

# **HEC FSIO Session 8: Metadata Breakout**

**Lee Ward, Sandia National Lab**

August 2009



# 2006 HECURA/CPA Projects

- Techniques for Streaming File Systems and Databases; Michael Bender, SUNY at Stony Brook and Martin Farach-Colton, Rutgers University New Brunswick

The performance of many high-end computing applications is limited by the capacity of memory systems to deliver data. As processor speeds increase, I/O performance continues to lag. Thus, I/O is likely to remain a critical bottleneck for high-end computing.

The researchers propose to address core problems on how to organize data on disk to optimize I/O, thus re-examining decades-old questions in the face of new applications, new technology, and new techniques. Specifically, the researchers propose to build prototypes of their streaming B-tree and variants for a file system or database. Streaming B-trees index and scan data at rates one-to-two orders of magnitude faster than traditional B-trees; they use cache-oblivious techniques to achieve platform independence. Several

issues remain to be addressed, specifically, how to deal with different-sized keys, how to support transactions, how to scale to multiple disks and processes, and how to provide O/S support for cache-obliviousness and memory-mapped massive data.

The proposed work represents a promising new direction for manipulating massive data and overcoming classic I/O bottlenecks. In HEC file systems and databases, this technology will permit rapid streaming of data onto and off of disks for high-throughput processing of data. This work will result in the transfer of recently developed algorithmic techniques to other areas of computer science, engineering, and scientific computing and is intended to transform how scientists and engineers manipulate massive data sets.

# 2006 HECURA/CPA Projects

- Petascale I/O for High End Computing; Maccabe, Arthur, UNM and Schwann, Karsten, Georgia Tech Research Corporation – GA Inst of Tech

The purpose of the new functionality inherent in SSDS is to help developers

carry out complex I/O tasks. Technical topics to be addressed to realize this goal include the development of automated methods for deploying graph nodes to the physical sites that perform I/O functions, of dynamic management methods that maintain desired

levels of QoS for those I/O functions that require it (e.g., when accessing remote sensors). A key aspect of this work is the automation of I/O Graph creation and deployment. XML-based interfaces will make it easy for developers to provide information about the

structure of I/O data, and to specify useful data manipulations. Efficient representations of metadata will enable both in-band and out-of-band data manipulation, to create I/O Graphs that best match current I/O needs and available machine resources. New offline techniques will derive metadata that can be used to enrich I/O graphs and more generally,

meta-information about the large data volumes produced and consumed by MPP applications. Finally, this work will improve flexibility for I/O in future MPP machines, where virtualization techniques coupled with new chip (i.e., multicore) and interconnect technologies will make it easier to construct multi-use MPP platforms capable of efficiently performing both computational and I/O tasks.

# 2006 HECURA/CPA Projects

- Applicability of Object-Based Storage Devices in Parallel File Systems; Pete Wyckoff, NetApp (formerly Ohio State University Research Foundation)

We will examine multiple aspects of the mismatch between the needs of a parallel file system, in particular PVFS2, and the capabilities of OSD. Topic areas include mapping data to objects, metadata, transport, caching and reliability. Trade-offs arise from the mapping of files to objects, and how to stripe files across multiple objects and disks, in order to obtain good performance. A distributed file system needs to track metadata that describes and connects data. OSDs offer automatic management of some critical metadata components that can be used by the file system. There are transport issues related to flow control and multicast operations that must be solved. Implementing client caching schemes and maintaining data consistency also requires proper application of OSD capabilities.

Our work will examine the feasibility of OSDs for use in parallel file systems, discovering techniques to accommodate this high performance usage model. We will also suggest extensions to the current OSD standard as needed.

# 2006 HECURA/CPA Projects

- SAM<sup>2</sup> Toolkit: Scalable and Adaptive Metadata Management for High-End Computing; Hong Jiang, University of Nebraska-Lincoln and Yifeng Zhu, University of Maine

# 2006 HECURA/CPA Projects

- Improving Scalability in Parallel File Systems for High End Computing; Walt Ligon, Clemson University

The issues we would study are scalable metadata operations, small, unaligned data accesses,

reliability through redundancy, and management of I/O resources. Techniques we expect to

employ include active caching and buffering, server-to-server and client-to-client communication, and autonomies. We intend to employ middleware whenever possible in order to enhance portability and control complexity. A major theme of the proposal is that file systems that provide everything all of the time are at a disadvantage in terms of scalable performance because features, like strict consistency and parity-based redundancy, are hard to implement with good scalability. A file system that can configure itself to match the needs of the application can get the best performance possible. Thus, PVFS2 was developed to allow a large degree of configurability, and the proposed research intends to enhance that file system so that it will scale to very large sizes.

# 2006 HECURA/CPA Projects

- Microdata Storage Systems for High-End Computing; Charles Leiserson, MIT

The research focuses on three promising technologies:

Microdata storage structures, such as buffered repository B-trees, which can improve the performance of insertions and range queries of microfiles by orders of magnitude over traditional B-trees, while still preserving high performance on macrofiles.

Cache-oblivious data structures, which provide passive self-tuning of the file organization and may actually outperform tuned cache-aware data structures for disk file systems.

Virtual-memory-based transactional memory, which allows programmers to implement complex file structures in a straightforward manner, while providing lock-free programming and automatic crash recovery.

# 2006 HECURA/CPA Projects

- High Throughput I/O for Large Scale Data Repositories; Ali Saman Tosun, University of Texas at San Antonio

Decustering has attracted a lot of interest over the last few years and has applications in many areas including high-dimensional data management, geographical information systems and scientific visualization. Most of the declustering research have focused on spatial range queries and finding schemes with low worst-case additive error. This research investigates various aspects of declustering including novel declustering schemes, replicated declustering, heterogeneous declustering, adaptive declustering and declustering using multiple databases. The investigators approach every issue both theoretically and practically, study what is theoretically possible, what can be achieved in practice and try to close the gap between the two. The investigators study novel declustering schemes with solid theoretical foundations including number-theoretic declustering and design-theoretic declustering. Replication strategies for various types of queries including spatial range queries and arbitrary queries are studied. Retrieval algorithm for design-theoretic replication has linear complexity and guarantees worst-case retrieval cost. The investigators study tradeoffs in retrieval between complexity and retrieval cost and develop a suite of protocols for retrieval. This research involves adaptive declustering schemes that adapt to disk failures, disk additions and changing query types by moving buckets between disks during idle

# 2009 HECURA

- A New Semantic-Aware Metadata Organization for Improved File-System Performance and Functionality in High-End Computing; Yifeng Zhu, University of Maine

# 2009 HECURA

- Scalable Data Management Using Metadata and Provenance; Ethan Miller, UCSC

# Current R&D Gaps

- **Scaling (scored 56)**
  - Recovery methods/self recovery for dealing with failures at scale – should be noted/emphasized
  - Transaction management ?
  - User-extensible metadata performance
- **Extensibility and Name Spaces (scored 35)**
  - Efficient search
  - Alternate indexing schemes research is needed
- **Hybrid devices exploration (scored 21)**
  - Specific to metadata
  - Novel algorithms (beyond caching) and data structures that utilize these devices
- **Data transparency and access methods (scored 10)**
- **File system/archive metadata integration (scored 7)**
  - Replace with cross discipline metadata integration?

## 2008 Metadata Gap Area

Area	Researcher	FY 07	FY 08	FY 09	FY 10	FY 11	FY 12	Rankings
Scaling	Bender/Farach-Colton	■	■	■				 All existing work is evolutionary. What is lacking is revolutionary research; no fundamental solutions proposed.  This category includes archive metadata scaling.  More research in reliability at scale is needed
	Jiang/Zhu	■	■	■				
	Leiserson	■	■	■				
	Maccabe/Schwann	■	■	■				
	SciDAC - PDSI	■	■	■				
	HECEWG HPC Extensions	■	■	■	■	■	■	
	UCSC's Ceph	■	■	■	■	■	■	
	CEA/Lustre	■	■	■	■	■	■	
	CMU/ANL – Large Directory	■	■	■	■	■	■	
	PVFS	■	■	■	■	■	■	
Panasas	■	■	■	■	■	■		
Extensibility and Name Spaces	Bender/Farach-Colton	■	■	■				 All existing work is evolutionary.  Extensibility includes provenance capture
	Jiang/Zhu	■	■	■				
	Leiserson	■	■	■				
	Tosun	■	■	■	■	■		
	Wyckoff	■	■	■				
	UCSC – LIES/facets	■	■	■				
	CMU/ANL - MDIFS	■	■	■				
SciDAC PDSI	■	■	■					
File System/ Archive Metadata Integration	Lustre HSM	■	■	■	■			 Extended Attributes, although not standardized, could solve problem.
	UMN Lustre Archive	■	■	■				
Hybrid Devices Exploitation	CMU – Flash Characterization		■	■				 Research is being done, but little research focused on metadata
Data Transparency and Access Methods	<i>None</i>							 No research focused on metadata

