Scientific Innovation Through Integration

EMSL

Active Storage

Evan Felix

e@pnl.gov

EMSL is located at PNNL

Pacific Northwest
NATIONAL LABORATORY

PNNL-SA-63248

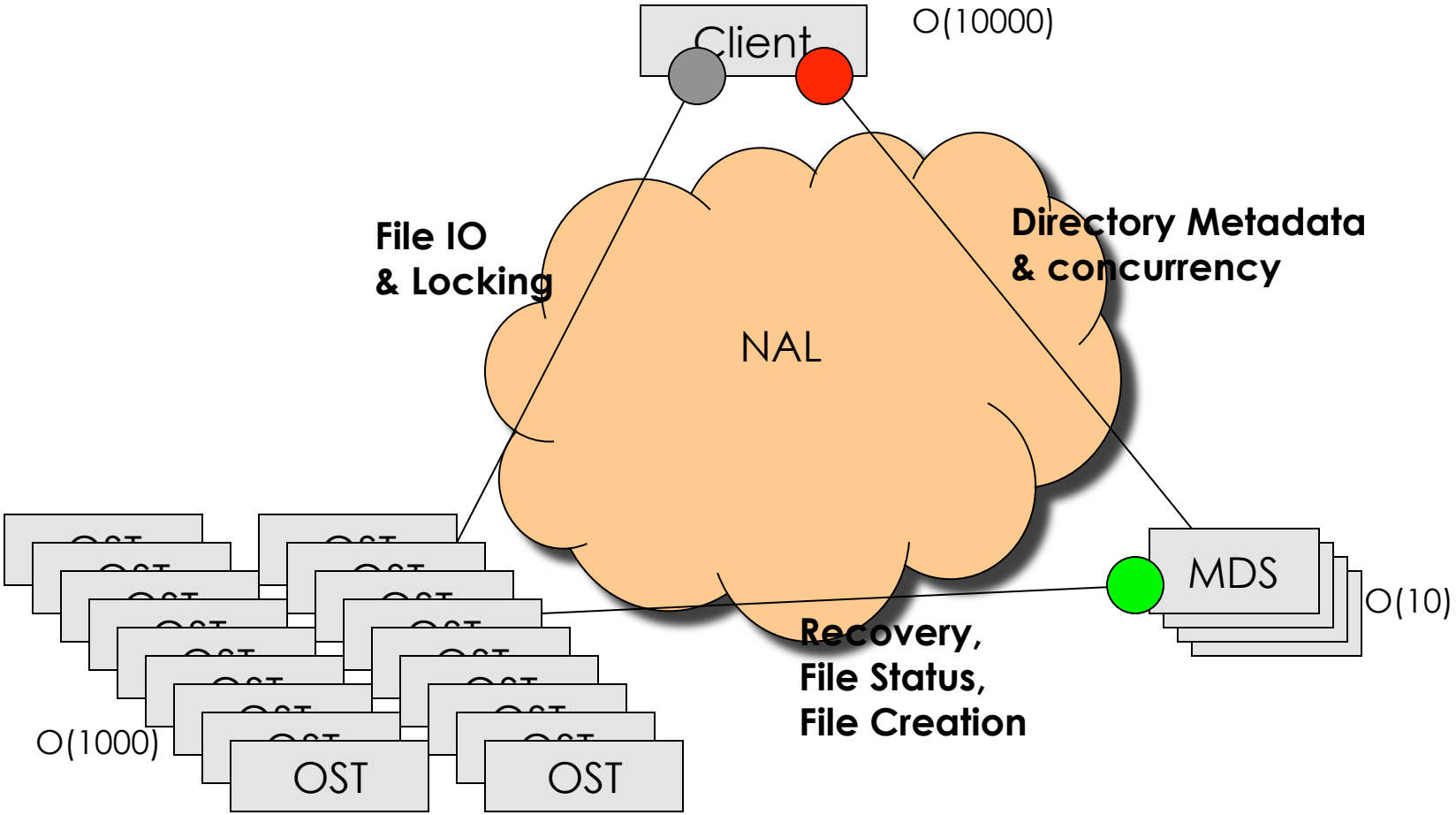
Previous work on Active Disks/Storage

EMSL

- Aims to use Processing resources 'Near' the disk.
  - On the Disk Controller.
  - On Processors connected directly to disks.
  - Reduce network bandwidth/latency limitations.
- Other Research
  - DiskOS Stream Based model (ASPLOS'98: Acharya, Uysal, Saltz)
  - Evolving RPC for Active Storage(ASPLOS'02: Sivathanu, A. Arpaci-Dusseau, R. Arpaci-Dusseau
- Research proved its possible, but:
  - Vendors are not providing supporting hardware
  - Specialized hardware still expensive
  - No comprehensive solution available

Pacific Northwest
NATIONAL LABORATORY

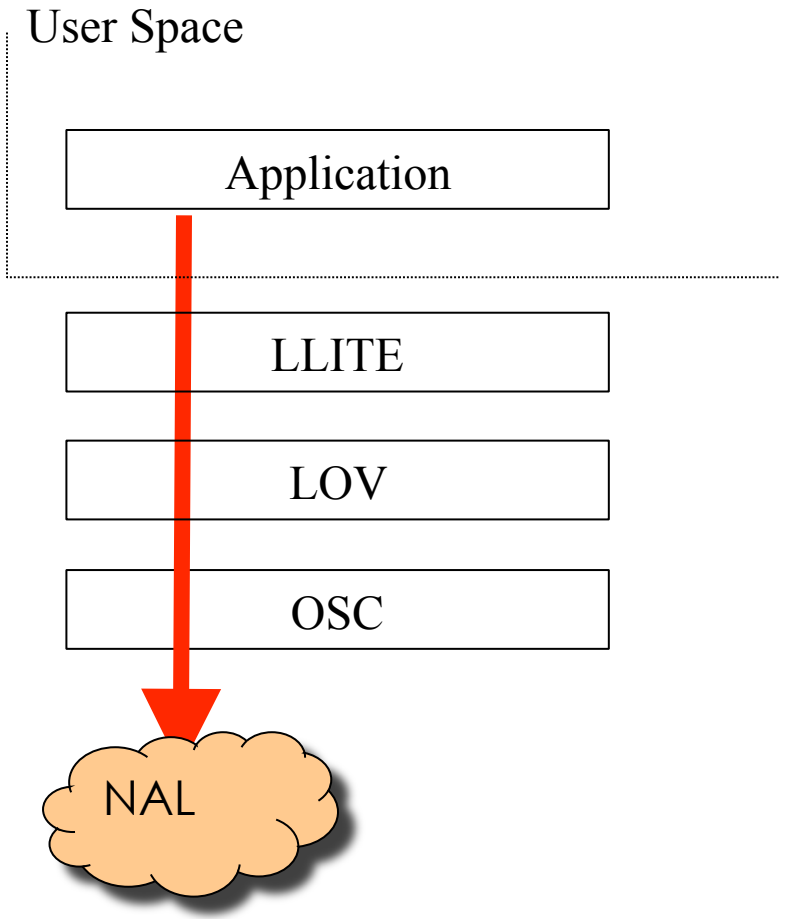
Lustre Overview



More info at www.lustre.org

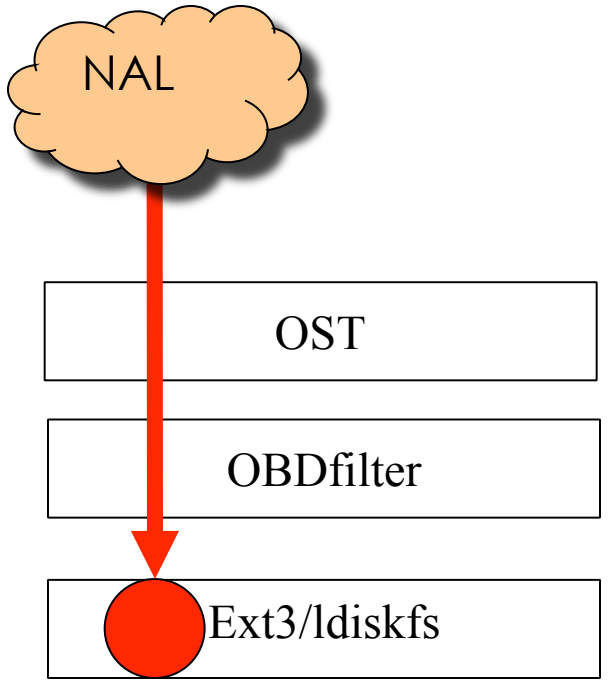
Lustre Client

- Application IO requests
- LLITE module implements Linux VFS layer
- LOV stripes object and targets IO to correct Object Client
- OSC packages up request for transmission over the NAL



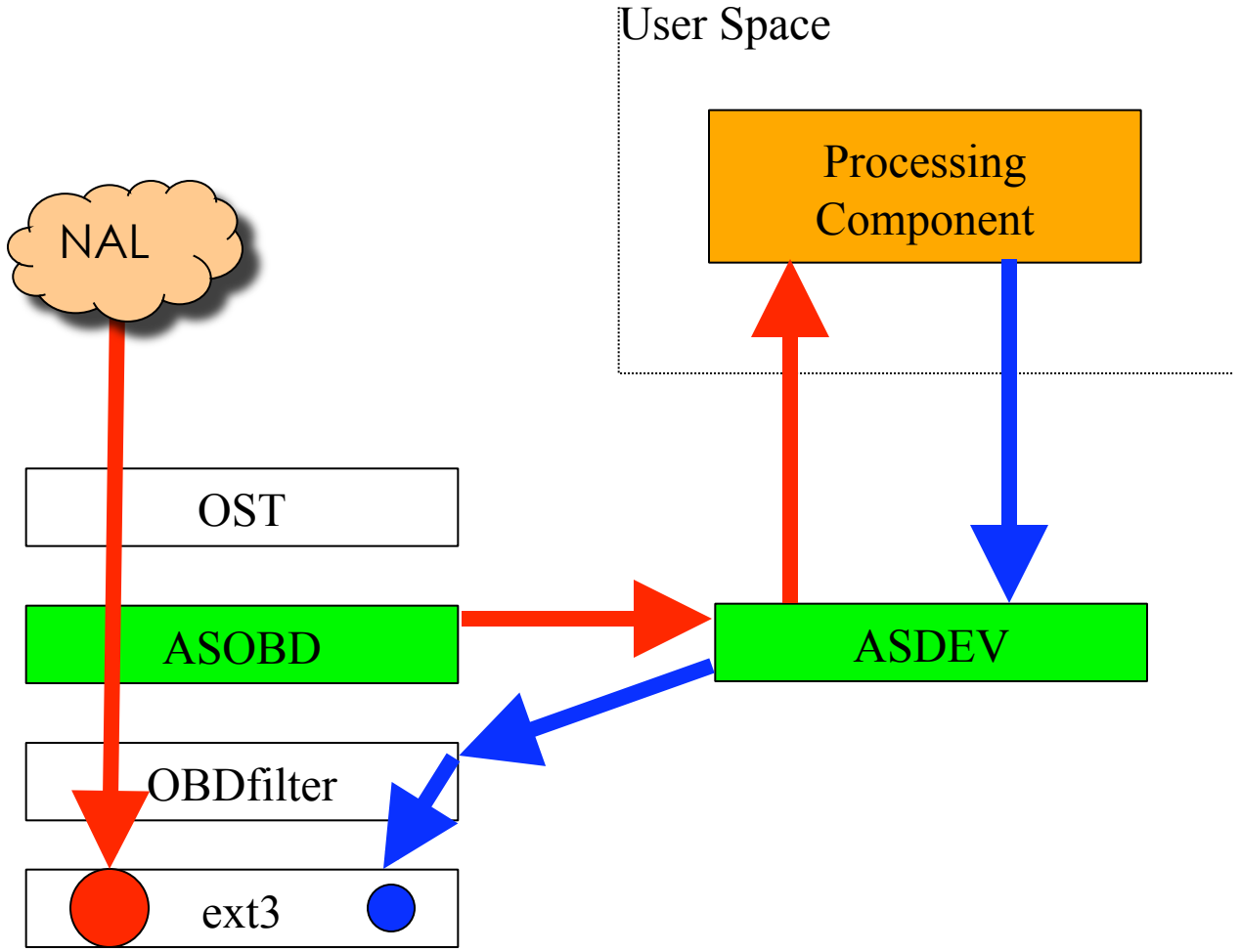
Lustre Object Storage Server

- Requests arrive from Portals NAL
- Object Storage Target directs Request to appropriate lower level OBD
- OBDfilter presents ext3/ldiskfs as Object Based Disk



Active Storage: As Implemented

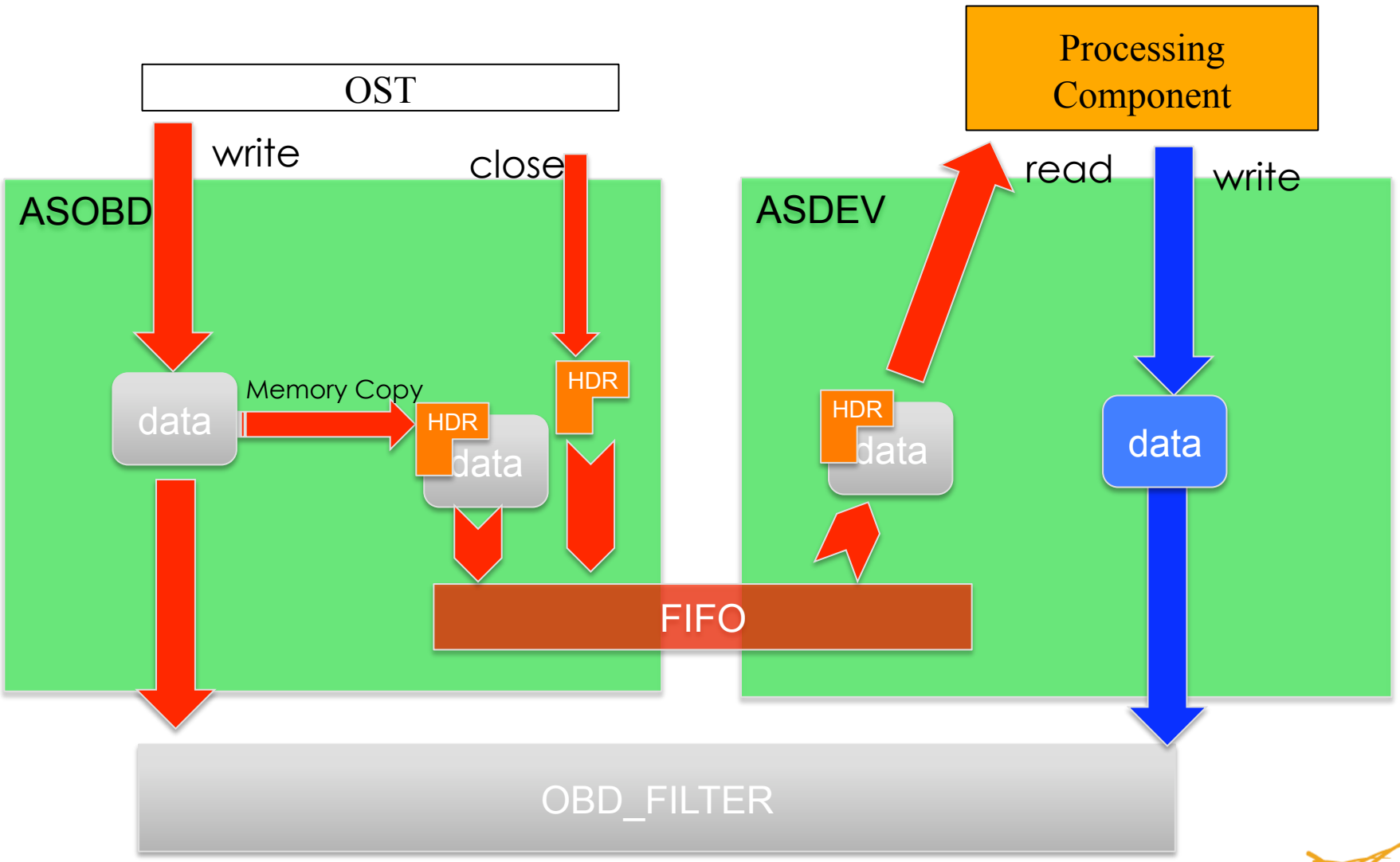
EMSL



- Extra Module passes data, until told to pipe data elsewhere.
- Data is sent to userspace process through Unix Character Device File.
- Processed Data is written back to disk.
- Pattern: 1W->2W

Pacific Northwest
NATIONAL LABORATORY

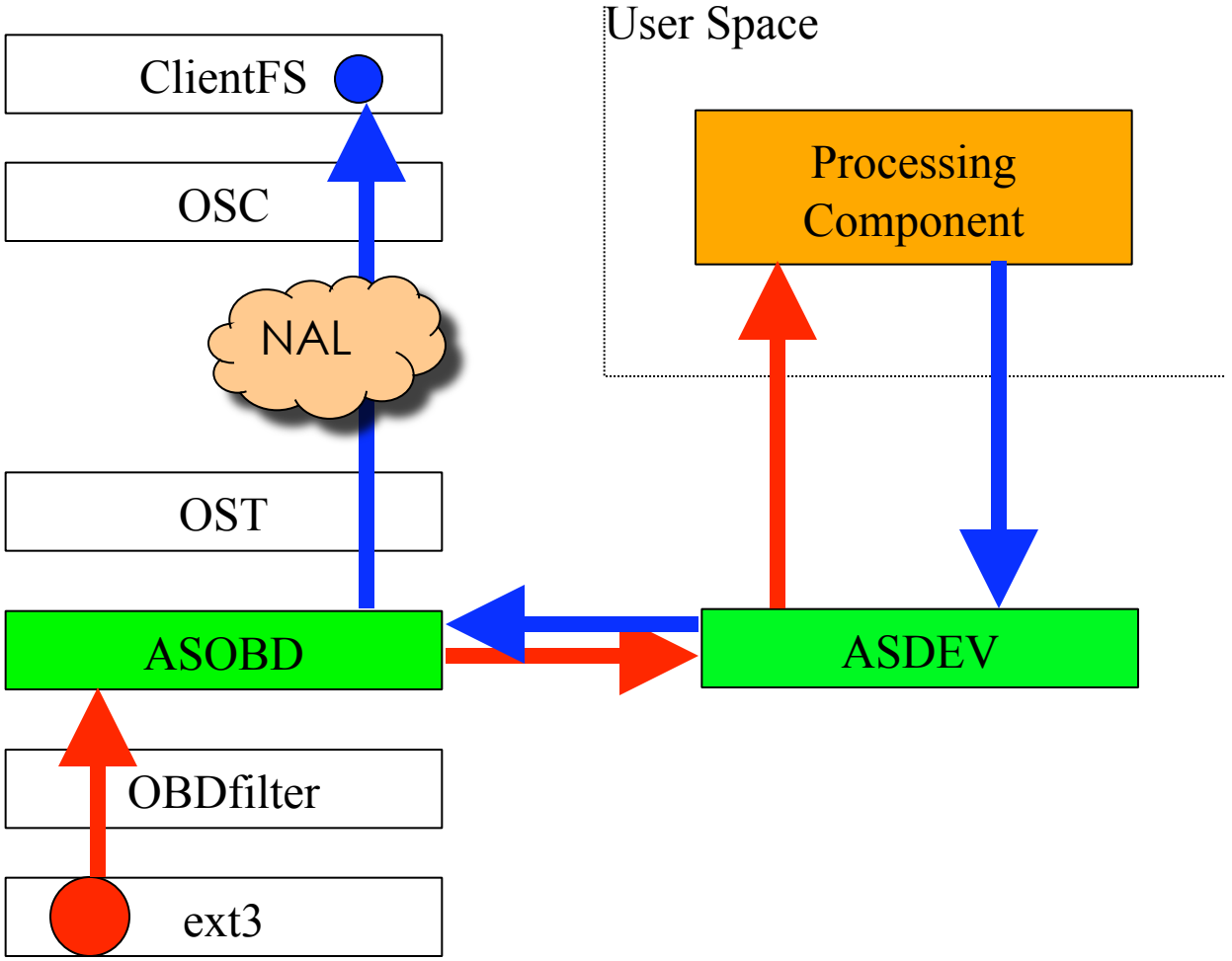
Active Storage: Inside



PNNL-SA-63248

Active Storage: Possibilities

- Extra Module Reads Data off the disk
- Data is sent to userspace process through Unix Character Device File.
- Processed Data is sent back to reader process
- Pattern: 1R->1R



Active Storage Processing Patterns

EMSL

Pattern	Description
1W->2W	Data will be written to the original raw file. A new file will be created that will receive the data after it has been sent out to a processing component.
1W->1W	Data will be processed then written to the original file
1R->1W	Data that was previously stored on the OBD can be re-processed into a new file.
1W->0	Data will be written to the original file, and also passed out to a processing component. There is no return path for data, the processing component will do 'something' with the data.
1R->0	Data that was previously stored on the OBD is read and sent to a processing component. There is no return path
1W->#W	Data is read from one file and processed, but there may be many files that are output from
#W->1W	There are many inputs from various files being written as outputs from the processing component.
1R->1R	Data is read from a file on disk, sent to a processing component, then the output is sent to the reading process.

Pacific Northwest
NATIONAL LABORATORY

StorCloud Initiative – SC2004

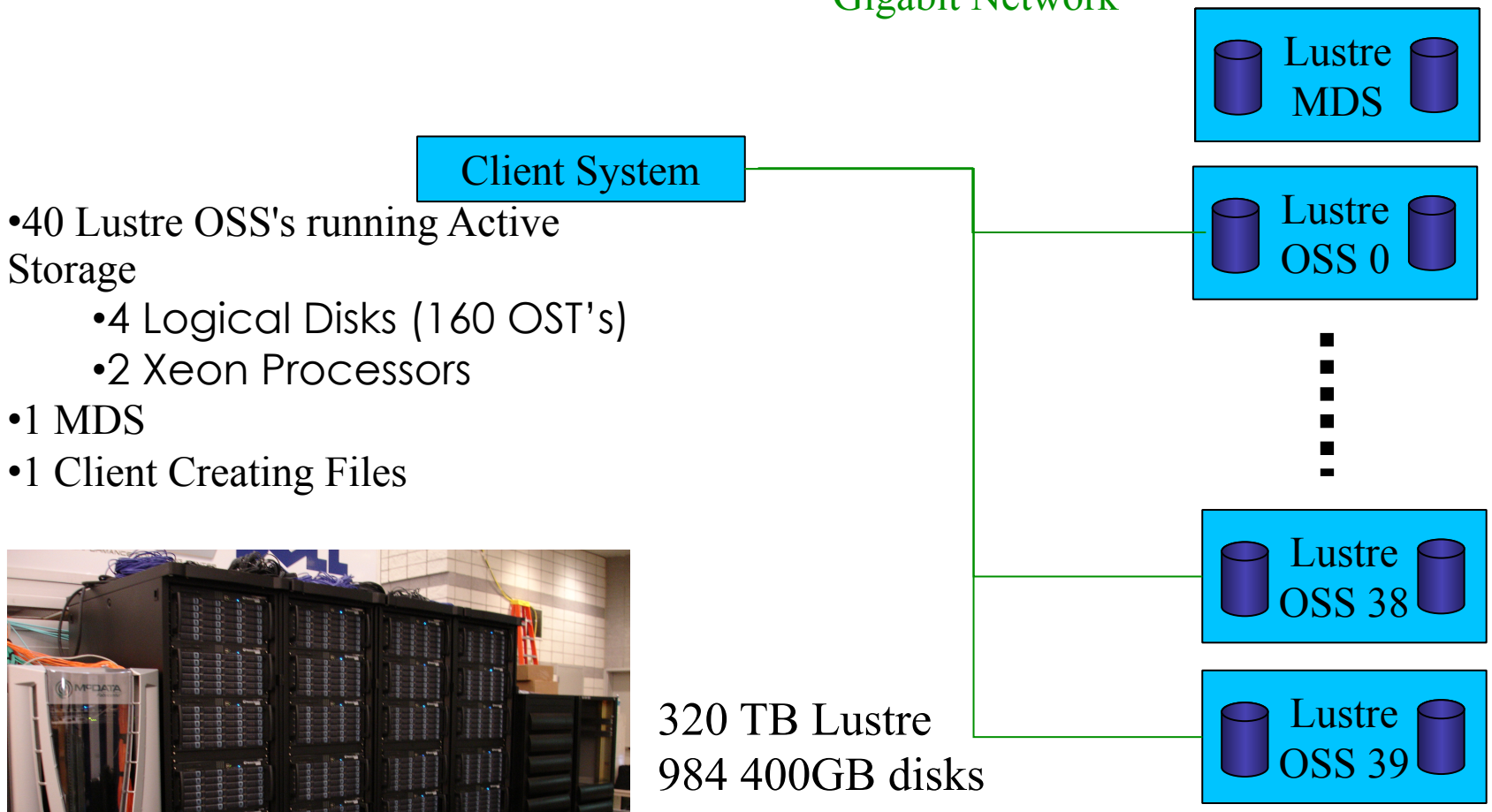
EMSL

- Goals of StorCloud
  - Petabyte-scale
  - Terabyte/s speeds
  - High performance storage capability on the conference exhibits floor
  - Highlight storage technologies
  - Create a virtual, on-site "storage on request" system to support researchers in demonstrating high bandwidth applications.
- Comprised of state of the art
  - SAN, NAS
  - File Systems
  - emerging technologies

Pacific Northwest
NATIONAL LABORATORY

SuperComputing 2004: StorCloud

EMSL



- 40 Lustre OSS's running Active Storage
  - 4 Logical Disks (160 OST's)
  - 2 Xeon Processors
- 1 MDS
- 1 Client Creating Files

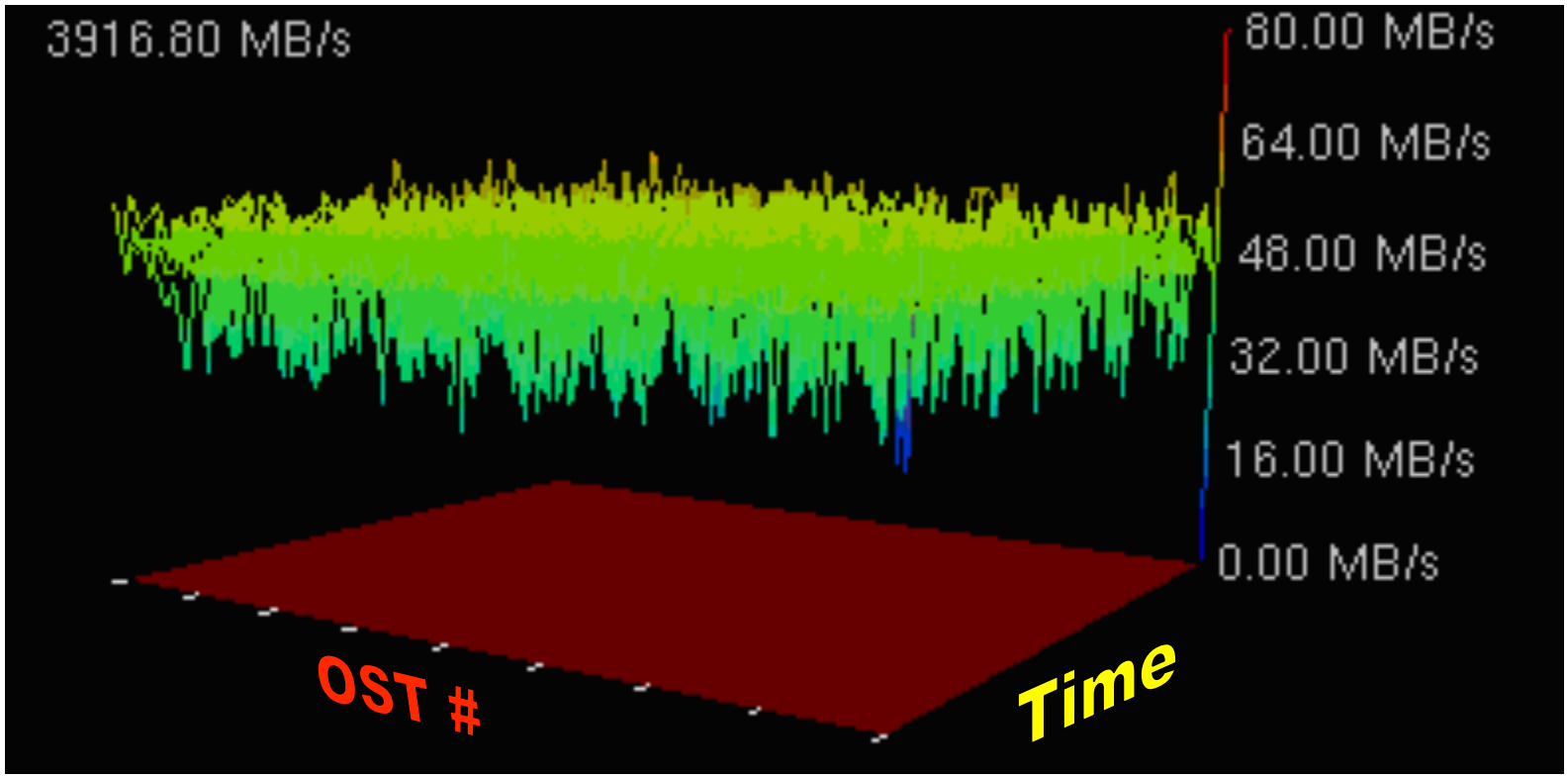
320 TB Lustre
984 400GB disks
Sustained 4.0GB/s Active Storage write

Awarded: Most Innovative use of Storage

Pacific Northwest NATIONAL LABORATORY

Real Time Visualization

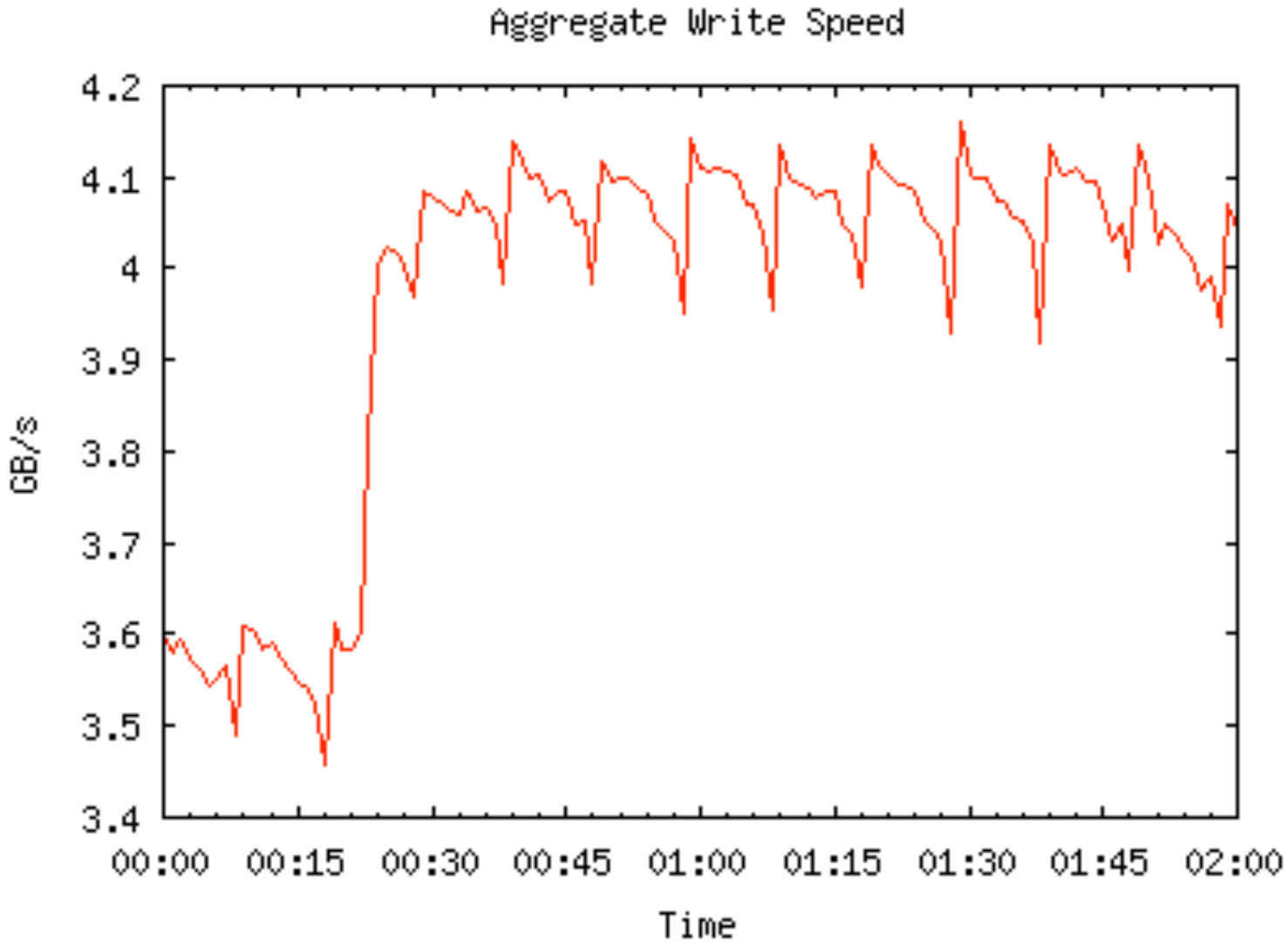
EMSL



Pacific Northwest
NATIONAL LABORATORY

StorCloud Challenge Run

EMSL



Pacific Northwest
NATIONAL LABORATORY

Status of the implementation.

EMSL

- Status
  - Proof of concept 1W->2W code works.
  - Difficult to Administer and Use.
  - Possible Memory pressure problems.
  - Possible Scheduling Issues.
  - Lustre IO interference.

Pacific Northwest
NATIONAL LABORATORY

Conclusions

EMSL

- Active Storage within the Lustre file system can work
- Early Bioinformatics applications have show viability of the approach
- Work to extend and provide other types of processing in progress.

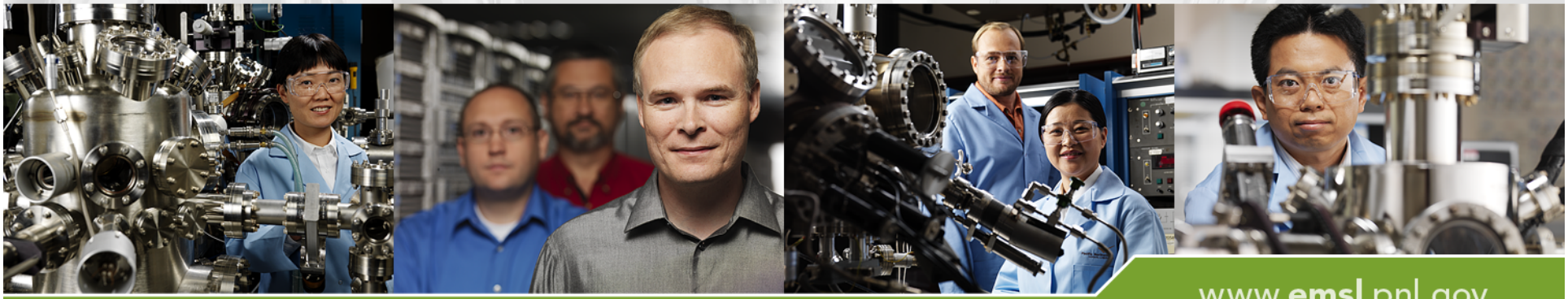
- We NEED a defined API
- Striping Alignment Issues.

Pacific Northwest
NATIONAL LABORATORY


EMSL
Scientific Innovation Through Integration
Questions?
Evan.felix@pnl.gov
EMSL is located at PNNL
Pacific Northwest
NATIONAL LABORATORY

Environmental Molecular Sciences Laboratory

A national scientific user facility integrating experimental and computational resources for discovery and technological innovation



www.emsl.pnl.gov

EMSL | U.S. DEPARTMENT OF ENERGY

Pacific Northwest NATIONAL LABORATORY