

Active Storage Processing in a Parallel File System

Evan Felix, Kevin Fox, Kevin Regimbal,
Jarek Nieplocha

evan.felix@pnl.gov

Outline

- ▶ Introduction to The Environmental Molecular Research Laboratory
- ▶ Introduction to Active Storage
- ▶ Lustre Overview
- ▶ Active Storage
 - Lustre Module
 - Possibilities
 - Experience @ Supercomputing 2004
 - Results
- ▶ Conclusions



William R. Wiley Environmental Molecular Sciences Laboratory



► Who are we?

- A 200,000 square-foot U.S. Department of Energy national scientific user facility
- Operated by Pacific Northwest National Laboratory in Richland, Washington

► What we provide for users

- Free access to over 100 state-of-the-art research instruments
- A peer-review proposal process
- Expert staff to assist or collaborate

► Why use EMSL?

- EMSL provides - under one roof - staff and instruments for fundamental research on physical, chemical, and biological processes.

Battelle

WWW.EMSL.PNL.GOV

Pacific Northwest National Laboratory
Operated by Battelle for the U.S. Department of Energy

3



Molecular Science Computing Facility

Provides a high-performance computer with Intel Itanium2 processors, Quadrics interconnect, and HP RX2600 nodes, which supports a wide range of environmental molecular science.

► Instrumentation

- Allocated to users accomplishing peer-reviewed Grand Challenge science in support of DOE missions
- 1,976-processor system in 978 nodes, with 11.8 Teraflop peak, 6.8 Terabytes of memory and 500 Terabytes of disk
- Silicon Graphics 3400 graphics and visualization server with integrated video and audio editing system
- 300 TB Lustre Archive System



Battelle

WWW.EMSL.PNL.GOV

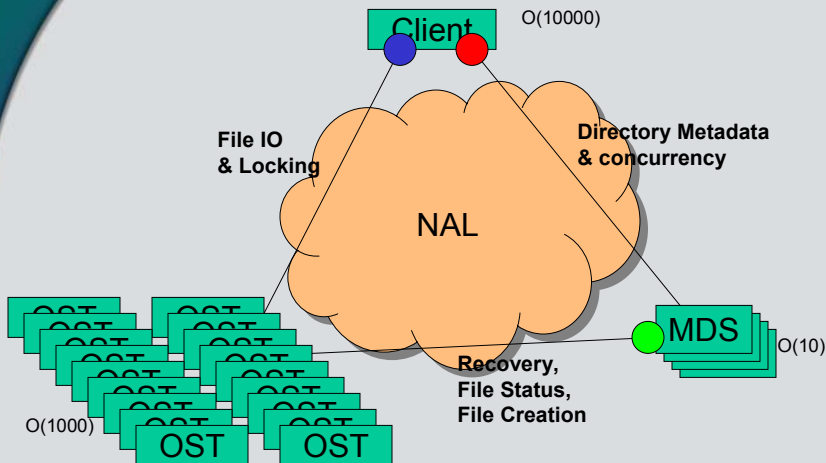
Pacific Northwest National Laboratory
Operated by Battelle for the U.S. Department of Energy

4

Previous work on Active Disks/Storage

- ▶ Aims to use Processing resources 'Near' the disk.
 - On the Disk Controller.
 - On Processors connected to disks.
 - Reduce network bandwidth/latency limitations.
- ▶ Other Research
 - DiskOS Stream Based model (ASPLOS'98: Acharya, Uysal, Saltz)
 - Simulation of Parallel Disks (VLDB '98: Riedel, Gibson, Faloutsos)
- ▶ Research proved its possible, but:
 - Vendors are not providing frameworks
 - Specialized hardware still expensive
 - Difficult to take advantage of

Lustre Overview

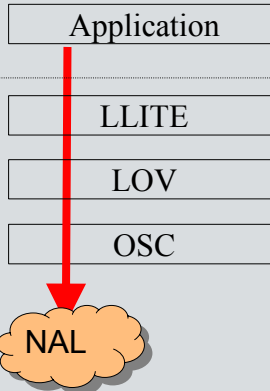


More info at www.lustre.org

Lustre Client

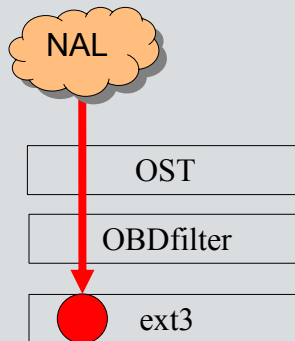
- Application IO requests
- LLITE module implements Linux VFS layer
- LOV stripes object and targets IO to correct Object Client
- OSC packages up request for transmission over the NAL

User Space



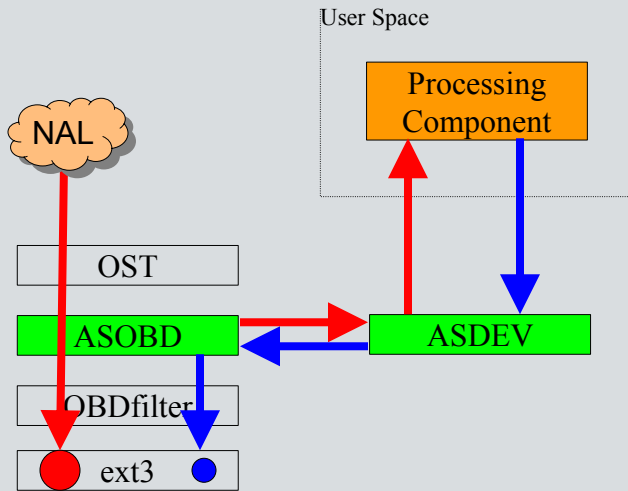
Lustre Object Storage Server

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



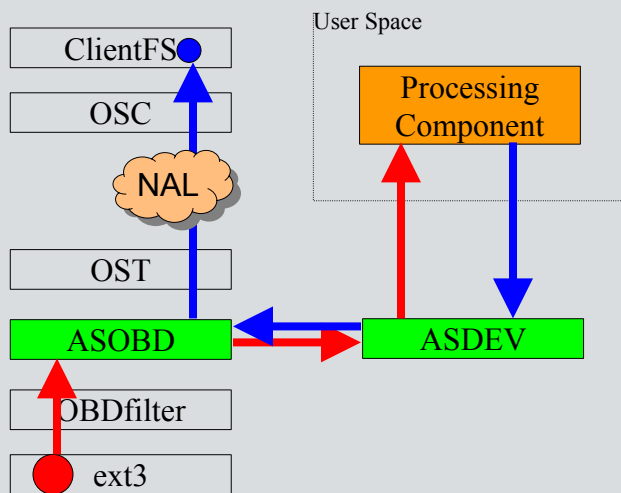
Active Storage: Today

- 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



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

| 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. |

StorCloud Initiative – SC2004

► 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

Gigabit Network

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



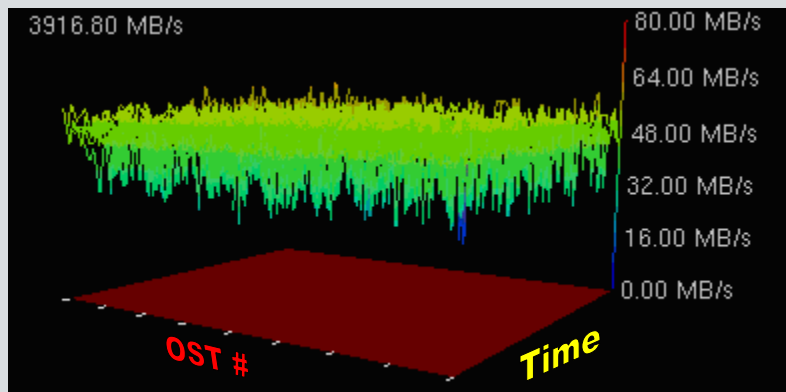
320 TB Lustre
984 400GB disks
Sustained 4.0GB/s Active
Storage write

Client System

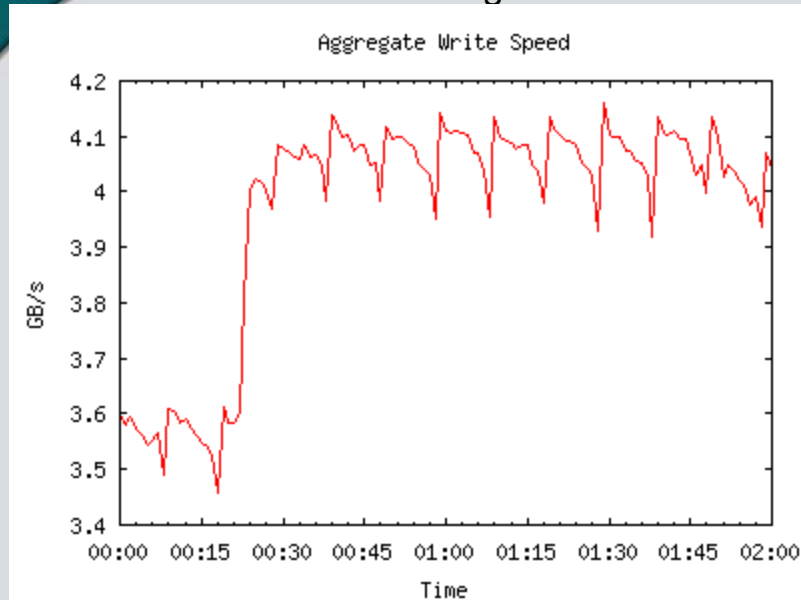


Awarded: Most Innovative use of Storage

Real Time Visualization



StorCloud Challenge Run



Status of the implementation and future direction

► Status

- Proof of concept 1W->2W code works now.
- Difficult to Administer and Use.
- Possible Memory pressure problems.

► Future

- Provide Active Storage Client tools.
- Define OST configuration better.
- Implement more flexible Streaming code to support other processing patterns.

Conclusions

- ▶ 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.

Questions?

Evan.felix@pnl.gov