LAWRENCE
LIVERMORE
NATIONAL
LABORATORY

UCRL-TR-226526

Purple L1 Milestone Review Panel: GPFS Functionality and Performance

W. E. Loewe

December 4, 2006

This work was performed under the auspices of the U.S. Department of Energy by University of California, Lawrence Livermore National Laboratory under Contract W-7405-Eng-48.

Purple L1 Milestone Review Panel
GPFS Functionality and Performance
Bill Loewe, 12/1/2006

Deliverable

The GPFS deliverable for the Purple system requires the functionality and performance necessary for ASC I/O needs. The functionality includes POSIX and MPIIO compatibility, and multi-TB file capability across the entire machine. The bandwidth performance required is 122.15 GB/s, as necessary for productive and defensive I/O requirements, and the metadata performance requirement is 5,000 file stats per second.

Criteria

To determine success for this deliverable, several tools are employed. For functionality testing of POSIX, 10TB-files, and high-node-count capability, the parallel file system bandwidth performance test IOR is used. IOR is an MPI-coordinated application that can write and then read to a single shared file or to an individual file per process and check the data integrity of the file(s). The MPIIO functionality is tested with the MPIIO test suite from the MPICH library.

Bandwidth performance is tested using IOR for the required 122.15 GB/s sustained write. All IOR tests are performanced with data checking enabled. Metadata performance is tested after "aging" the file system with 80% data block usage and 20% inode usage. The fdtree metadata test is expected to create/remove a large directory/file structure in under 20 minutes time, akin to interactive metadata usage. Multiple (10) instances of "ls -lR", each performing over 100K stats, are run concurrently in different large directories to demonstrate 5,000 stats/sec.

Results

In November, 2005, the Purple acceptance test was performed on a 1024-node system with 3 metadata servers and 101 I/O servers. The functionality testing was completed successfully.

In April, 2006, on the full 2.0 PB Purple file system with 3 metadata servers and 125 I/O servers, IOR showed file-per-process performance rates of 73 GB/s for write and 115 GB/s for read on 512 nodes to 16GB-files. The single-shared-file performance was 129 GB/s (W) and 153 GB/s (R) for 1024 nodes to an 8TB-file.

In September, 2006, a single 1.4 PB GPFS file system (/p/gscratch3) using the upgraded I/O subsystem with 3 metadata servers and half the available I/O servers was tested. For the metadata performance, the "ls -lR" performance was 10,000 stats/sec in aggregate on 10 nodes. The fdtree test completed in 16 minutes as well. This satisfied the metadata performance requirements.

In addition, application testing was performed under load using UMT2K on 1024 nodes (4096 processors) writing 1.4TB of file data per timestep while IOR was repeatedly writing shared 40TB-files from 300 nodes (300 processors). While running simultaneously, the lowest reported UMT2K performance was ~59 GB/s, while the lowest IOR reported was ~19 GB/s to the same file system.

Further, on this file system using half of the total I/O servers, IOR showed file-per-process performance rates of 71 GB/s for write on 1024 nodes to 40GB-files. The single-shared-file performance was 71 GB/s for writing from 512 nodes to a shared 20TB-file.

In November, 2006 the second 1.4 PB GPFS file system (/p/gscratch4) using the other half of the I/O servers was complete. Running across both of these file systems (/p/gscratch3 and /p/gscratch4) in a file-per-process pattern (even tasks writing their files to one file system, odd to the other) achieved 131 GB/s write performance on 500 nodes writing 10GB files. 800 nodes reading 10GB files achieved 139 GB/s. This satisfied the 122 GB/s requirement.

Conclusion

The functionality and performance requirements were met for Purple's GPFS. The file system passed the POSIX and MPIIO functionality requirements. As well, the metadata performance results satisfied the requirement of 5,000 stats per second and fdtree completion in under 20 minutes. For the bandwidth performance, no data integrity issues were encountered during the testsing, and the 122.15 GB/s was satisfied with over 130 GB/s write performance.