

High End Computing File System and I/O R&D Gaps Roadmap

**James Nunez
Los Alamos National Lab
August 2008**

LA-UR-08-2876

HEC FSIO Accomplishments

- Accomplishments in 3 years:
 - 3 national workshops
 - \$15M+ in NSF HECURA and CPA I/O and File Systems research awards - 30 projects
 - \$25M SciDAC2 I/O and File Systems related research 5 year awards – 2 projects (SDM Center and PDSI)
 - Simulation resources – Incite and NSF infrastructure
 - \$1M DOD ACS I/O – 3 awards
 - Massive amount of failure, usage, event, and parallel trace data released
 - Progress on relevant standards – pNFS and POSIX HECEWG
 - Help Universities with storage, file system, and I/O programs – ISSDM

HEC FSIO 2008 Goals

- HECURA ending soon, we need to spin up a new set of research
 - Address gaps that are left or new gaps
 - Good portfolio of short/long evolutionary/revolutionary
 - Develop new HECURA call for FY08-09 NSF solicitation
 - CPA FY08
- We also need to update roadmaps for each gap area with
 - Current/future/needed research
 - Current/future needed productization
 - Make it easy for Agencies to know what gaps they care about and what investment makes sense for them

HEC FSIO Current Information

- Categories of needed research:
 - Metadata
 - Measurement and Understanding
 - Quality of Service
 - Security
 - Next generation I/O architectures
 - Communication (and) protocols
 - Management and RAS
 - Archive

2007 Measurement and Understanding Gap Area

Area	Researcher	FY 07	FY 08	FY 09	FY 10	FY 11	FY 12	Rankings
Understanding system workload in enterprise environment	Arpaci-Dusseau	■	■	■				 A comprehensive tool is nowhere in sight; problem is complex.
	Reddy	■	■	■				
	Zadok	■	■	■				
	SciDAC - PDSI	■	■	■				
	SciDAC - SDM	■	■	■				
Standards for HEC I/O benchmarks	<i>None</i>							 Low on agencies priorities; over simplifies problem and could drive vendors to incorrect solutions. Gap should really be replaced by release of traces, workload characterization, etc.
Testbeds for I/O Research	Ligon	■	■	■				 Simulators are being developed. No real testbeds being built. This problem will only get worse over time, i.e. as systems get bigger.
	Thottethodi	■	■	■				
Applying cutting edge visualization/analysis tools to large scale I/O traces	Reddy	■	■	■				 More traces are becoming available from Labs. Many opportunities to evaluate this research.
	Zadok	■	■	■				

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

2007 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.
	Leiserson	■	■	■				
	Maccabe/Schwann	■	■	■				
	SciDAC - PDSI	■	■	■	■	■	■	
	HECEWG HPC Extensions	■	■	■	■	■	■	
	UCSC's Ceph	■	■	■	■	■	■	
	Lustre	■	■	■	■	■	■	
	ANL/CMU – Large Directory PVFS	■	■	■	■	■	■	
Extensibility and Name Spaces	Bender/Farach-Colton	■	■	■				 All existing work is evolutionary.
	Leiserson	■	■	■				
	Tosun	■			■			
	Wyckoff	■	■	■				
	UCSC – LIFS/facets	■	■	■				
	ANL/CMU - MDFS	■	■	■				
	SciDAC PDSI	■	■	■	■	■	■	
File System/ Archive Metadata Integration	Lustre HSM	■	■	■	■			 Extended Attributes, although not standardized, could solve problem.
	UMN Lustre Archive	■	■					
Hybrid Devices Exploitation	<i>None</i>							 Research is being done, but no research focused on metadata
Data Transparency and Access Methods	<i>None</i>							 No research focused on metadata

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

2007 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.
	Chiueh	■	■	■				
	Ganger	■	■					
Standard API for QoS	SciDAC - PDSI	■	■	■	■	■		 Very partially addressed by proposed HEC POSIX Extensions. Will be driven by above "End to End QoS in HEC".
	POSIX HPC Extensions	■	■	■	■	■	■	
	PVFS	■	■	■	■	■	■	

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

2007 Next Generation I/O Architectures Gap Area

Area	Researcher	FY 07	FY 08	FY 09	FY 10	FY 11	FY 12	Rankings
Understanding file system abstractions - File system architectures	Choudhary	█	█	█				 Good work, but much of research is in infancy. A small portion ready for commercialization.
	Dickens	█	█	█				
	Maccabe/Schwan	█	█	█				
	Reddy	█	█	█				
	Shen	█	█	█				
	Thain	█	█	█				
	Wyckoff	█	█	█				
	SciDAC – PDSI	█	█	█				
PNNL	█	█	█	█				
Understanding file system abstractions - naming and organization	Bender/Farach-Colton	█	█	█				 Very hard problem. More researchers need to attack this problem.
	Thain	█	█	█				
	Tosun	█	█	█				
	Zhang/ Jiang	█	█	█				
	SciDAC – SDM	█	█	█				
	SciDAC - PDSI	█	█	█				
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	█	█	█				
Architectures using 10 ⁶ storage components	Ligon	█	█	█				 Very little work being done here for a very near term problem. Simulators will/must play a role here
	PNNL	█	█	█	█			
Hybrid architectures leveraging emerging storage technologies	Gao	█	█	█				 Big potential reward, but very little work being done in the HPC area.
	PNNL	█	█	█	█			
HEC systems with multi-million way parallelism doing small I/O operations	Choudhary	█	█	█				 Good initial research; needs to be moved into testing. More fundamental solutions being pondered including non-volatile solid state store.
	Dickens	█	█	█				
	Gao	█	█	█				
	FASTOS – I/O Forwarding	█	█	█				

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

2007 Management and RAS Gap Area

Area	Researchers	FY 07	FY 08	FY 09	FY 10	FY 11	FY 12	Rankings
Automated problem analysis and modeling	Reddy							 More researchers need to look at this problem.
Formal Failure analysis for storage systems	Arpaci-Dusseau							 Good research done here. Will people use this work?
Improved Scalability	Ganger							 More research is needed here. Testbed is 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.
Reliability	<i>None</i>							 Industry is working on this problem

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

Unclassified

2007 Security Gap Area

Area	Researchers	CY 06	CY 07	CY 08	CY 09	CY 10	CY 11	Rankings
Long term key management	Odlyzko							 Current researcher need data to validate designs
End-to-end encryption	Odlyzko							 Current researcher need data to validate designs
Performance overhead and distributed scaling	Sivasubramaniam							 Problem reasonably well understood, unclear if enough demand for product
Tracking of information flow, provenance, etc.	<i>None</i>							 Industry will help some, but not in HPC context. Nothing to commercialize yet.
Ease of use, ease of management, quick recovery, ease of use API's	Sivasubramaniam							 Current researchers need data to validate designs Nothing to commercialize yet.

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

2007 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 Tosun							 Current research is in terms of file systems, not archive. API merging with POSIX and API for searching lacking
	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 commercialization is low because of other drivers, i.e. HIPPA and others will drive this. Redundancy techniques reasonably sufficient for archives
	Narasimhan							
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							
	Leiserson							
	Ganger							
	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 |  Needs Commercialization |
|  Low Importance |  Does Not Need Research |  Does Not Need Commercialization |
|  Full Calendar Year Funding |  Partial Calendar Year Funding |  On-Going Work |

2007 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.
Alternative I/O transport schemes	Sun							   Most aspects are being addressed.
	Wyckoff							
	Lustre pNFS							
Coherent Schemes	ANL/CMU							   No consensus on how to do this correctly, but some solutions are in products.
	UCSC's Ceph							
	Lustre							
	Panasas PVFS							

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

2007 Assisting with Standards, Research and Education

Area	FY07	FY 08	FY 09	FY 10	FY 11
Standards:					
POSIX HEC	PDSI U <u>Mich</u> CITI patch pushing/ <u>maint</u> Revamp of man pages	First Linux full patch set			
ANSI OBSD	V2 nearing pub	Some file system pilot test			
IETF <u>pNFS</u>	V 4.1 nearing pub Assistance in testing may be needed	Initial products			
Community Building	<i>HEC FSIO 2007 HEC presence at FAST and IEEE MSST</i>	<i>HEC FSIO 2008 HEC presence at FAST and IEEE MSST</i>	<i>HEC FSIO 2009 HEC presence at FAST and IEEE MSST</i>	<i>HEC FSIO 2010 HEC presence at FAST and IEEE MSST</i>	<i>HEC FSIO 2011 HEC presence at FAST and IEEE MSST</i>
Equipment	<i>Incite and NSF Infra Need scale CS disruptive facility</i>	<i>Incite and NSF Infra Need scale CS disruptive facility</i>	<i>Incite and NSF Infra Need scale CS disruptive facility</i>	<i>Incite and NSF Infra Need scale CS disruptive facility</i>	<i>Incite and NSF Infra Need scale CS disruptive facility</i>
Simulation Tools	<u>Ligon</u> <i>PDSI Felix/Farber</i>	<u>Ligon</u> <i>PDSI Felix/Farber</i>	<u>Ligon</u> <i>PDSI Felix/Farber</i>		
Education	<i>LANL Institutes as one example PDSI</i>	<i>Other Institute like <u>activites</u></i>			
Research Data	<i>Failure, usage, event data</i>	<i>Many more traces, FSSTATS, more disk failure data</i>			

Unclassified

Resources

- HEC FSIO planning site
 - <http://institute.lanl.gov/hec-fsio/>
- Send comments to:
jnunez@lanl.gov