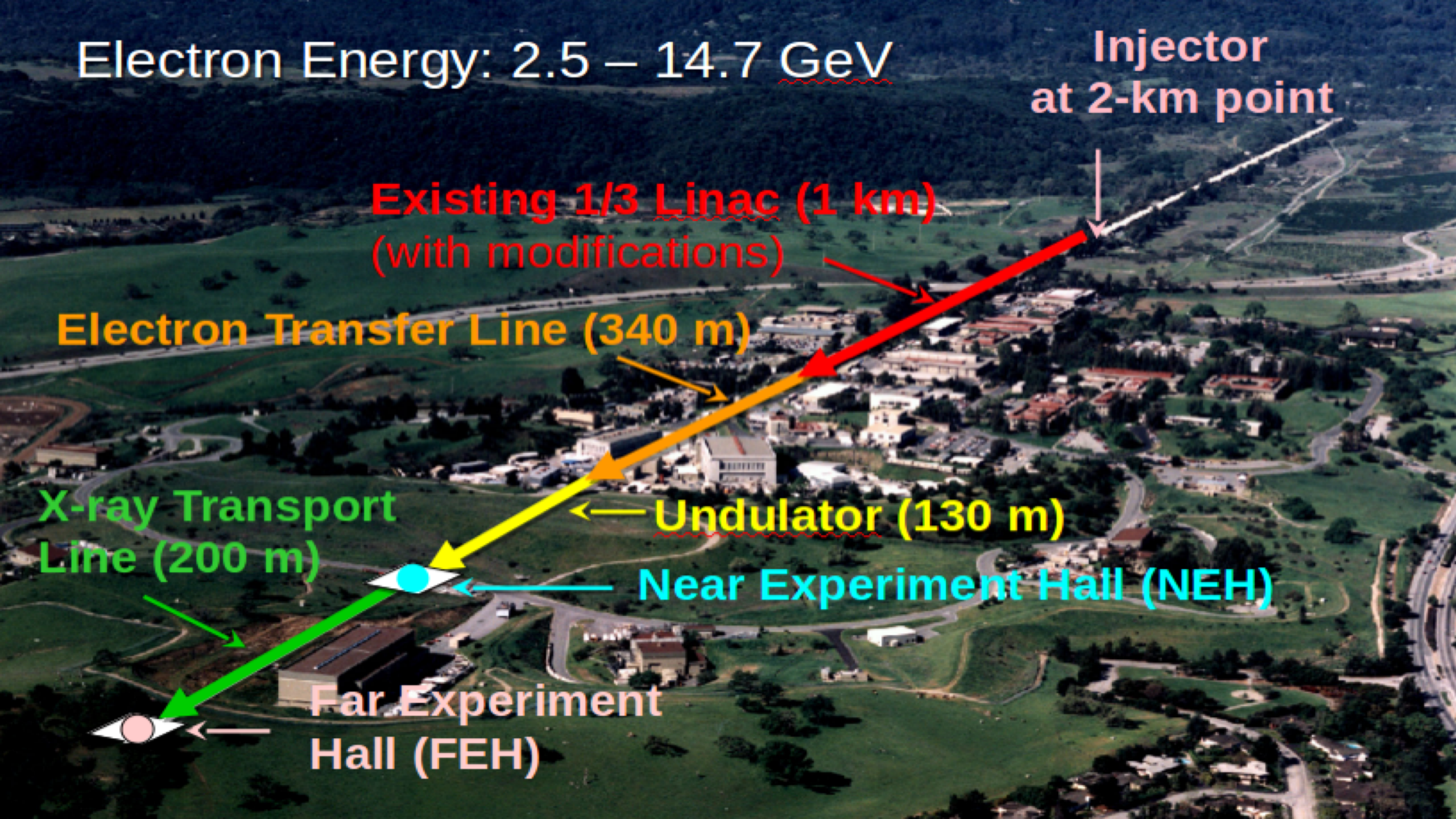
LCLS

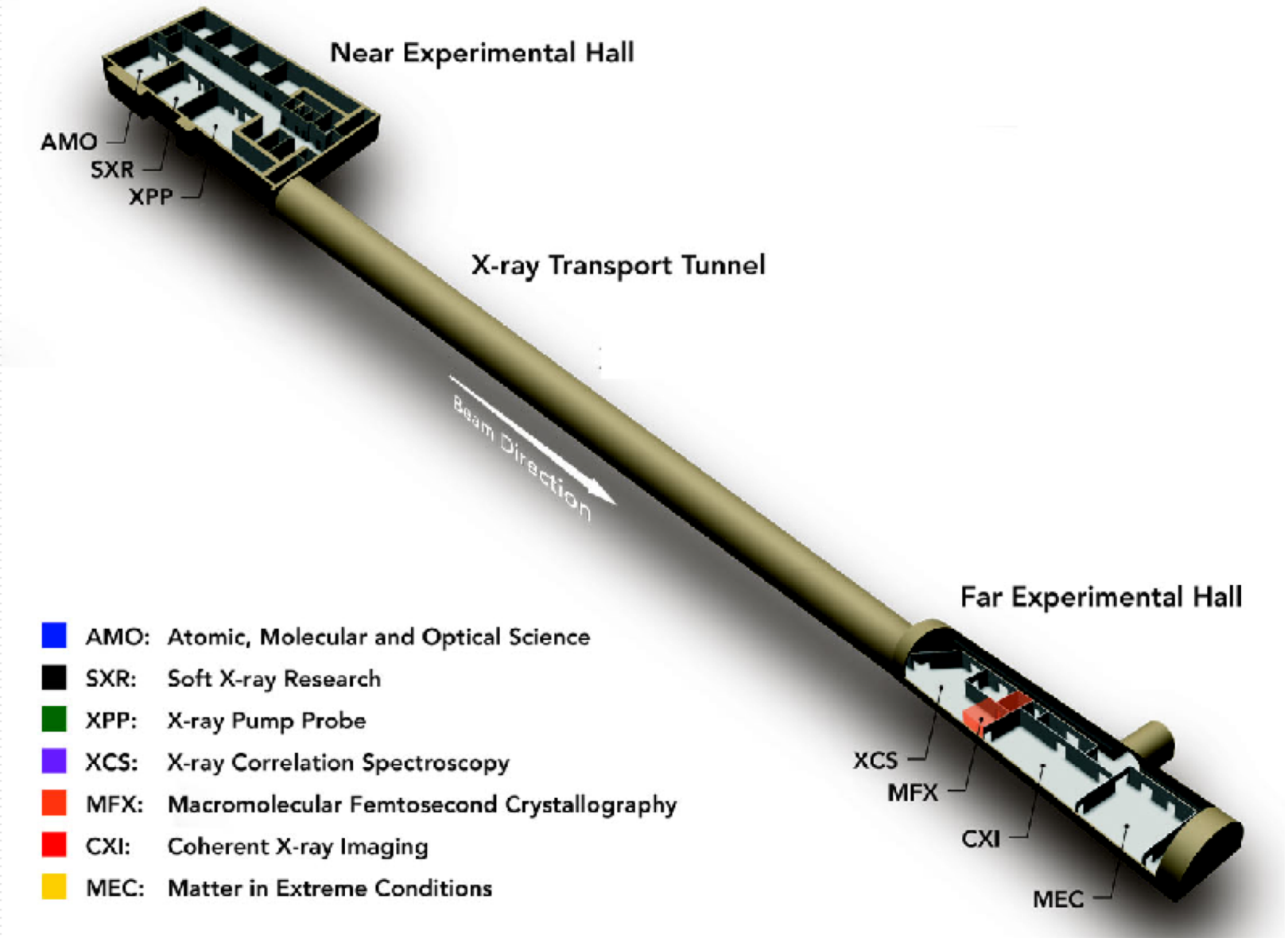
SLAC NATIONAL ACCELERATOR LABORATORY

Scalable High Performance Storage based on Lustre/ZFS over NVMe SSD @LCLS/SLAC

Riccardo Veraldi


Electron Energy: 2.5 – 14.7 GeV
Injector
at 2-km point
Existing 1/3 Linac (1 km)
(with modifications)
Electron Transfer Line (340 m)
X-ray Transport
Line (200 m)
Undulator (130 m)
Near Experiment Hall (NEH)
Far Experiment
Hall (FEH)

LCLS Instruments



LCLS has already had a significant impact on many areas of science, including:

- ➔ Resolving the structures of macromolecular protein complexes that were previously inaccessible
- ➔ Capturing bond formation in the elusive transition-state of a chemical reaction
- ➔ Revealing the behavior of atoms and molecules in the presence of strong fields
- ➔ Probing extreme states of matter

LCLS

SLAC NATIONAL ACCELERATOR LABORATORY

Data Analytics for high repetition rate Free Electron Lasers

FEL data challenge:

- **Ultrafast X-ray pulses** from LCLS are used like flashes from a high-speed strobe light, producing stop-action movies of atoms and molecules
- Both **data processing** and **scientific interpretation** demand intensive computational analysis

LCLS-II will increase **data throughput by three orders of magnitude** by 2025, creating an exceptional scientific computing challenge

LCLS-II represents SLAC's largest data challenge

Computing Requirements for Data Analysis: a *Day in the Life of a User* Perspective

- During **data taking**:
  - Must be able to get real time (~1 s) **feedback** about the **quality of data taking**, e.g.
    - Are we getting all the required detector contributions for each event?
    - Is the hit rate for the pulse-sample interaction high enough?
  - Must be able to get **feedback** about the **quality of the acquired data** with a latency lower (~1 min) than the typical lifetime of a measurement (~10 min) in order to optimize the experimental setup for the next measurement, e.g.
    - Are we collecting enough statistics? Is the S/N ratio as expected?
    - Is the resolution of the reconstructed electron density what we expected?
- During **off shifts**: must be able to run **multiple passes** (> 10) of the full analysis on the data acquired during the previous shift to optimize analysis parameters and, possibly, code in preparation for the next shift
- During **4 months** after the experiment: must be able analyze the raw and intermediate data on **fast access storage** in preparation for publication
- **After 4 months**: if needed, must be able to **restore** the archived data to test new ideas, new code or new parameters

The Challenging Characteristics of LCLS Computing

1. **Fast feedback** is essential (seconds / minute timescale) to reduce the time to complete the experiment, improve data quality, and increase the success rate
2. **24/7 availability**
3. **Short burst** jobs, needing very short startup time
4. **Storage** represents significant fraction of the overall system
5. **Throughput** between storage and processing is critical
6. Speed and flexibility of the **development cycle** is critical - *wide variety of experiments, with rapid turnaround, and the need to modify data analysis during experiments*

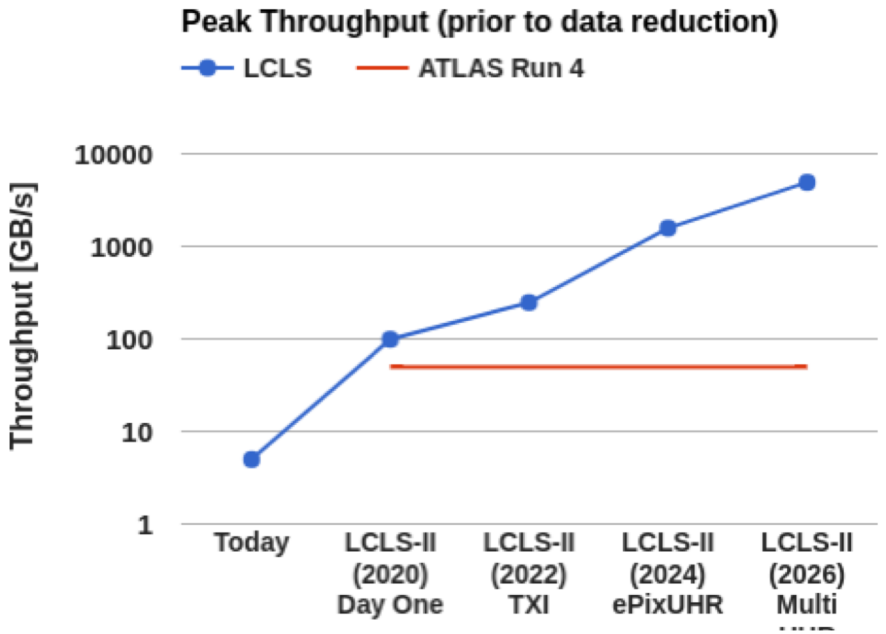
Example data rate for LCLS-II (early science)

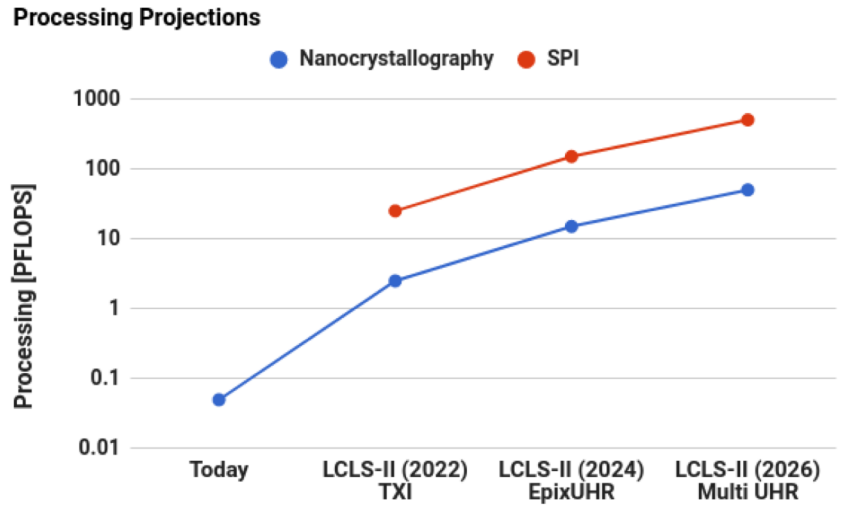
- 1 x 4 Mpixel detector @ 5 kHz = **40 GB/s**
- 100K points fast digitizers @ 100kHz = **20 GB/s**
- Distributed diagnostics 1-10 GB/s range

Example LCLS-II and LCLS-II-HE (mature facility)

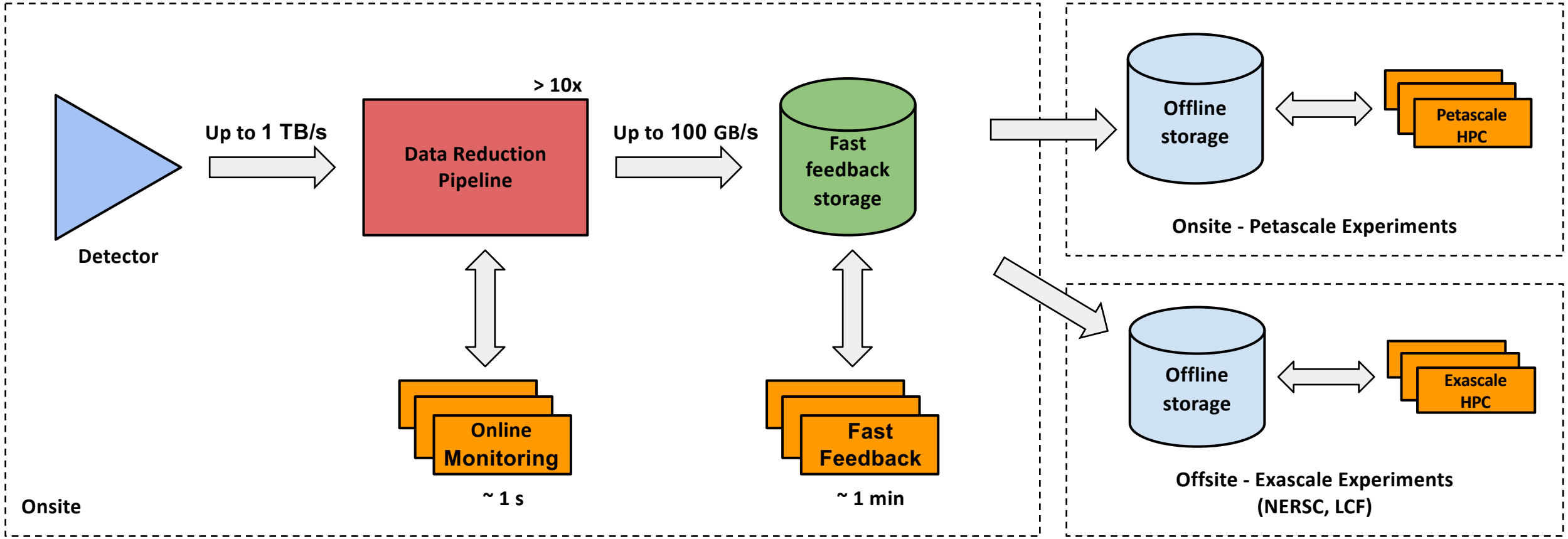
- 2 planes x 4 Mpixel ePixUHR @ 100 kHz = **1.6 TB/s**

Sophisticated algorithms under development within ExaFEL (e.g., M-TIP for single particle imaging) will require exascale machines





LCLS-II Data Flow



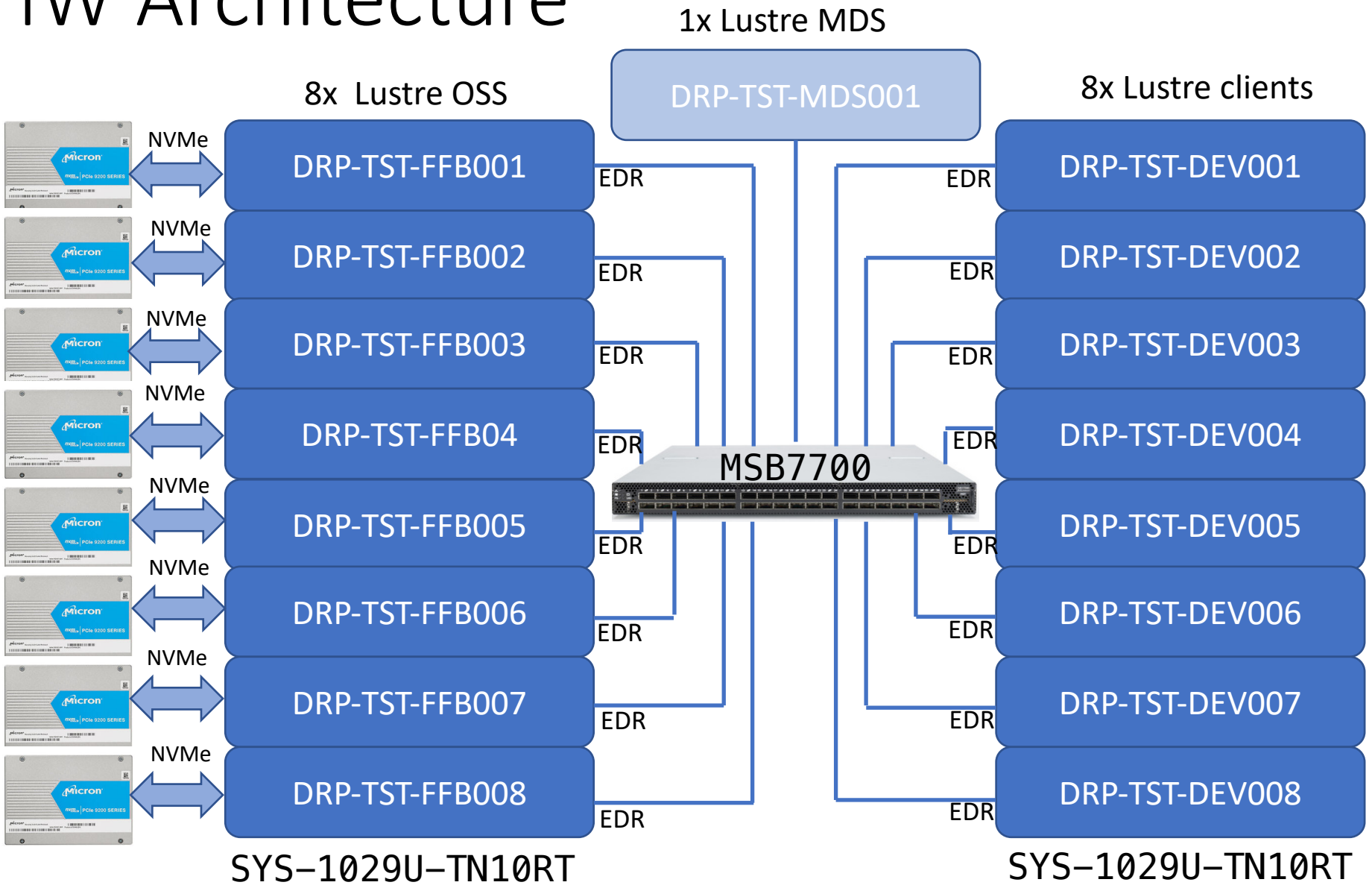
Data reduction mitigates storage, networking, and processing requirements

FFB HW Architecture

NAME	STATE	READ	WRITE	CKSUM
ffb01-ost01	ONLINE	0	0	0
raidz1-0	ONLINE	0	0	0
nvme0n1	ONLINE	0	0	0
nvme1n1	ONLINE	0	0	0
nvme2n1	ONLINE	0	0	0
nvme3n1	ONLINE	0	0	0

Micron NVMe 9300 MAX

65T



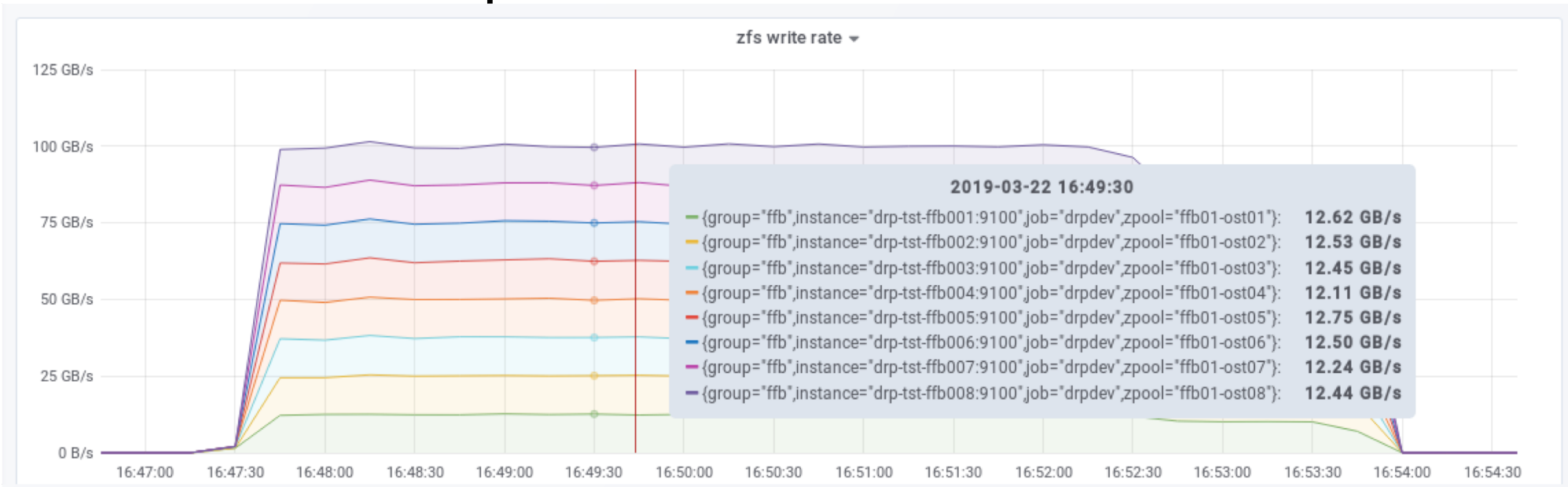
FFB SW Architecture

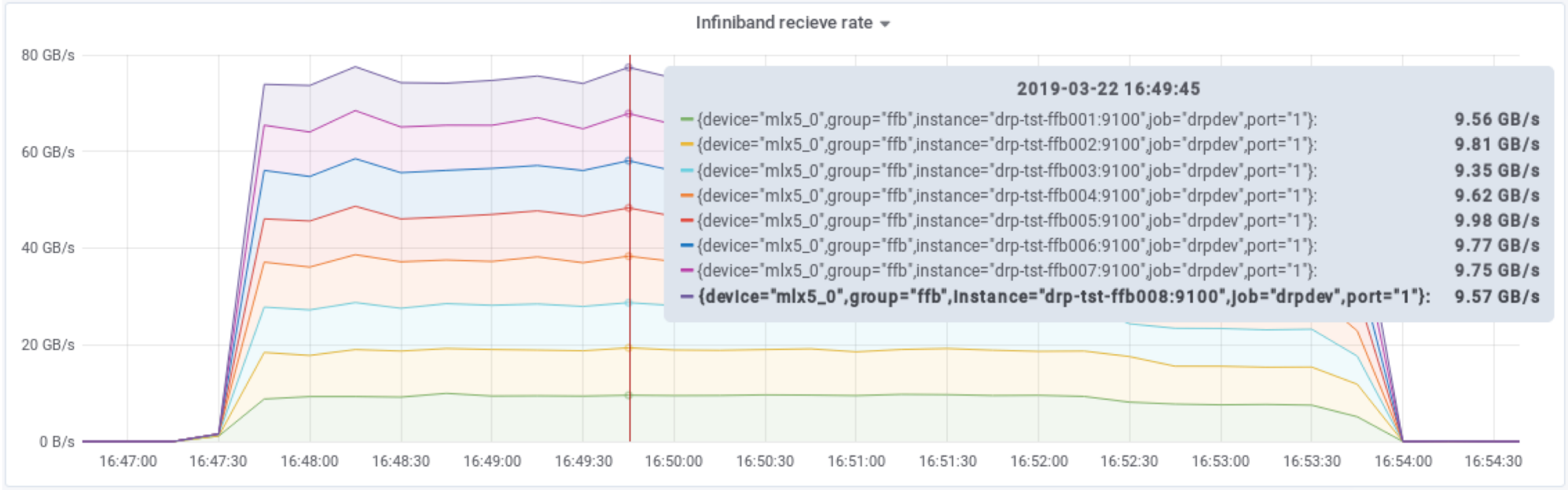
- RHEL 7.6 3.10.0-957.5.1.el7.x86_64
- Lustre Server/Client 2.12.0
  - 8x Lustre Servers
    - 1x OST per OSS
    - 1x raidz ZPOOL per OST (each OST is a raidz), each OSS has ONE OST
  - 8x Lustre Clients
- ZFS 0.7.12 (ZoL)
- Several ZFS kmod optimizations:
  - zfs_vdev_sync_write_max_active
  - zfs_vdev_sync_read_max_active
  - zfs_vdev_async_write_max_active
  - zfs_vdev_async_read_max_active
  - …
- Several Lustre parameters optmizations
  - max_pages_per_rpc
  - max_rpcs_in_flight
  - …
- Mounted partition
  - **172.21.52.149@o2ib:/ffb01 65T 45T 21T 68% /ffb01**

Testing performance

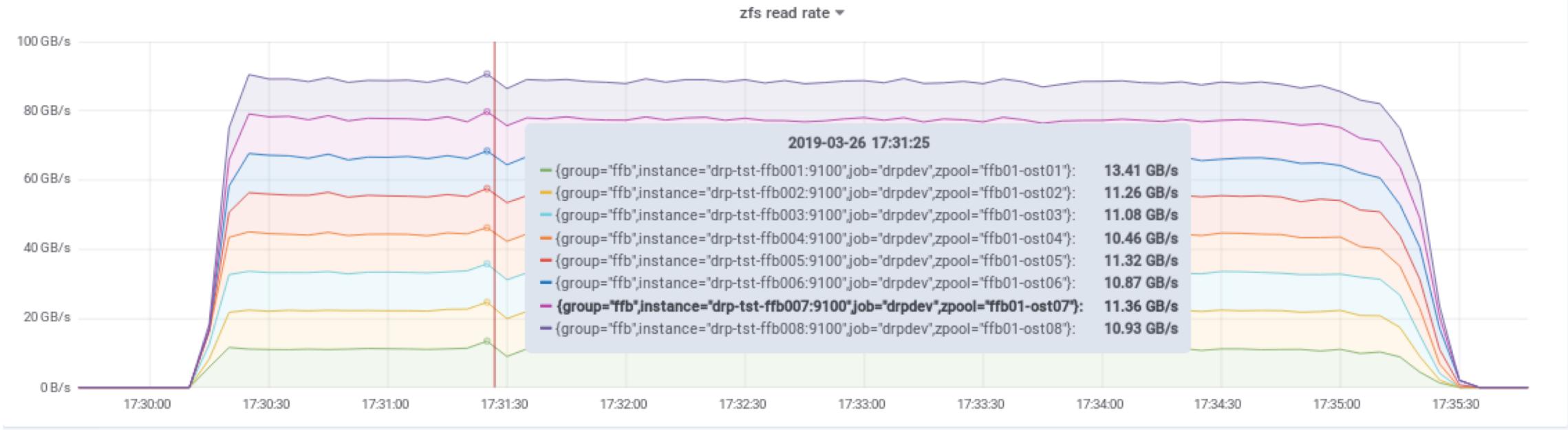
- Parallel write from each client 4 sequential instances per OSS/OST
  - 32 parrallel writes for each client
- Parallel read from each client 4 sequential instances per OSS/OST
  - 32 parrallel reads for each client
- Used custom code for testing performance: *https://github.com/rveraldi/ccff*
  - Validated then by well know tools (fio, iozone)

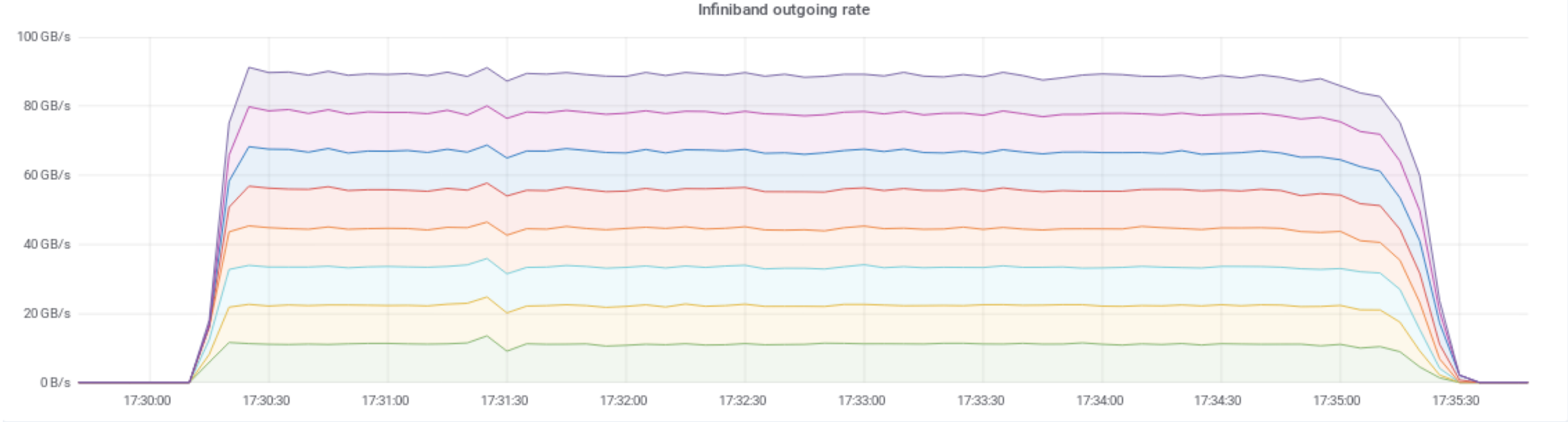
Lustre wr performance





Lustre read performance





Considerations

- ZFS raidz hw/sw capabilities are saturated (12GB/s per each OSS server)
- IB data rate close to 80GB/s
- More performance can be gained using mirror instead of raidz
  - Need more NVMe/SSD devices
    - Waste of storage space
  - No ZFS raid/mirror would imply no redundancy, no data protection
- Performance scales up linearly adding more OSS servers