

High End Computing Interagency Working Group (HECIWG) Sponsored File Systems and I/O 2009 Roadmaps

Marti Bancroft DOD/NRO
John Bent DOE/NNSA LANL
Evan Felix DOE/Office of Science PNNL
Gary Grider DOE/NNSA LANL
James Nunez DOE/NNSA LANL
Steve Poole DOE/Office of Science ORNL
Robert Ross DOE/Office of Science ANL
Ellen Salmon NASA
Lee Ward DOE/NNSA SNL

Executive Summary	2
Metadata Roadmap	3
Measurement and Understanding Roadmap	5
Quality of Service Roadmap.....	6
Next-generation I/O Architectures Roadmap	7
Communication and Protocols Roadmap.....	9
Archive Roadmap	10
Management and RAS Roadmap.....	11
Security Roadmap.....	12
Assisting with Standards, Research and Education Roadmap.....	13
Conclusion	15

Executive Summary

The need for immense and rapidly increasing scale in scientific computation drives the need for rapidly increasing scale in storage capability for scientific processing. Individual storage devices are rapidly getting denser while their bandwidth is not growing at the same pace. In the past several years, initial research into highly scalable file systems, high level Input/Output (I/O) libraries, and I/O middleware was conducted to provide some solutions to the problems that arise from massively parallel storage. To help plan for the research needs in the area of File Systems and I/O, the inter-government-agency published the document titled “HPC File Systems and Scalable I/O: Suggested Research and Development Topics for the Fiscal 2005-2009 Time Frame” which led the High End Computing Interagency Working Group (HECIWG) to designate this area as a national focus area starting in FY06. To collect a broader set of research needs in this area, the first HEC File Systems and I/O (FSIO) workshop was held in August 2005 in Grapevine, TX. Government agencies, top universities in the I/O area, and commercial entities that fund file systems and I/O research were invited to help the HEC determine the most needed research topics within this area. All HEC FSIO workshop reports can be found at <http://institute.lanl.gov/hec-fsio/docs/>. All presentation materials from all HEC FSIO workshops can be found at <http://institute.lanl.gov/hec-fsio/workshops/>

The workshop attendees helped and continue to gather each year to help

- catalog existing government funded and other relevant research in this area,
- list top research areas that need to be addressed in the coming years,
- determine where gaps and overlaps exist, and
- recommend the most pressing future short and long term research areas and needs necessary to help advise the HEC to ensure a well coordinated set of government funded research

The recommended research topics are organized around these themes: metadata, measurement and understanding, quality of service, security, next-generation I/O architectures, communication and protocols, archive, and management and RAS. Additionally, university I/O center support in the forms of computing and simulation equipment availability, and availability of operational data to enable research, and HEC involvement in the educational process were called out as areas needing assistance.

Metadata Roadmap

Investigation into metadata issues is needed, especially in the areas of scalability, extensibility, access control, reliability, availability, and longevity for both file and archival systems. Additionally, consideration for very revolutionary ideas such as new approaches to name spaces and use of novel storage devices needs to be explored.

2009 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. File system research will be fast enough for archive. More research in reliability at scale is needed
	Jiang/Zhu							
	Leiserson							
	Maccabe/Schwann							
	Zhu/Jiang							
	Bender/Farach-Colton/Leiserson/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							
	Panda (formerly Wyckoff)							
	Miller/Seltzer							
UCSC – LiFS/facets								
CMU/ANL - MDIFS								
SciDAC PDSI								
Cross Discipline (file system/archive/DB) Metadata Integration	Lustre HSM							 Extended Attributes, although not standardized, could solve problem.
	UMN Lustre Archive							
Non Traditional Device Exploitation	CMU – Flash Characterization							 Research is being done, but little research focused on metadata Caching is already well funded
Data Transparency and Access Methods	None							 No research focused on metadata



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

Measurement and Understanding Roadmap

Research tools for measurement and understanding of parallel file system and end-to-end I/O performance are needed for advances in future file systems including evolutionary ideas such as layered performance measurement, benchmarking, tracing, and visualization of I/O related performance data. Also, more radical ideas like end-to-end modeling and simulation of I/O stacks and the use of virtual machines for large scale I/O simulation need to be explored.

2009 Measurement and Understanding Gap Area

Area	Researcher	FY 07	FY 08	FY 09	FY 10	FY 11	FY 12	Rankings
Measurement and understanding of system workload in HEC environment	Arpaci-Dusseau	■	■	■	■			<p>A comprehensive tool is nowhere in sight; problem is complex.</p> <p>This gap area includes monitoring.</p>
	Narasimhan			■				
	Reddy				■			
	Smirni	■	■			■		
	Zadok	■	■		■			
	Riska			■	■	■	■	
	He			■			■	
Zadok (2009 HECURA)			■		■	■		
Standards and common practices for HEC I/O benchmarks	Zadok/Miller		■	■	■			<p>Danger of over simplifying problem and could drive vendors to incorrect solutions.</p>
	Ma/Shen/Winslett			■	■	■	■	
Modeling, simulation and test environments.	Ligon	■	■	■	■			<p>Simulators are being developed. No real testbeds being built. This problem will only get worse over time, i.e. as systems get bigger.</p>
	Thottethodi	■	■	■	■			
	Maltzahn			■	■	■		
Applying cutting edge analysis tools to large scale I/O	Reddy	■	■	■	■			<p>Data are becoming available from Labs including I/O traces. Many opportunities to evaluate this research.</p> <p>This includes applying analysis and visualization tools to I/O traces</p>
	Zadok	■	■	■	■			
	LANL/CMU – Trace replay and Visualizer		■	■	■			
	Ma/Iskra			■	■			

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

Quality of Service Roadmap

Quality of service (QoS) can be defined as features of a storage architecture that allow a user or administrator to recommend policies for data movement during I/O operations. QoS is a ripe topic for research especially in the area of providing prioritized, deterministic performance in the face of multiple, complex, parallel applications running concurrently with other non-parallel workloads. More revolutionary ideas such as dynamically adaptive end-to-end QoS throughout the hardware and software I/O stack are desirable.

2009 QoS Gap Area								
Area	Researchers	FY 07	FY 08	FY 09	FY 10	FY 11	FY 12	Rankings
End to End QoS in HEC	Brandt	■	■	■	■			 Good research, but much work needed to get a standards based solution. Scale and dynamic environments have to be addressed at some point in time. Some progress in single node/disk not much on distributed QoS, need a demo of distributed QoS in the next few years
	Chiueh	■	■	■	■			
	Ganger	■	■	■	■			
	Zhao/Figueiredo			■	■	■	■	
	Kandemir/Dennis			■	■	■	■	
	Burns			■	■			
Interfaces for QoS	SciDAC - PDSI	■	■	■	■			 Very partially addressed by proposed HEC POSIX Extensions. Will be driven by above "End to End QoS in HEC". We Should pursue getting info from resource managers, maybe an API from the RMS is in order and leverage SLA thinking
	POSIX HPC Extensions	■	■	■	■	■	■	

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

Next-generation I/O Architectures Roadmap

Until recently, I/O stacks and architectures have been static forcing developers to adopt awkward solutions in order to achieve target I/O rates. There is great need for research into next-generation I/O architectures, including evolutionary concepts such as extending the POSIX I/O API standard to support archives in a more natural way, access awareness, and HEC/high concurrence. Studies into methods to deal with small, unaligned I/O and mixed-size I/O workloads as well as collaborative caching and impedance matching are also needed. Novel approaches to I/O and file systems also need to be explored including redistribution of intelligence, adaptive and reconfigurable I/O stacks, user space file systems, data-aware file systems, and the use of novel storage devices. This area may be well-served by delving into and applying the research from the modeling community.

2009 Next Generation I/O Architectures Gap Area

Area	Researcher	FY 07	FY 08	FY 09	FY 10	FY 11	FY 12	Rankings
Storage abstractions and Scalable file system architectures	Choudhary/Kandemir							 Good work, but much of the research is in its infancy. A small portion ready for commercialization.
	Dickens							
	Ligon							
	Maccabe/Schwan							
	Reddy							
	Shen							
	Sun							
	Thain							
	Panda (formerly Wyckoff)							
	SciDAC – SDM							
	SciDAC – PDSI							
	Sarkar/Dennis/Gao							
	Rangaswami							
Choudhary (2009 HECURA)								
PNNL								
Self-assembling, Self-reconfiguration, Self-healing storage components	Ganger							 Good work being done, but it's a hard problem that will take more time to solve.
	Ligon							
	Ma/Sivasubramaniam/ Zhou							
	SciDAC - PDSI							
	SciDAC - SDM							
Non Traditional architectures leveraging emerging storage technologies	Gao							 Big potential reward, but very little work being done in the HEC area. Includes power consumption.
	Urgaonkar							
	Szalay/ Huang							
	He							
	Rangaswami							
	Arpaci-Dusseau (2009 HECURA)							
	PNNL							
HEC systems with multi-million way parallelism doing small I/O	Choudhary/Kandemir							 Good initial research; needs to be moved into testing. More fundamental solutions
	Dickens							
	Gao							
	Sun							
	Zhang/ Jiang							

2009 Next Generation I/O Architectures Gap Area

Area	Researcher	FY 07	FY 08	FY 09	FY 10	FY 11	FY 12	Rankings
operations	Sun							being pondered including non-volatile solid state storage.
	FASTOS – I/O Forwarding							
	CMU – Log Structured FS							

-  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

Communication and Protocols Roadmap

In the area of file system related communications and protocols, evolutionary items such as exploitation of Remote Direct Memory Access (RDMA), Object Based Secure Disk (OBSD) extensions, Network File System Version 4 (NFSv4) extensions, and parallel Network File System (pNFS) proof-of-concept implementations as well as more revolutionary exploration of server to server communications are needed.

2009 Communication and Protocols Gap Area

Area	Researchers	FY 07	FY 08	FY 09	FY 10	FY 11	FY 12	Rankings	
Active Networks	Chandy							 Novel work being done, but not general enough.	
	Maccabe/Schwan								
Alternative I/O transport schemes	Sun	This gap area is best represented in Next Generation I/O Architectures. Thus, the gap sub area will be removed from the Communications and Protocols Roadmaps.							 Most aspects are being addressed.
	Wyckoff Lustre pNFS								
Coherence Schemes	ANL/CMU							 No consensus on how to do this correctly, but some solutions are in products.	
	UCSC's Ceph								
	Lustre								
	Panasas PVFS								
Topology aware storage layout	None								
Wide area storage protocols	None								

- 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

Archive Roadmap

In the area of archive, the interfaces to the file systems and I/O stacks in HEC systems and long term care for the massive scale of an archive in the HEC environment are difficult areas needing more research than they have received before.

2009 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, deduplication prediction, etc.	Ma/Sivasubramaniam/ Zhou							 Current research is in terms of file systems, not archive. API merging with POSIX and API for searching and management lacking API could assist with helping us find out if deduplication would help us.
	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
	Odlyzko							
Long term data reliability and management	Arpaci-Dusseau							 Need for research and commercialization is low because HIPPA and others will drive this. Redundancy techniques reasonably sufficient for archives
Policy driven management	None							 Sarbanes-Oxley Act is solving this problem. If we were collecting xattrs that could help us manage files then we might need some research in this area but we don't have any information on which to manage beyond what we know how to manage with

-  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

Management and RAS Roadmap

In the area of management, reliability and availability at scale, management scaling, continuous versioning, and power management are all needed research topics. Additionally, more revolutionary ideas like autonomies, use of virtual machines, and novel devices exploitation need to be explored.

2009 Management and RAS Gap Area

Area	Researchers	FY 07	FY 08	FY 09	FY 10	FY 11	FY 12	Rankings
Proactive Health Methods	None							
Problem detection, reporting, analysis and modeling	Reddy							 More researchers need to look at this problem.
	Narasimhan							
Formal Failure analysis and tools for storage systems	Arpaci-Dusseau							 Good research done here. Will people use this work?
Improved Scalability	Ganger							 More research is needed here. Test beds are probably needed for this work.
	Ligon							
Power Consumption and Efficiency	Qin							 Industry is working on this problem. Storage is not a large consumer of energy at HEC sites.
	Zadok (2009 HECURA)							
	Khuller							
Scalable replication, relocation, failure detection, and fault tolerance	CMU – Diskreduce							 Industry is working on this problem
	IBM – Perseus							

- 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

Security Roadmap

Aspects of security such as usability, long term key management, distributed authentication, and dealing with security overhead are all topics for research. There is also room for more difficult research topics such as novel new approaches to file system security including novel encryption end-to-end or otherwise that can be managed easily over time. The need for standardization of access control list mechanisms is also needed and investigation into a standard API for end-to-end encryption could be useful.

2009 Security Gap Area

Area	Researchers	FY 07	FY 08	FY 09	FY 10	FY 11	FY 12	Rankings
Performance overhead and distributed scaling	Sivasubramaniam	■	■	■	■			 Problem reasonably well understood, unclear if enough demand for product
End-to-end confidentiality and tracking of information flow, provenance, etc.	Odlyzko	■	■	■	■			 Industry will help some, but not in HEC context.
	McDaniel/Sion/Winslett			■	■	■	■	
Use and management, quick recovery.	Sivasubramaniam	■	■	■	■			 Current researchers need data to validate designs Nothing to commercialize yet. Note: NSF should incorporate this into a call for security research; this topic is larger than FSIO.
Alternative Architectures for Authentication and Authorization	None							 Supporting Cloud Computing makes this HEC FSIO.

-  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

Assisting with Standards, Research and Education Roadmap

At the HEC FSIO 2005 workshop, there was a recognition that the HEC FSIO community should find ways of supporting students working in the general area of I/O as well as students working more specifically on I/O within HEC. Investment to support the research of these students was considered worthwhile both because they may provide important research while still in school as well as by cultivating these students such that they may continue to work on HEC I/O problems following their graduation and, with any luck, become the next generation of HEC I/O experts.

Over the past decade, the HEC community has had a role in the formation and adoption of various FSIO related standards. The most notable are the ANSI T10 1355D specification for Object Based Storage Devices (OBSD), the IETF NFSv4 standard including the new pNFS portion of the NFSv4.1 minor revision of the NFSv4 specification, and the newly formed Open Group HEC Extensions to the POSIX standards work has also been an outcome of HEC FSIO and the HEC I/O community work. Past years are status, future years are identified needs or desires

2009 Assisting with Standards, Research and Education					
Area	FY07	FY 08	FY 09	FY 10	FY 11
Standards:					
POSIX HEC	PDSI UM CITI patch pushing/maintenance Revamp of manual pages	First Linux full patch set	Layout Query going into POSIX		
ANSI OBSD	V2 nearing publication	Some file system pilot test	V2 ratified		
IETF pNFS	V 4.1 nearing pub Assistance in testing may be needed	Initial products	NFS v4.1 final voting ("last call") Linux Server is somewhat stalled		
Community Building	HEC FSIO 2007 HEC presence at FAST and IEEE MSST	HEC FSIO 2008 HEC presence at FAST and IEEE MSST	HEC FSIO 2009 HEC presence at FAST and IEEE MSST	HEC FSIO 2010 HEC presence at FAST and IEEE MSST	HEC FSIO 2011 HEC presence at FAST and IEEE MSST
Equipment	Incite and NSF Infra Need scale CS disruptive facility	Incite and NSF Infra Need scale CS disruptive facility	Incite and NSF Infra Need scale CS disruptive facility	Incite and NSF Infra Need scale CS disruptive facility	Incite and NSF Infra Need scale CS disruptive facility
Simulation Tools	Ligon PDSI Felix/Farber	Ligon PDSI Felix/Farber	Ligon PDSI Felix/Farber Updated Disksim including MEMS simulation SNL releasing kernel I/O tracing tool		
Education	LANL Institutes PDSI	Other Institute-like activities			
Research Data	Failure, usage, event data	Many more traces, FSSTATS, more disk failure data	More data released; I/O traces, Cray event logs, work		

2009 Assisting with Standards, Research and Education

Area	FY07	FY 08	FY 09	FY 10	FY 11
			station file system statistic data		

Conclusion

The petascale and, soon, the exascale supercomputers will be commonplace at HEC environments; sites will deploy supercomputers with hundreds of thousands processors routinely. Million-way parallelism is around the corner and, with it, bandwidth needs to storage will go from tens of gigabytes/sec to terabytes/sec. Online storage requirements to support work flows for efficient complex science will begin to approach the exabyte range. The ability to handle a more varied I/O workload ranging seven orders of magnitude in performance characteristics, extremely high metadata activities, and management of trillions of files will be required. Global or virtual enterprise wide-area sharing of data with flexible and effective security will be required. Current extreme-scale file system deployments already suffer from reliability and availability issues, including recovery times from corruption issues and rebuild times. As these extreme-scale deployments grow larger, these issues will only get worse. It will possibly be unthinkable for a site to run a file system check utility, yet it is almost a given that corruption issues will arise. Recovery times need to be reduced by orders of magnitude, and these types of tools need to be reliable, even though they may rarely be used. The number of storage devices needed in a single coordinated operation could be in the tens to hundreds of thousands, requiring integrity and reliability schemes that are far more scalable than available today. Management of enterprise-class global parallel file/storage systems will become increasingly difficult due to the number of elements involved, which will likely approach 100,000 spinning disks with widely varying workloads. The challenges of the future are formidable.