Introduction to the HPCS I/O Scenarios

John Carrier

2011-12-08

DARPA HPCS Program Award

In November 2006, DARPA awarded Cray a $250 million development contract under its High Productivity Computing Systems (HPCS) program.

HPCS Goals:
Provide a new generation of economically viable high productivity computing systems for the national security and industrial user community in the 2010 timeframe

- **Performance** (time-to-solution):
  speed up critical applications by factors of 10 to 40
- **Programmability** (idea-to-first solution):
  reduce cost and time for developing application solutions
- **Portability**:
  insulate application software from system specifics
- **Robustness**:
  protect applications from hardware faults and system software errors

The result will be greater productivity for users and administrators (not just faster machines)

DARPA HPCS I/O

- To demonstrate improvements in I/O productivity, DARPA provided its HPCS vendors with a set of 14 representative application workloads, the "HPCS Mission Partner File I/O Scenarios"
  - Most of the Scenarios must run against three different storage configurations, where the limiting I/O resource is doubled between each configuration
  - Other Scenarios (see backup slides for details) must demonstrate repeatable results
- For all Scenarios, scaling performance is more important than absolute throughput for a single configuration
  - The Scenarios are designed to be run on a large percentage of the machine to demonstrate large system performance
  - The Scenarios focus on storage scalability rather than peak performance generated with a subset of the machine during system acceptances
- The primary focus of the Scenarios is to provide an objective tool to demonstrate how a storage solution will scale when deployed with real workloads

HPCS I/O Scenarios

- DARPA organized the 14 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

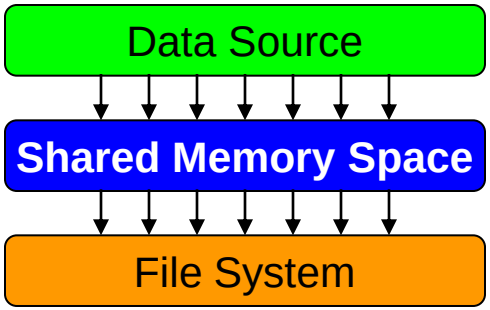
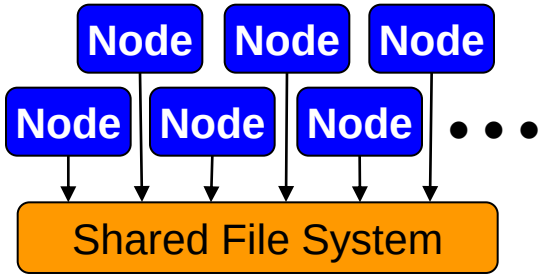
1. Single stream with large data blocks operating in half duplex mode
2. Single stream with large data blocks operating in full duplex mode
3. Multiple streams with large data blocks operating in full duplex mode
4. Extreme file creation rates

Capture Environment

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

HPCS I/O Environments

Capture Environment	Parallel Environment
System dedicated to moving external data to a large shared memory space where an application will analyze the data streams, create files and then write data to disk	System with many client nodes connected to a single shared file system or name space. Both the data and metadata operations are important for efficient use.
	
DARPA's goals: ▪ Streaming I/O at 30 GB/sec ▪ 32K file creates per second	DARPA's goals: ▪ 30K nodes ▪ One trillion files in a single file system ▪ 10K metadata operations per second

- In 2011, Lustre deployments in Parallel Environments approach DARPA's expectations
- Lustre is not yet an option for Capture Environments because of limits in Lustre client

HPCS I/O Scenarios Tests

- Cray's HPCS program has implemented the Scenarios and published the source under Cray's BSD-compliant Open Source License on SourceForge

  http://hpcs-io.cray.com/

  - The Scenarios and Cray's implementation are file system agnostic
  - Other proprietary implementations are known to exist, but Cray's version of the HPCS I/O Scenarios tests are the first to be made available as open source
- Instead of building new benchmarks, Cray proposes that the community start with the HPCS I/O Scenarios
  - Lustre should continue its focus on scalability, not just peak performance
  - If necessary, create additional Scenarios to generate relevant workloads to meet customer requirements

BACKUP

Streaming I/O Scenarios

Scenario	Expected Results
1. Single stream with large data blocks operating in half duplex mode	Provide information on the scalability of increasing the size of the storage system using at least three (3) data points taken from three (3) different size storage configurations. Desired performance is at least *Peak*% of theoretical peak of hardware performance. Likewise desired scaling is at least *Scaling*% from a small storage configuration to a larger configuration.
2. Single stream with large data blocks operating in full duplex mode	
3. Multiple streams with large data blocks operating in full duplex mode	

Scenario	*Peak*%	*Scaling*%
1	90%	95%
2	85%	90%
3	75%	80%

The Streaming I/O Scenarios require three different hardware configurations.

Parallel I/O Scenarios

Scenario	Expected Results
5. Checkpoint/Restart to single file with large I/O requests	Achieve 75% of theoretical peak of the bandwidth of the slowest hardware component in the data path of the system, where the slowest component is defined as the total bus bandwidth, the total bandwidth to storage, or the total storage bandwidth. Run the tests on at least three (3) different hardware configurations to understand the scalability of this sequential I/O test.
6. Checkpoint/Restart to single file with small I/O requests	
7. Checkpoint/Restart to multiple files with large I/O requests	
8. Checkpoint/Restart to multiple files with small I/O requests	
13. Small block random I/O to multiple files	Run the tests on at least three (3) different hardware configurations to understand the scalability of this random I/O test. Understand the performance as a percentage of hardware performance.
14. Small block random I/O to a single file	

The Parallel I/O Scenarios require three different hardware configurations.

Metadata I/O Scenarios

Scenario	Expected Results
4. Extreme file creation rates	Provide information on the scalability of increasing the amount of time this Scenario is run and the number of files created per second. At least three (3) data points should be provided, for example one (1), five (5), and ten (10) minute file creation runs. Must execute a full file system check for each run and report the time to complete the full file system check.
9. Walking through directory trees	Execute at least three (3) runs using cold cache and three (3) runs using warm cache. The variance in reported performance should be low especially for the warm cache results.
10. Parallel walking through directory trees	
11. Random stat() system call to files in the file system by one process	
12. Random stat() system call to files in the file system by multiple processes	

The Metadata I/O Scenarios require three runs on the same configuration.