Storage Scale @ DESY

Stefan Dietrich, on behalf of IT-Systems
Hamburg, 2025-08-14

HELMHOLTZ

DESY.

Storage Scale @ DESY

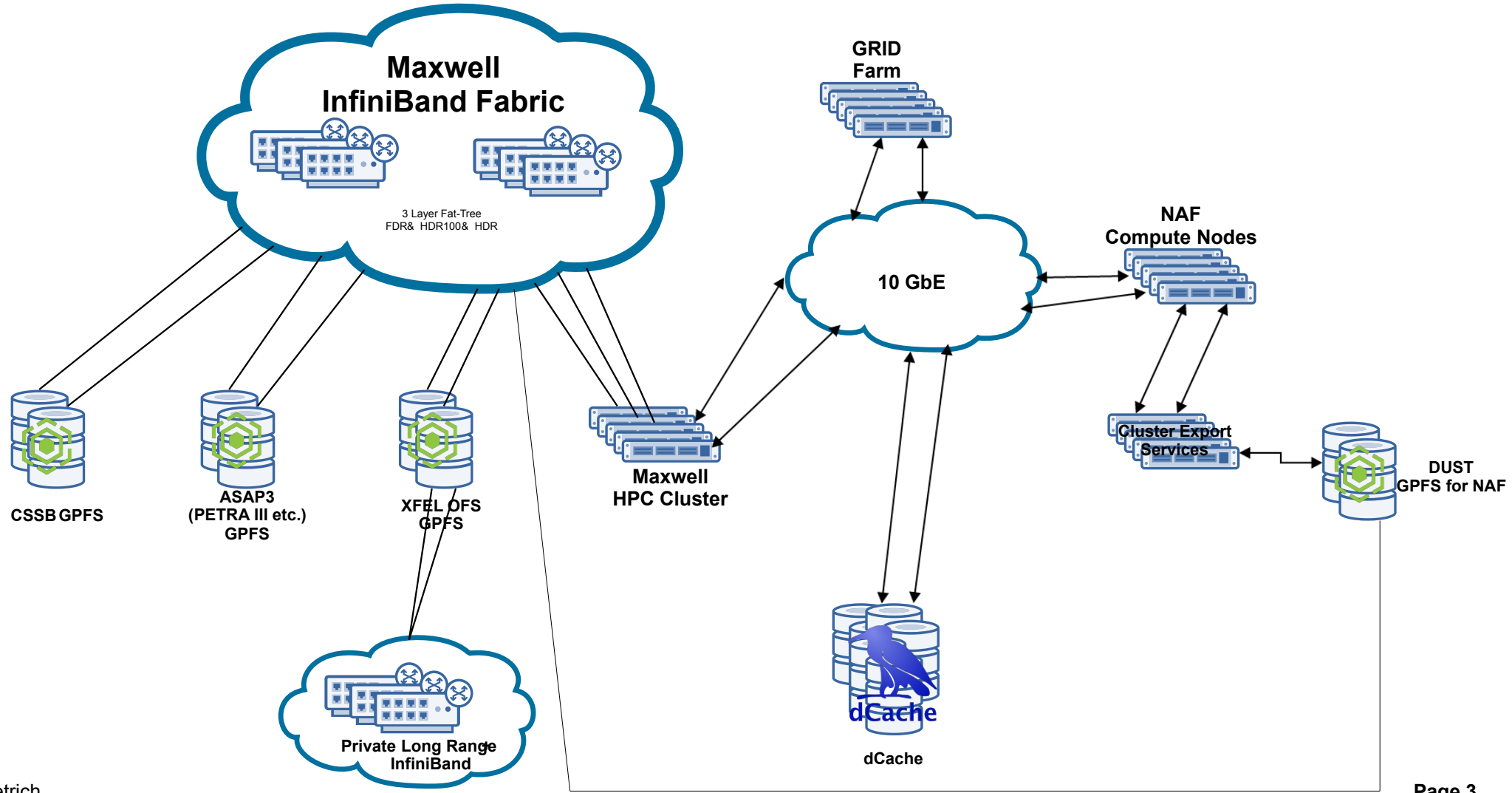
Overview

- Operating GPFS since ~2015
- No „traditional“ GPFS:
  always with GPFS Native RAID
- GPFS Native RAID: Declustered Software RAID
  - Erasure Coding
  - End-to-end checksumming
  - Fast rebuild times
  - IBM Storage Scale System or Lenovo DSS-G
- Several GPFS storage clusters in operation
  - core GPFS administration: 2 people

- XFEL
  - Data Storage for Eu.XFEL
  - ~60 PiB in OFS, ~11 PiB in ONS (Schenefeld)
- ASAP3
  - Primary storage for DESY Photon Science
  - ~19 PiB on HDD + ~450 TiB on NVMe
- CSSB (Centre for Structural Systems Biology)
  - Data from Cryogenic electron microscopy
  - ~11 PiB
- DUST (**D**ESY **U**ser **St**orage)
  - User/Project Space for IDAF (Maxwell & NAF)
  - ~5.3 PiB

Overview

A High Level View of Storage Systems and Compute Facilities



Storage Scale @ DESY

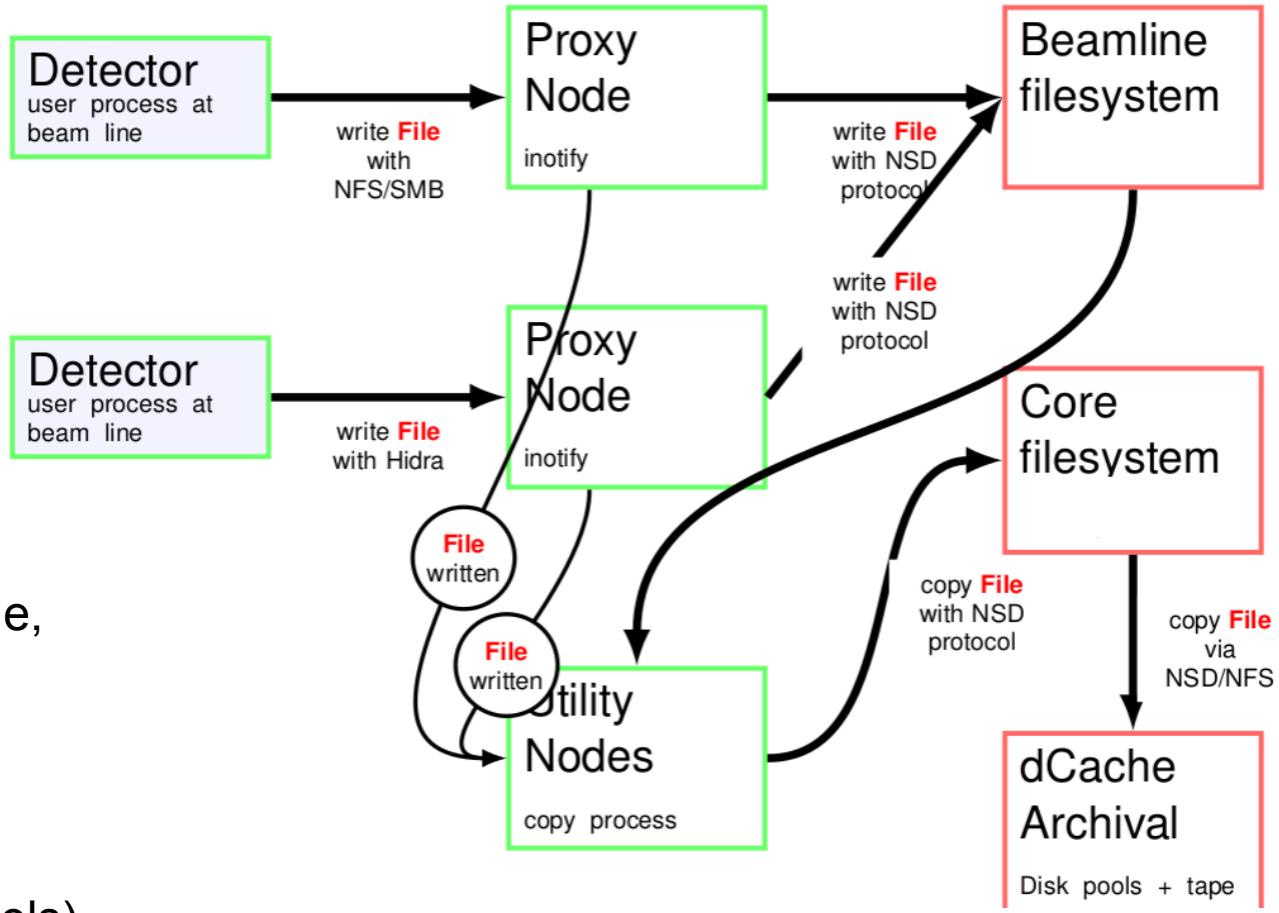
Noteworthy details

- GPFS access always with InfiniBand & RDMA
  - No native GPFS access with just TCP/IP!
- No InfiniBand? Cluster Export Services.
  - DUST → 6xCES Nodes with 2x100GbE serving NAF
- DUST is mounted on both Maxwell and NAF
  - Maxwell: Higher bandwidth due InfiniBand
  - NAF: NFSv4 via CES
  - Allow users to move between Maxwell and NAF
    → same path & data, just different worlds
- Currently running Scale 5.2.2 & 5.2.3
  - Storage blocks are patched 1-2x per year
    → automation provided by IBM/Lenovo
  - Clients: Bump version due to bugs or new RHEL minor release

ASAP3 – Data Storage for Photon Science @ DESY

DESY – PETRA III

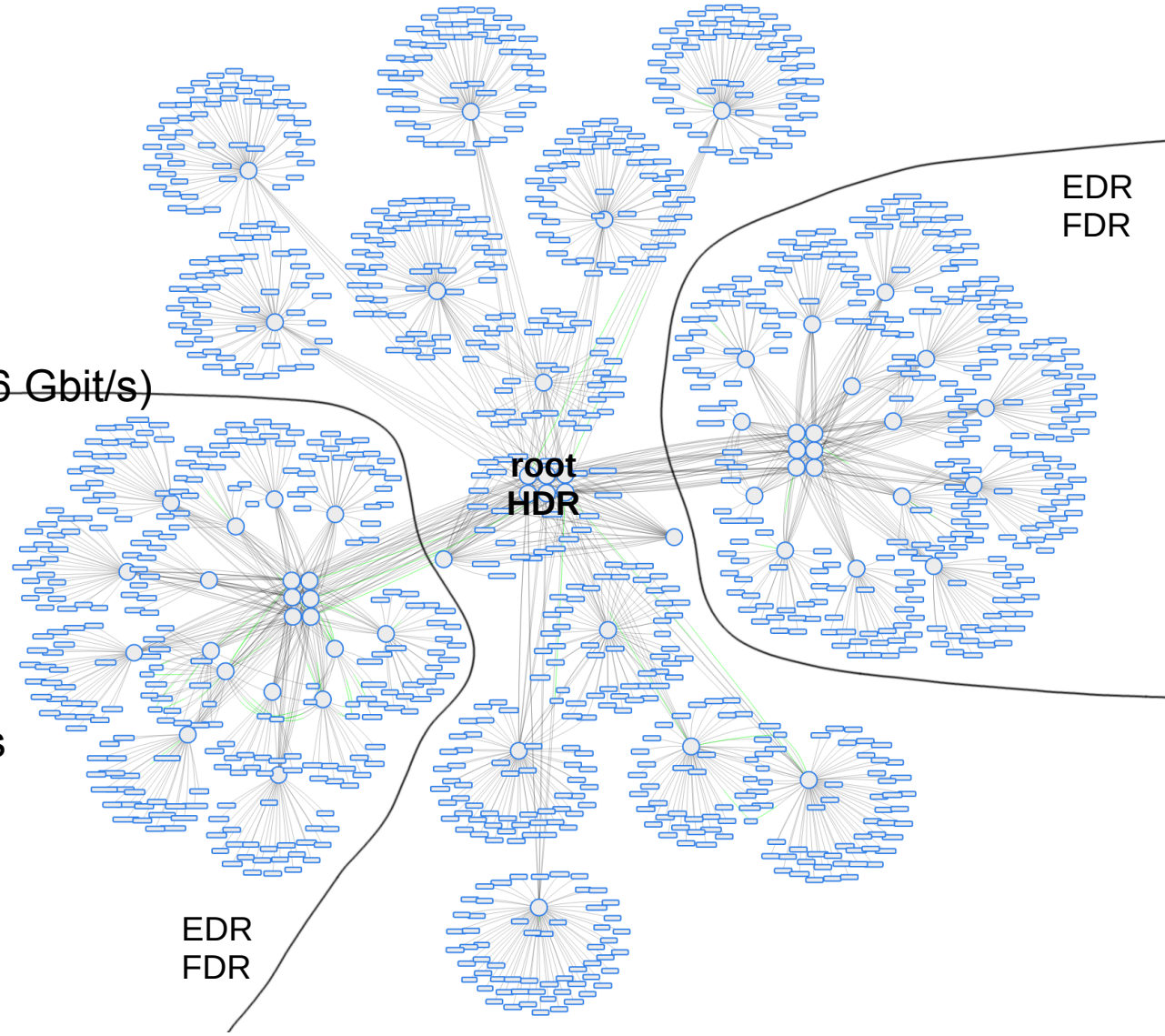
- See talk from Andre Rothkirch from Tuesday
- Data acquisition: HiDRA (ZMQ), NFS, SMB
- Detectors at beamline
  - Different vendors
    → Commercial: Dectris
    → Home-grown: Lambda
  - Example: Dectris Eiger 2 16M:
  - 100 GbE Ethernet, ~5 GiB/s in continues mode,
    ~8-9 GiB HDF5 container files
  - ~100 TiB generated in first beamtime
- Beamline filesystem: ~415 TiB NVMe (ESS 6000)
- Core Filesystem: ~19 PiB (various ESS HDD models)



InfiniBand

Maxwell InfiniBand Fabric

- 2/3 layered InfiniBand fabric
  - Root: 6xHDR (200 Gbit/s)
  - Old Top/Leaf: 12xEDR (100 Gbit/s), 26xFDR ~~(56 Gbit/s)~~
  - Leaf: 18xHDR
- Currently >=1300 active links
- Compute clients: 1xHDR100
- GPFS storage blocks: multiple HDR100 or HDR links
- Planning for an upgrade to NDR (400 Gbit/s)



Vielen Dank!
Fragen?