BEST PRACTICES FOR PARALLEL FILE SYSTEM SETUP

The Cray oil and gas team and Taming Traces Consulting profiled the application I/O requirements for Landmark's SeisSpace® seismic processing software in a variety of workflow scenarios and produced a series of recommendations for parallel file system setup.

Problem

Seismic processing organizations don't have time to ask exploratory questions of their parallel file systems. Could a different type of parallel file system be used for better performance? What about combining primary and secondary storage requirements on the same system? Is Lustre or GPFS better? What's important to keep in mind when designing a new parallel file system?

Cray's oil and gas team and Taming Traces Consulting worked together to answer these questions. They profiled the application I/O requirements for Landmark's SeisSpace® seismic processing software and established recommendations for setting up your next seismic processing parallel file system.

Specifically, they looked at workflow performance under the following circumstances:

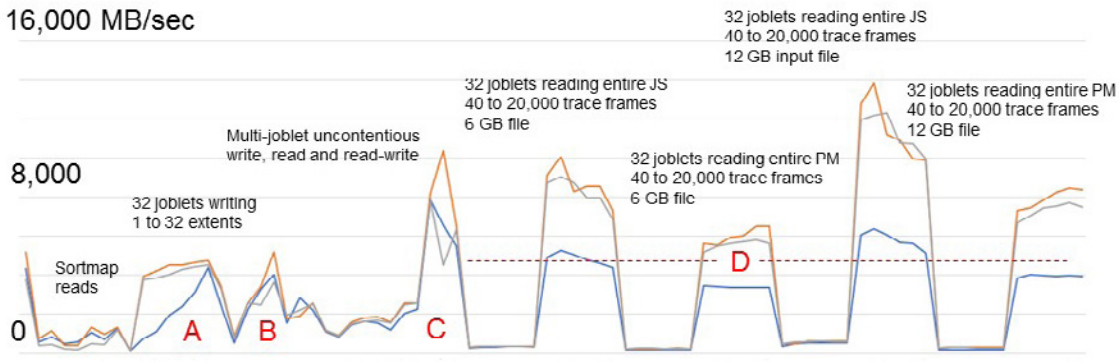
- Combining primary and secondary data on the same parallel file system
- Using GPFS instead of Lustre®
- Modifying file system block size
- Using Lustre pools for primary and secondary separation

Solution

Using the Cray® CS400™ cluster supercomputer and three separate storage platforms — the Cray® Sonexion® with a Lustre file system, IBM GL6 with a GPFS file system and DDN GS14KX™ with a GPFS file system — the team addressed I/O problems by looking closely at SeisSpace I/O requirements.

> Keep things simple when designing your parallel file system.

They ran ran hundreds of production-level read, write and read-then-write SeisSpace jobs with both GPFS and Lustre. They varied the input for block size, tested the same exact workflows on each storage platform, and tested primary and secondary on the same parallel file system for all runs.



Lustre, IBM and DDN Averages for All Jobs

Results

The team ran SeisSpace jobs in groups while varying joblets, extents, contention and read/write file size. Each workflow was extremely consistent in results, with some critical differences.

The graph above illustrates the results using the average of the runs. The blue line is Lustre on the Sonexion system, the gray line is GPFS on the DDN system and the gold line is GPFS on the IBM system. Each point on the graph represents an accumulated average of all the jobs run for each workflow. Areas A through C is where the majority of SeisSpace workflows would be run in a normal sequence of processing steps. Here, I/O is categorized as random. Area D workflows read and write all of the data within an extent at one time, called streaming I/O, similar to what is done in most migrations.

Looking at the graph as a whole, it appears GPFS is the performance winner. But when closely analyzed the situation could be seen differently.

In Area A, 32 joblets write data using 1 to 32 extents. Here, GPFS functions better with lower extents and evens out with Lustre when extents and joblets become equal. In Area B the team observed some minor read-only and write-only issues that will require further investigation. Then area C revealed a difference between the systems as a result of a 784 GB write, a read, and then a read-write with load; this area shows GPFS performing better.

For area D, the team tested both JavaSeis and ProMAX data formats. The series of troughs and peaks show the writing and reading of a 6 GB JavaSeis file and 6 GB ProMAX file. Next, the team doubled the file sizes to 12 GB. The workflows all use 32 joblets with varying trace frames. GPFS handles reads with higher loads better in these cases than Lustre. Writes (troughs) are all essentially the same.

KEY FINDINGS

- Little performance reduction from placing primary and secondary on the same file system
- Lustre and GPFS perform equally well under production workflows
- Newest versions of Lustre and GPFS work extremely well out of the box
- Block size is most critical parameter = 4M is optimum for combined file system
- Pooling in Lustre provides no performance improvement
- Bottom line? Keep things simple

Cray Inc.
901 Fifth Avenue, Suite 1000
Seattle, WA 98164
Tel: 206.701.2000
Fax: 206.701.2500
WWW.CRAY.COM