

High End Computing Interagency Working Group (HECIWG) Sponsored File Systems and I/O 2011 Road Maps

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
Roadmap Legend	3
Metadata Road Map	4
Measurement and Understanding Road Map.....	6
Quality of Service Road Map	8
Next-generation I/O Architectures Road Map.....	9
Communication and Protocols Road Map	11
Archive Road Map.....	12
Management and RAS Road Map	14
Security Road Map	16
Assisting with Standards, Research and Education	17
Conclusion	19

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. The HEC FSIO 2005 workshop report 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

- 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.

Roadmap Legend

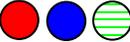
The following road maps use the legend below to rank each FSIO gap area in importance, does the area still need research, is the research ready for commercialization and duration of funding.

 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

Metadata Road Map

Investigation into metadata issues is needed 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.

2011 Metadata Gap Area

Area	Researcher	Fiscal Year								Rankings	
		07	08	09	10	11	12	13	14		
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 – Large Directory (Giga+)										
	PVFS/Orange FS										
Panasas											
Extensibility, Access Methods and Name Spaces	Bender/Farach-Colton										 All existing work is evolutionary. Extensibility includes provenance capture
	Jiang/Zhu										
	Leiserson										
	Tosun										
	Panda (formerly Wyckoff)										
	SciDB										
	Miller/Seltzer										
	UCSC – LiFS/facets										
	CMU/ANL - MDFS										
	SciDAC PDSI										
Non Traditional Device Exploitation	CMU – Flash Characterization										 Research is being done, but little research focused on metadata Caching is already well funded
	UCSD – NVM Characterization										



Very Important



Greatly Needs Research



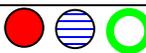
Greatly Needs Commercialization

 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 Road Map

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.

2011 Measurement and Understanding Gap Area

Area	Researcher	Fiscal Year								Rankings
		07	08	09	10	11	12	13	14	
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>
	Reddy									
	Smirni									
	Zadok									
	Narashimhan									
	Riska									
	He									
	Zadok (2009 HECURA)									
	SciDAC - PDSI									
	SciDAC – SDM									
	Darshan – ANL									
	Power Management – Curry – SNL/LANL/ Clemson									
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>
	High Productivity Computing Systems (HPCS) Benchmarks									
	Ma/Shen/Winslett									
Modeling, simulation and test environments.	Clemson - Ligon									 <p>Simulators are being developed. PROBE's testbeds for use are retired clusters. No real testbeds being built.</p> <p>This problem will only get worse over time, i.e. as systems get bigger.</p>
	CODES – ANL/RPI									
	PROBE – LANL/CMU									
	Thottethodi									
	UCSC - Maltzahn									
	DiskSim in SST – Oldfield - Sandia									
	DMD – SNL/LBNL /UMD /Columbia									
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</p>
	Zadok									
	LANL/CMU – Trace replay and Visualizer									

2011 Measurement and Understanding Gap Area

Area	Researcher	Fiscal Year								Rankings
		07	08	09	10	11	12	13	14	
	Ma/Iskra									this research. This includes applying analysis and visualization tools to I/O traces

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

Quality of Service Road Map

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.

2011 Quality of Service Gap Area

Area	Researchers	Fiscal Year								Rankings
		07	08	09	10	11	12	13	14	
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. Someone needs to take the existing QoS pieces and demo an end-to-end solution.
	Chiueh				■					
	Ganger	■	■	■						
	Zhao/Figueiredo			■	■	■	■			
	Kandemir/Dennis			■	■	■	■			
	Burns			■	■					
	FairIO - Teller				■					
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 Road Map

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.

2011 Next Generation I/O Architectures Gap Area

Area	Researcher	Fiscal Year								Rankings
		07	08	09	10	11	12	13	14	
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)									
	DAMSEL – NCSU/ NWU/ ANL									
	Damasc – UCSC/LLNL									
	Long/Miller - UCSC									
PNNL										
GoofyFS – SNL/UMinn /Clemson/UAB/ANL/ORNL										
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	Gao									 Big potential reward, but very little work
	Urgaonkar									
	Szalay/ Huang									

2011 Next Generation I/O Architectures Gap Area

Area	Researcher	Fiscal Year								Rankings
		07	08	09	10	11	12	13	14	
storage technologies	He									being done in the HEC area. Includes power consumption. Traditional block-based solutions ready for commercialization. Alternative interfaces not yet well explored.
	Rangaswami									
	Arpaci-Dusseau (2009 HECURA)									
	UCSD (Swanson/Gupta) - NVTM									
	NoLoSS - ANL/LLNL									
	Blackcomb – ORNL/ HP/ UM/ Penn State									
	PNNL									
HEC systems with multi-million way parallelism doing small I/O operations	Choudhary/Kandemir									 Good initial research; needs to be moved into testing. More fundamental solutions being pondered including non-volatile solid state storage.
	Dickens									
	Gao									
	Sun									
	Zhang/ Jiang									
	Sun									
	FASTOS – I/O Forwarding									
	PLFS - LANL/CMU									
	SCR/PLFS – LANL/LLNL									
Alternative I/O Transport Schemes	Sun									 Most aspects are being addressed.
	Panda (formerly Wycoff)									
	Lustre									
	pNFS									

- | | | |
|--|---|---|
|  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 Road Map

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.

2011 Communication and Protocols Gap Area

Area	Researchers	Fiscal Year								Rankings
		07	08	09	10	11	12	13	14	
Active Networks	Chandy									 Novel work being done, but not general enough.
	Maccabe/Schwan									
Coherence Schemes	UCSC's Ceph									 There's no consensus on how to do this correctly, but some solutions are in products.
	Lustre									
	Panasas									
	PVFS									
Topology aware storage layout	Panasas									
Wide area storage protocols	ORNL - xdd									

- 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 Road Map

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.

2011 Archive Gap Area											
Area	Researchers	Fiscal Year								Rankings	
		07	08	09	10	11	12	13	14		
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										
	UMN/CRIS – Multi-Dimensional File System										
	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
Cross Discipline (file system /archive/DB) Metadata Integration	Lustre HSM										 Extended Attributes, although not standardized, could solve problem.
	UMN Lustre Archive										
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

2011 Archive Gap Area

Area	Researchers	Fiscal Year								Rankings
		07	08	09	10	11	12	13	14	
										but we don't have any information on which to manage beyond what we know how to manage with

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

Management and RAS Road Map

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 device exploitation need to be explored.

2011 Management and RAS Gap Area

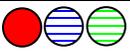
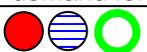
Area	Researchers	Fiscal Year								Rankings	
		07	08	09	10	11	12	13	14		
Proactive Health Methods	None										
Problem detection, reporting, analysis and modeling	Reddy										 More researchers need to look at this problem.
	Narasimhan										
Formal Failure analysis and tools for storage systems	Arpaci-Dusseau										 Good research done here. Will people use this work?
Improved Scalability	Ganger	This gap area was combined with “Scalable replication, relocation, failure detection, and fault tolerance” in Management and RAS. Thus, this gap sub area will be removed from the Road Map.								 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										
	Miller - UCSC										
Improved Scalability, scalable replication, relocation, failure detection, and fault tolerance	Ganger - CMU										 Industry is working on this problem More research is needed here. Test beds are probably needed for this work.
	Ligon - Clemson										
	CMU – Diskreduce										
	IBM – Perseus										
	GoofyFS – Sandia/UMinn/Clemson/UAB/ANL/ORNL										
	Ceph - UCSC										
	PVFS (Replication) - ANL										

- 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 Road Map

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.

2011 Security Gap Area										
Area	Researchers	Fiscal Year								Rankings
		07	08	09	10	11	12	13	14	
Performance overhead and distributed scaling	Sivasubramaniam	■	■	■	■					 Problem reasonably well understood, unclear if enough demand for product
	OrangeFS - Clemson				■	■				
End-to-end confidentiality and tracking of information flow, provenance, etc.	Odlyzko	■	■	■	■					 Industry will help some, but not in HEC context.
	McDaniel/Sion/Winslett			■	■	■				
	Miller/Seltzer				■	■				
	Horus - Rajendran/Miller/Long - UCSC				■					
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	<i>None</i>									 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

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

2011 Assisting with Standards, Research and Education

Area	FY07	FY 08	FY 09	FY 10	FY 11	FY12
Standards:						
POSIX HEC	PDSI UM CITI patch pushing/maintenance Revamp of manual pages	First Linux full patch set	Layout Query going into POSIX	HEC Extensions are finding their way into the kernel or experimental settings.		
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	Ratified and pNFS demonstrations by BlueArc at SC10		
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/Testbeds	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 LANL, CMU and NSF proved PROBE as a disruptive facility for CS systems research	
Simulation Tools	Ligon PDSI Felix/Farber	Ligon PDSI Felix/Farber	Ligon PDSI Felix/Farber	PFS Sim from U Florida and		

2011 Assisting with Standards, Research and Education

Area	FY07	FY 08	FY 09	FY 10	FY 11	FY12
			<p>Updated Disksim including MEMS simulation</p> <p>SNL releasing kernel I/O tracing tool</p>	<p>Florida International [Zhao09]</p> <p>Disksim added to SST from Sandia National Lab</p>		
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 station file system statistic data		<p>Update of LANL Machine and Failure Data, Archive and file system listing data</p> <p>ANL released Darshan data</p>	

Conclusion

Today, we are seeing sites deploying supercomputers with hundreds of thousands processors. 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, to tolerate extremely high metadata activities, and to efficiently manage 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 is approaching 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 1,000,000 spinning disks with widely varying workloads. In short, the challenges of the future are formidable.