A Review on the Paper

PVFS: A Parallel File System for Linux Clusters

by Sughosh Kadkol

PVFS is proposed to be an open source solution available for download and use in research for parallel file systems and parallel I/O. The paper discusses the motivation, techniques and experimental results in developing an alternative to parallel file systems dominated by commercial parallel machines. PVFS is designed to provide high bandwidth concurrent I/O, support multiple API sets along with basic UNIX interoperability, be robust, scalable with a relative ease in installation and use. The tool described should provide a simple and cost-effective solution for data intensive research projects.

Platform specific commercial clusters and the lack of suitability of distributed file systems for large parallel scientific applications presented the need for a robust and scalable solution to PFS. To allow simple operation of the PFS, it is designed to include a wrapper in a custom kernel module replacing the standard UNIX wrapper with logic for both kernel and PVFS I/O support. The MPI-IO API is introduced to handle I/O operations to handle a custom data storage specification.

Additional overheads are avoided by maintaining request ordering and using standard low level stream I/O. Multiple I/O daemons and client library instances together handle file I/O while a single manager daemon controls file access and permissions and other metadata operations. The manager daemon communicates with applications only, the locations of I/O nodes associated with requested files, playing no role in read/write operations. This daemon may be pointed out to be a single point of failure in the system.

Tests showed, for concurrent read/write operations with native PVFS , maximum speeds of approximately 220 Mbytes/sec with fast Ethernet and 650 Mbytes/sec on Myrinet were reached with the latter configuration showing better consistency on variation of the compute-to-I/O node ratio. MPI-IO had comparable results despite a slight overhead while the BTIO benchmark confirmed the better performance with PVFS on Myrinet. Limitations imposed by TCP on Myrinet were an indication of a requirement for improved communication and tuning support.

The merits of PVFS are clearly seen to be ease of use, simple installation and cost effectiveness being its major strengths. These advantages also meant that PVFS was not designed for dedicated systems and thus I/O nodes might not boast a high processing throughput. Fault tolerance was not addressed, clear guidelines redundancy support were lacking. Data striping and partition mechanisms in PVFS seem to affect performance while fragmentation has no discernable effect. With better I/O description formats and better scheduling algorithms in daemons though, PVFS promises to be a simple and effective PFS alternative in future research.