Rethinking I/O in High-Performance Computing Environments

Ohio Supercomputer Center

Nawab Ali and P. Sadayappan

The Ohio State University • www.cse.ohio-state.edu/~alin

Introduction

Application Trends

- HPC applications increasingly require processing of large data sets: Sloan Digital Sky Survey, CERN LHC, NEES

Technological Trends

- Introduction of smart peripherals such as Object-based Storage Devices (OSDs)
- Bandwidth of WAN and Internet backbones growing at a rate that makes it comparable to local interconnect speed
  - TeraGrid, Lambda Rail: ~40Gb/s. InfiniBand: ~10 Gb/s

Research Challenges

- How do we manage the petabyte-scale data generated by HPC applications?
- How do we provide applications with high-bandwidth access to data?

Proposed Solutions

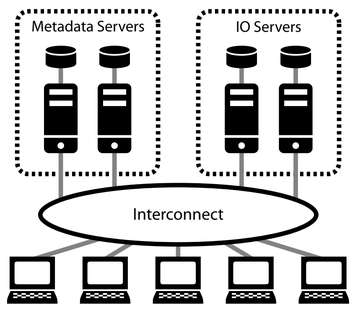
- Redesign parallel file systems using Object-based Storage Devices
- Parallel I/O over Wide Area Networks

Redesigning Parallel File Systems Using Object-based Storage Devices

- HPC application performance is often bottlenecked by I/O throughput.
- Current parallel file system designs, with dedicated I/O and metadata servers are unable to handle the requirements of data-intensive scientific applications.
- OSD is a logical extension to a disk. It abstracts data layout and low-level management from the file system.
- This work examines the feasibility of using OSDs to design a high-performance, scalable parallel file system.

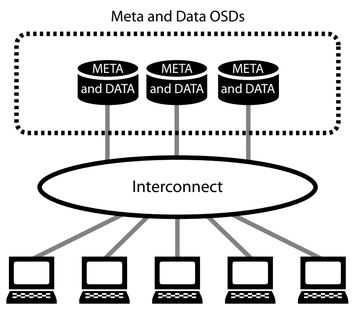
Parallel File System

Segregated data and metadata

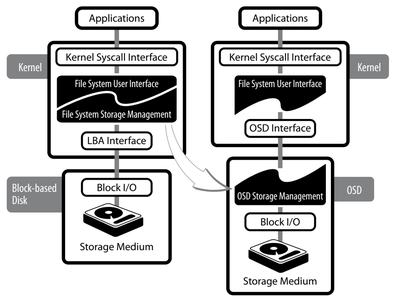


Parallel File System

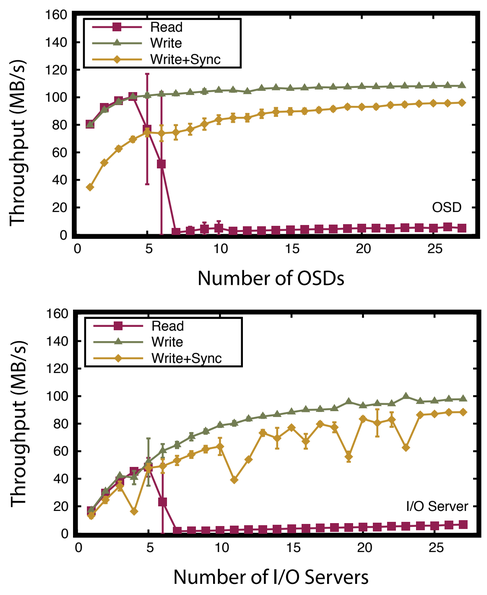
OSD-based integrated file system



Comparison of Block-based and Object-based Disks



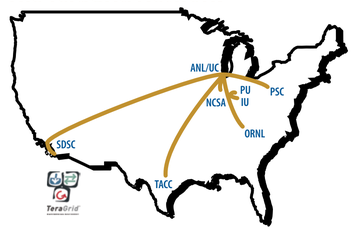
perf I/O throughput



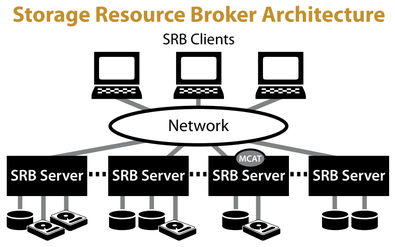
Parallel I/O Over Wide Area Networks

- HPC applications increasingly access data stored in remote locations.
- This paradigm shift has been brought about by the availability of high-speed wide area networks and the large amounts of data generated by these applications.
- SEMPLAR is a scalable, high-performance, remote I/O library that performs I/O over the Internet.
- SEMPLAR is based on the SDSC Storage Resource Broker (SRB).

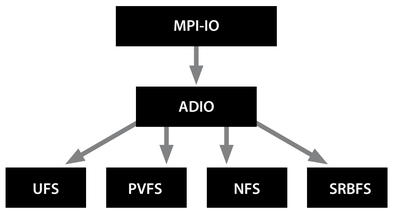
The TeraGrid Network



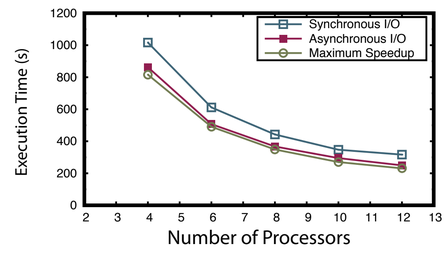
Storage Resource Broker Architecture



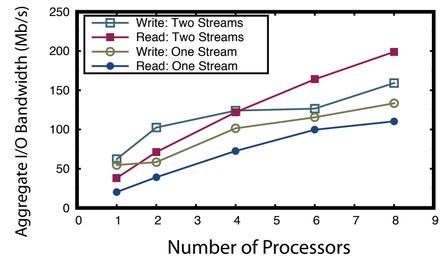
The ADIO Framework for Parallel I/O



MPIBLAST: NCSA TeraGrid



perf: NCSA TeraGrid



Future Work

- Parallel file systems over WANs
- Journaling using OSDs
- Data redundancy across multiple concurrent I/O streams
- Dynamic degree of data stream parallelism

Related Publications

[1] N. Ali and M. Lauria. SEMPLAR: High-performance remote parallel I/O over SRB. In 5th IEEE/ACM International Symposium on Cluster Computing and the Grid, Cardiff, UK, 2005.

[2] N. Ali and M. Lauria. Improving the performance of remote I/O using asynchronous primitives. In 15th IEEE International Symposium on High Performance Distributed Computing, Paris, France, 2006.

[3] A. Devulapalli, D. Dalessandro, N. Ali, and P. Wyckoff. Attribute storage design for Object-based Storage Devices. In MSST'07, San Diego, CA, Sept. 2007.

[4] A. Devulapalli, D. Dalessandro, P. Wyckoff, N. Ali, and P. Sadayappan. Integrating parallel file systems with Object-based Storage Devices. In Proceedings of SC'07, Reno, NV, Nov. 2007.

Empower. Partner. Lead.