



NCCS Lustre Workshop

April 16, 2008

Presented by

**Oleg Drokin
Wang Di**

Sun Microsystems Inc.



Author	Date	Description of Document Change	Client Approval By	Client Approval Date
Nirant Puntambekar	4/18/08	Created document with input from Oleg and WangDi		

Summary

NCCS and Sun Microsystems hosted an all day workshop focussed on helping application scientists get the most from the Lustre File System, which forms the data storage layer for the Oak Ridge National Laboratory's Jaguar Supercomputing Cluster. The workshop was well attended with about 15-20 participants. This report summarizes the topics discussed and the questions posed to the Sun Engineers who conducted the workshop.

Agenda

Morning Agenda

- 1) Topic: Application I/O lib with Lustre
 - a) HDF5(a popular scientific I/O lib) with lustre.
 - b) Lustre ADIO driver.
 - c) Several ORNL applications(POP, flashIO, wrf-mode) analysis and improvements to I/O
- 2) Topic: How to optimize lustre I/O performance in application layer.
 - a) Description about lustre I/O process. How does I/O work in a Lustre File System. Similarities/Differences with other common file systems
 - b) What you should avoid in your application I/O parts with lustre, and why, including practical examples

Afternoon Agenda

- 3) Open Discussion about Lustre I/O with Scientific applications – please be prepared to describe your applications so that Lustre engineers can offer possible methods for improving performance.
- 4) Open Discussion to discuss possible future application IO needs – audience to describe what they see for future application needs in regards to Lustre
- 5) Demonstration of Lustre as run on the Jaguar system.

Note: Presentation Slides used will accompany this report.



Questions Received from the Participants

- How to debug data corruption a user is seeing that looks like part of file data was not written out.
 - Oleg explained that the likely cause is eviction of clients from overloaded OST's. Users load situation was discussed and the conclusion was that ost's were overloaded by some others users job. User asked to file a bug for this.
- How to profile Lustre and find out about i/o patterns application produces and where time is spent.
 - Use the Cray supplied pats tool.
- A user from a German University had several questions about the future Lustre roadmap, some sysadmin questions, ways to support failover without actually having real shared storage, advice on why after migrating some of their clusters (in Germany) from gige to o2ib, they did not see much performance benefit.
 - Explained Lustre roadmap, directed user to the LLNL Lustre monitoring tool, explained why there may not be speedup when upgrading to faster networks due to disk contention and suggested use of dual port firewire hdd enclosures.
- For a new user with a new application, what suggestions do you have for using Lustre.
 - Use parallel io instead of single io for improved performance
- What is the threshold for switching from single io to parallel io ? What are the usual tuning parameters?
 - Always use parallel io if possible. Tuning parameters are stripe_size and stripe_count.
- Why is the new Lustre ADIO driver not mainstream yet ?
 - WangDi explained that the communication overheads are still a bit high which we are currently investigating. The driver should be introduced into the Jaguar system when complete and this transition should be seamless to the end users.
- What should one care about when choosing stripe_size and stripe_count ?
 - For stripe_count, it depends on the Application IO requirements and IO bandwidth the OST can provide. Stripe_size selection depends on the IO pattern and disk IO_size requirement. Usually on jaguar/jaguarCNL, the default values are stripe_size 1M, stripe_count 4.
- Question on whether lustre is posix compatible. For IO performance, except IO size what else do we need to care about.
 - a) stripe_size/stride_count. b) right IO way (posix, independent, or collective). c) parallel IO.
- Chris (Chris.kerr@noaa.gov) has a performance problem for the FSM application. (2k cores process and write its own NetCDF file, and each one is about 100M), which cost 3hrs. But he only tried that in catamount (liblustre), not tried on CNL yet.
 - Need to create ticket and perform further investigation.
- Question was asked about the system configuration for examples in the presentation.
 - Wang Di mentioned it is in JaguarCNL, 256 nodes. (Seastar 7Gbit/Sec, 300M for DDN8850 disk system).



- Question was posed to the audience: Which is the preferred io library out there, netcdf or hdf5 ?
- Most people said netcdf, also nobody has tried pnetcdf yet.
- What is the Support Model for nccs systems in case of filesystem or other questions
- Sarp explained that there is a support organization within nccs to help out end users, difficult issues will be forwarded by nccs support team to Lustre Engineers on site.

