GPFS at DESY

Stefan Dietrich, Martin Gasthuber, Jürgen Hannappel
Hamburg, 2022-07-14

HELMHOLTZ

DESY.

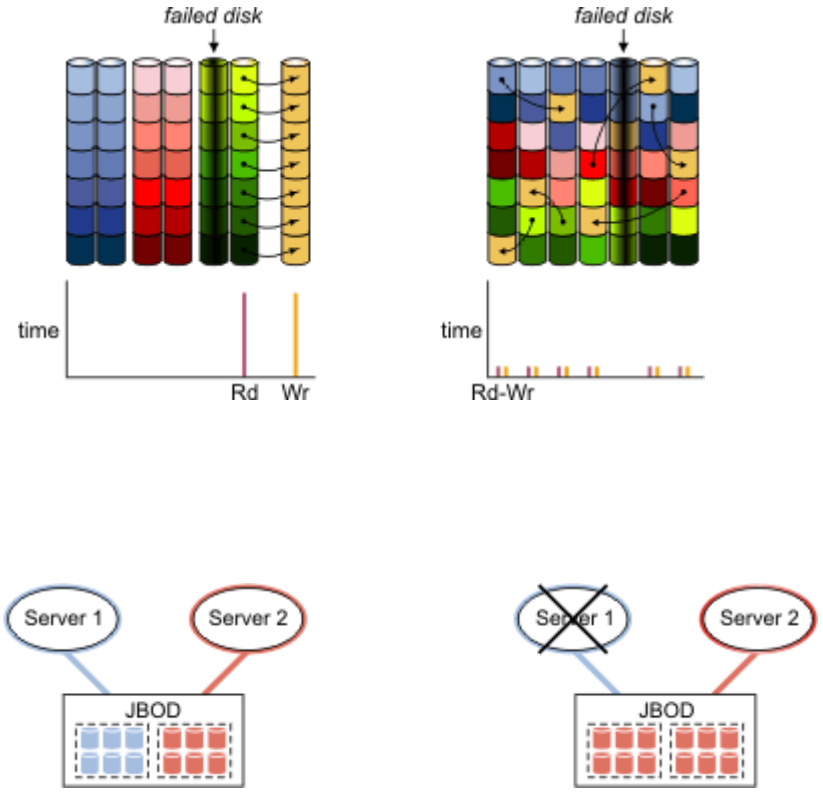
GPFS@DESY in a Nutshell

Some Numbers

- Started in ~2015, SPEED Project with IBM
- All storage systems based on GPFS Native RAID
  - 28xIBM ESS, 1xLenovo DSS-G
  - 1 cluster without GNR; NVMe drives
  - No Erasure Code Edition
- Licensing: Helmholtz ULA for clients/servers
  - GPFS Native RAID and Erasure Code Edition not included
- ~62 PiB of GPFS
- 8 Storage + 12 Remote Clusters
- InfiniBand in use as cluster interconnect
  - Mix of FDR, EDR, HDR100 and HDR
  - 1 cluster with 100 GbE and RoCEv2
- Spectrum Scale >=5.1.2
- All filesystems on format >=18.00 with increased subblock size
  - Migrated by copying data to filesystem with new format
- Separation between system and data pool
  - SAS SSD or NVMe for system pool
  - TRIM not yet enabled on NVMe

GPFS Native RAID (GNR)

- Software RAID implementation
  - No RAID controller, JBOD
  - Declustered RAID
  - Reed Solomon Erasure Coding or n-way replication
- Advantages over traditional GPFS setup
  - End-to-end checksumming, eliminates silent data corruption
  - Fast rebuilds
- Available in IBM ESS and Lenovo DSS-G
  - 2 servers for redundancy, 1-8 disk enclosures with redundant paths
- Erasure Code Edition
  - Same technology, but erasure coding over multiple servers

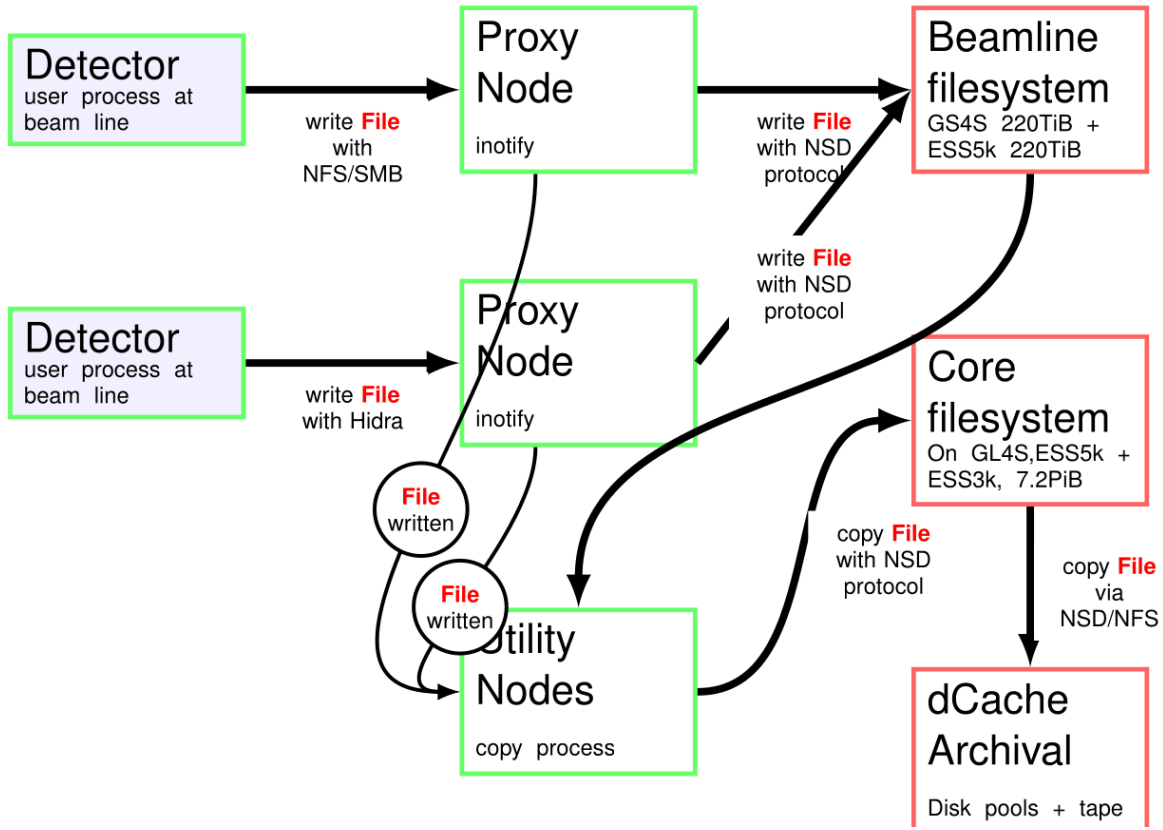


GPFS Use-Cases

- Storage for Photon Science at DESY
  - ASAP3 Data Storage System
  - Storage for data generated by local faciltities
    - PETRA III, FLASH, PETRA IV (preparation)
    - Research Groups
- Storage for European XFEL
  - Online Storage for data ingest
  - Offline storage for data analysis from Maxwell cluster
    - Long term storage: dCache
- „Scratch“ space for Particle Physics
  - Scratch space in the context of National Analysis Facility (NAF)
  - For german users, data analysis of LHC data
  - Access only via NFSv4 through Cluster Export Services
  - HTC compute cluster with ~400 nodes

ASAP3 – Data Storage for Photon Science

- Data Ingest from Beamlines with demanding detectors
  - Protocols: NFS, SMB, HiDRA or ASAP::O
  - 10 or 100 GbE connection
  - Initial landing point: Beamline filesystem based on SSD
  - Detector variety: >=250 Hz with 2,4 MiB files or ~9 GiB HDF5 files@5 GiB/s
- Core Filesystem
  - Data copy beamline -> core filesystem with custom copy tools, very small delay
  - Data resides on core filesystem for data analysis
- Long term storage: dCache
  - Data removed after 180 days from GPFS
  - Restage to GPFS on user request



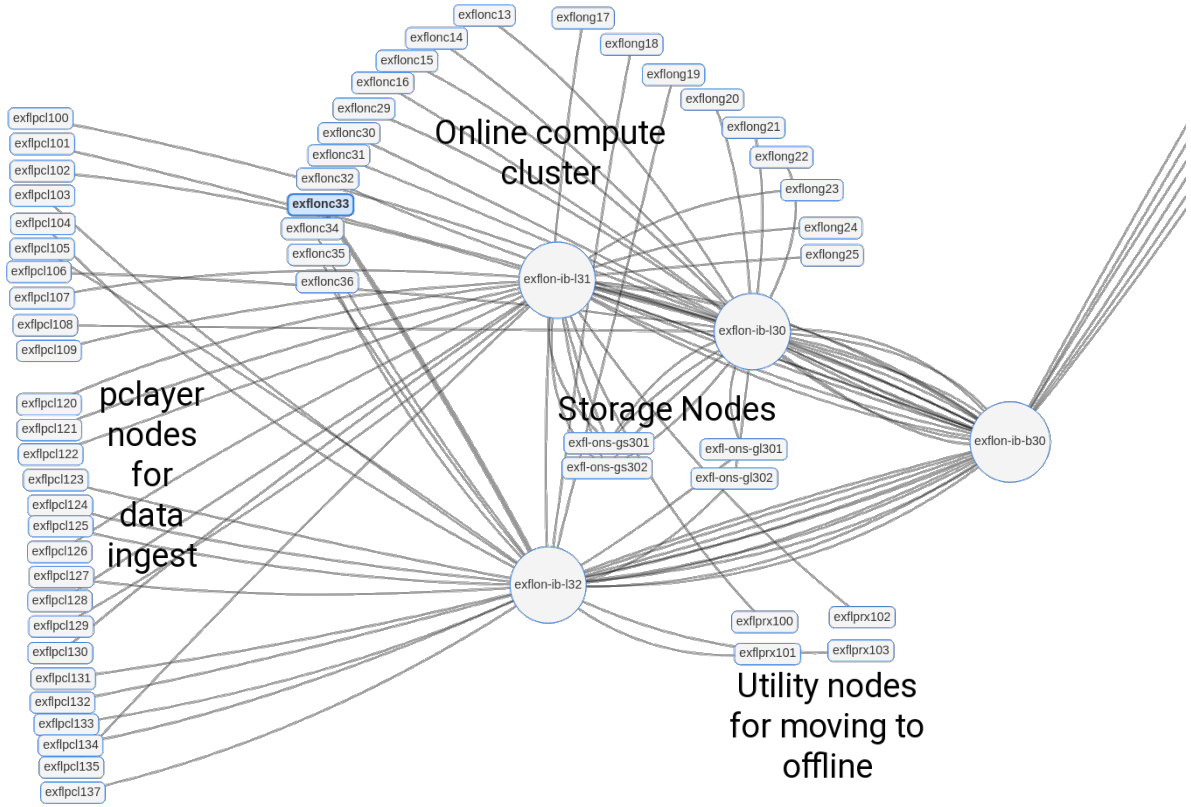
ASAP3 – Relevant GPFS Features

- Dependent Filesets
  - Beamtime represented as dependent fileset, for easy data management (~3500 filesets)
- GPFS Policy Runs
  - Generating lists, e.g. for copy processes to dCache
- kNFS for Beamline Filesystem
- Cluster Export Services for data access from desktops
- Snapshots for core filesystem

- Core Filesystem: ~11 PiB, ~652 Mio. files
  Blocksize: 1 MiB system + 8 MiB data
- Beamline Filesystem: ~440 TiB
  Blocksize: 2 MiB system + 4 MiB data

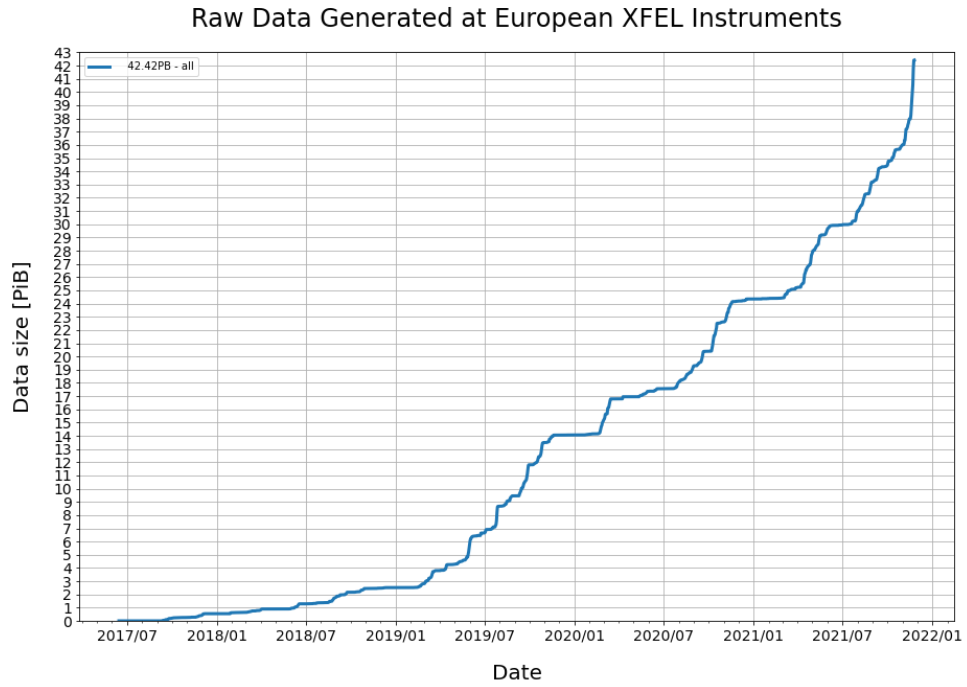
Data Storage for European XFEL

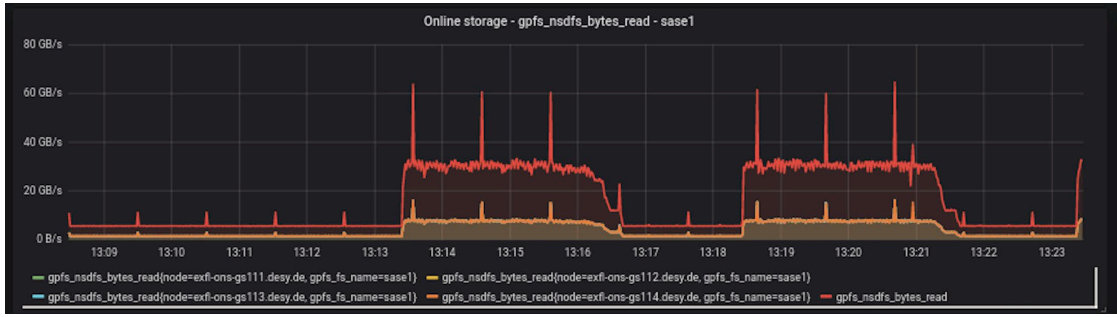
- Data Ingest in Schenefeld (~3.5km)
- Multiple SASEs, each with own storage and compute resources
  - 1 BB with SSD (~220 TiB), 1 BB with HDD (~1.6 PiB)
  - Data migrated from SSD -> HDD via Policy Engine
  - SASE1: Higher demand -> double storage resources
  - >=40 GiB/s for data ingest,
    async copy to Offline with ~30 GiB/s
- Offline Storage
  - Data migrated from Online -> Offline
    - Connection via long range InfiniBand
    - HDR switches with 20xEDR links

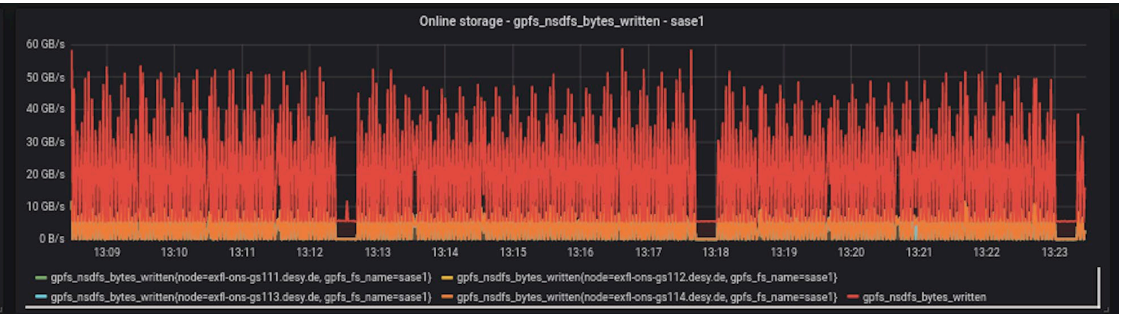


Data Storage for European XFEL

- Offline storage: biggest GPFS instance
  - 11 building blocks, ~40 PiB
  - Up to ~175 GiB/s reads observed from Maxwell compute cluster
- Custom copy process between Online -> Offline
- >=1000 dependent filesets, per proposal and type (raw, proc)
- Snapshots
- AFM for small data transfer Offline -> Online
- Watchfolder Evaluation: Identify hot files
  - Discovered multiple issues, still ongoing

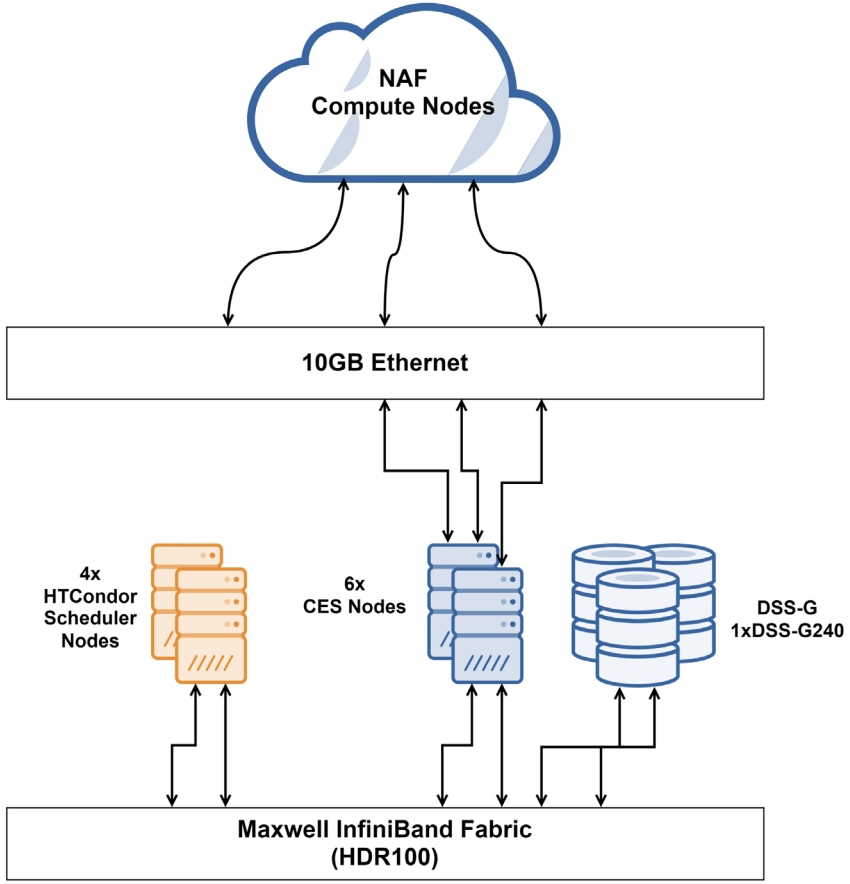






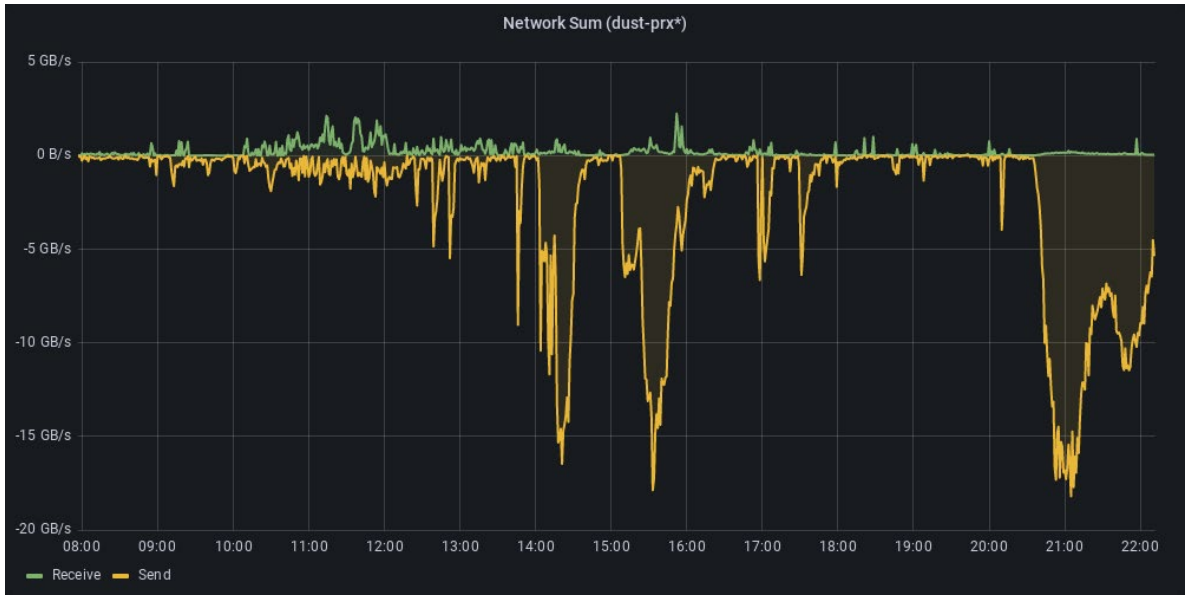
DUST – „scratch“ for Particle Physics

- „scratch“ space for NAF users
- DUST: Access only via NFSv4
  - ~400 NFS clients with 1/10GbE
  - 6 CES nodes with 4x10GbE (2x100GbE upgrade planned)
- …also provides $HOME and $SOFTWARE for Maxwell
  - Future: IDAF, option to provide DUST in Maxwell with native GPFS
- DUST: ~1.6 PiB, ~956 Mio. files, 1MiB + 16 MiB
  Maxwell $HOME: ~15 TiB, ~85 Mio. files, 1 MiB + 4 MiB



DUST – Ganesha Performance

- NAF batch jobs can saturate the network for DUST
- HTCondor schedulers require **very stable** storage acccess
  -> native GPFS access instead of NFSv4
- 6 CES nodes: 4 for non-interactive and 2 for interactive NFS clients
- ~10 months for Ganesha 2.3 -> 2.7 migration
  - …Ganesha 2.7 -> 3.5: no issues ¯_(ツ)_/¯
  - Failover seems to be broken in 3.5



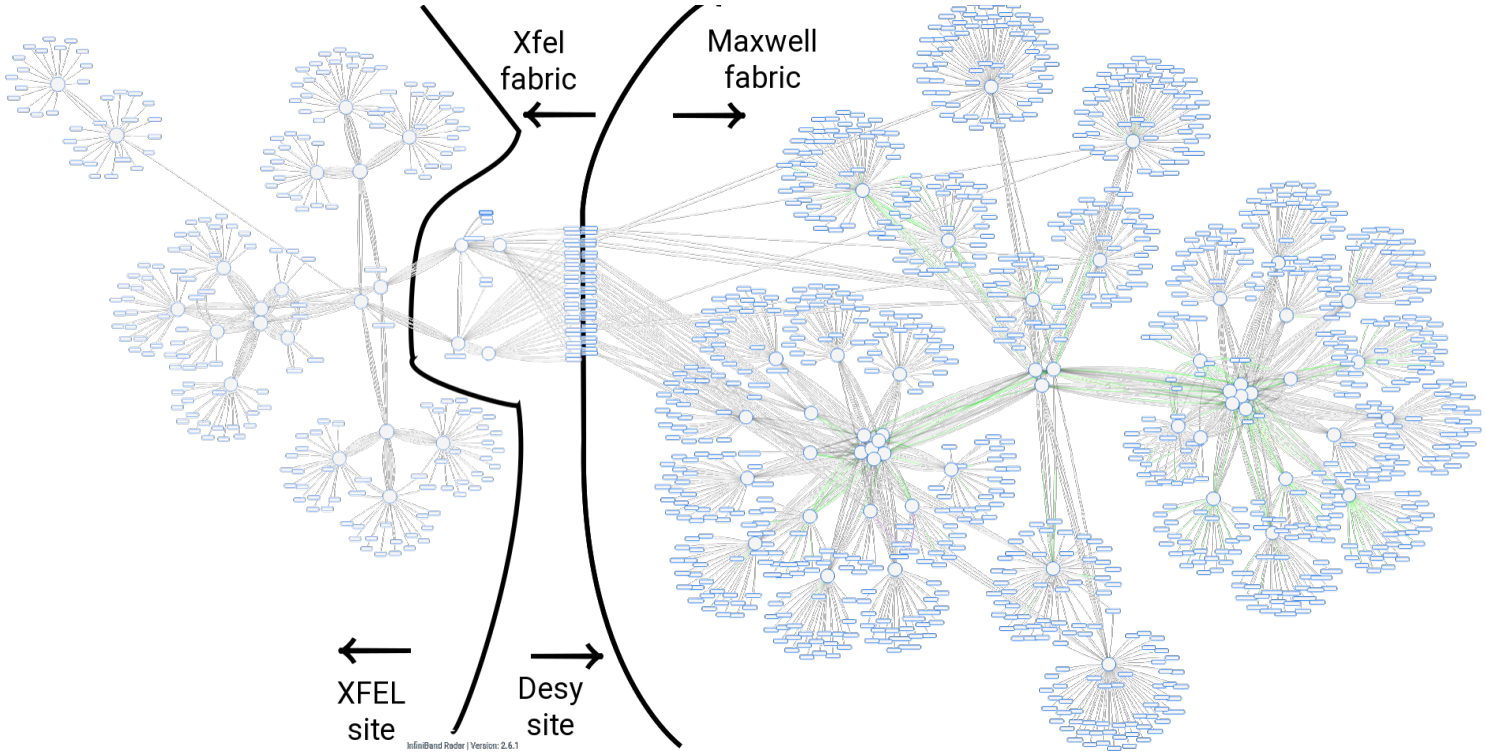
Maxwell Cluster

https://maxwell.desy.de

- Computing Platform provided by IT for
  - Photon Science Data Analysis
  - GPU accelerated computations (AI)
  - General HPC and scientific computing
- Fast dedicated storage
  - GPFS & BeeGFS
- Fast, low latency InfiniBand
  -> Maxwell InfiniBand fabric
- ~800 CPU+GPU nodes
- SLURM for job scheduling
  - Core- and group specific partitions
- Buy-in model for groups
- Both offline and near-realtime data analysis available

Maxwell InfiniBand Fabric

- 2/3 Layer Fat-Tree Topology
- Grown over time
  - From 2 to 3 layer and FDR -> EDR -> HDR
- Today
  - ~1400 ports available for clients
  - # switches: 17xHDR, 12xEDR, 29xFDR
- XFEL: ESS nodes act as gateway
  - Connected to Maxwell and XFEL via long range InfiniBand



Issues and Experiences

- Overall, GNR systems perform well and run stable
- ESS/DSS-G specific issues
  - Deployment buggy, require special (read: useless) deployment networks
  - Big chunks of storage, buy-in model not feasible
  - Odd hardware selection by IBM for certain ESS generations, no BMC or USB Ethernet
- Deadlocks
- Filesystem copy required for new filesystem enhancements
- Stable GPFS operation depends on very stable network
  - …also check the arp cache settings for Linux…
- Ganesha Issues
  - No easy debugging, often asked for FULL_DEBUG logs
- Fileset limits: more independent filesets beneficial
- Important configurations at filesystem creation time
- Performance tuning
  - Arcane amount of settings, often undocumented or outdated
  - Slow file open: Impossible to diagnose without support case
  - To be investigated: >=250 Hz file creation rate issues

Thank you