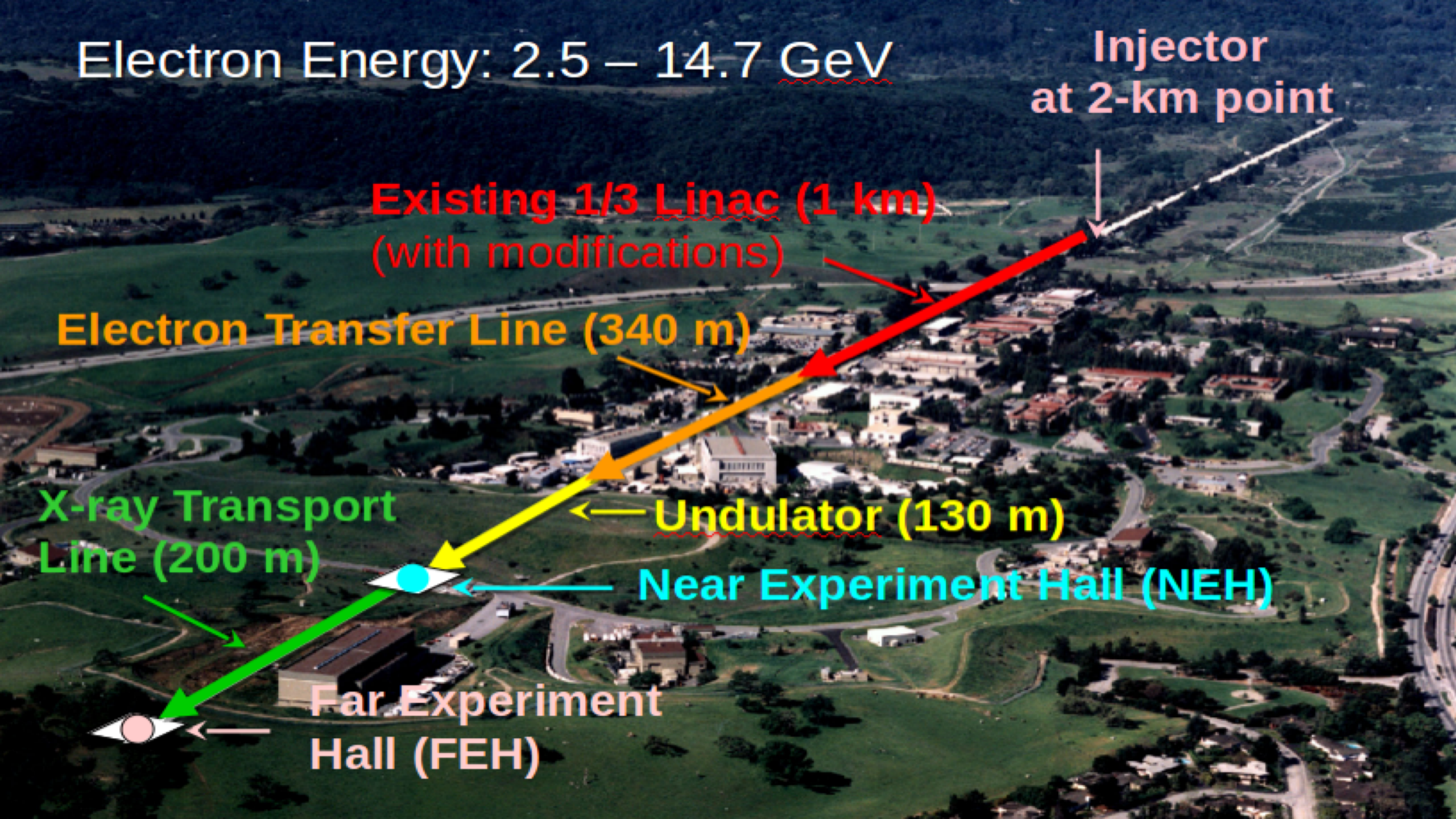
Scalable High Performance Storage based on Lustre/ZFS over NVMe SSD @LCLS/SLAC

Riccardo Veraldi

Lustre HA cluster deployment with Kubernetes

Federico Fornari


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: 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

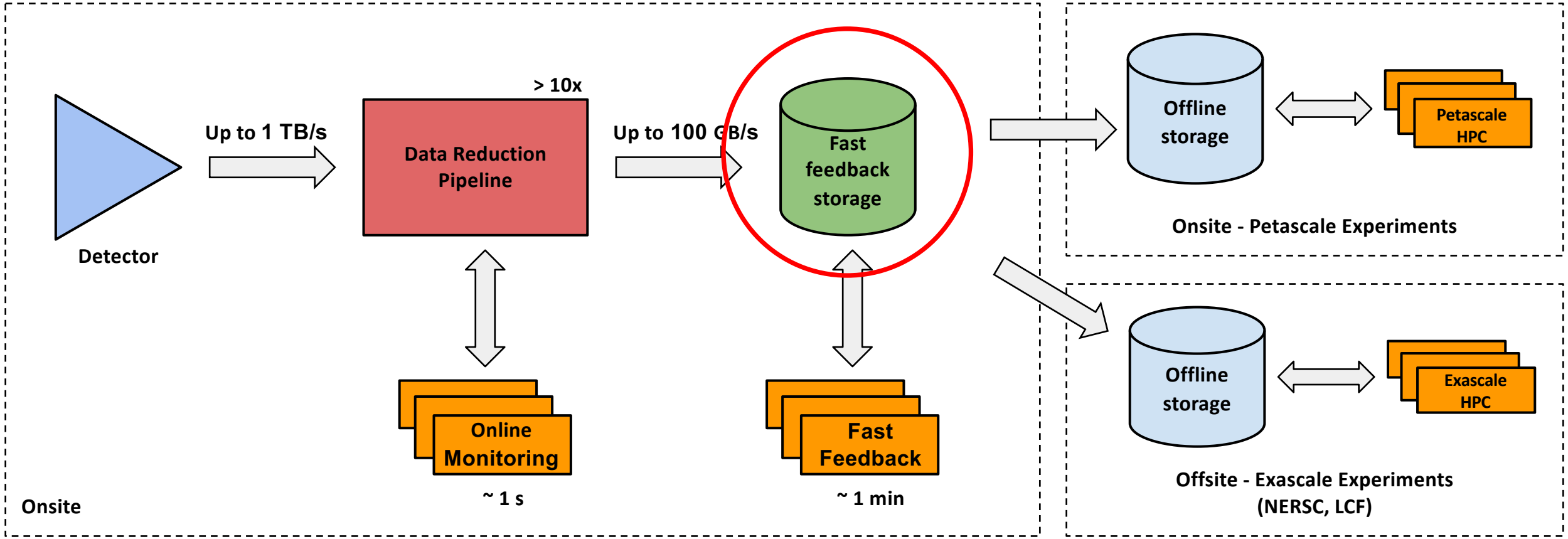
LCLS-II will increase **data throughput by three orders of magnitude** by 2025, creating an exceptional scientific computing challenge

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.
  - 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.
- 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

LCLS-II Data Flow



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

NVMe

1x Lustre MDS

DRP-TST-MDS001

8x Lustre OSS

DRP-TST-FFB001
DRP-TST-FFB002
DRP-TST-FFB003
DRP-TST-FFB04
DRP-TST-FFB005
DRP-TST-FFB006
DRP-TST-FFB007
DRP-TST-FFB008

EDR

MSB7700

8x Lustre clients

DRP-TST-DEV001
DRP-TST-DEV002
DRP-TST-DEV003
DRP-TST-DEV004
DRP-TST-DEV005
DRP-TST-DEV006
DRP-TST-DEV007
DRP-TST-DEV008

SYS-1029U-TN10RT

SYS-1029U-TN10RT

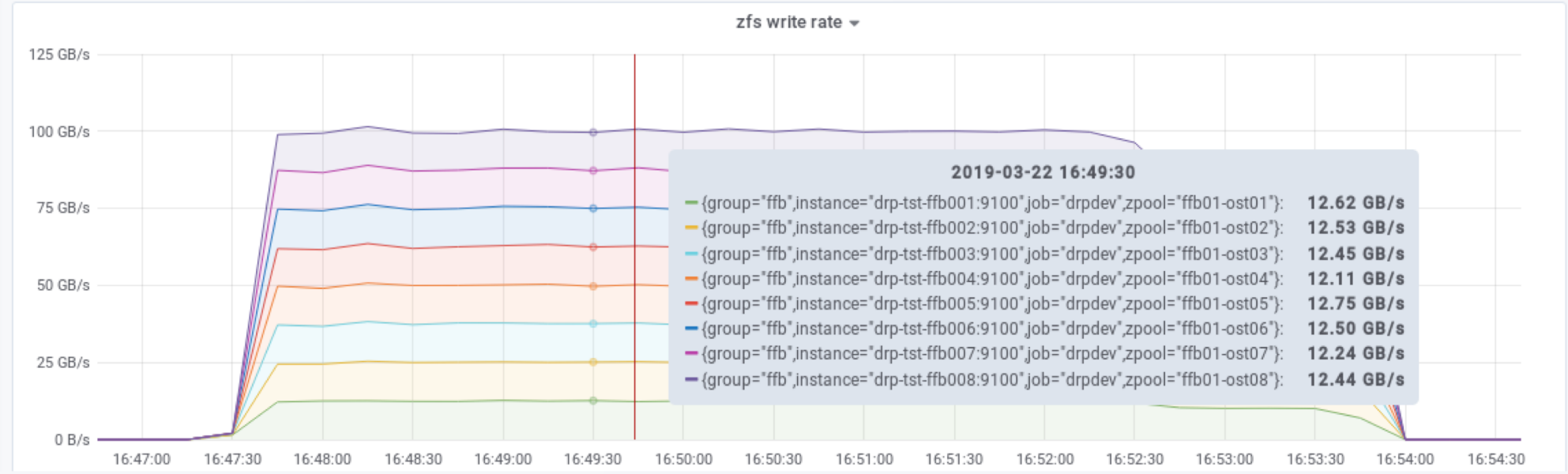
FFB SW Architecture

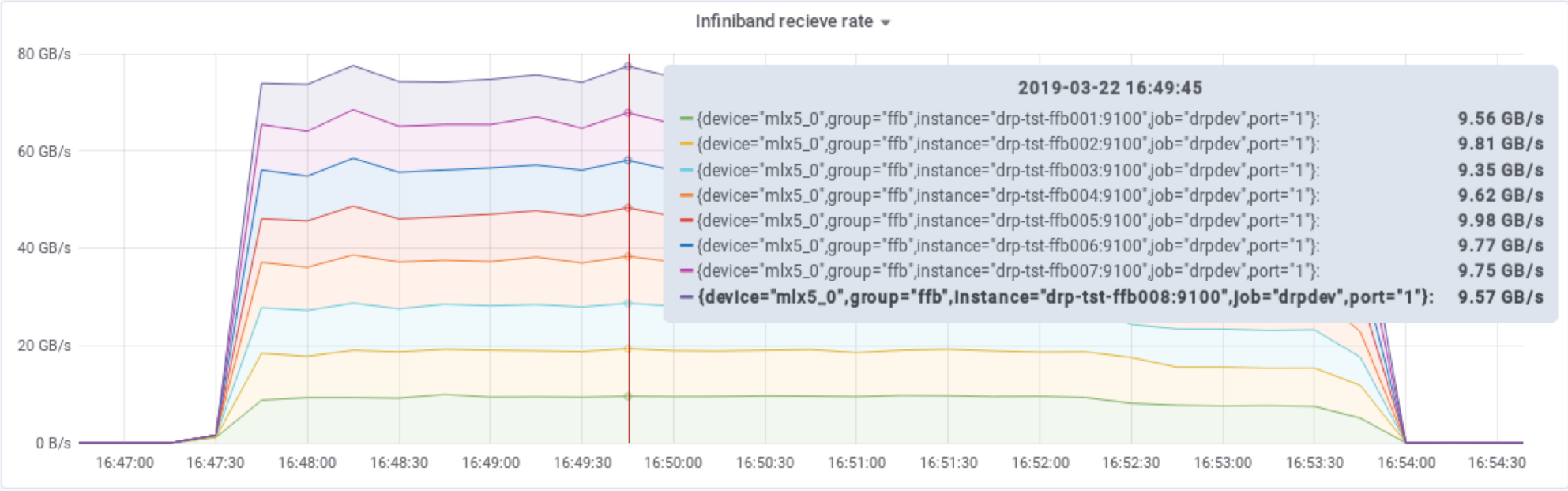
- RHEL 7.6 3.10.0-957.5.1.el7.x86_64
- Lustre Server/Client 2.12.0
  - 8x Lustre Servers
    - 1x OST per OSS
    - 1x raidz ZPOOL per OST (each OST is a raidz), each OSS has ONE OST
  - 8x Lustre Clients
- ZFS 0.7.12 (ZoL)
- Several ZFS kmod optimizations:
  - zfs_vdev_sync_write_max_active
  - zfs_vdev_sync_read_max_active
  - 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

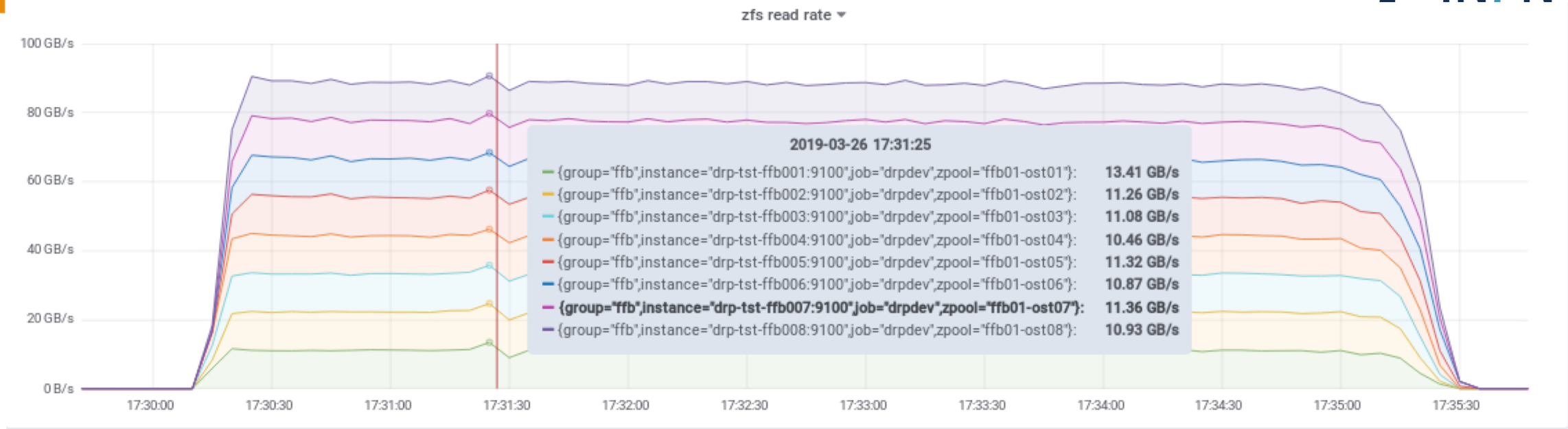
- Parallel write from each client 4 sequential instances per OSS/OST
  - 32 parrallel writes for each client
- Parallel read from each client 4 sequential instances per OSS/OST
  - 32 parrallel reads for each client
- Used custom code for testing performance: *https://github.com/rveraldi/ccff*
  - Validated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

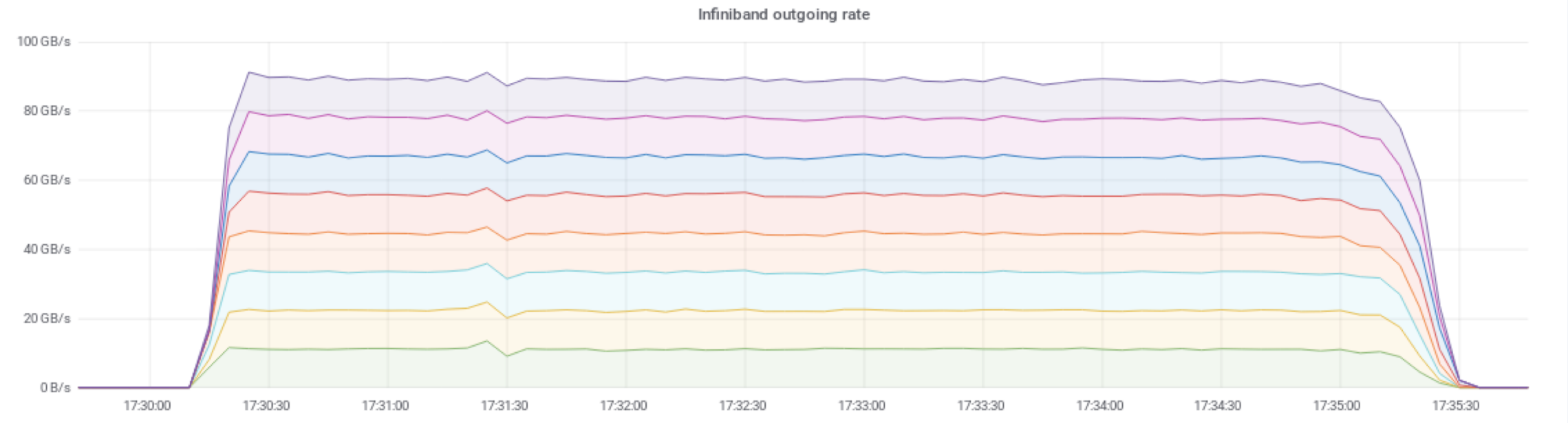
Lustre wr performance





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

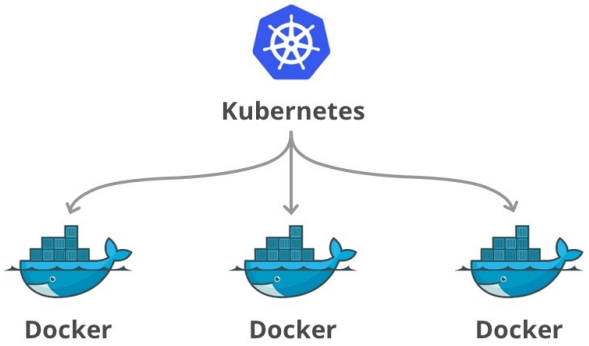
Lustre HA cluster deployment with Kubernetes

K8s Lustre Cluster Setup

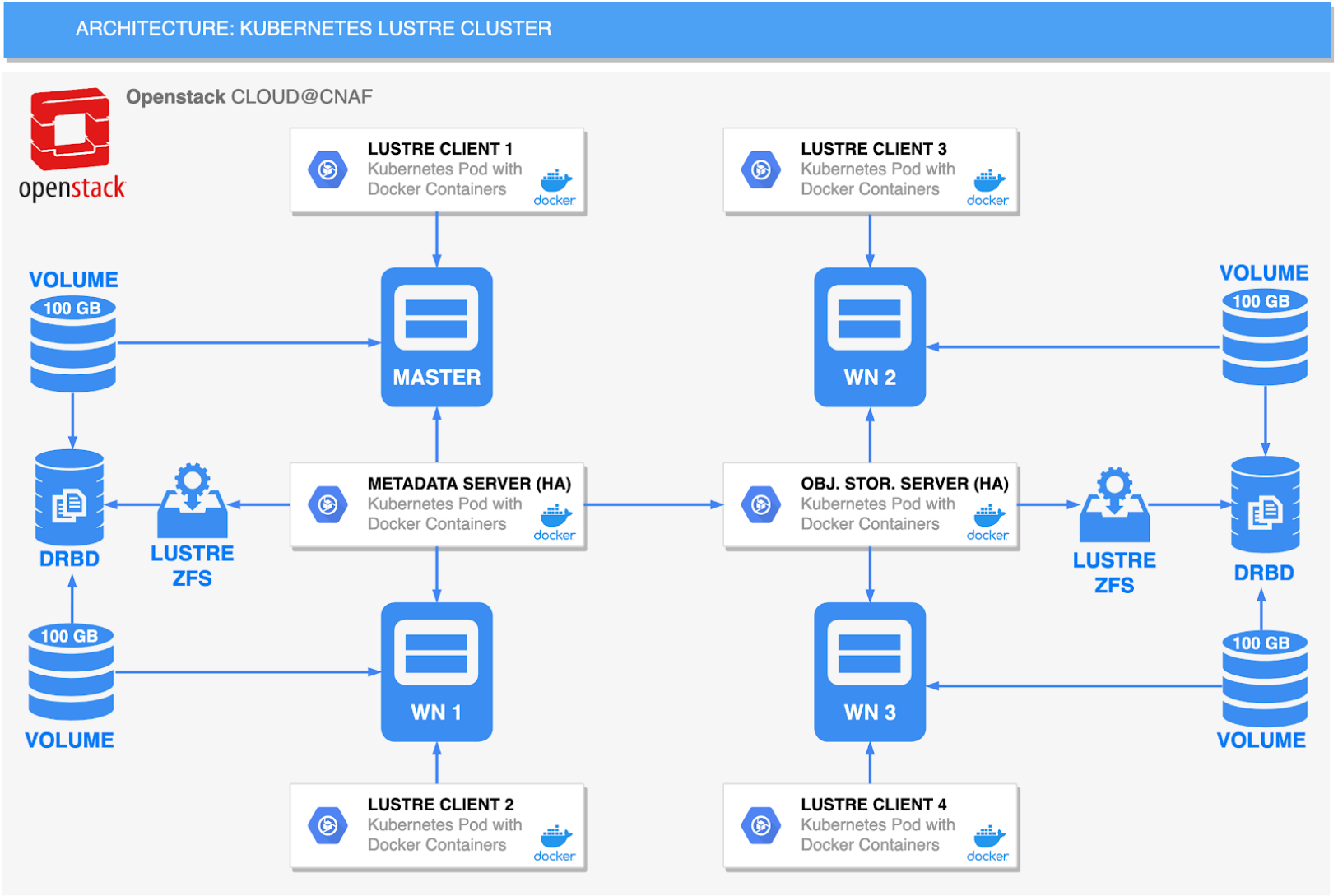
4 Openstack CentOS 7.8.2003 VMs:

Instance Name	VCPUs	RAM (MB)	Disk (GB)	Usage (hours)	State
k8s-worker-2	4	8196	80	27,99	Active
k8s-worker-1	4	8196	80	30,01	Active
k8s-master	8	16384	160	44,56	Active
k8s-worker-3	4	8196	80	27,96	Active

Orchestrator: **Kubernetes 1.19.2**
Virtualizer: **Docker 19.03.13**



Filesystem: **Lustre 2.12.5 (ZFS 0.7.13)**



K8s Lustre GitLab Project

- Link to project: https://baltig.infn.it/fornari/kube-lustre
- K8s YAML files for services configuration (to be edited by end users):
  - kube-lustre-configurator.yaml to select which services to run (servers, clients)
  - kube-lustre-config.yaml to adjust servers/clients configurations, enable/disable HA
- Docker images for Docker containers (orchestrated by Kubernetes):
  - kube-lustre-configurator reads kube-lustre-config.yaml, then generates templates and assign resources to specific Kubernetes nodes
  - lustre makes lustre target, then imports ztool and mounts lustre target
  - lustre-client mounts lustre filesystem
  - lustre-install installs lustre and zfs packages and dkms modules
  - drbd makes and runs distributed replicated block device resource (for HA)
  - drbd-install installs drbd packages and dkms modules

K8s Lustre Cluster Statistics

Filesystem Status & Info

```
POD_NAME                              POD_STATUS   POD_NODE
kube-lustre-configurator-cntw7        Succeeded    k8s-worker-3.novalocal
lustre1-client-j5jv4                  Running      k8s-worker-1.novalocal
lustre1-client-jcd4k                  Running      k8s-worker-3.novalocal
lustre1-client-pcbjc                  Running      k8s-master.novalocal
lustre1-client-xtgkl                  Running      k8s-worker-2.novalocal
lustre1-mdt0-mgs-0                    Running      k8s-worker-1.novalocal
lustre1-mdt0-mgs-drbd-6754h           Running      k8s-master.novalocal
lustre1-mdt0-mgs-drbd-rw7bq           Running      k8s-worker-1.novalocal
lustre1-ost0-0                        Running      k8s-worker-2.novalocal
lustre1-ost0-drbd-mwt5h               Running      k8s-worker-3.novalocal
lustre1-ost0-drbd-wbwjl               Running      k8s-worker-2.novalocal
```

```
[centos@k8s-master ~]$ kubectl exec -n lustre lustre1-client-pcbjc -- chroot /host-root lfs check servers
lustre1-OST0000-osc-ffff9956eeb07800 active.
lustre1-MDT0000-mdc-ffff9956eeb07800 active.
[centos@k8s-master ~]$ kubectl exec -n lustre lustre1-client-xtgkl -- chroot /host-root lfs osts
OBDS:
0: lustre1-OST0000_UUID ACTIVE
[centos@k8s-master ~]$ kubectl exec -n lustre lustre1-client-jcd4k -- chroot /host-root lfs df -h
UUID                       bytes        Used   Available Use% Mounted on
lustre1-MDT0000_UUID        95.6G        3.0M       95.6G   1% /stor/lustre1[MDT:0]
lustre1-OST0000_UUID        95.6G        3.0M       95.6G   1% /stor/lustre1[OST:0]
filesystem_summary:         95.6G        3.0M       95.6G   1% /stor/lustre1
[centos@k8s-master ~]$ kubectl exec -n lustre lustre1-client-j5jv4 -- chroot /host-root lfs df -i
UUID                      Inodes       IUsed       IFree IUse% Mounted on
lustre1-MDT0000_UUID    22805719         329    22805390   1% /stor/lustre1[MDT:0]
lustre1-OST0000_UUID     3134085         357     3133728   1% /stor/lustre1[OST:0]
filesystem_summary:      3134057         329     3133728   1% /stor/lustre1
```

PODs Status & Location

K8s Lustre Cluster Usage - Write Test

```
[centos@k8s-master ~]$ kubectl exec -n lustre lustre1-client-j5jv4 -- chroot /host-root \
dd if=/dev/zero of=/stor/lustre1/file2 count=1 bs=1073741824
1+0 records in
1+0 records out
1073741824 bytes (1.1 GB) copied, 4.92709 s, 218 MB/s
[centos@k8s-master ~]$ kubectl exec -n lustre lustre1-client-jcd4k -- chroot /host-root \
ls -lrth /stor/lustre1
total 1.1G
-rw-r--r--. 1 root root 1.0G Oct  7 13:46 file2
[centos@k8s-master ~]$ kubectl exec -n lustre lustre1-client-pcbjc -- chroot /host-root lfs \
df -h
UUID                       bytes        Used   Available Use% Mounted on
lustre1-MDT0000_UUID        95.6G        3.0M        95.6G   1% /stor/lustre1[MDT:0]
lustre1-OST0000_UUID        95.6G        1.0G        94.6G   2% /stor/lustre1[OST:0]
filesystem_summary:         95.6G        1.0G        94.6G   2% /stor/lustre1
[centos@k8s-master ~]$ kubectl exec -n lustre lustre1-client-xtgkl -- chroot /host-root lfs \
getstripe -v /stor/lustre1/file2
/stor/lustre1/file2
lmm_magic:          0x0BD10BD0
lmm_seq:            0x200000402
lmm_object_id:      0x1
lmm_fid:            [0x200000402:0x1:0x0]
lmm_stripe_count:   1
lmm_stripe_size:    1048576
lmm_pattern:        raid0
lmm_layout_gen:     0
lmm_stripe_offset: 0
       obdidx           objid           objid           group
            0               6             0x6               0
```