

HEC FSIO Session 7: Archive Gaps Roadmap

Gary Grider, Los Alamos National Lab
August 2009

Current R&D Gaps

- APIs/Standards for interface, searches, and attributes, staging, etc.
- Long term attribute driven security
- Long term data reliability and management
- Metadata scaling
 - This gap area has been integrated into Metadata “scaling” since work in file system scaling is related and applicable to archive.
- Policy driven management

2006 HECURA/CPA Projects

- David Du, University of Minn/NSF
 - “Integrated Infrastructure for Secure and Efficient Long-Term Data Management”
 - This project will develop a high-performance long-term data management system that will ensure the necessary levels of security throughout the lifecycle of a data set. The goal is a hierarchical cluster-based archival storage solution that will provide: (i) transparent backup, restore, and data access operations that will allow individual application programs and business entities to securely and efficiently archive data for decades; (ii) high-performance data access in a cluster computing environment; and (iii) innovative techniques for efficiently insuring long-term data security and accessibility, including long-term key management. The solution will be suitable for heterogeneous computing environments, including the extremely high-throughput ones of the high-performance computing (HPC) community.

2006 HECURA/CPA Projects

- Patrick McDaniel for Anand Sivasubramaniam, PSU
 - “Asymmetry in Performance and Security Requirements for I/O in High-end Computing”
 - The motivation for our DataVault project is driven by the need to secure storage systems which cater to the demands of high-end applications, while meeting their stringent performance requirements. Rather than have a one-solution-fits-all approach, we propose to investigate the rich design space - threats, storage architecture, enforcement mechanism, performance – to offer insightful choices that can be useful when deploying/customizing storage systems. DataVault will also include a usable objective-driven policy interface to configure the system for a given set of security and performance needs, while offering a convenient visualization dashboard for security management.

2009 HECURA Projects and Presentations

- None

2008 Archive Gap Area

Area	Researchers	FY 07	FY 08	FY 09	FY 10	FY 11	FY 12	Rankings	
API's/Standards for interface, searches, and attributes, staging etc.	Ma/Sivasubramaniam/Zhou							 Current research is in terms of file systems, not archive. API merging with POSIX and API for searching lacking	
	Tosun								
	UCSC – Facets Work								
	SciDAC – SDM								
	SciDAC – PDSI								
Long term attribute driven security	Ma/Sivasubramaniam/Zhou							 Current research is in terms of file systems, not archive. Current researchers need data supporting proposed solutions usefulness	
	Odlvzko								
Long term data reliability and management	Apaci-Dusseau							 Need for research and commercialization is low because HIPPA and others will drive this. Redundancy techniques reasonably sufficient for archives	
Metadata scaling	Bender/Farach-Colton							 Current research is in terms of file systems, not archive, but this work can be applied to archive. File system research will be more than fast enough for archive.	
	Jiang/Zhu	This gap area has been integrated into Metadata “scaling” since work in file system scaling is related and applicable to archive.							
	Leiserson								
	Panasas								
	Lustre								
	ANL/CMU								
Policy driven management	None							 Sarbanes-Oxley Act is solving this problem	

- | | | |
|--|---|---|
|  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 |

Breakout Discussion

- Interfaces to archives may be inefficient or incapable of getting policy information, maybe we need entirely new interfaces that enable metadata richness
- Should we consider looking at Dedup for archive. We can possibly leverage industry products, but they are largely for backup. How can we tell if Dedup will help us. Maybe a project to fingerprint data as it is stored to the archive so we can collect data somewhat painlessly so that in a year or two we might know if Dedup might help us.

Breakout Discussion (Cont)

- Policy based management with more extensible metadata is needed that is indexed for easy use. If Xattrs were indexed, that might open up new things we could do in management of the huge amount of files.
- Will we be able to leverage cloud storage capabilities or any technologies they might develop?

New R&D Gaps

- APIs/Standards for interface, searches, and attributes, staging, etc. $3\ 1\ 3\ 3 = 10$
 - Richer interfaces for metadata
- Long term attribute driven security $1 = 1$
- (added) Analysis of what is in archive for Dedup etc. in HEC environments (see if we can find out how much this would help without having to read the entire archive) $2\ 2\ 3\ 2\ 2 = 11$
- Long term data reliability and management $3\ 2 = 5$
- Metadata scaling $3\ 3 = 6$
 - This gap area has been integrated into Metadata “scaling” since work in file system scaling is related and applicable to archive.
- Policy driven management $3\ 1\ 1\ 1 = 6$