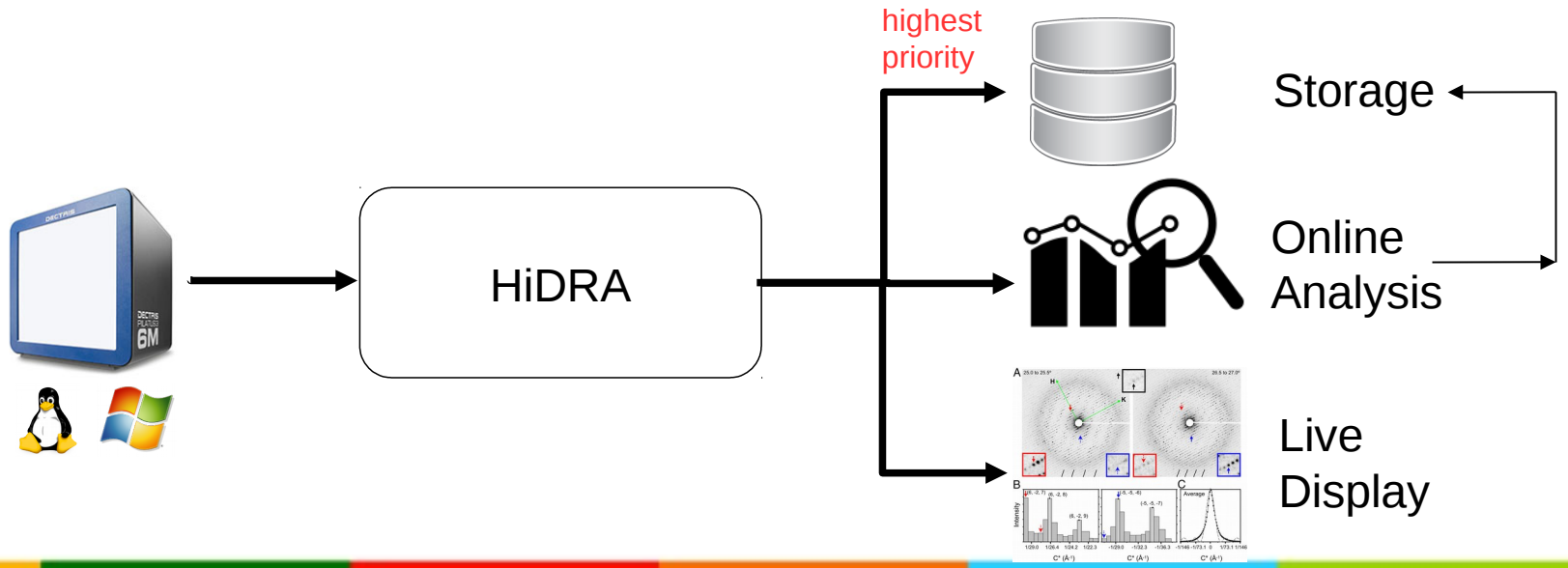
LSDMA All-Hands Meeting, October 2016

Manuela Kuhn



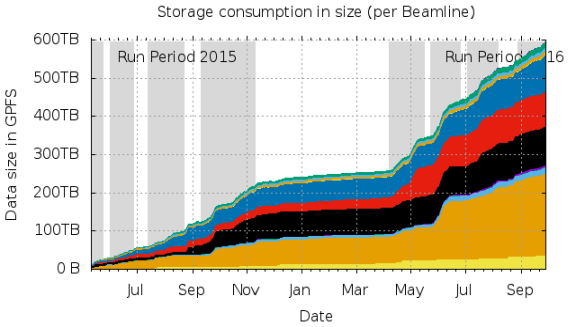
HiDRA – Petra III & Flash @DESY

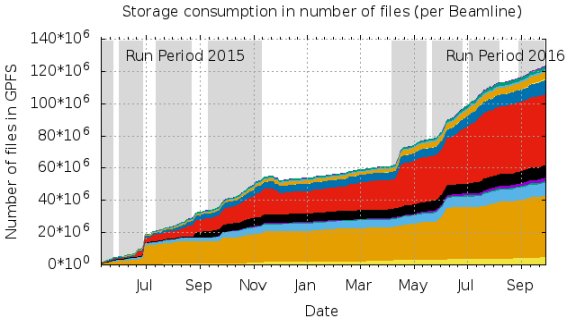
- Problem:
  - data has to be drained from the detectors fast enough (>30Gb/sec)
  - experimental conditions have to be monitored and analyzed in close to real time to prevent the collection of unfavorable data, which also helps with preserving the valuable sample (online analysis)
- HiDRA is a generic tool set for high performance data multiplexing with different qualities of service based on Python and ZeroMQ



Current Status and Outlook

- ASAP3 running successfully for one and a half years
- Currently joining:
  - PETRA III extension
  - Flash I + II
  - Other DESY labs:
    - Detector development
    - Microscopy labs
    - Nanolab
  - XFEL: similar architecture + components (ASAP3 as blueprint)





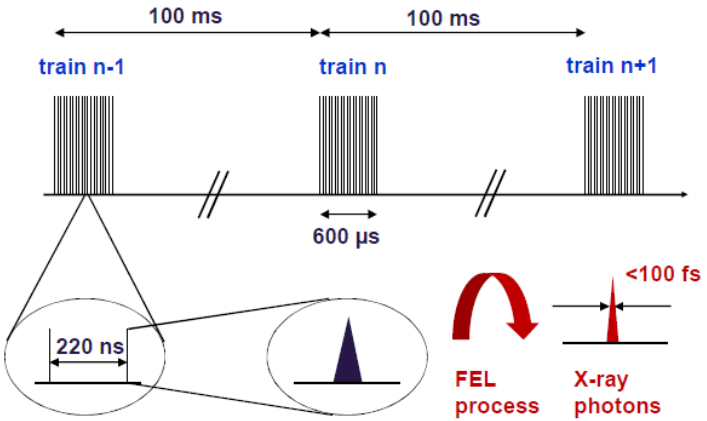
→ will become the only system for the DESY photon science light sources and labs

ASAP3 system expansion

- Doubling of the installed SSDs
- Doubling of the InfiniBand connections
- Additional proxy nodes for new beamlines of P3-extensions, special instruments (detektor development) and Flash I+II
- Expansion of the disk based GPFS installation (gain of 1.5 PB resp. 2 PB depending on disk capacity)
- Expansion/modification of the beamline-fs
- Tests for all Flash solution for beamline-fs (25 GB/s by 250 TB capacity)

European XFEL: DAQ challenges

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

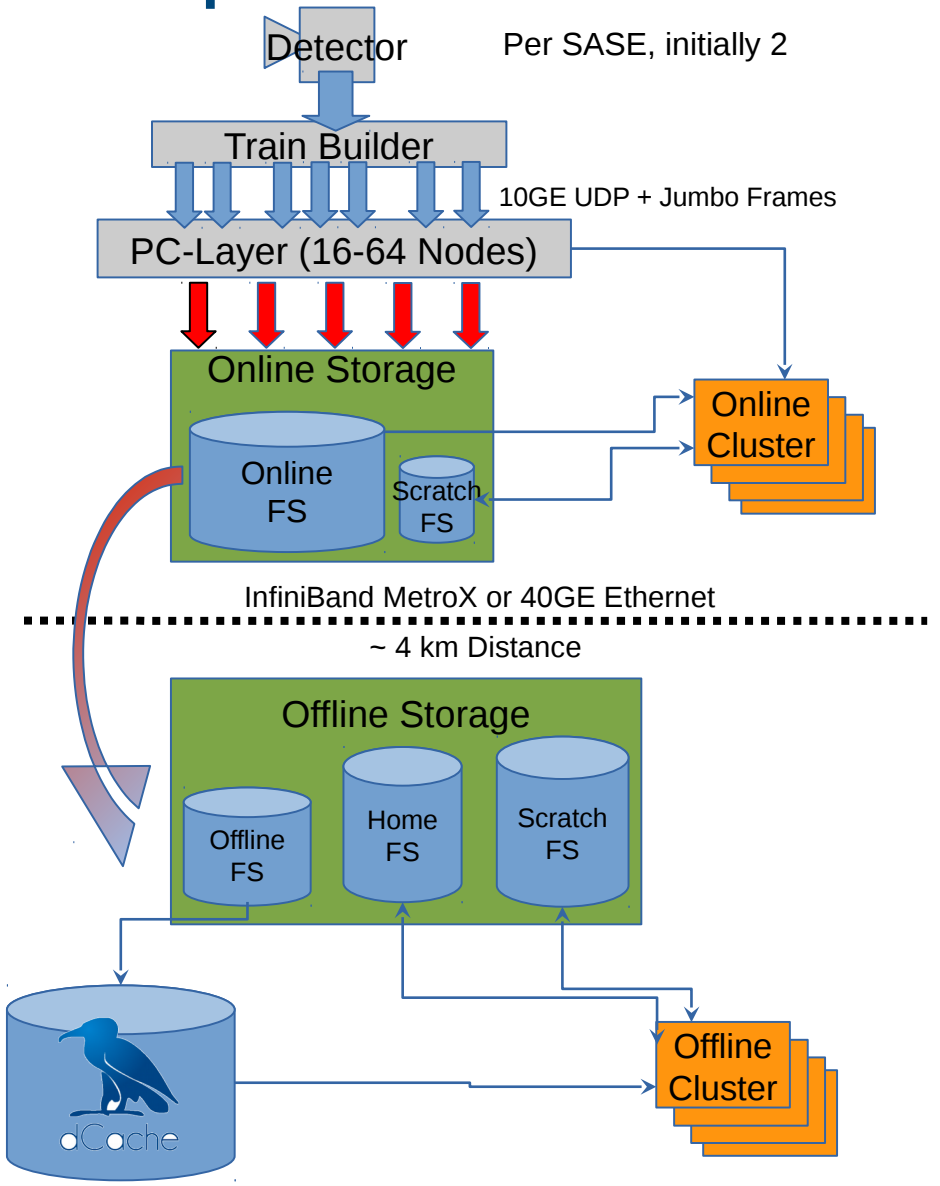


Detector type	Sampling	Data/pulse	Data/train	Data/sec
1 channel digitizer	5 GS/s	~2 kB	~6 MB	~60 MB
1 Mpxl 2D camera	4.5 MHz	~2 MB	~1 GB	~10 GB
4 Mpxl 2D camera	4.5 MHz	~8 MB	~3 GB	~30 GB*

Detector data rates are huge

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

European XFEL: Online and Offline Data Flow



Train Builder

- Reshuffles picture modules into whole picture
- Pictures shuffled in trains
- Sends single trains per channel

PC-Layer

- Data analysis for monitoring
- Data Reduction, e.g. FPGA compression
- Veto
- File creation in memory and online filesystem

Online Cluster

- 10-80 nodes
- Online data analysis and re-calibration

Transfer Online → Offline Storage

- Evaluation: MetroX or 40GE Ethernet
- Evaluation: GPFS AFM or custom scripts

Offline Storage

- Shared across multiple SASE
- Data arrives after delay, stored on GPFS
- Copy data to dCache for long term archival
- Raw data access only from dCache (TBD)
- Offline cluster stores calibrated data on GPFS
- Additional analysis from calibrated data

European XFEL: Initial setup

- PC-Layer: currently 19 nodes in one cluster, diskless
- Online storage in Schenefeld:
  - One cluster with 2 servers (1x GL4) and 4 clients
- Connection online ↔ offline: Mellanox MetroX TX6100
  - Evaluation equipment provided by Mellanox
  - 3x long range fibre uplinks
  - 6x IB FDR links to local switch
- Offline storage in DESY computer center
  - 1 cluster with 2 servers (1x GL4) and 4 clients
- EDR InfiniBand infrastructure
  - Clients will stay on FDR for now

European XFEL: First tests

- Setup from ASAP3
  - 2 filesystems in 2 clusters + cluster for PC-Layer
  - Testing filesystems with different block sizes
    - 8MB → ~10GBps
    - 16MB → ~12 GBps
- Strechted cluster
  - One data-storage-cluster strechted over two sites + cluster for PC-layer (MetroX needed)

  → Tests in progress
- All Flash – 512TB@22GB/sec system (with GPFS integration)
- First tests with QoS from GPFS

ESS GL4 in Schenefeld

Balcony Rooms (2x visible)

European XFEL: Current Plans

- 5 PB dCache
- 6000 cores for computing
- Additional Storage:
  - 1 GS1 and 2 GL4 for online
  - 1 GS1 and 1 GL4 for offline

→ Till end of 2016: 260 additional units in computer center

DLCL Structure of Matter

FAIR, XFEL and PETRA III

M. Gasthuber, K. Schwarz
DESY, GSI

DLCL Structure of Matter: FAIR

- GSI hosts the German ALICE T2 centre, providing 7% of the ALICE T2 resources
- ALICE T2 jobs run in a multi-purpose HPC centre. Data to/from the HPC environment are tunneled through an XrootD forward proxy
- Storage Element: XrootD data servers on top of Lustre file system - an XrootD client plugin can provide direct access to Lustre for jobs running at GSI. Server side plugin is planned.
- ALICE T2 has successfully been moved to GC
- ALICE T2 jobs now in Hyperthreading mode
- experience gained in context of ALICE T2 provide an important guideline for the forthcoming distributed computing environment of FAIR