The Lustre Centre of Excellence at ORNL

Makia Minich
Clustre Monkey, HPC Software Stack
Lustre Group

May 2008

Introduction

- Lustre Centre of Excellence (LCE)
  - Established Nov 2006 by Cluster File Systems, Inc. (acquired by Sun in Sept 2007) in the National Center for Computational Sciences (NCCS) at Oak Ridge National Lab (ORNL).
- ORNL deploying a peta-scale supercomputer at by the end of 2008, needs a matching filesystem.
- Scientific application teams could benefit from closer interaction with filesystems architects to increase I/O performance.
- Performance and scalability as systems keep growing larger.

Goals for the LCE

- Enhance the scalability of the Lustre File System to meet the performance requirements of petascale systems
- Build Lustre expertise through training and workshops
- Assist scientific application teams in getting the maximum I/O performance from their applications.

LCE Resources

- Several resources allocated to the LCE
  - Three senior engineers on site at ORNL
  - Other senior engineers and architects from Lustre Team provide guidance as and when required.
  - Quality Engineering resources.
  - Support Engineering resources.
  - Program Management resources.

LCE: 2008 Milestones

- January – June 2008
  - Mitigate risk of Cray supplied Lustre
  - Organize a Lustre Summit at ORNL
  - Establish a baseline peak and delivered performance numbers for a scalable unit.
  - Complete implementation and verify improvements for scalability studies from the previous contract period
  - Organize an application workshop (early 2008)
  - Ongoing Lustre support and I/O optimisations for applications
  - Provide early access to a ZFS based release
  - Assist in identifying and correcting deficiencies in Lustre and LNET encountered at ORNL

LCE: 2008 Milestones

- July – December 2008
  - Support the deployment of 1PF system
  - Ongoing Lustre support and I/O optimizations for applications
  - Provide Lustre Internals training
  - Demonstrate the delivery of at least 85% of the peak aggregate I/O bandwidth across the entire PF storage system to Lustre clients.
  - On-going operational support in deploying a center wide file system based on Lustre at ORNL
  - Address the goal of taking scalability, performance, and robustness of Lustre to the level required by multi-petaflop systems.
  - November 2008 – develop milestones for the third year

LCE Summit

- Held February, 2008 in Burlington, MA
- Attendees from most of our customers.
- “Achievements and Vision Going Forward” was the theme of the summit

Lustre – Achievements so far

Issue	Result
The most scalable HPC FS	Good – 5 years in a row now, 7 of the top 10
Offering high product quality	Improving, but far from a Skype or OS X like experience
Broad adoption	Not yet, not on track for it

Lustre Vision going forward

Facet	Activity	Difficulty	Priority	Timeframe
Product Quality	Major work is needed, except on networking	High	High	2008
Performance fixes	Systematic benchmarking & tuning	Low	Medium	2009
More HPC Scalability	Clustered MDS, Flash cache, WB cache, *Request Scheduling*, Resource management, *ZFS*	Medium	Medium	2009 - 2012
Wide area features	*Security*, WAN performance, proxies, replicas	Medium	Medium	2009 - 2012
Broad adoption	Combined pNFS / Lustre exports	High	Low	2009 - 2012

Note: These are visions, not commitments

LCE Summit: Users Top 5 Priorities

- System and File System Administration
- Improved support for multi-clustered environments
- Data Integrity
- Evolve Lustre towards a more community driven development model
- Support for ultra-large clusters and WAN

Enhancing I/O Efficiency

- As system size and filesystem size grow, applications need to modify their I/O handling.
- Case Study on improving the performance for the Parallel Ocean Program (POP) on the Jaguar system at NCCS in Oak Ridge National Laboratory.
- Results of paper submitted by:
  - Wang Di (Sun Microsystems)
  - Galen Shipman (ORNL)
  - Sarp Oral (ORNL)
  - Shane Canon (ORNL)

POP Background

- “POP is an ocean circulation model which solves the three-dimension primitive equations for fluid motions on the sphere.”
- Grid dimension for this testing: 3600x2800x42
  - > 42 is the depth of the ocean chosen for this testing.

http://climate.lanl.gov/Models/POP/

POP I/O Pattern

- POP is an ocean circulation model for resolving the three-dimensional primitives equation.
  - Creates 4 files: history, movie, restart and tavg.
  - Only restart and tavg file are relatively big. (tavg 13G, restart 28G).
  - In most cases, the I/O size is 65M from each client
    - 3600 * 2400 * *byte-length of the element*
- It was seen that the history file dominated most of the I/O, so work focused on the I/O for this file.
  - File is segmented by horizontal layering of the ocean.
  - 42 Segments for our configuration.

POP

- POP IO model
  - General Scientific application IO layer

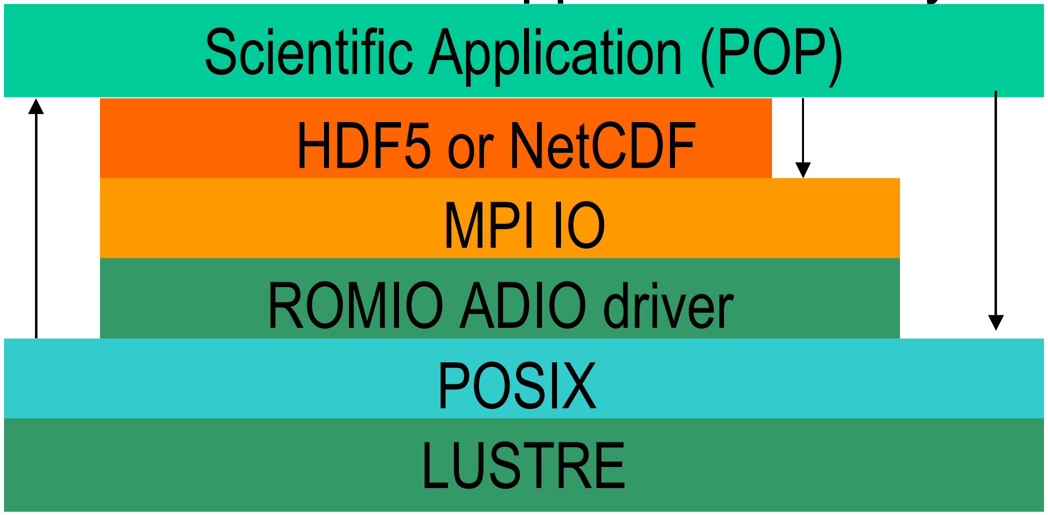

Figure1. HPC application software stack

POP

- POP originally implements I/O in one of two ways
  - POSIX(Fortran Record)
    - 42 clients, the performance is ok, but not very convenient.
  - NetCDF
    - Does not support parallel I/O. And the performance is very bad.

POP

- HDF5 porting
  - HDF5 is one of the most popular scientific I/O LIB right now.
  - It supports parallel I/O by MPI-IO.
  - Re-implement POP with HDF5 for investigating performance of POP + HDF5 + Lustre.

POP

- HDF5 performance investigation
  - HDF5 manages data and metadata in the single file by setting different data_set.
  - Writing extra metadata block for each HDF5 file. (overhead)
  - HDF5 support different I/O API. (POSIX, Independent, collective)

POP

- Several HDF5 parallel I/O features.
  - Open existing file (TRUNC flags) will cause all the clients to call MPI_Set_file_size(truncate) at the same time.
  - If open HDF5 file with write flag, then it will call flush when close the file.
  - Improper using data-sieving(read-modify-write) in HDF5 collective write mode.
    - Read-modify-write is very expensive for liblustre, since no client cache there.

POP

- Performance Results

I/O Method	I/O Processes	Time Step Length (mins)	Duration of I/O (mins)	Overhead %
NetCDF	1	60	26	43
Fortran record	1	60	9	15
HDF5 Collective	42	60	12	20
HDF5 Independent	42	60	2	3

POP

- Lustre ADIO driver
  - The final target is to resolve all the improper I/O problems in Lustre ADIO driver
  - For POP
    - Fix that improper read-modify-write in ADIO driver.
    - Split big I/O size to stripe size I/O, because application could achieve best I/O performance with stripe size I/O.

Links

- ORNL's LCE Site
  - > http://ornl-lce.clusterfs.com
- LCE Summit Slides
  - > http://ornl-lce.clusterfs.com/images/c/c6/LCESummitSlides.pdf

Thank You

The Lustre Centre of Excellence at ORNL

Makia Minich (makia@sun.com)
Clustre Monkey, HPC Software Stack
Lustre Group

May 2008