HPCS I/O Scenarios Tests

John Carrier*, Jeff Garlough*, John Dawson+, Mike Booth+, Ed Giesen+

*Cray Inc. +Routing Dynamics

Introduction

To facilitate evaluation of storage system scalability, the Defense Advanced Research Projects Agency (DARPA) provided its High Productivity Computing Systems (HPCS) vendors with a set of 14 representative workloads.

Capture Environment

1. Single stream with large data blocks operating in half duplex mode
2. Single stream with large data blocks operating in full duplex mode
3. Multiple streams with large data blocks operating in full duplex mode
4. Extreme file creation rates

Parallel Environment

5. Checkpoint/restart with large I/O requests
6. Checkpoint/restart with small I/O requests
7. Checkpoint/restart large file count per directory large I/Os
8. Checkpoint/restart large file count per directory small I/Os
9. Walking through directory trees
10. Parallel walking through directory trees
11. Random `stat()` system call to files in the file system (one process)
12. Random `stat()` system call to files in the file system (multiple proc's)
13. Small block random I/O to multiple files
14. Small block random I/O to a single file

DARPA organized these Scenarios into two groups based on the target usage. Scenarios 1-4 represent a *Capture* environment where I/O depends on the ability of a few nodes to create files from streaming data. Scenarios 5-14 represent usage in a more typical *Parallel* environment, where thousands of nodes access the file system using file per process (N-to-N) or shared file (N-to-1) access patterns. Scaling performance, rather than absolute throughput, is important to all Scenarios

Although there have been several proprietary implementations of the Scenarios, none have been made available as open source. Cray initiated its program with the explicit goal of making the tests available to the broader HPC community. Cray is releasing the source and scripts for its tests on SourceForge under Cray's BSD-compliant Open Source License. This poster announces the general availability of these tests at **http://hpcs-io.cray.com/**.

Test Configuration

In March 2011, after decommissioning their Cray XT4, Oak Ridge National Laboratory (ORNL) graciously provided dedicated access to the supercomputer and its file system for Cray to validate our initial implementation of the HPCS Scenarios tests.

ORNL's XT4 had 18 DDN 9550 storage controllers connected to 72 Lustre object storage servers (OSS) operating on XT4 I/O nodes. Each XT4 OSS had four Lustre object storage targets (OST). Eight OSSs were connected to two DDN controller couplets via 16 FC4 links to create a scalable storage unit (SSU) capable of ~5 GB/s (figure 1).

Cray reconfigured the storage to create three file systems, each with its own metadata server, using one, two and four SSUs (called FS1, FS2, and FS4, respectively). If the SSU was a bottleneck in a Scenario test, then repeating the test on a file system with twice the number of SSUs should double the performance of the Scenario. The XT4 had enough Lustre client nodes and sufficient network bandwidth that we could run tests against all three file systems simultaneously.

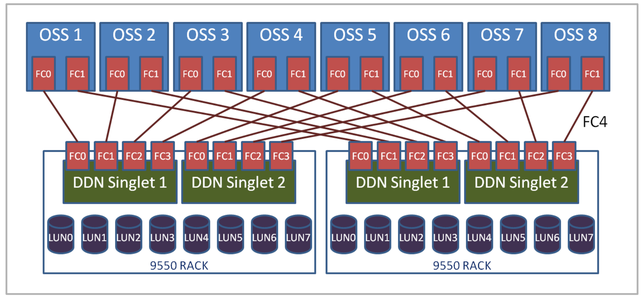

Figure 1. Scalable Storage Unit (SSU) attached to ORNL's Cray XT4.

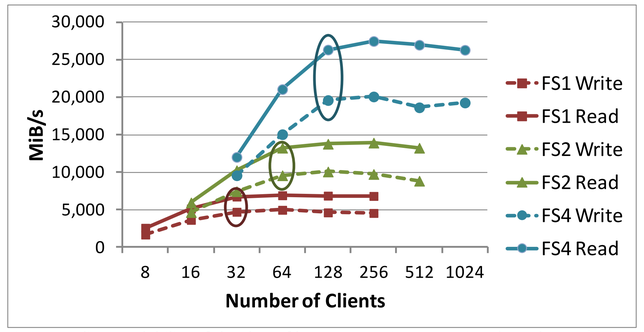

Figure 2. Client scalability of the three file systems using Scenario 7 (N-N, large I/Os).

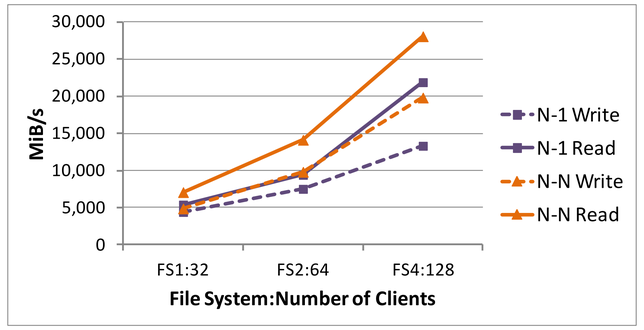

Figure 3. File system scalability on Large I/O Checkpoint Scenarios 5 (N-1) and 7 (N-N).

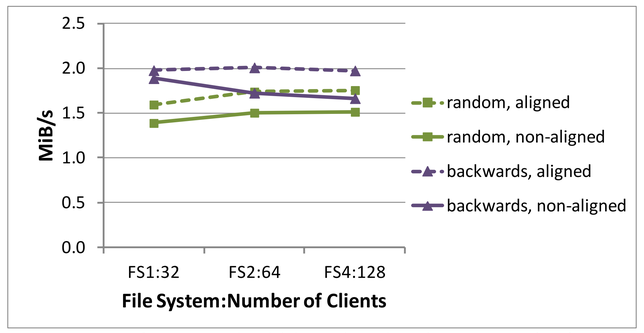

Figure 4. Poor performance and scalability with random I/O Scenario 13 (N-N).

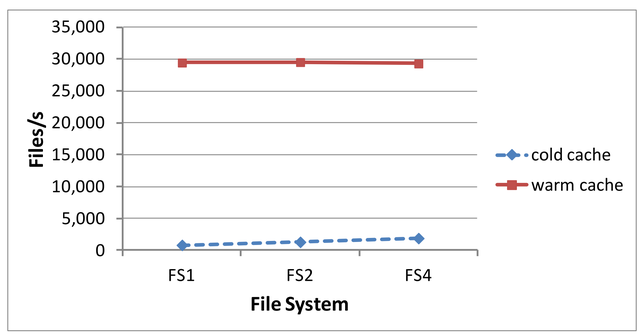

Figure 5. Expected flat metadata performance with Scenario 9.

Preliminary Results

We used Scenario 7 to determine the number of clients to use on subsequent tests with each file system (figure 2). Reads were consistently faster than writes. Doubling the number of SSUs in the file system doubled the peak bandwidth observed.

Each SSU had eight Lustre OSTs. The peak performance for reads or writes occurred with four clients per OST. FS4, for example, with four SSUs, had 32 OSTs and needed 128 clients to reach the performance plateau during the test. In contrast, FS2 needed just 64 clients and FS1 only 32. We used these client counts in subsequent tests of each file system.

Scenarios 5-8 represent sequential checkpoint workloads that vary across two dimensions: file access pattern (N-N vs. N-1) and I/O size (small vs large). For N-1, the two I/O sizes correspond to strided and segmented shared file access patterns. Sample results are shown for large I/O (figure 3). File per process (N-N) performance (Scenario 7) was greater than shared file (N-1) performance (Scenario 5), although both workloads scaled well.

Such scaling was not seen with the random workloads defined in Scenarios 13 and 14. Performance was significantly worse for random N-N workloads (figure 4) than the sequential N-N workloads (figure 3) and showed no scaling across the three file systems.

Lustre currently has a single metadata server, regardless of the size of the file system. As a result, metadata performance shows no scaling as the capacity of the file systems increased (figure 5).

Conclusions

The HPCS Scenarios expose the strengths and weaknesses of a storage subsystem. Our hope is that storage vendors will use the tests to demonstrate the effectiveness of their storage solutions in addressing HPC workloads. As such, the Scenarios create a level playing field for comparing the capabilities of different storage systems.

Acknowledgements

This material is based upon work supported by the Defense Advanced Research Projects Agency under its Agreement No. HR0011-07-9-0001. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Defense Advanced Research Projects Agency.

Our tests were performed at the Oak Ridge National Laboratory, which is managed by UT-Battelle, LLC. The authors gratefully acknowledge the assistance of Buddy Bland, Don Maxwell, Galen Shipman and Sarp Oral of ORNL's National Center for Computational Science (NCCS); Tom Griffith and Dick Sandness of Cray's benchmarking group; and Jeff Becklehimer, Kim Kafka, and, especially, John Lewis from Cray's ORNL Support team.

For more information, contact John Carrier (carrier@cray.com).

CRAY® THE SUPERCOMPUTER COMPANY