

HEC FSIO Session 3: Communication and Protocols Talks and Roadmap

Rob Ross, Argonne National Lab

August 2009



Reminder of Current R&D Gaps

- Active Networks
- Alternate I/O transport schemes
- Coherence schemes

2006 HECURA/CPA Projects

- The Server-Push I/O Architecture for HEC; Xian-He Sun, Illinois Tech
 - Unlike traditional I/O designs where data is stored and retrieved by request, a new I/O architecture for High End Computing (HEC) is proposed based on a novel "Server-Push" model where a data access server proactively pushes data from a file server to the compute node's memory. The objective of this research is two fold: 1) increasing fundamental understanding of data access delay, 2) producing an effective I/O architecture that minimizes I/O latency. The PIs plan to increase the fundamental understanding through the study of data access pattern identification, prefetching algorithms, data replacement strategy, and extensive experimental testing. The PIs will verify the performance improvement with their file server design for various critical I/O intensive applications by using a combination of simulation and actual implementation in the PVFS2 file system.

2006 HECURA/CPA Projects

- Scalable I/O Middleware and File System Optimizations for High-Performance Computing; Alok Choudhary, Northwestern University and Mahmut Kandemir, Pennsylvania State University University Park
 - The main goals of this project are to design and implement novel I/O middleware techniques and optimizations, parallel file system techniques that scale to ultra-scale systems, design and development of techniques that efficiently enable newer APIs and flexible I/O benchmarks that mimic real and dynamic I/O behavior of science and engineering applications. Specifically, the objectives are to (1) design and develop middleware I/O optimizations and cache system that are able to capture small, unaligned, irregular I/O accesses from large number of processors and uses access pattern information to optimize for I/O; (2) incorporate these optimizations in MPICH2's MPI-IO implementation to make them available to a large number of users; (3) design and evaluate enhanced APIs for file system scalability, and (4) develop flexible, execution oriented and scalable I/O benchmarks that mimic the I/O behavior of real science, engineering and bioinformatics applications.

2006 HECURA/CPA Projects

- Active Storage Networks for HEC; John Chandy , U. Conn.
 - In this project, we suggest a new approach called active storage networks (ASN) - namely putting intelligence in the network along with smart storage devices to enhance storage network performance. These active storage networks can potentially improve not only storage capabilities but also computational performance for certain classes of operations. The main goals of this project will include investigation of ASN topologies and architectures, creation of ASN switch from reconfigurable components, studying HEC applications for ASNs, protocols to support programmable active storage network functions, and storage system optimizations for ASNs.

2006 HECURA/CPA Projects

- Active Data Systems, A.L. Narasimha Reddy , TAMU
 - This project plans to address several issues related to broadening the practicality of active storage. More specifically, this project plans to study and investigate:
(1) The impact of mixed workloads (both active and normal requests) at the active devices. (2) The impact of multiple active applications at the active devices. (3) The resource scheduling and QOS policies for a diverse set of workloads. (4) The impact of intelligent allocation in active storage systems.
 - In order to address these issues, the project plans to develop (a) an "active data" model to allow flexible processing of data, either at devices or at the requester. (b) QOS algorithms and security mechanisms for mixed workloads. (c) Algorithms and prototypes for exploiting the nature of data to develop content-based active storage.

2009 HECURA Projects and Presentations

- None -

2008 Communication and Protocols Gap Area

Area	Researchers	FY 07	FY 08	FY 09	FY 10	FY 11	FY 12	Rankings
Active Networks	<u>Chandy</u>							   Novel work being done, but not general enough.
	<u>Maccabe/Schwan</u>							
Alternative I/O transport schemes	Sun							   Most aspects are being addressed.
	Wyckoff							
	<u>Lustre</u>							
	<u>pNFS</u>							
Coherent Schemes	ANL/CMU							   No consensus on how to do this correctly, but some solutions are in products.
	<u>UCSC's Ceph</u>							
	<u>Lustre</u>							
	<u>Panasas</u>							
	<u>PVFS</u>							

- | | | |
|--|---|--|
|  Very Important |  Greatly Needs Research |  Greatly Needs Commercialization |
|  Medium Importance |  Needs Research |  Ready and Needs Commercialization |
|  Low Importance |  Does Not Need Research |  Not Ready for Commercialization |
|  Full Calendar Year Funding |  Partial Calendar Year Funding |  On-Going Work |

Discussion

- New Areas (two more in a moment)
 - Scalable failure detection, replication, and relocation
 - Note: There is overlap here with self-* topic in next-generation I/O breakout
 - Wide-area storage access protocols
 - Multi-fabric issues considered related to this topic
 - Topology aware access and layout
 - Common system for in-network checksumming
- Considered removing “Alternate I/O transport schemes”

Expanding Comm. and Protocols Area?

- Storage abstractions for scientific data
- Might include:
 - Blurring line between memory and storage
 - Non-POSIX
 - APIs for small object manipulation (e.g. `setattrlist`)
- Note: I/We thought (without looking) that something like this would already exist in the next-generation I/O area, but it really doesn't.

Development to Achieve Critical Mass

- Need for development work to attain critical mass for certain technologies
 - pNFS and OSD are possible storage examples
 - Reference implementations as enabling use and fostering adoption (example: MPI)
- Observation that this isn't a research gap
- Related suggestion: Reword circle on gap area table for R to R&D, or new D circle to specifically call out need.

Votes

Storage abstractions for scientific data	46
Scalable replication, relocation, failure detection, and fault tolerance	27
Need for development work to attain critical mass for certain technologies	21
Topology awareness, layout in storage access	18
Wide area storage access protocols	10
Active networks	6
Common system for in-network checksumming	4
Alternate I/O transport schemes	3
Coherence schemes	0

Shameless Workshop Plug



Friday September 4, 2009
New Orleans, LA, USA

Call for Papers

High-performance computing simulations and large scientific experiments such as those in high energy physics generate tens of terabytes of data, and these data sizes grow each year. Existing systems for storing, managing, and analyzing data are being pushed to their limits by these applications, and new techniques are necessary to enable efficient data processing for future simulations and experiments.

The purpose of this workshop is to provide a forum for engineers and scientists to present and discuss their most recent work related to the storage, management, and analysis of data for scientific workloads. Emphasis is placed on forward-looking approaches to tackle the challenges of storage at extreme scale or to provide better abstractions for use in scientific workloads.

IMPORTANT DATES

Paper Submission Deadline:
June 12, 2009

Author Notification:
July 10, 2009

Final Manuscript:
July 31, 2009

Workshop:
September 4, 2009

SUBMISSION INFO

<http://www.mcs.anl.gov/events/workshops/iasds09/>