High End Computing Interagency Working Group (HECIWG) Sponsored File Systems and I/O Workshop HEC FSIO 2011

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

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" [Appendix C] 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 advice 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.

With the information from the HECIWG I/O Document [Appendix C] and the HEC FSIO 2005 workshop, a number of activities occurred during 2006:

- In the area of R&D
    - a National Science Foundation (NSF) HEC University Research Activity (HECURA) solicitation for university research in the FSIO area was written based on the eight areas of research identified during the HEC FSIO 2005 workshop
    - the NSF/HECURA solicitation was conducted resulting in 62 proposals from over 80 Universities

    - from a careful analysis of the proposals, 23 HECURA awards were made
    - the DOE Office of Science awarded two SciDAC2 FSIO projects
- In the area of providing computational resources
    - the DOE Office of Science INCITE program for supplying computing clusters
    - NSF infrastructure program for providing computing infrastructure
- In the area of providing operational data to enable research
    - LANL release of failure, event, and usage data
    - Other sites and industry including the Library of Congress, and HP working on data release
    - Consortia for failure data release is forming up.

Additionally, due to the success of the HEC FSIO 2005 workshop and subsequent activities, a permanent HEC FSIO advisory group was formed to help continue to advise the HEC in how best to coordination FSIO activity.

The HEC FSIO advisory group held the HEC FSIO 2006 workshop on August 20-22 in Washington DC. The location was picked to encourage more HEC agency participation, and indeed three more HEC agencies were represented. The workshop was again, attended by top university, government HEC, and industry FSIO R&D professionals. The goals for this workshop were:

- To update everyone on the
    - 23 HECURA and two SciDAC2 FSIO research activities
    - programs available to get computing resources
    - activities to make HEC site operational data available to enable research.
- To solicit input on
    - remaining gaps in the needed research areas
    - gaps in providing center support such as providing
        - computational resources
        - operational data available for research
        - getting the HEC community involved in the educational process.

The information gathered at this and the previous workshop was used to advise the HEC on how to facilitate better coordinated government funded R&D in this important area in the coming years.

The HEC FSIO advisory group held the HEC FSIO 2007 workshop on August 5-8 in Arlington, VA at the headquarters of the National Science Foundation.

From the 70 original attendees of the 2005 workshop, the attendance grew to 100 in 2006, and stayed at about 100 in 2007. The workshop was well received and accomplished its goals; to showcase the 23 HECURA projects, to continue to foster the development of the HEC FSIO community, to provide a venue for information sharing, to update everyone on related standards, data releases, and other support activities; and to revisit the gap areas. The presentations from the workshop are located at the HEC FSIO Workshop web site http://institute.lanl.gov/hec-fsio .

New items for 2007 included

- In the area of R&D
    - Five I/O related projects funded from the NSF CPA 2007 program which are being coordinated with the 23 HECURA projects. The abstracts from these projects appear in Appendix A of this document.
- In the area of providing computational resources
    - the NSF infrastructure program has provided some computing infrastructure to FSIO projects
- In the area of providing operational data to enable research
    - USENIX provided a web page that indexes a large number of research data release sites for failure, usage, event, placement, and trace data
    - Many sites and industry started to release research data, more are coming
- In the area of assisting education
    - LANL formed two FSIO related institutes to assist with collaboration and HEC site involvement at universities

The HEC FSIO advisory group held the HEC FSIO 2008 workshop on August 4-6 in Arlington,VA at the Westin Hotel, 1 block from the NSF headquarters.

The attendance was at an all time high of 105. The workshop was well received and accomplished its goals: to remind the attendees about current research activities, allow for community building, and to introduce and solicit input on the HECFSIO Roadmaps which represent the HECFSIO R&D portfolio of needs and research addressing those needs. The presentations from the workshop are located at the HEC FSIO Workshop web site http://institute.lanl.gov/hec-fsio .

New and continuing items for 2008 include

- In the area of R&D
    - Two new NSF CPA 2008 I/O related projects that will be coordinated with all the other HECFSIO projects, details of which are in Appendix A of this document.
    - Five 2007 I/O related projects funded from the NSF CPA program that continue to be coordinated with the 23 HECURA projects. The abstracts from these projects appear in Appendix A of this document.
    - the 23 HECURA projects have produced exciting results presented at mid year 2008 status meetings with the HEC FSIO team and at the 2008 workshop

    - the DOE Office of Science SciDAC FSIO projects are also making progress
    - the DOE Office of Science/NNSA FASTOS I/O forwarding scalable layer project was introduced
- In the area of providing computational resources
    - the DOE Office of Science INCITE program has supplied access to computing clusters for FSIO researchers
    - the NSF infrastructure program has provided some computing infrastructure to FSIO projects
- In the area of providing operational data to enable research
    - The DOE SC SciDAC2 Petascale Data Storage Institute (PDSI) began providing trace data, file systems stats, and other data to researchers
    - The USENIX web page that indexes a large number of research data release sites for failure, usage, event, placement, and trace data
    - Many sites and industry started to release research data, more are coming
- In the area of assisting education
    - LANL has two FSIO related institutes to assist with collaboration and HEC site involvement at universities
    - Many joint university/HEC site FSIO related projects are springing up

2008 was a pivotal year for the HEC FSIO area. Many of the HECURA research projects are nearing their end in 2008-2009. Two important things happened at the workshop.

- the HECFSIO Roadmaps were introduced which is the method the HECFSIO advisory group will use to manage the portfolio of problems/issues and R&D in progress to address these issues

and

- NSF announced the intent to solicit (using the HECFSIO Roadmaps as input) R&D for a HECURA 2009 call to continue the momentum from the HECURA 2006 efforts.

Attendees were overall pleased with the workshop, although they were not as euphoric as they were at previous workshops. There was not as much new research and other support announced in 2008. Attendees were quite happy to hear that the NSF will solicit a follow on to the HECURA 2006 call. Additionally, attendees were vocal about needed adjustments to the HECFSIO Roadmaps which was one of the main goals for the 2008 workshop.

New items for 2009 included

- In the area of R&D
    - a National Science Foundation (NSF) HEC University Research Activity (HECURA) solicitation for university research in the FSIO area was once again written based on the eight areas of research identified during the HEC FSIO 2005 workshop

- the NSF/HECURA 2009 solicitation was conducted resulting in 109 proposals from over 54 Universities and National Labs
- from a careful analysis of the proposals, $17 M was awarded in over 20 HECURA awards

The focus of this year's conference was, once again, to get input from the FSIO community on what gap areas remain with an emphasis on planning for exascale and beyond and to update our road maps to reflect these gaps. Although successful in updating the road maps, the discussion was limited on if we are prepared for exascale computing and, if not, what we need to do to prepare for it. This will be a focus for the 2010 meeting.

The HEC FSIO advisory group held the HEC FSIO 2010 workshop on August 2-4 in Arlington,VA at the Westin Hotel, 1 block from the NSF headquarters.

The attendance was at an all time high of 114. The workshop was well received and accomplished its goals: to remind the attendees about current research activities, allow for community building, and to introduce and solicit input on the HECFSIO R&D portfolio of needs and research addressing those needs.

New items for 2010 include:

- The DOE Advanced Scientific Computing Research (ASCR) Program awarded 27 FSIO and data analytics projects for exascale computing under the Advanced Architectures and Critical technologies for Exascale Computing, Scientifice Data Management and Analysis at Extreme Scale, and X-Stack Software Research programs

The focus of this year's conference was to inform and get input from the community on the expected problems that exascale compute systems will pose for storage and file systems. Through the invited talks and panels, the participants were presented with a variety of views on what the FSIO world will look like and what existing technologies will have to be left behind to scale for tomorrow's exascale machines.

In addition, it is clear that data intensive (super) computing is and will be influencing how HEC FSIO is done in the future. Two of the four breakout sessions were to understand what tools and best practices we can adopt from existing data intensive computing tools and the pros and cons of their reliability and redundancy schemes.

Analysis of the 2010 workshop identified gaps by the HEC is summarized in the revised roadmaps below.

The HEC FSIO advisory group held the HEC FSIO 2011 workshop on August 8-10 in Arlington, VA at the Westin Hotel, 1 block from the NSF headquarters.

The attendance remained the same over last year at 114 attendees. The format of this year's workshop was based on a small number of talks and panels to allow for updates and viewpoints from more researchers. The format change decreased the time for discussion of research and gap areas.

New items for 2011 include:

- The NSF funded Parallel Reconfigurable Observational Environment (PRObE) for data intensive super-computing and high end computing project was announced; an initiative to offer large scale compute clusters to systems researchers for low-level systems research and, possibly, destructive testing.

Roadmaps 2011

Metadata

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

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	■	■	■	▒					● ⊜ ○ A comprehensive tool is nowhere in sight; problem is complex. This gap area includes monitoring.
	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		🟨	🟨	🟨					⊜ ○ ○ Danger of over simplifying problem and could drive vendors to incorrect solutions.
	High Productivity Computing Systems (HPCS) Benchmarks				■	■				
	Ma/Shen/Winslett			▒	■	■	▒			
Modeling, simulation and test environments.	Clemson - Ligon	■	■	■	▒					● ⊜ ○ Simulators are being developed. PROBE's testbeds for use are retired clusters. No real testbeds being built. This problem will only get worse over time, i.e. as systems get bigger.
	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	■	■	■	▒					● ● ○ Data are becoming available from Labs including I/O traces. Many opportunities to evaluate this research. This includes applying analysis and visualization tools to I/O traces
	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

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

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 /Clemsom/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 storage technologies	Gao	▒	■	■	▒					🔴 ⊜ ⊜ Big potential reward, but very little work being done in the HEC area. Includes power consumption. Traditional block-based solutions ready for commercialization. Alternative interfaces not yet well explored.
	Urgaonkar		▒	■	■	▒				
	Szalay/ Huang			▒	■	■	▒			
	He			▒	■	■	▒			
	Rangaswami				▒	▒				
	Arpaci-Dusseau (2009 HECURA)			▒	▒					
	UCSD (Swanson/Gupta) - NVTM				■	■				
	NoLoSS - ANL/LLNL				■	■	■			
	Blackcomb – ORNL/ HP/ UM/ Penn State				■	■	■			
	PNNL	🟨	🟨	🟨	🟨					

2011 Next Generation I/O Architectures Gap Area

Area	Researcher	Fiscal Year								Rankings
		07	08	09	10	11	12	13	14	
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

2011 Communication and Protocols Gap Area

Area	**Researchers**	**Fiscal Year**								**Rankings**
		07	**08**	**09**	**10**	**11**	**12**	**13**	**14**	
Active Networks	Chandy	■	■	■	▒					⊜ ⊜ ○
	Maccabe/Schwan	■	■	■	▒					Novel work being done, but not general enough.
Coherence Schemes	UCSC's Ceph	🟨	🟨	🟨	🟨					⊜ ⊜ ⊜
	Lustre	🟨	🟨	🟨	🟨					There's no consensus on how to do this correctly, but some solutions are in products.
	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

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 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

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									⊜ ⊜ ○ 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				🟨	🟨				

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.

● 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

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	*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

Past years are status, future years are identified needs or desires

2011 Assisting with Standards, Research and Education

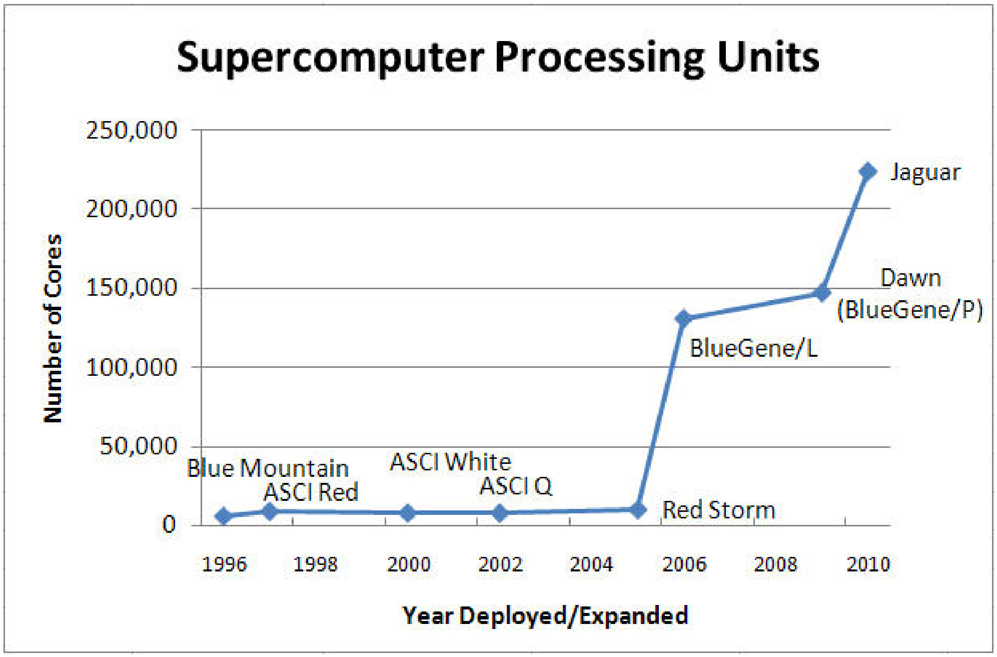
Area	FY07	FY 08	FY 09	FY 10	FY 11	FY12
Standards:						
POSIX HEC	PDSI UM CITI patch pushing/ maintenance Revamp of manual pages	First Linux full patch set	Layout Query going into POSIX	HEC Extensions are finding their way into the kernel or experimental settings.		
ANSI OBSD	V2 nearing publication	Some 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 Updated Disksim including MEMS simulation SNL releasing kernel I/O tracing tool	PFS Sim from U Florida and Florida International [Zhao09] Disksim added to SST from Sandia National Lab		
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		Update of LANL Machine and Failure Data, Archive and file system listing data	

2011 Assisting with Standards, Research and Education

Area	FY07	FY 08	FY 09	FY 10	FY 11	FY12
					ANL released Darshan data	

Compelling Case Information and Background

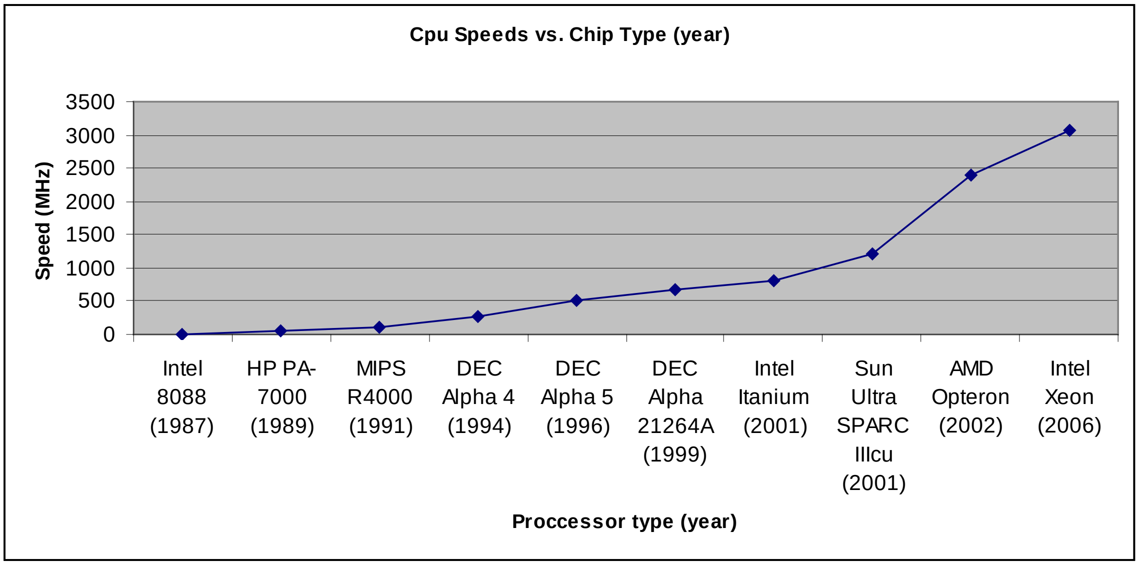
Historical and projected future trends can be used to explain why I/O and file system research is needed and which particular areas of research need to be pursued.



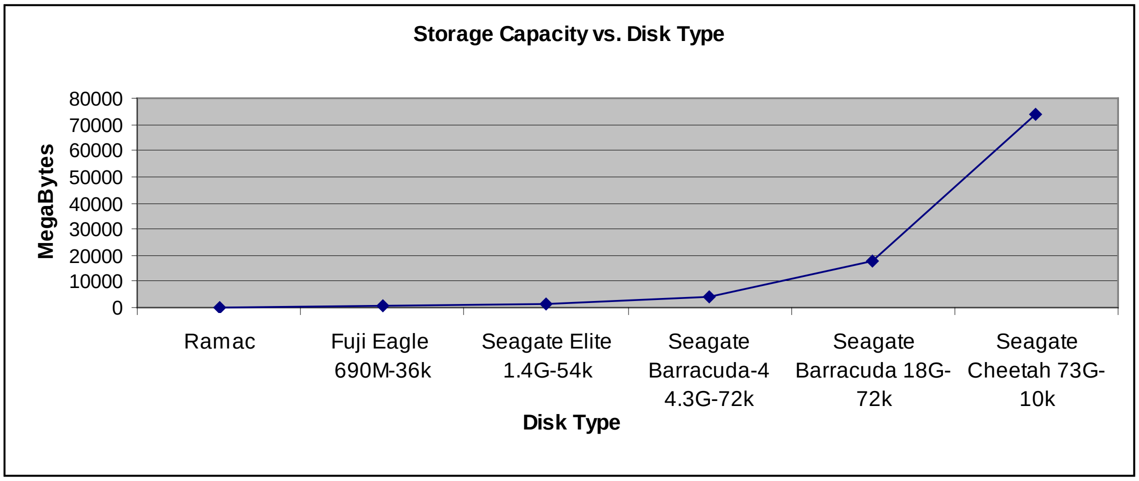
Although processor clock speeds have grown drastically over the last two decades, the rapid increases in processor clock rates have ended in favor of many processing units per chip and for hybrid processing unit architectures. The microprocessor industry is deploying processor architectures that have many more processing units per chip or board to continue to meet the processing power growth demand. This implies that scientific applications have to rely more heavily on multi-process/task parallelism at a greater scale than ever before. When single processors were getting much faster each year, applications could gain advantage over time by keeping a constant number of processes per task but this is no longer true. In order for applications to continue to gain speed up, it now requires the use of more processing elements over time. The compute capabilities of machines anticipated for scientific computing is growing rapidly. The following graph illustrates the corresponding growth in the number of processing units we have used to build these large scientific computers.

Additionally, it is important to note that the amount of memory per processor has gone down over the last few years from 2-4 GB/processor on most ASC machines to .25 GB/processor on BG/L. Memory per Teraflop (TF) has gone down from around 1 Terabyte (TB) per TF to about .1 TB per TF.

In the past 50 years, the performance of individual processors has gone up by 4-5 orders of magnitude.

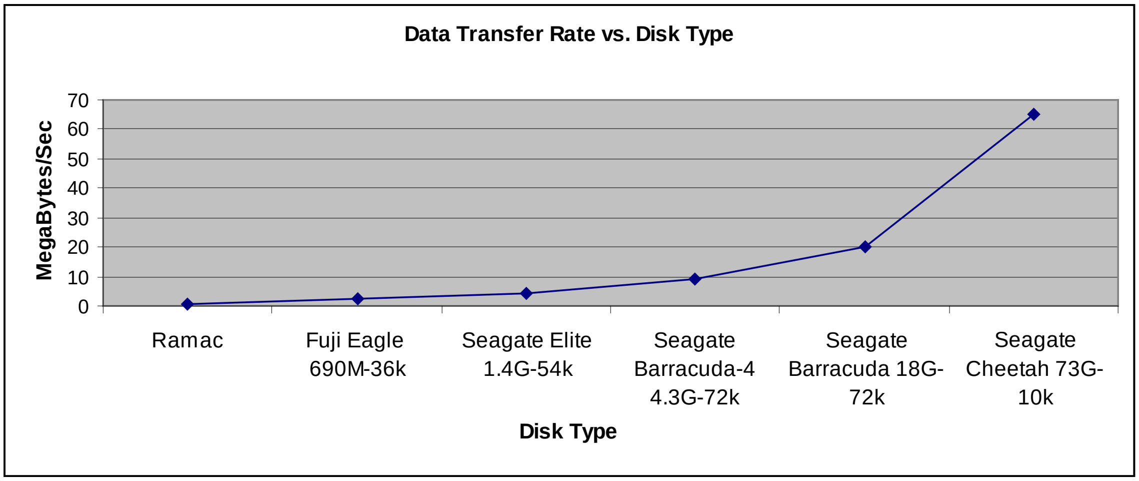


Disk drives have become far more dense over the past five decades. The following graph shows the relative density of the original disk drive, the IBM RAMAC 1956 versus a recent state of the art Seagate 15k RPM 2.5 inch 73 gigabyte disk drive.

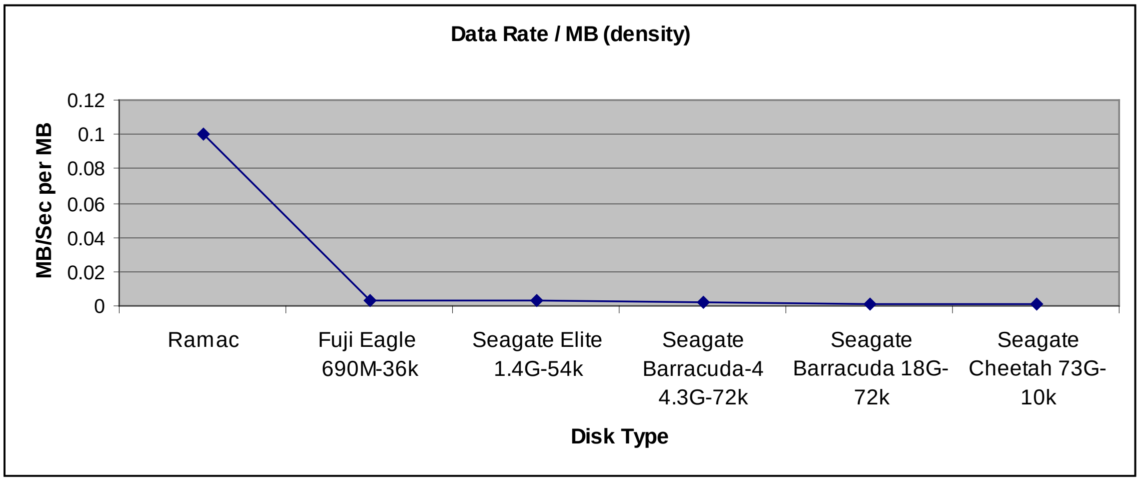


Disk capacity has grown at an amazing rate: five orders of magnitude in 50 years; approximately the same change as in CPU speeds.

In the following graph, the disk drive data transfer rate speed up over the past 50 years is shown.

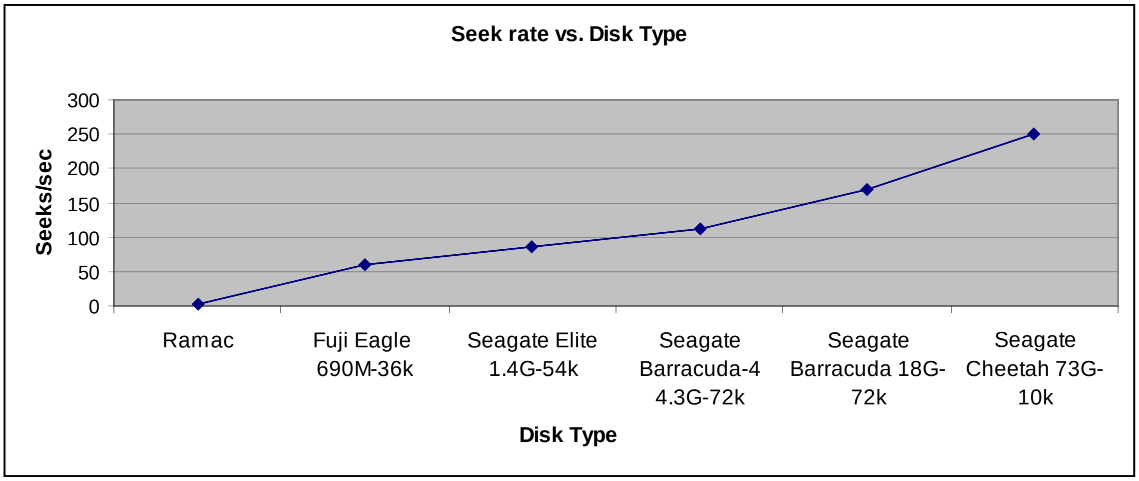


Data transfer rates have increased only about two orders of magnitude. This is an important observation. Another way to look at this phenomenon is in the graph below which shows disk transfer rates normalized to byte of storage capacity (which would be roughly equivalent to normalizing to processing power as well given the similarities between the growth of disk capacity and CPU speed).

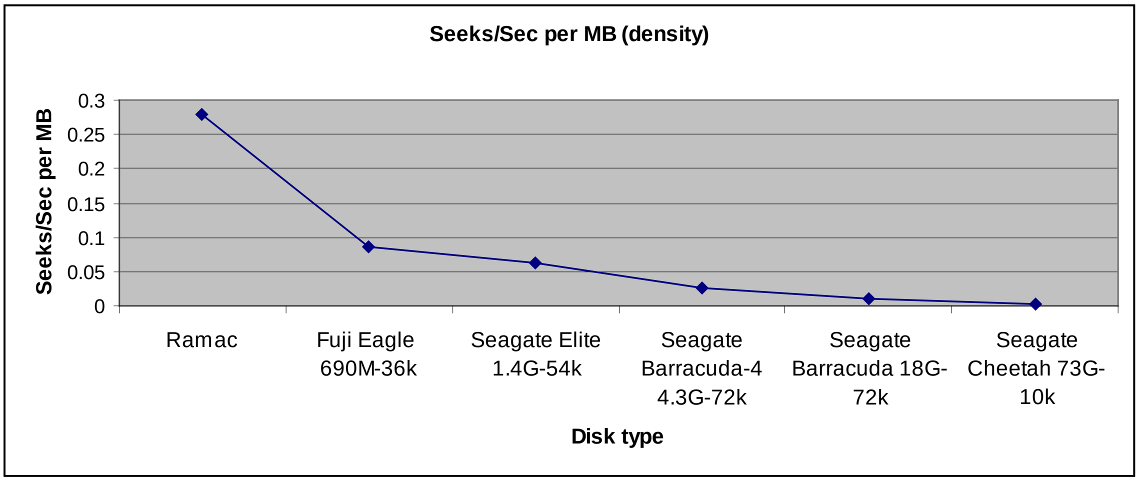


As you can see, the data rate performance of disk drives has gone down by about two orders of magnitude**. Just as in the case of transfer rate to density, this means that it takes two orders of magnitude more disk drives per CPU to do the same relative workload in a balanced way than 50 years ago!**

Disk agility has also gotten much better over the last 50 years. The following graph shows the seek performance increase achieved.



Again, the seek performance has gone up by about two orders of magnitude. This is also an important observation. Another way to look at this phenomenon is in the graph below which shows disk seek rates normalized to byte of storage capacity (which would be roughly equivalent to normalizing to processing power as well given the similarities between the growth of disk capacity and CPU speed).



The seek rate per drive density performance of disk drives has gone down by about two orders of magnitude. **This means that it takes two orders of magnitude more disk drives per CPU to do the same relative workload in a balanced way than 50 years ago!**

These changes in component characteristics have some interesting effects on the I/O and File Systems services:

- The number of disk drives needed to build a balanced system over time has increased drastically. The largest sites today currently have in the 20,000 disk drive range. It’s likely we will see sites with well over 100,000 disk drives soon and perhaps over a million when machines reach tens of petaflops.
- Since the number of processing elements is going up rapidly and the amount of memory per processing element is doing down, the I/O system must orchestrate

collecting memory from far more memories to be sent to far more disk drives, due to the slow growth in disk bandwidth.

- As the numbers of processing elements goes up, the required file system metadata operations per second goes up, which again implies orchestrating metadata requests from more clients to more disks than ever due to the slow growth in disk agility.

Two major questions sum up most of the I/O and File System's challenge:

1) How do you manage 1,000,000 mechanical disk drive devices and their associated environment, both hardware and software? This includes, RAS, QoS for usage that varies by seven orders of magnitude, management without requiring an army of administrators, security, etc.

2) How do you productively use 1,000,000 mechanical disk drive devices and their associated environment? This includes middleware, high level libraries, file systems, dealing with massive in-flight data, etc.

It is no surprise that the above trend information describes well most of the existing and new I/O and file systems issues.

Despite these troubling trends in processor and storage, an examination of the quantity of computer science research in I/O and file systems compared to other areas of the HEC environment reveals that the I/O and file systems area has been greatly neglected. This neglect helps explain why investments in research in this area are needed so acutely and why the HECURA, CPA, and SciDAC2 I/O projects were so well received by the HEC FSIO community. The area of I/O has really been overlooked in both in how to manage enormous scale I/O systems and how to productively use such systems.

The Eight Areas of Needed R&D

The HEC FSIO 2005 workshop recommended investment in both evolutionary and revolutionary research in eight areas. Although these areas are revisited each year, the original eight categories remain. Many of the areas are cross cutting.

- Metadata - Metadata operations will involve the orchestration of more clients to more storage devices. 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.
- Measurement and understanding - Measuring and understanding performance becomes more difficult as more and more storage devices are involved. Research into measurement and understanding of end-to-end I/O performance is needed 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 to deal with the scale of future deployed systems.

- Quality of Service - Mixed workloads in an environment comprised of hundreds of thousands of processors will make providing determinism increasingly difficult. 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 equally important.
- Security - Even without scale, security is a difficult unsolved problem; complicated in a HEC environment by need to know issues. Aspects of security such as usability, long term key management, distributed authentication, and dealing with security overhead are all good 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.
- Next-generation I/O architectures - 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.
- Communications and protocols - 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, topology-aware, and wide area protocols are needed.
- Archive – 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 previously.
- Management and RAS - Management and RAS in an environment with 100,000 or more disks and million way parallelism will be extremely difficult. Adding to the difficulty of management is the added difficulty of having to deal with long term persistent data. Failure of processors typically means a job is re-run. Failure of a storage device may mean loss of valuable data or information. 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 autonomics, use of virtual machines, and novel devices exploitation need to be explored.

In addition to the eight research areas, more focused and complete government research investment needs to be made in the file systems and I/O middleware area of HEC, given its importance and its lack of sufficient funding levels in the past, as compared to other elements of HEC.

Scalable I/O is perhaps the most overlooked area of HEC R&D, and given the information generating and processing capabilities being installed and contemplated. It is a mistake to continue to neglect this area of HEC. One of the primary purposes of this document is to present the areas in need of new and continued investment in R&D and standardization in this crucial area of HPC file systems and scalable I/O that should be pursued by the government.

Frequently Used Terms

ASC – The Advanced Simulation and Computing program supports the Department of Energy’s National Nuclear Security Administration simulation based stockpile stewardship.

File system – A combination of hardware and software that provides applications access to persistent storage through an application programming interface (API), normally the Portable Operating System Interface (POSIX) for I/O. The file system provides an abstraction from the hardware make up.

Global – refers to accessible globally (by all), often implies all who access a given resource see the same view of the resource

HECRTF – High End Computing Revitalization Task Force, an effort to make the US more competitive in high end computing

HEC – High End Computing Inter-agency Working Group, an inter US government agency working group to help coordinate government funding of R&D activities in the HEC area

HECURA – High End Computing University Research Activity – A government funded university research activity in the HEC area

FSIO – File Systems and I/O

Higher level I/O library – software libraries that provide applications with high level abstractions of storage systems, higher level abstractions than parallelism, examples are the Hierarchical Data Formats version 5 library (HDF5) (http://www-rcd.cc.purdue.edu/~aai/HDF5/html/RM_H5Front.html) and parallel Network Common Data Formats library (PnetCDF) (http://www-unix.mcs.anl.gov/parallel-netcdf/sc03_present.pdf)

I/O – input/output

I/O Middleware – software that provide applications with higher level abstractions than simple strings of bytes, an example is the Message Passing Interface – I/O library (MPI-IO) (http://www-unix.mcs.anl.gov/romio)

Metadata – information that describes stored data; examples are location, creation/access dates/times, sizes, security information, etc.

Parallel – multiple coordinated instances, such as streams of data, multiple computational elements, etc.

POSIX - Portable Operating System Interface (POSIX), the standard user interfaces in the UNIX based and other operating systems (http://www.pasc.org/plato/)

QoS – Quality of Service

SAN – Storage Area Network, network for connecting computers to storage devices

Scalable – decomposition of a set of work into an arbitrary number of elements, the ability to subdivide work into any number of parts from 1 to infinity

WAN – Wide Area Network, refers to connection over a great distance, tens to thousands of miles

Research Themes Identified From the Workshops

Throughout the HEC FSIO workshops, a number of research themes emerged. 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. The following subsections describe the research appropriate to each theme area. Each subsection will cover both evolutionary and more revolutionary research topics that were identified as needing additional research attention. Input from the 2005, 2006, 2007, 2008, 2009 and 2010 workshops are provided for a complete picture of the area. Additionally, the list of HECURA, CPA, and SciDAC2 I/O awards are described in Appendix A of this document.

Metadata

HEC FSIO 2005 Problem Definition

The 2005 workshop identified a number of research needs for metadata management. The 2006 HECURA call funded four strong proposals in response. These proposals addressed direct scalability, augmented metadata for IO scalability, applicability and benchmarking of alternative approaches, roles, semantic awareness, extensibility, and fundamental algorithms of implementation.

Scalability

Evolution

Problem Definition

Clustered file systems seem to be converging on an architecture that employs a centralized metadata service to maintain layout and allocation information among multiple, distinct movers. While this has had significant, positive impact on the scalability in the data path, it has been at the expense of the scalability of the metadata service. The transaction rate against the metadata service has increased, as has the amount of information communicated between the metadata service component and its clients has increased. These trends indicate that distributed metadata storage is key to successful scalability.

2005-2006 Progress

The 2006 HECURA call produced a strong response in this area. As was noted, at the 2005 workshop, performance is critical for any petaFLOPS attempt and scalable performance is paramount. All four, funded, proposals directly addressed this issue in various ways: peering and distribution of metadata information, on-store layouts, applicability of object-based storage to the problem, enabling scalability performance in consistency and coherency guarantees, and fundamental algorithms for efficient implementation of classic name spaces.

Revolution

Problem Definition

In the near-term we are likely to see relatively simple metadata distribution schemes used to allow for more concurrency in metadata operations for cluster file systems with a moderate number of metadata storage devices. In the longer term, it will be necessary to extend these schemes in order to facilitate the use of very large numbers of metadata storage devices. Techniques for discovery of file objects and mapping of tree-based name spaces onto in a very large metadata space are just two potential research areas for long-term study of metadata scalability. Continued scaling implies an increase in in-flight data and metadata and adapting to this change is an area of great interest for scaling research as well.

2005-2006 Progress

This topic was partially addressed in the responses to the 2006 HECURA call; however, not in any real, direct, way. Autonomics will provide for the management of a large number of storage devices but no direct examination of those devices with respect to file system or user metadata was proposed. Similarly, efficiency improvements for generic metadata was proposed in the fundamental algorithms response but scalability was not discussed. Many of the proposals discussed scalability but none offered novel architectural approaches. It is probably true that the existing approaches haven't been sufficiently, critically, examined. New approaches, offering other options would be welcome also.

Extensibility

Evolution

Problem Definition

HEC applications are increasingly creating, storing, and relying on derived data and provenance information as part of the discovery process. At the moment, additional databases or files are used to store this information. At the same time, extended attribute support is becoming a common feature of local file systems. Additional work is needed to understand appropriate storage mechanisms for these user-created attributes within cluster and parallel file systems and to create interfaces to this data.

2005-2006 Progress

One response in the FY'06 call addressed part of this extensibility issue. It proposes to examine whether and how object-based devices are to be leveraged in this area. Another call proposing work in fundamental implementation will also be applicable, although the researchers did not make specific note of this.

Revolution

Problem Definition

As extendible metadata and data transparency become more common, it is likely that the file system will know increasingly more about the data being stored. With respect to metadata, an important issue will be how to store semantic information alongside file data in a manner that is accessible and understandable by the file system itself.

2005-2006 Progress

One funded proposal in the FY'06 HECURA call did mention semantic indexing. However, only as to how content might be efficiently indexed and searched. Application at scale, in a large distributed machine, or environ was not contemplated. The proposal was more fundamental, an algorithm. Perhaps a follow-on proposal will be able to hybridize two or more approaches to extend the work into a distributed environment.

Metadata and Archiving

Evolution

Problem Definition

Archiving of cluster file systems is problematic for a number of reasons. One key factor is the vast number of individual items (e.g. files) that must be archived. Efficiently storing the metadata for these objects in a manner that is also efficiently accessed on streaming storage is an unsolved challenge.

2005-2006 Progress

There were proposals in the FY'06 call addressing the wedding of file systems with archival storage. It is difficult to predict the suitability and fortunes of the described approaches as this problem is, first, not new and, second, has been worked on many times previously without significant effect. Work always looks promising and good results are produced but these demonstrated benefits always seem to have issues when the work is contemplated for incorporation into existing and contemplated product from industry.

Revolution

Problem Definition

While simply archiving large volumes of data stored on cluster file systems is a challenge in itself, tighter coupling of storage system and archival system is desirable. One challenge in bringing file systems and archival systems together is the need for additional metadata describing locations of data, which might in fact not even exist on the file system at that point in time. Additional challenges to archiving of file system data come when name space changes occur. Capturing novel file system organizations on archival storage will require new approaches.

2005-2006 Progress

No response produced the revolutionary approach discussed here. No new name space organization and, therefore, no issue introduced with respect to on-line versus archival name space integration appeared.

Access Control Lists

Evolution

Problem Definition

Access Control Lists (ACLs) are a widely-implemented mechanism for limiting access to file system objects. The challenge in applying ACLs to cluster file systems is in their distributed nature. As we move away from centralized metadata storage to distributed metadata storage, efficiently verifying permissions will become more complicated and communication-intensive. Novel approaches to distributing ACLs and maintaining consistency of ACLs in a distributed environment are necessary to prevent these checks from becoming an artificial I/O bottleneck.

This item is closely related to security, of course. Here, however, we are talking about the issue in terms of performance and scalability. No proposal targeted at metadata management directly responded. While a couple of the funded proposals for security discuss researching the topic, none do so in terms of performance. The issue may be unavoidable, so progress in all of the metadata and the security responses should be tracked with this in mind.

2005-2006 Progress

One response did attempt to address the usability of security which begins to get at the heart of this issue. Also the POSIX HECEWG, an attempt to enhance the POSIX I/O API by the HEC community, effort is starting to address this.

Revolution

2005-2006 Progress

No response produced a revolutionary approach to this issue.

Data Transparency

Evolution

Problem Definition

Traditionally, file systems have operated in terms of streams of bytes. However, today’s file systems are accessed by numerous, often heterogeneous systems. In order to store data in a platform-independent manner, high-level libraries are used to convert data prior to storage on the file system. Augmenting file systems to understand basic data format semantics would allow these types of operations to be moved into the file system proper allowing the file system to make decisions on the best format for physical storage and likely reducing the overhead of data recoding.

2005-2006 Progress

No response in the list of funded activities for the 2006 HECURA call addressed this issue. Further, without researchers directly addressing it, it seems unlikely that any progress will be forthcoming. The topic itself is difficult to grasp and, given the lack of any response, it is probably best if the research community is educated and a specific request for research in the area is made, perhaps, in a later call.

Name Spaces

Revolution

Problem Definition

In order to assist applications with managing enormous amounts of data, application programmers and data management specialists are calling for the ability to store and retrieve data in organizations other than the age-old file system tree-based directory structure. Data format libraries currently provide some of this function but are not at all well-mated to the underlying file system capabilities. Databases are often called upon to provide these capabilities but they are not designed for petabyte or exabyte scale stores with immense numbers of clients. Exploratory work in providing new metadata layouts and finding data in these new layouts is vital to address this identified need.

2005-2006 Progress

One response in the 2006 HECURA call briefly mentioned indexing and incorporation of semantic awareness. How that might be exposed or used is unclear. No response offered to examine the utility of such a mechanism. No response contemplated examination in the high-end computing arena with its scaling needs.

The revolutionary approach that the 2005 workshop thought was needed has not, so far, been proposed.

Hybrid Devices

Revolution

Problem Definition

New storage technologies such as MEMS, MRAM, FLASH, and others all provide storage that is faster than spinning disk but at a higher cost. While it will be some time before such technologies supplant disk (if ever), these technologies are very amenable to use as metadata storage or metadata cache spaces. Integrating these devices into file system infrastructure holds the promise of increased metadata rates.

2005-2006 Progress

The HECURA 2006 call produced no responses in this area. Perhaps the right input is not being solicited. Attendees at the 2005 and 2006 workshops leaned heavily toward file system research. While the storage industry was represented, no presentation or discussion mentioned any examination of the utility of hybrid devices.

Overall

There were a number of funded HECURA projects that will attempt to address metadata issues.

- Collaborative Research: Petascale I/O for High End Computing [Maccabe]
- Collaborative Research: Techniques for Streaming File Systems and Databases [Bender06]
- Applicability of Object-Based Storage Devices in Parallel File Systems [Wyckoff]
- Collaborative Research: SAM^2 Toolkit: Scalable and Adaptive Metadata Management for High-End Computing [Jiang 2006]
- Improving Scalability in Parallel File Systems for High End Computing [Ligon06]

HEC FSIO 2006 Analysis and Update

Additional research is needed to address longer-term concerns in the area of metadata storage and management.

Neither the HECURA call nor the 2006 workshop produced much in the way of truly revolutionary approaches for high-end computing. This is not to say that the evolutionary work is somehow deficient or lacking. However, one would expect that a vibrant research community would produce occasional fresh examinations of old problems. This has not occurred. Maybe this should not be surprising. The HECURA call marks something of a restart of research in high-end I/O and so, researchers probably have not had sufficient time or regained sufficient understanding of the issues to begin thinking outside of "the box". Additionally, it’s questionable that revolutionary research will be considered for funding.

It should be very useful to look, again, at all of these proposals in a year or more for new issues brought to light and contemplated actions. It's been a long time since this research community has been offered the opportunity to remove existing constraints and contemplate radical change.

Identified Gaps

- Revolutionary scaling
- Revolutionary extensibility
- Revolutionary name spaces
- Revolutionary File System/Archive metadata integration
- Revolutionary Hybrid devices exploitation
- Revolutionary Data Transparency

HEC FSIO 2006-2007 Analysis and Update

One of the five CPA FSIO proposals deals with metadata and was funded during this period; Ali Saman Tosun from the University of Texas at San Antonio “High Throughput

I/O for Large Scale Data Repositories". His adaptive declustering research might prove useful in determining when it is advisable to replicate and distribute metadata to improve lookup rates and could be applied to multi-dimensional data/indexing.

During the 2007 Workshop, the topic of metadata was discussed and many participants agreed that the general issue of metadata is being worked and progress has been made, but there is still a lot of work ahead in making fundamental changes to how metadata is stored and manipulated and how to manage large scale data, analyze it, move it and all of the metadata associated with it is an unsolved problem. The complexity and amount of data that is currently produced and will be produced in the coming years demands a jump in the level of metadata scaling and will demand fundamental changes, i.e. revolutionary ideas in the management of metadata.

No new gaps were identified. Extensibility and understanding and defining tags were singled out as areas that need attention.

HEC FSIO 2007-2008 Analysis and Update

Metadata continues to be an area that, while good work is being done, demands more work. On example is the work that Michael Bender from SUNY Stony Brook and his team from Rutgers and MIT have taken their research in Cache-Oblivious Streaming B-Trees and has transferred the technology to and started up a company called Tokutek. Work is being done with Hybrid devices, but most of this work does not target metadata. CMU and UCSC have started work to try and speed up and scale metadata operations with hybrid devices. It was noted that all work in extensibility and name spaces is still evolutionary and that revolutionary work is necessary.

At the 2008 HECFSIO workshop, the metadata session was supplemented with a panel comprised of researchers concerned with storage and tracking of metadata. Garth Gibson from Carnegie Mellon University spoke on "GIGA+: Scalable Directories for Shared File Systems" extending the current limitations on directory size to allow for millions of files in a single directory and speeding up metadata rates within these huge directories. Ethan Miller from the University of California at Santa Cruz spoke on "Highly Scalable Metadata Search and Indexing" which lays out their plan for the Spyglass Design that will break up the file system into subtrees allowing for faster or incremental indexing and allow for faster searches. Margo Seltzer from Harvard University spoke on "Provenance: Meta-data or Not?" in which she explains why provenance data is important to keep with data, security concerns unique to provenance data, and why it is unique and can't be treated just like extended attributes. The metadata panel was very well received and generated good discussion.

Identified Gaps

- Metadata integrity is missing from this area and becomes more important as we scale to hundreds of millions of files
- More work needs to be done on metadata scaling
- A renewed call for defining metadata
- Use cases and workloads would be very useful to guide research in this area

HEC FSIO 2008-2009 Analysis and Update

There is a feeling that high petaflop and exaflop machines are getting closer and focus to address the needs of those machines was highlighted in the discussions. Similar to classic HPC, larger data intensive efforts using smaller machines but equivalent or greater I/O sub-systems seem to be headed our way. Both of these tracks, then, highlight the necessity of a renewed focus on support for metadata.

The issue, as always, has been scaling. Many feel that the classic solution, using an increasing amount of storage hardware, will continue to dominate in the short term at least. The obvious implications are that the metadata management sub-system will require greater capabilities in order to manage the associated aggregations and service the greater number of requests. In particular, the group felt that the "scaling" gap needed the greatest focus in the short term. Highlighted during the discussions, were fault management, automatic recovery, and approaches leveraging transactional semantics. Support on the service side for transactional semantics is novel and attractive since it was discussed in this context of how we might enable larger scale. Such semantics could be used to support highly efficient access to file system and user-defined metadata; the latter being something that has not been practical to date at scale.

The metadata breakout group also felt that the "extensibility and name spaces" gap should include an emphasis on efficient search. It was felt that that the semantic power of something like Apple Corporation "spotlight" could be useful to our HEC user-base if it could be successfully made to scale. For definition, this function indexes, or leverages pre-existing indexes, capturing attribute and **content** for all files that the user has access to. When asked, it can perform many different kinds of searches, including keyword search of the indexed content.

There was also discussion around how hybrid and emerging hardware storage devices might enable metadata management and scalability. It was noted that the classic proposal in this area leveraged these devices as just yet another cache tier and that it would be particularly attractive, instead, to see new data structures and algorithms that leveraged their particular advantages.

New HECURA 2009 projects in the Metadata area

- A New Semantic-Aware Metadata Organization for Improved File-System Performance and Functionality in High-End Computing
    - Yifeng Zhu at the University of Maine and Hong Jiang at the University of Nebraska-Lincoln
    - This project investigates the potential of using file metadata to enhance scalability of content access and provide efficient queries.
- Scalable Data Management Using Metadata and Provenance
    - Ethan Miller at the University of California at Santa Cruz and Margo Seltzer at Harvard University

    - This project will investigate an alternative name space solution based on indexed content and provenance information.
- Multidimensional and String Indexes for Streaming Data
    - Charles E. Leiserson at the Massachusetts Institution of Technology, Martin Farach-Colton at Rutgers University New Brunswick, and Michael A. Bender at the State University of New York in Stony Brook
    - This project will investigate methods for efficiently supporting streaming, superlinear indexes. The thought is that algorithms that successfully accomplish this may form the organizational basis for a file system.

2008 – 2009 Identified Gaps

There were no new gaps identified. Similarly, it was felt that the existing gaps needed to remain. However, it was decided that:

- Extensibility and namespaces should be modified to highlight a desire for highly efficient search at scale.
- File system and archive metadata integration should be changed to something reflecting a more generic need. A suggested wording was “cross-discipline metadata integration”. The fault-tolerance bullet and discussion should be changed to emphasis recovery and self-recovery at scale.

2009 – 2010 Analysis and Update

Metadata remains an unsolved problem that will only become aggravated with storage systems growing to accommodate exascale systems. It is clear from discussions at the workshop and from panelists that much richer metadata/indexing is required now and for the requirements of exascale. Elements of metadata issues are strewn across most of the identified research areas.

2010 – 2011 Analysis and Update

The format for this discussion was changed in 2011 to include a panel session prior to the open discussion. Panelists were Michael Bender (Stonybrook/Tokutek), Walter Ligon (Clemson), Aleatha Parker-Wood (UC Santa Cruz), and Galen Shipman (ORNL). Panelists were asked to comment on the Metadata roadmap and present their thoughts on exactly how metadata is most likely to grow in size and how users might change their interactions with their datasets.

The discussion of finding data lead to an intense discussion of data models in general, including how key-value stores, tools such as BigTable and Cassandra, and libraries such as HDF5 provide alternative views on the data model and data storage problem. Dimensionality of data and ability to search were considered important considerations in how to best organize datasets. It was noted that the same data structures that we are using to store data may also be used for what we might term metadata (e.g., indices for the dataset itself), and that there is a blurring of what is data and what is metadata, at least from the storage system's perspective. There was a call for experiments looking at storing and accessing large HEC datasets, via alternative interfaces and at scale, to help the community better understand this important component of the metadata, and next-generation storage, challenges. Based on this component of the breakout discussion, there was a suggestion to rename the "Scaling" area to "Data Structures".

More robust search capabilities are considered one promising approach to alternative name spaces. However, it is not clear how to present a search-based interface to users or that the types of searches (e.g., popularity) used in the Internet services space are relevant in the HEC context. More research is needed to identify how search might be made part of the storage interface and what sort of searches are most critical, keeping in mind prior work in the HEC space (e.g., semantic file systems).

Overall, exascale is considered an opportunity in that there is consensus that change, of some sort, will be required. Better understanding of how individual application and application groups behave is needed in order to better specify critical metadata challenges. This information, coupled with a better understanding of the methods of data discovery users might employ, could better focus our activities.

2010 – 2011 Identified Gaps
No new gaps were identified this year.

Measurement and Understanding

HEC FSIO 2005 Problem Definition
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. In parallel application building and tuning, there are a multitude of correctness and performance tools available to applications. In the area of scalable I/O and file systems, however, there are few generally applicable tools available. Tools and benchmarks for use by application programmers, library developers, and file system managers would be an enormous aid.

There is a need for research into evolutionary ideas such as layered performance measurement, benchmarking, tracing, and visualization of I/O related performance data. More radical ideas to be explored include end to end modeling and simulation of I/O stacks and the use of virtual machines for large scale I/O simulation. With research into these ideas, a future generation of high performance file systems could be understood and more efficiently pursued.

2005-2006 Progress
There were several success stories within the HEC FSIO community in measurement and understanding. First, the NSF HECURA program has funded many proposals targeting the measurement and understanding of parallel file systems and end-to-end I/O performance. From this funding, we expect exciting results from the following:

- A collaborative venture between Alok Choudhary’s group at Northwestern and Mahmut Kandemir at Penn State exploring scalable I/O middleware [Choudhary]
- The Arpaci-Dusseaus’ group at Wisconsin providing an analysis of formal failure models for storage systems [Arpaci-Dusseau 2006]

- Tzi-Cker Chiueh's work on quality of service guarantees for scalable parallel storage systems will be predicated on a strong initial foundation of measurement and understanding of this problem area [Chiueh]
- Priya Narasimhan at Carnegie Mellon has been funded to research automated problem analysis of large scale storage systems [Narasimhan]
- Purdue's Mithuna Thottethodi will be investigating performance models and systems optimizations for disk-bound applications [Thottethodi]
- Erez Zadok at SUNY – Stony Brook is exploring many aspects of HEC file systems such as tracing, replaying, profiling and analyzing [Zadok 2006]

Additionally, a group from Carnegie Mellon University has recently completed important work [Mesnier] in this area by developing a trace replay tool which automatically discovers causal events within HEC parallel applications. Replay tools such as this are important as they can drastically speed-up the research cycle by allowing researchers access to a wider variety of application behaviors and can potentially allow the I/O patterns of private applications (e.g. classified government codes) to be publicly released and studied.

HEC FSIO 2006 Analysis and Update

Discussion at the workshop indicates that there remain several important, currently unaddressed, gaps in this space. The attendees felt that current work, although highly important, is not looking at system workload in a realistic enterprise environment in which additional consideration must be given to questions arising from aging, reconfiguration, and workloads consisting of multiple heterogeneous applications. Further, the group feels that developing standards for HEC I/O benchmarks is important and that these benchmarks must account for the realistic enterprise environment challenges listed above such as aging. Also, testbeds for I/O research should be made available so that the cost of entry to do research in this area can be lowered. One final identified gap is a lack of cutting edge visualization tools to analyze large-scale I/O traces. Such a tool could help identify complex causal dependencies and would prove highly valuable for analysis of HEC I/O.

Identified Gaps

- Understanding system workload in enterprise environment
- Standards for HEC I/O benchmarks
- Testbeds for I/O research
- Applying cutting edge visualization/analysis tools to large scale I/O traces

HEC FSIO 2006-2007 Analysis and Update

There is tremendous interest in capturing trace data from real systems for analysis and replay of observed problems. Broadly, the gaps identified are (1) capturing data at scale and (2) providing tools that can analyze and visualize the data. Some progress has been made in capturing trace data, but the problem is far from being solved. Examples of progress include the work at Stony Brook University in capturing traces on their 250 node cluster which has minimal performance impact (stated as < 4% overhead). UCSC is mirroring a SNIA repository that contains some gigantic traces. Unfortunately, different people want to measure different characteristics of their systems and some data costs more to collect in terms of performance overhead. Further, metadata capture and inference, privacy issues, analysis tools and visualization are only a few of the many gaps still remaining. We are still not able to simulate or perform analysis on many large HEC systems and it is unclear if it will be possible to collect data on the largest systems due to the sheer volume of data that must be captured and reluctance of the users of these systems to accept any performance degradation.

From the length and level of interaction of the participants, more questions were raised and gaps identified than were answered. Although the workshop attendees are happy with the funded proposals and the work accomplished thus far, there is unanimous agreement that this problem is challenging and there are trade-offs that must be considered between the value of the data and performance impacts of collecting the data.

2007 Identified gaps:

- Availability of modern I/O traces with valid input and problems seen on I/O servers and Meta-Data Servers
- Visualization tools for traces
- A standard Tracing Format
- Distributed Multi-level tracing

HEC FSIO 2007-2008 Analysis and Update

There is ongoing interest in traces. Although there are traces becoming available, there are still more traces needed for a good understanding of application file system use. As traces started to become available, there was new interest in understanding the difference between traces, workloads and benchmarks. There is a growing interest in how researchers and vendors will use these traces, as many of them make assumptions about the nature of the IO, such as assuming a block based implementation, or striping patterns.

A continuing look at refining the tracing tools and methodologies is needed. Research is needed for classifying which application or workload traces show useful information for analyzing the usability of a given file system. Research for determining a set of data collection templates that will provide interesting traces is also needed. Work on dynamic tracing at SUNY may yield information on the appropriate data points.

Interest in understanding the failure of parts of very complex systems that are currently being built, and of future systems is going to be increasingly important. These systems will soon have tens of thousands of components that are all expected to work together to provide file system service, and understanding the effect of a production level system that is always running in a degraded mode from part failure will provide key research for building these systems.

2008 Identified Gaps:

- Understanding partial failures in a large system
- Trace refinements.

2008-2009 Analysis and Update

The 2009 HEC FSIO Conference had a goal to get feedback on the road map gap areas from the participants and, thus, had focused breakout sessions for each of the FSIO Gap Areas. Most people agreed with the problems defined in the Measurement and Understanding gap area with two exceptions; tracing formats and simulation and modeling.

The general area of tracing was a major topic during the Measurement and Understanding break out session. There was general agreement, although not unanimous, that we would probably not be able to standardize on trace formats. One reason for this opinion is the Storage Networking Industry Association's (SNIA) Input/Output Traces, Tools, and Analysis effort's failure to standardize trace formats. Although many people would still like some kind of standard for traces, what you collect really depends on what you want to do with the data. Since there is overhead with collection of trace data, it's not realistic to collect everything that everyone could ever want. Some people thought that a more realistic goal would be to set requirements on a standard or minimum set of data to collect for I/O traces. Thus, it was recommended that "trace formats" be removed from the problem area "Standards and common practices for HEC I/O benchmarks and trace formats".

In terms of benchmarks, people acknowledged a lack of and a need for locality measuring benchmarks, meaning benchmarks that measure the memory hierarchy do not exist; SPECFS is the only useful one, but it's outdated.

In terms of providing testbeds for I/O research, several people acknowledged that getting any hardware is helpful to fulfilling their needs; in order for the research to be relevant, they really need current or cutting edge hardware for their research. Since many clusters and machines installed at HPC sites are not only current, but on the cutting edge of hardware architectures, if we want the results from their research to be relevant to HPC sites, researchers need modern hardware. We are still trying to understand how virtual machines will help with testing at scale.

Since testbeds will not be plentiful to Universities any time soon, a new problem area of "Simulation and Modeling" was recommended under Measurement and Understanding This area includes work on "normalization" or how to compare vastly different systems. Work is already being done in this area with Walt Ligon's HECURA work [Ligon06] and research being done at UC Santa Cruz by Carlos Maltzahn.

Of the existing problems in the Measurement and Understanding Gap Area, the majority of participants rated "Understanding system workload in HEC environment" and "Applying cutting edge analysis tools to large scale I/O" as the most important problems in this area that needs research. The recommended new research area of "simulation and modeling" came in third as the most important work that needs to be worked on.

The 2009 HECURA solicitation awarded three projects with a primary focus on Measurement and Understanding, which are:

- Automatic Extraction of Parallel I/O Benchmarks from HEC Applications; [Ma 2009]
- Visual Characterization of I/O System Behavior for High-End Computing; [Kwan-Liu]
- RUI: Automatic Identification of I/O Bottleneck and Run-time Optimization for Cluster Virtualization; [He 2009]

In addition, another project was awarded under the HECURA 2009 solicitation that has a strong Measurement and Understanding component to it:

- Performance- and Energy-Aware HEC Storage Stacks [Zadok 2009]

2008 – 2009 Identified Gaps

- Simulation and modeling

2009 – 2010 Analysis and Update

Measurement and Understanding continues to be an issue with at HEC sites and for vendors. In particular, the issue of testing at scale is a problem that is continuously brought up. When HEC sites procure their storage systems, vendors frequently find it difficult to test and understand the performance issues and bugs that come with scale before delivery. Measurement and Understanding was not an area of discussion targeted at this year's workshop.

2010 – 2011 Analysis and Update

The format for this discussion was changed in 2011 to include a panel session prior to the open discussion. Panelists were Remzi Arpaci-Dusseau, University of Wisconsin-Madison, who was unable to attend due to illness; Phil Carns, Argonne National Laboratory; Xiasong Ma, North Carolina State University and Frank Mueller, North Carolina State University. Panelists were asked to review the relevant research in the area and consider whether the roadmaps were up to date and to help identify remaining gaps. Panelists also were asked to present deeper technical discussions of their specific research projects.

During the discussion, it was agreed that capturing relevant data is still extremely challenging. The choice is between sampling techniques, such as Darshan, which can miss behavioral phase shifts, or blunt force approaches, such as strace, which introduce overwhelming amounts of overhead. Of course, exascale class machines will drastically increase this challenge. There is a need for technologies such as MR-Net to be used to help aggregate and compress measurements. Machine learning with dynamically adjusted granulaties of tracing would be a valuable contribution in this space. A productization of the black-box debugging project from CMU could potentially be extremely helpful as well.

Another gap area is making correlations between application behavior and observed behavior at the storage devices. An ability to tag application requests and remember the tags as the requests wend down through the storage stack would be tremendously useful. Of course, another challenge is a lack of resources available to researchers to study these problems at scale. PROBE is a huge step in the right direction but it is also a gap area to provide a useful exascale system simulator. Getting traces to the community is something that Los Alamos, Sandia and other National Labs have been working on very hard for the past few years but more traces, and a more diverse set of traces, are always requested.

Continuous tracing which allows for early failure detection is an important gap area; even better, and more challenging to build, would be such a system that can identify performance faults in addition to correctness faults.

2010 – 2011 Identified Gaps

Candidates for new gap areas under Measurement and Understanding are

- Correlations between application behavior and storage devices
- Continuous tracing

Quality of Service

HEC FSIO 2005 Problem Definition

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. These QoS policies can reach a broad range of integration into software, file systems, and hardware devices. Policies such as guaranteed I/O performance, specific redundancy requirements, or I/O priority settings will allow the system to perform optimally for a given work profile. Further research into areas such as adaptive QoS systems, end-to-end solutions, hardware support, and cross-system integration will revolutionize storage systems that will be created in the next few years. These research topics will bring the storage systems to a point where users, systems, or entire clusters can be insulated from each other, while using the same storage infrastructure. This will also allow for predictable I/O performance and response time for the users.

2005-2006 Progress

There were a few successes based on this problem statement in the HEC FSIO community. Four projects were funded that address the QoS needs of the community. These projects are discussed in more detail in Appendix A

- Quality of Service Guarantee for Scalable Parallel Storage Systems [Chiueh]
- Active Data Systems [Reddy]

- Exploiting Asymmetry in Performance and Security Requirements for I/O in High-end Computing [Sivasubramaniam]
- End-to-End Performance Management for Large Distributed Storage [Brandt]

HEC FSIO 2006 Analysis and Update

The topic of quality of service was discussed again at the 2006 workshop, focusing mainly on the areas that were not addressed in current research and in the HECURA funded and research. The high priority areas were robust availability, defining QoS as it relates to HEC, and policy management.

Robustness in the QoS area is an emerging new emphasis in this area and consequently has not been worked on extensively. For this work to move forward research needs to be done in the area of providing feedback from components of the system to the different layers of QoS subsystems. Protocols for this work will also need to defined and prototyped to determine their usability.

The High End Computing community needs to better define what QoS means in relation to the specific needs of the HEC community. QoS requirements are very different for HEC as opposed to real-time visualization, data capture, and multi-media applications and systems. Research in this area is required to define what is needed to define QOS for HEC and then determine ways to evaluate the systems. A set of benchmarks testing each area of QoS for HEC is one way to do this evaluation.

Policy management is a key component to a useable and robust QoS system. It must support the security, authorization, and policy control. These systems need to implement a system that can be managed automatically without high operator overhead. Research needs to be done in the areas of policy implementation, authorization, distributed QOS needs, and using computing bounds, estimated or measured.

Identified Gaps

- Robust availability
- Defining QoS as it relates to High End Computing
- Policy Management for Quality of Service

HEC FSIO 2006-2007 Analysis and Update

Although there is still work to be done, the HECURA projects that deal with QoS have made great strides in the area of disk quality of service and performance insulation, but an end-to-end solution is still desired. In addition, a goal for QoS is in an environment with several clusters sharing a global parallel file system. One topic of discussion was the need for a standard QoS API, i.e. how to specify a desired level of QoS from an application. A report was given about the OSD-1 standard and how it allows for specifying QoS at the object level.

Other points discussed:

- Need to determine what we can guarantee before we offer this to an application
- The other approach is to request what is needed (not what is offered)
- The application needs to be able to ask the system what types of capabilities are available and map back to application.

- Need to be able to specify attributes of job for QoS in terms of different size runs (few nodes, large node counts, etc).
- Multimedia, backup requirements, virtual machine, service-level agreements characterization is available; with these classes, distill into more general
  - Need policy management
    - How to limit the max QoS request?
    - How to provide a minimum for apps?
  - QoS heuristics, learning and dynamically adjusting policy
  - Common format for storing information on heuristics
  - How to use the non-seeking storage for this?

2006 – 2007 Identified Gaps

- Need a standard API for QoS
- End-to-End QoS solution for HEC

HEC FSIO 2007-2008 Analysis and Update

There was broad consensus that the ongoing research is successfully addressing the difficult question of how to partition resources and provide Quality of Service guarantees in a single server system. This work is considered ready for commercialization and several vendors have begun looking at this. However, this larger area is not yet finished as several important challenges remain. First, scaling is still very much an open research area which needs demonstrations that these QoS techniques are applicable in a multi-server environment. In fact, it was observed that the current model in Brandt's research is not scalable by design as it uses a central broker to manage reservations. This is the remaining challenge to figure out how to make this broker scalable through some sort of hierarchical distribution of its responsibility.

Several participants also felt that this work can not be considered complete until it addresses the question of how the user will interact with the QoS system. It must be easy for the user to express their needs and this is predicated on the assumption that the user will be able to easily identify their resource needs. Workloads with changing resource needs also provide additional challenges. It was pointed out however that in HEC environments, different scheduling queues often have different priorities and this can be mined as a valuable first hint in discovering workload needs.

In essence, the current work solves the simple base problem where a well-understood and static workload can receive QoS on a single server static system. Scale, dynamic workloads, changing environments, multi-server environments, user interface, and workload characteristic discovery are all open additional research areas.

Much discussion in the workshop suggested a reorganization of the problem areas into four categories, from [Brandt][Ganger]:

- End-to-End QoS for Storage
  Good research has been done, but additional work is needed to integrate QoS

mechanisms across different resources and to develop mechanisms for coping with dynamic changes to access patterns, workload demands, and storage system resources.

- Standard Interfaces for QoS (Name Change) - Existing description is fine.
- Storage QoS at Scale
  Storage QoS mechanisms to date assist with single servers and small-scale clusters. Research is needed to scale these mechanisms to the large storage infrastructures of HEC and data center environments.
- QoS-based Management of Storage Systems
  Effective QoS requires usable administrative interfaces and internal mechanisms to assist with the phases of managing shared storage performance. Doing so includes initial planning and provisioning. It also includes runtime metrics, monitoring and handling of changes in workloads and system states (e.g., individual disk failures and other performance blips).

2007 - 2008 Identified Gaps:

- User interface
- QoS at scale
- Dynamic workloads and resources

HEC FSIO 2008-2009 Analysis and Update

Progress has been made in the QoS area, mostly in the area of server side QoS, mostly at the disk level. Some network based QoS has been attempted. A large amount of funding has been invested in this area over the last 3 years. No distributed QoS capability has been shown. A good goal for this area would be to demonstrate distributed/parallel QoS in the next few years. Using control theory came up as an opportunity, and there is a new HECURA 2009 project in that area. There was concern about how we get input into a QoS system and the thinking was to attack the problem of getting information from the resource management system. There was also discussion of how important measurement and understanding is to QoS and a new gap for QoS in this area was suggested. Also, interactivity versus batch as input to QoS was discussed. Another area that was touched on was how scheduling of data intensive computing or cloud resources and scheduling for compute intensive workloads are two completely separate fields of study and that there is room for optimization for scheduling of the entire computing/network/storage resource which considers both data and compute intensive capabilities.

New HECURA 2009 projects in the QoS area

- QoS-driven Storage Management for High-end Computing Systems [Zhao09]
  - This NSF HECURA project tackles the challenges in quality of service (QoS) driven HEC storage management, aiming to support I/O bandwidth guarantees in PFS.

- Interleaving Workloads with Performance Guarantees on Storage Cluster [Riska09]
  - This research focuses on the design and implementation of a lightweight, yet, versatile middleware framework that provides effective and scalable solutions to the problem of interleaving storage workloads with a wide spectrum of demands.
- Adaptive Techniques for Achieving End-to-End QoS in the I/O Stack on Petascale Multiprocessors [Kandemir09]
  - This project investigates a revolutionary approach to the QoS-aware management of the I/O stack using feedback control theory, machine learning, and optimization.
- CRAM: A Congestion-Aware Resource and Allocation Manager for Data-Intensive High-Performance Computing [Burns09]
  - This project will develop a job scheduling and resource allocation system for data-intensive high-performance computing (HPC) based on the congestion pricing of a systems' heterogeneous resources. This extends the concept of resource management beyond processing: it allocates memory, disk I/O, and the network among jobs. The research will overcome the critical shortcomings of processor-centric resource management, which wastes huge portions of cluster and supercomputer resources for data-intensive workloads, e.g. I/O bandwidth governs the performance of many modern HPC applications but, at present, it is neither allocated nor managed.

2008 – 2009 Identified Gaps

- Need a distributed demonstration of QoS
- Need to pursue RMS systems providing input to QoS
- Need to evaluate a possible measurement and understanding for QoS gap

2009 – 2010 Analysis and Update

Quality of Service was not a target research area discussed this year, but the general consensus is that good work is being done and progress has and is being made in this research area between the 2006 HECURA and 2009 HECURA QoS projects. We need to watch the research published by these research teams and, when appropriate point vendors to the work for integration into products. What is really needed at this point is a demonstration of true end-to-end quality of service.

2010-2011 Analysis and Update

Progress has been made in the QoS area. Four nice presentations on work being done in this area were presented by: Scott Brandt, UCSC; Greg Ganger, CMU; Mahmut Kandemir, PSU; Ming Zhao, FIU. Missing was a talk by Alma Riska, WMU who is also working in this area. A large amount of funding has been invested in this area over the last 5 years. No distributed QoS capability has been shown. By and large, the research has shown that the various parts of a single data transfer to a single storage device. This part of the problem appears to be making major progress. Some progress has been made to glue together the various QoS layers (network, disk, cache, etc.). The area that is lacking is parallel/scalable QoS. There appears to be little progress in this area. Commercialization of some of the single node/stream/device QoS is happening by NetApp and others. Additionally, the interface to QoS area appears to have been largely not addressed to date. Some discussion of why we can and can't use our HPC resource managers to tell applications when they can or cant write a checkpoint emerged. The realization that scheduling is indeed the QoS problem and whether we do it at a low level at devices/networks/etc. in a very fine grained way (miliseconds) or a high level at a resource management level in minutes to hours. Also, virtualization techniques like "stunning a virtual machine" to slow its IO injection rate came up, which is an interesting reuse of cloud ideas. Since jitter is not an issue during an IO, that may be of help. Largely people felt the big elephant in the room in the QoS area is parallel/scale.

2010 – 2011 Identified Gaps

- Some artifacts are being commercialized.
- There may be things that we can learn from the cloud community on protecting infrastructure performance and responsiveness by telling clients to slow down. Of course if server side pull was available, most of this would be moot.
- The largest challenge seems to be applying QoS techniques to scaled out parallelism

Security

HEC FSIO 2005 Problem Definition

In the 2005 workshop, the topic of security was recognized as one of growing importance with several areas needing attention. The bulk of these issues were recognized to relate to security usability, functionality, and overhead.

In the area of usability of security, it was recognized that if security is not easy to both use and understand, it will not be used. Ease-of-use of APIs and interfaces is an area of research for file system and I/O researchers to study, simplifying the use of security features for file systems and I/O. Standardization and validation, in turn, would become important as APIs and interfaces emerge.

In the area of security functionality, the topics of key management, distributed authentication/authorization, and end-to-end encryption APIs were noted:

- Long-term key management is an area in security that needs research and offers opportunities in industry. Security that incorporates encryption of data at rest requires carefully considered long-term key management schemes. Some issues include handling encrypted data at rest for which the encryption algorithm has become easily breakable, the protection of keys that allow for flexible use, and the management and longevity of keys over long periods of time and throughout natural events.

- Security systems must be able to handle authentication and authorization issues in a completely distributed world, often with hundreds of thousands of entities needing protection. Further, collaboration between multiple sites and organizations is becoming more prevalent where security solutions are necessary to flexibly handle these distributed file system security situations in terms of scale, distance, and flexibility.
- End-to-end encryption needs a robust, accepted application program interface that enables all data on the storage service and moving across the network to reside and move in encrypted form. Such an interface would provide for the encryption and decryption of data within the client-side operating system.

In the area of performance overhead, security within file systems and I/O comes at a price. There will always be an overhead associated with providing security, of course, but given the massive scale that HEC environments represent, as well as geographically-dispersed HEC sites, building security solutions that have acceptable overhead is a difficult, but important, task. There is a need for research into security overhead for HEC security applications.

2005-2006 Progress

To address some of the work needed in the area of security, two proposals were funded in the 2006 NSF HECURA Awards:

The areas of security usability and the trade-offs of security overhead versus performance are being pursued through Pennsylvania State University research on security framework [Sivasubramaniam]. The aim of this work is to provide a security solution that can easily accommodate different points in the security-performance space, offering different levels of security for clustered SAN-based architectures for different environments. Rather than attempt to accommodate each environment with a customized system, the framework would be tunable for performance or security based on site policies.

Another area of security being researched is that of long-term protection of stored data. Research at the University of Minnesota addresses scalable, global, and secure (SGS) online storage, building on the existing Lustre and Panasas object-based file systems [Du]. The work will investigate transparent, end-to-end encryption for high-performance backup and archival functions. Further, this work will investigate the area of long-term protection of cryptographic keys, including loss/recovery of keys, user/group membership changes, and retrieval of old data.

HEC FSIO 2006 Analysis and Update

The aforementioned research activities tackle several areas (usability, performance, end-to-end encryption, and key management) from the original problems noted in the 2005 workshop. There is certainly more work to be done in the area of security, however, as security issues are an ever-growing challenge. In a world with the issue of data stored on mobile laptops and portable hard drives, as well as the certainty of malicious parties bent on circumventing security measures, there are significant opportunities for novel designs and approaches.

In the 2006 workshop, additional areas of security research were discussed with existing topics from the previous year's workshop. The top ideas for research from the 2006 workshop included:

- Tracking the path of information as it flows through the system, determining where data has been and whether it has been compromised in-flight or at-rest
- Additional research into long-term key/algorithm management issues. In particular, as long-term persistence of security for storage holds different challenges than other types of security which can be more transient or sessioned in nature
- Performance and scalability overhead issues with security features
- Research on resilient security, including quick recovery from compromise
- Usability, addressing improved interfaces/APIs for ease-of-use
- A need for standards for secure deletion, balancing performance versus disk overwrite for deletion
- Understanding composition of security as it applies to end-to-end security
- Exploring the capability of quality of security as it pertains to HEC environments
- Developing methods for searching and indexing encrypted data

Identified Gaps

- Tracking the path of information as it flows through the system
- Additional research into long-term key/algorithm management issues
- Impact of security overhead on performance and scalability

HEC FSIO 2006-2007 Analysis and Update

In the 2007 workshop, these areas of security research needs were discussed:

- Storage system support for data provenance, secure scheduled destruction, and data privacy.
- Support for forensics and audit-ability, how the data has been processed over time.
- HEC high performance encryption capabilities, hardware assists etc.
- Making security easy to use for the HEC application programmer/analyst so that security is used and used correctly.

2006 – 2007 Identified Gaps

- Tracking the path of information as it flows through the system
- Impact of security overhead on performance and scalability
- Security ease of use

HEC FSIO 2007-2008 Analysis and Update

In the 2008 workshop, these areas of security research needs were discussed:

- Data Provenance, tracking what has happened to data, how, and by whom. The discussion was mostly about how multiple communities like the OS, network, and storage community need to work together to gather and track this information. It appears much research is spinning up in this area, including Penn State, Minnesota, and Harvard.

- A passionate plea for key management products was made by a government agency. It is unclear why products have simply not emerged in this area. Long term key management is not a solved problem especially for massive archive data at rest.
- The attendees felt that end-to-end encryption was not a research topic anymore, it is a solved problem. There are few products that do this however.

2007 – 2008 Identified Gaps

- Data Provenance
- Long Term Key Management

HEC FSIO 2008-2009 Analysis and Update

In the 2009 workshop, these areas of security research needs were discussed:

- End-to-End Confidentiality – This area was merged with "Tracking of Information Flow and Data Provenance" and is now "End-to-end Confidentiality and tracking of information flow, provenance, etc." Although not included this category in the "top gaps" ranking, the overall sense was that it must be dealt with differently in the HEC arena as compared to servers). Others contended that although encryption at scale is a difficult problem but not unique to HEC. Provenance can cover aspects of this area, but is more focused on authenticity of the data. This problem also focuses on who can or cannot access data, which is not typically encompassed by provenance.
- Performance Overhead and Distributed Scaling rankings were the same as from the 2008 workshop: of medium importance, needs research, and ready for commercialization. Some felt this was not a gap in itself, but instead an aspect of all the gaps.
- Tracking of Information Flow and Data Provenance: This area was merged with "End-to-End Confidentiality" and is now "End-to-end Confidentiality and tracking of information flow, provenance, etc." This was the second highest ranked gap at the 2009 workshop, with status same as in 2008 (very important, needs research, not ready for commercialization). At present we do not know how provenance will scale. It will be important to develop a calculus to understand costs and tradeoffs for different granularities of metadata, and to consider how to handle the provenance record for files that last for decades.
- Mechanisms and APIs to Facilitate Usability/Ease of Use, Ease of Management, Quick Recovery (previously named Ease of use, ease of management, quick recovery, ease of use API's). While there was some discussion of removing this gap, the consensus agreed that desired future lighter-weight alternative security systems may expose more of the underlying "stack," and make this gap more critical. It was noted that it is very important to consider usability at the beginning of such new alternative systems.
- Alternative Architectures for Authentication and Authorization was a new gap identified in the 2009 workshop, and it received the highest gap ranking. Current

authentication and authorization mechanisms at the file system level are very heavy-weight. The group advocated for alternative scalable architectures, such as lightweight asymmetric authentication/authorization systems similar to those currently used in web commerce. These alterative models will require APIs and middleware to support the lightweight security mechanisms, and may need to accommodate anonymous and very transient accesses.

- Distributed Persistent Storage/Memory was also a new gap identified in 2009, and was cited as a particular concern for emerging technologies such as of nonvolatile memory, key/value anonymous stores, and cloud computing. Use of persistent memory is a change in paradigm for HEC systems. Confidentiality and revocation issues are of particular importance. This gap area was deemed a special case of "Alternative Architectures for Authentication and Authorization."

2008 – 2009 Identified Gaps

- Alternative Authentication and Authorization Architectures
- Data Provenance
- Distributed Persistent Storage/Memory

2009 – 2010 Analysis and Update

In the 2010 HEC FSIO Workshop, security was not a research area targeted for discussion, but remains an area of interest.

2010 – 2011 Analysis and Update

Security was discussed in a breakout this year. The researchers asked right away, can you present a security use case. The organizers presented the following case:
Clusters are very complex in their software stacks. It is difficult to keep clusters completely up to date on security patches due to software dependency issues. This means that clusters are naturally vulnerable from root elevation attacks. If you become root you defeat UNIX permissions and other protections for data access including enormous amoungs of data at rest. It is important in some environments to enforce need to know and UNIX permisions can be defeated to easily perhaps. If we assume root elevation is the norm, is there a way to enforce need to know? More importantly is there a way to do this in a scalable way. Many solutions require centralized authentication, which is problematic because of the bursty nature of HPC workloads. Further, most large HPC sites use a batch model for workflow, so it is not possible for a person to be available to provide credentials all the time.

In general people felt that there are point solutions to solve some of this problem/use case. There was agreement that there are no complete solutions and that there are some fundamental scaling and batch oriented issues that still require fundamental research. NFSv4 and labeling are things that are being commercialized. Also, the feeling was that provenance was needed and that work is far from commercializable.

2010 – 2011 Identified Gaps

The primary gaps identified in the security session were provenance capture and scaling/batch operation at scale assuming root elevation.

Next-Generation I/O Architectures

HEC FSIO 2005 Problem Definition

I/O stacks and architectures have been static for some time now forcing developers to adopt awkward solutions in order to achieve target I/O rates. Changes in the storage model could result in significant gains in performance and usability; however, there is little incentive for vendors to make major changes given that most new acquisitions require interoperability with legacy codes and systems.

Short and medium term efforts in this area should include the definition of extensions to the POSIX I/O API to support high-end computing (including support for access awareness, small and unaligned access, and mixed workloads), mechanisms to better integrate archival storage with on-line storage, research into system-wide collaborative caching, and impedance matching across the I/O stack. In the longer term, additional effort could help redefine the I/O stack itself, such as moving intelligence lower into the I/O stack, eliminating independent client-side caches, integrating application-specific capabilities, or incorporating semantic awareness into the I/O architecture. At the same time, user-space interfaces, novel devices and hybrid architectures, and peer-to-peer technologies provide challenges and opportunities in this space.

Interfaces

Evolution

Problem Definition

The POSIX file access calls were simply not designed for high-end computing outside of an explicitly shared memory model. The environment for which it was designed assumed that file descriptors, synchronization, and buffer management were all supported in local, directly accessed, memory by very low-latency operations. In today's high-end computing, the dominant solution is a cluster or multi-programmed parallel machine, and these architectures directly expose a distributed memory. Instead of the current situation (vendors receiving exemptions from the POSIX standard, and acquisitions therefore having to be non-POSIX compliant for performance reasons), we need to move to a POSIX standard set that supports typical HEC file systems and file systems I/O access patterns.

One type of enhancement would address the lack of support for collective I/O. For instance, a collective open could mitigate considerable startup times on very large clusters. Locking at the process group level might allow coherence to be maintained between applications without the overhead of traditional locking interfaces and implementations.

Another area for enhancement is in how applications describe accesses. Currently applications are very limited in their ability to describe access to disjoint regions in the “stream of bytes” that make up a file’s data. By allowing applications to describe more complex accesses, we can significantly reduce the number of small transfers, converting them into a smaller number of large transfers instead.

2005-2006 Progress

An effort has begun under the Open Group to define HEC Extensions to the POSIX interface, an important step in adoption of extensions in vendor products. However, none of the funded proposals from the 2006 HECURA call addressed this area.

2007 Identified Gaps

- POSIX mandated name space, the directed acyclic graph, provides insufficient support for recalling the identities of enormous numbers of related files [Miller]

2007-2008 Analysis and Update

Researchers are having increasing difficulty finding their files because of current naming practices. Increasingly large numbers of related files require additional information to distinguish them, and their relationships. Some form of alternate indexing could be explored, perhaps? Classic path names could be augmented by allowing regular expressions and qualifiers based on file metadata? Something else? In any case, it was noted that the supplied, traditional, interface is becoming increasingly insufficient to the task of organizing these large scientific datasets.

While the Open Group project remains officially "active", little or no action has occurred. An initial, intense, foray produced a fairly complete set of application programmer interface documents and proposed semantics. Subsequent, harsh, criticism by the Linux file systems community demonstrated a lack of appreciation with respect to high-performance I/O in the multi-programmed, parallel universe. The entire topic, now, appears to be a "hot button", polarizing both the HEC I/O community and the Linux file systems community. Resolution of this issue would seem to be a pre-requisite to further action in the project.

Revolution

Problem Definition

Active disk, the ability to move part of an application near and on to the disk, has been an active area of research. However, no good interface and set of semantic rules has come along that would make it generally useful. Currently, it would seem that all solutions in this arena are restricted to modifications to support specific applications. A design that provides a “sandbox” so that multiple, unique applications can leverage the promise of active disk simultaneously would be welcome.

2005-2006 Progress

One funded proposal addressed this area specifically and is a good start at defining the operating environment for processing on disk in conjunction with traditional workloads. More follow-on work will be necessary to fully understand how active disks will fit into the larger I/O picture.

Other revolutionary work related to interfaces is covered in the I/O Software Stacks section, below.

Access Patterns

Evolution

Problem Definition

File system designs tend to require tuning to efficiently support either large or small transfers. Unfortunately, they do not seem amenable to supporting both simultaneously. Worse, some applications attempt both during different phases of their processing. Something adaptive is clearly called for to avoid performance penalties both for the bulk streaming I/O (where bandwidth dominates) and for the smaller transactions that are latency sensitive. Modern self-describing data organizations such as are employed by HDF and CDF/netCDF sometimes scatter attributes throughout the file, interposed with the data, which can result in these mixed workloads. It is insufficient to simply handle small and large transfers. We must also be able to handle these in a directed, or vectored, fashion.

2005-2006 Progress

This 2006 HECURA call produced a very strong response in this area. Funded proposals cover data layouts for more efficient access, small and mixed I/O optimizations, and access pattern recognition. Further concepts and proposals addressing these challenges are addressed in the Caching and Coherence section, below.

2008 Progress

Dr. Pete Wyckoff has shown results addressing the issue of varying latencies in the I/O path in WAN environments. However, with his departure from the Ohio Supercomputing Center it seems unlikely the work will continue. With the up and coming cloud-computing concept, the issue could become increasingly emergent.

I/O Software Stacks

Evolution

Problem Definition

The existing IO software stack is deep and composed of different, sometimes disparate, modules. For instance, any distributed file system will rely on networking components. A fresh look at this stack, end-to-end, could address bottlenecks, especially when disparate modules call on each other. At the least, it would be highly desirable to have a standard method to indicate a long latency path was in use so that normal timeouts were not employed. This is a problem with today's hierarchical storage management systems when part of the path is over a WAN. Object based storage systems will have a similar concern because their logical evolution is to have objects appear "equal" regardless of physical constraints (unequal file system behavior as well as distance).

2005-2006 Progress
Funded proposals specifically address how object storage can best fit into the I/O stack, the use of anticipatory I/O scheduling to better manage resources, and how caching at various I/O layers could be coordinated to improve overall performance. No proposals directly address this issue of varying latencies in the I/O path in WAN environments.

2008 Progress
Dr. Pete Wyckoff has shown results addressing the issue of varying latencies in the I/O path in WAN environments. However, with his departure from the Ohio Supercomputing Center it seems unlikely the work will continue. With the up and coming cloud-computing concept, the issue could become increasingly emergent.

Revolution

Problem Definition
Many large engineering and physics simulations are burdened by CPU data cache coherency semantics. It's always been known that the ability to turn these off, where possible, enhances observed performance. Similarly, in the I/O world, a distributed application often does not require the services of a local buffer cache. Changes to the file system could be made that would react to or enable an application to remove these high-overhead but low value components from the control or data path could be usefully leveraged.

In the HEC space, we have clearly outgrown the decades old initiator-target I/O paradigm, yet the requirements for performance, data integrity, and coherency remain. Achieving a revolutionary approach to I/O is constrained within the respected paradigm.

File system solutions in the high end have relied on a core stack from commodity file systems design for workstations and servers. This could be redesigned in order to add to, remove from, or alter the placement of existing file systems components in the software stack. For instance, a local buffer cache could be removed in favor of a collaborative cache maintained by a distributed application for its own use. Early research indicates a benefit here; however, the only implementation has been in the presence of the local host buffer cache. Other components could be reexamined, in order to find potentially better placement within the stack.

2005-2006 Progress

Revolutionary concepts in I/O stacks were partially covered by the 2006 HECURA responses. Work is funded in the areas of I/O graphs as part of the I/O stack, collective I/O, and tunable consistency. Additional work in communication protocols also addresses I/O stack concerns.

File Systems

Revolution

Problem Definition

While parallel file systems are a common component of HEC systems, very little work in recent years has focused on parallel file system architectures or optimizing these systems for HEC workloads.

Existing solutions utilize the network as a communications channel but there is power in the network far beyond that. While some solutions go so far as to generate multiple simultaneous transfers, there is not much that is fundamentally different from the classic initiator-target model. This method of using the network is highly portable but ignores the real power in high-end networks. Fresh approaches, such as peer-to-peer solutions, use new paradigms to reorganize storage. Such solutions would directly incorporate geographical distance in their cost function and significantly lower the bar that prevents massive replication, enhancing fault characteristics, or directly leverage other attractive properties in the network.

While many research file systems utilize components in user space, the practice is uncommon for production file systems. Performance data in the high end, at least, would seem to suggest that user-space file systems are a practical approach in general. Benefits from the eased development and debugging effort might, conceivably, offset the slightly higher call latencies. One of the historic reasons for preventing user space file system activity has been the data integrity concern – research into mechanisms (shared secrets, others) that could allay those concerns and permit user space and system space file system behaviors to co-exist is desirable.

File systems move bytes. Some government agencies believe that a file system augmented with knowledge of the data stored and transported could go much further. For instance, a file system that understood the machine word format used on the machine where the data was originally deposited, could reformat for a heterogeneous network of machines when required. As well, understanding the relationships between records accessed by an application could allow the file system to do a better job when storing the data or use much more intelligent prefetch strategies when retrieving it. Providing a method for the definition of such associations could be useful. Then, researching how changes within the file system might leverage such information would be appropriate.

Long-lived data will need to convey format many years into the future, potentially. The concern is not just word lengths and endian issues but such things as floating point formats (mostly resolved) and the representations of complex math components. A generic API that could stand the test of time is desirable. Efficiency in performance, CPU cycle consumption, and storage will remain issues.

2005-2006 Progress

Work such as the PVFS project and the Light Weight File System (LWFS) project continue to push the boundaries of what is possible with user-space file systems and provide necessary infrastructure for additional research in parallel file systems.

Work funded by the 2006 HECURA call addresses the use of advanced network capabilities in the I/O system, server-to-server communication, autonomics, server co-scheduling, and content addressable storage. These concepts cover a wide space of possible file system designs and promise to uncover viable new architectures with characteristics better suited to HEC. No proposals specifically addressed this concept of long-lived data. This could be construed as an archiving issue, however, and not specifically a file system issue.

Caching and Coherence

Evolution

Problem Definition

Modern operating systems inevitably buffer I/O transfers. While this has proved optimal in performance for locally attached storage, it presents problems when the goal is to efficiently use remote storage that is byte granular. The client operating system will typically employ buffers of fixed size. An application that does not fill a buffer when writing, can place the operating system in the uncomfortable position of having to read a full buffer, update the content and then prematurely flush the modified content back to the stable store. If a buffer system was available that was variable length or naturally supported modified sub-regions then byte-granular stable stores such as object-based disk could be efficiently leveraged.

Another related goal is to minimize the number of buffering steps required for data. It is not always done in current software stacks and that is a concern both from a performance and a memory usage perspective.

A file's address space is unqualified even when accessed by the several clients in a cluster or MPP machine. While this is an accepted well-understood paradigm, it amounts to globally shared memory without any hardware assists; something long-ago recognized as sub-optimal. The core problem appears to be that a globally coherent view must be maintained. Many high-performance applications do not require such a thing but, the service section in a high-end machine would.

The metadata service relies upon lock services to accomplish the globally coherent views. By definition, such a thing is provided by the cluster, or MPP, service section. Usually, these are a magnitude, or more, smaller in size than the compute client section as they are simple overhead. They enable the compute section but do not directly contribute. Worse, the available lock algorithms do not seem to scale, so even if a larger service section was available, it would consume itself while trying to manage locks. Clearly, a renewed interest in scalable lock services is needed. Solutions involving the compute partition, to augment the power of the metadata service and relaxing the API semantics with respect to coherency, could lessen the load on the metadata service section.

2005-2008 Progress

A number of 2006 HECURA proposals cover aspects of caching and coherence, from collaborative caching (using the caches of multiple clients in concert) and multi-level caching (efficiently leveraging caches at multiple levels in the I/O stack) to enhanced pre-fetching algorithms to better fill caches. These efforts fit well with other work in I/O software stacks and interfaces. Improvements to metadata services that leverage atomic operations were also covered but in general, no solutions were proposed that attempt to leverage transactions in file systems or otherwise attempt to move away from the lock-based coherence paradigm. Some work discussed under Metadata is also applicable to the problems addressed here. Additional work is called for in this area.

One project, exploring persistent file domains for MPI-IO has terminated and the work has been incorporated into the MPICH distribution from Argonne National Laboratory. This library serves as the reference and base implementation for many MPI libraries in the community.

Novel Hardware

Revolution

Problem Definition

Storage devices have not changed fundamentally in 50 years. The advancements possible in allocation policy and layout given a radical change in the access latency to bandwidth ratios seem attractive to explore. New devices with promise along these lines as well as hybrids combining existing storage solutions could spark renewed interest in, and value from, these core file system areas.

The focus, today, is on capacity. Obviously, there is a need. However, the increasing capacities also come at a price. The greater probability of faults on a single unit now jeopardizes RAID systems at rebuild time. One avenue that is being explored in depth is to use the aggregate to offset the error probabilities. Others could be in the individual disk units themselves.

2005-2006 Progress

Funded proposals cover new storage organizations to more efficiently use local storage in I/O systems and ways to integrate object storage and active storage into the I/O system. However, none of the funded proposals addressed the growing need for new redundancy schemes; additional work in this particular area is needed to fill critical caps.

2008 Identified Gaps

- Phase-change memory and MRAM solutions will potentially be exposed via non-block-oriented interfaces. Is research required in order to optimally leverage such interfaces by file systems?

2007-2008 Analysis and Update

Much discussion on the topic of Non-volatile storage solutions took place. Relevant to file systems, a new gap was identified; Non-traditional exposition of the media. It seems possible that the atomic unit for these could be very large. Will file system caches insulate us from these? Will the file system be required to wear-level and, if so, how does that effect use?

Archive

Evolution

Problem Definition

The directed graph mandated by the POSIX name space extends well into hierarchical storage systems. However, while solutions such as X/DSM allow a useful transition, for the user, from one part of the storage hierarchy to another, IO access to the file address space is problematic. Users unexpectedly encounter long delays while data is copied to or from the high-overhead portions of the storage hierarchy, for instance.

The presence of near-line and off-line storage encourages the user to view the storage system as infinite. This is in direct conflict with site policies normally. Inevitably, administrators require a gate or hurdle the user must encounter so that it is understood that scratch, ephemeral, and redundant data should not be placed into the deeper, or archival, layers of the hierarchy. In the past, this has been accomplished by adding manual steps to the process or imposing quotas. Both of those methods, however, are counterproductive in a name space that should, or could, span multiple layers or the storage hierarchy. This will become even more of a concern as the name space is expected to be global within an enterprise.

Many sites have archival mandates where some data lives forever. For these sites, moving from one product to another is difficult as a migration of this data from the old product to the new must be performed. Then, too, the amount of this data is forever growing which, as time goes by, makes the problem ever more difficult. Even the identification of such data sets is not incorporated. While migration is not addressed at all, industry does address mandated archival needs by providing special write-once file system and hardware solutions. This is unnatural, though, as the partitioning of the name space is necessarily tied to policy. A more natural solution would allow policy to be applied to storage without regard to file location in the name space.

2005-2006 Progress

Only one 2006 HECURA proposal covered archives. While archival storage sits at the edge of what we traditionally consider HEC I/O concerns, more effort is needed in this area to maintain the viability of archival storage in HEC environments.

Overall

As part of SciDAC2, the SciDAC2 Petascale Data Storage Institute was created and the SciDAC Scientific Data Management Center for Enabling Technologies was continued. Both of these entities are focused on key problems in scientific computing at scale.

Many of the NSF HECURA funded projects address problems in this space as well:

- The role of I/O graphs and metabots in I/O architectures [Maccabe]
- How new organizations such as streaming B-trees might impact storage organization [Bender06]
- Alternative I/O middleware organizations [Choudhary]
- Multi-level caching and alternative consistency semantics [Ma 2006]
- The potential role of active networks in the I/O stack [Chandy 2006]
- How tools might extract I/O patterns from applications for the purpose of prefetching and other optimizations [Chiueh]
- The role of server-to-server communication in parallel I/O systems [Ligon06]
- The real-world implications of integrating active storage with traditional storage systems [Reddy]
- Advanced scheduling schemes, including anticipatory scheduling and co-scheduling [Shen]
- The server-side push concept in parallel I/O systems [Sun 2006]
- Power-aware I/O architectures [Thottethodi]
- The convergence of object storage and parallel file systems [Wyckoff]

HEC FSIO 2006 Analysis and Update

The projects above hit upon a large number of key issues in this area. However, there are still a number of outstanding issues. The 2006 workshop identified a number of areas where additional work is warranted. Underlying file system abstractions topped the list of areas that are not adequately covered by existing work. This area includes next-generation virtual file system (VFS) layers for operating systems, alternative name spaces and file system organizations, and transactional file systems (and integration of database concepts in general). Advances in autonomic storage systems were seen as critical, particularly given the near-term construction of I/O systems incorporating tens to hundreds of thousands of devices. Novel devices and hybrid I/O architectures, noted as an important issue in the 2005 workshop, still need additional attention. Additionally, the specter of HEC systems having multi-million way parallelism with the need for very small I/O operations coming from all processes and how to best deal with this issue was brought up.

One area that did not receive much attention in the responses to the 2006 HECURA call and was identified as critical by government agencies was high-level I/O libraries. While one could consider the proposed work in I/O software stacks to address this indirectly, a more substantial effort focused on data formats and interfaces for HEC could improve performance, increase usability, and provide a long-lived data format, a combination that is very compelling.

Identified Gaps

- Underlying file system abstractions
    - Next-generation virtual file system interface
    - Alternative naming and organization schemes
    - Convergence of database and file system technologies, such as a transactional file system
- Self-assembling, self-reconfiguring, self-healing storage components
- Architectures using 10^4-10^5 storage components
- Hybrid architectures leveraging emerging storage technologies
- HEC systems with multi-million way parallelism doing small I/O operations

HEC FSIO 2006-2007 Analysis and Update

There has been a tremendous amount of work in this area, but because the challenges are substantial the majority of gaps still exist. We are still not able to actually simulate or perform analysis on the largest HEC systems. There are several researchers who are enabling modeling capabilities in lieu of having systems at the extreme scale. The areas for discussions in this section are data abstractions for applications, scalability, server push mechanisms, active storage networks, active data systems and the applicability of OBSD devices in parallel file systems. From the length and level of interaction of the participants, more questions were raised than answered. Although the workshop attendees were happy with the funded proposals and the work accomplished thus far, there was unanimous agreement that this problem space is particularly challenging and that larger amounts of effort must continue to be expended in this space. Specifically, the group feels that HEC researchers would be well-served by delving into the research in the modeling community. Perhaps being able to model individual elements of the entire system would be an effective approach. There are other projects in system-wide modeling that will add information and methods for a software system representing potential future extreme scale computing I/O systems.

In the 2007 workshop, the needs of Next Generation I/O Architectures discussed were:

- Most of the issues from 2006 are still open
  - large number of storage devices; 10^6 – 10^7 storage devices still not addressed
  - Fault tolerance
- An OBSD simulator/emulator capable of modeling OBSD's
- An extensible parallel file system simulation tool
- A dynamic pre-fetching architecture
- An Active Storage Network Architecture
  - Move the intelligence throughout the system
- Active Data Systems
  - Methods of solving the problem(s) when you run out of current file space

Identified Gaps

- The need for in-depth examination and analysis of the entire data path
  - Cost of pre-fetch at the disk in a batch system
  - Metrics and mitigation for pre-fetch, congestion
- Augmented job schedulers to assist with:
  - Data placement
  - Staging
  - Pre-fetching
  - Data hints ? Data compiler hints ?
- How can Flash or Hybrid Devices be better utilized
- What is the potential of commercial SSD's
  - What about the constraints of Flash devices
- Compiler/language extensions for I/O
  - In UPC, Fortress, Chapel, X10, etc.
- Data collection (traces of I/O at all levels in the system)
- Verification of data correctness over time
  - Proactive solutions
  - End to end solutions
  - Provably correct
  - Fault analysis
- In-depth stack visibility and definition
  - Fault analysis
- Improved / existent access methods
  - Content addressable

- Persistent store
- Global and shareable
- Semantics of indices

2007-2008 Analysis and Update

Once again, name space issues were discussed. Researchers are having increasing difficulty finding their files because of current naming practices. Increasingly large numbers of related files require additional information to distinguish them, and their relationships. Some form of alternate indexing could be explored, perhaps? Classic path names could be augmented by allowing regular expressions and qualifiers based on file metadata? Something else? In any case, it was noted that the supplied, traditional, interface is becoming increasingly insufficient to the task of organizing these large scientific datasets.

Some trivial progress has been made in trying to integrate I/O semantics into the HPCS programming languages. Benchmarks are being written along with a small team of researchers being involved in the language semantics for DARPA. This program is currently being funded by the DoD.

2008-2009 Analysis and Update

Not much progress in the area of I/O semantics has been made. There is still on-going work in adding I/O semantics to some of the DARPA languages (X10 and Chapel) but as for additional I/O semantics, little has been done, or funded. For once, the main topic of interest was not that of the name space issue. There were six gaps included on the roadmap. We had more than 40 people in this session. The points were added up points and found a reasonable distribution. Below are the rankings and the number of votes per area:

1.) Hybrid architectures leveraging emerging storage technologies (61)
2.) Understanding file system abstractions – file system architectures (36)
3.) Self-assembling, Self-reconfiguration, Self-healing storage components (35)
4.) Understanding file system abstractions – naming and organization (34)
5.) HEC systems with multi-million way parallelism doing small I/O operations (33)
6.) Architectures using 10^6 storage components (20)

Numbers four (4) and six (6) have been integrated into number two (2). So “understanding file system abstractions” will have a slightly broader area of coverage. There were a number of new areas discussed, but none of them were voted on to be “officially” included as areas of research. They are listed here for completeness and information at a later date.

•Several Simulator requests (we need to find out what currently exists)
•Power Aware
•POSIX Compliance
•End-to-End

Analysis (KDD)
Prediction / Optimization
Tools
Instrumentation

•Data Intensive Computing
•Database technologies (Hide from user)
•What about Next Gen Networks ?

In order to answer the Simulator requests, a call was given to the participants to send information to the committee on anything they are or will be working on in the Simulation arena. The committee will collect the information and disseminate it to the list. At the time of this report, no information has been received. The power aware topic was another on of interest. This is generally the case within the industry and those facilities where large-scale systems exist. This is most certainly of great interest to all of the large government sites. The place it seemed to most likely be a good fit would be in the self-* section of research. It was concluded that another topic of interest that belongs in the same subsection would be that of self-configuration. The recurring question of POSIX compliance arose again. Compliance as well as potentially novel and new approaches of integrating small enhancements to POSIX was discussed along with the understanding that without substantial funding, nothing will happen.

The PDSI project was cancelled by the DOE and the funding will end in September 2010. Several new projects were added. Below is a listing of the new projects added to this element:

•HaRD: The Wisconsin Hierarchically- Redundant, Decoupled Storage Project; [Arpaci-Dusseau 2009]

•A Dynamic Application-specific I/O Architecture for High End Computing; [Sun 2009]

•Cross-Layer Exploration of Non-Volatile Solid-State Memories to Achieve Effective I/O Stack for High-Performance Computing Systems; [He 2009]

•Streamlining High-End Computing with Software Persistent Memory; [Rangaswami]

•Programming Models and Storage System for High Performance Computation with Many-Core Processors; [Dennis]

• Balanced Scalable Architecture for Data-Intensive Supercomputing; [Huang]

• RUI: Automatic Identification of I/O Bottleneck and Run-time Optimization for Cluster Virtualization; [He and Scott]

At the time of the workshop, not all of these had been publically announced.

2009-2010 Analysis and Update

Since the 2010 HEC FSIO Workshop was focused on problems with systems at the exascale, exploration of alternate architectures was a major topic of discussion and was the subject of two breakout sessions. From these discussions and breakouts, it is clear that people expect there will be an additional level of storage, Flash or phase change memory, that will be necessary for the amount of data and acceptable amount of time for I/O of next generation machines.

The first break out dealing with advanced architectures was “Burst Buffer and Novel Scaling”. The subject for this breakout was centered around the evolution of existing parallel file systems for check pointing to include a solid state storage burst buffer capability to be used for staging data on and off of an exascale sized machine. The discussion explored the various issues around utilizing the burst buffer concept.
The following are the topics covered in the breakout session:

- Where should the burst buffer be located physically, in the compute node, in compute node rack, or both?

  - The topic of how placing the burst buffer in the compute node was explored, due to the long migration times between the burst buffer and the backing disk storage system. The conclusion was that jitter management needs to be better understood for exascale class machines.

- Who should manage the burst buffer, a relatively simple library tied to the resource manager, the file/storage system, or something else?

  - Industry representatives didn’t mind managing the burst buffer, if connected to/or managed by an IO node. They want to be able to own the IO stack for wherever they have to manage non-volatile storage. They were uncomfortable with managing this burst buffer resource on a compute node due to support issues for software in that compute node stack. It was noted that the operating system on the compute node client may not be conducive to even considering that environment for support.

  - If we let the file system manage the burst buffer, does this introduce risk beyond just managing it in a more simple way, less automated way. Industry didn’t seem to be bothered by this, but others were less convinced. It was noted that Ceph, Panfs, and GPFS all have concepts for storage pools and migration between storage pools, so many felt that file systems could be adapted to manage this resource in a straight forward way. Additionally, Panfs and Ceph, and perhaps other systems, have the concept of topology awareness which also could be leveraged to provide this management service for the burst buffer.

  - A very important concept was introduced to leverage the Cloud community to use virtual machines to hold the burst buffer manager which would allow the ability to put it wherever it makes sense.

- Will we need a JCL like language and other mechanisms for stage of data on and off the burst buffer?

  - Most felt that this would be required but there was some disagreement on how sophisticated this needed to be. Some felt that providing a simple way for staging on and off was reasonable, where others felt that data distribution information needs to be stored with the data on the disk storage system, so that intelligent recall of the data can be done enabling smart placement of the data onto the flash for an upcoming compute job. At a minimum, an interface to resource management will be needed.

  - Also, a discussion about how global the burst buffer should be seen in the storage system. This should inform the trade-off between convenience for stage in issues and increasing the failure domain. The majority seemed to lean towards small failure domains making the stage in of data from the disk storage system a bit harder to implement but making the failure domains much smaller and simpler.

- How reliability will this burst buffer have to be, will we have endurance issues?

  - After some analysis of writing over the flash every 3 hours, it was determined that we would not come close to the typical endurance expected for solid state storage for the burst buffer capability over the lifetime required (a few years). There was a question about whether we should even worry about duration management at all, but it was noted that solid state storage will likely come with this management no matter how you purchase it.

- Where should the burst buffer be connected, to the interconnect or elsewhere?

  - It was determined that the burst buffer doesn't need to connect to the interconnect of the Exascale machine, in fact given the small memory footprints of these future nodes, a very minimal data channel like SAS should keep up with the necessary speed and would be extremely cost effective. SAS parts are fast and extremely cheap, even support switches for muti-pathing/multi-porting to allow simultaneous access from multiple hosts (say all the compute nodes in a rack and a few IO nodes for reliable stage in/out capability.

  - If the decision is to put the burst buffer into the interconnect, topology awareness would be needed, and the more complex the interconnect technology the more difficult this task would be. Additionally, connecting to the interconnect would also be another source of possible jitter.

  - The group explored possibly routing the data through the IO Node to burst buffer. While this might add some value, it would add cost and was determined that it was not necessary.

- Can or will this move toward active storage concepts?

    - Since the burst buffer will likely be SSD, should we be considering doing active things like scatter gather to/from memory to/from disk, dedup, etc? The group did agree that this is possible and should be considered.

- Should the files that enter the burst buffer be made immediately globally accessible via the disk storage system (the name or the entire file)?

    - Once the checkpoint is written, the file could be made available via the name space at a minimum. Follow up use of the data in place could also be encouraged, like scheduling a set of data analysis nodes with access to the burst buffer could be considered. Scheduling would have to be done carefully to ensure that the Exascale platform will not stall.

- Will the compute or I/O node become too costly if you put an SSD burst buffer in it?

    - After some discussion, it was determined that you don't need to put the burst buffer in the compute or IO node but if you did, it could just be a chip on the mother board and would be very cheap and easy to get bandwidth to/from. It was also noted that because of the memory poor nodes in the Exascale system, extremely high bandwidth per burst buffer is not required.

In addition to the burst buffer breakout session, a concurrent session on "Combining Analytics and File Systems, Leveraging Google, etc." was held. The following questions were discussed:

- How much can we analyze, and how will we perform that analysis?
    - Applications are generating a "tsunami" of data, so much so that as much as 95% of the data produced is never analyzed. As a result, important discoveries may be missed. New techniques, such as in situ analysis (i.e., analysis of data while in memory during the simulation) are being investigated by the visualization and analysis communities in conjunction with application scientists. These techniques augment, but do not replace, the traditional "dump and post-process" model of analysis and visualization. That model will continue, with post-processing occurring either on the HPC system or on separate resources, perhaps the storage system itself (i.e., MapReduce style or active storage).

- How do we choose what data to keep for analysis?
    - In addition to the challenge of analyzing even a significant fraction of the data generated, we are decreasingly able to retain data over extended periods. While it doesn't appear to be a near-term issue, it is possible that some application teams will hit the point where they simply cannot save enough to do their science. This issue brings the question of prioritization to the forefront. Quantifying the cost of recomputing data versus storing and subsequently retrieving that data provides one metric, but the intrinsic value of data that hasn't been analyzed is difficult to assess. The number of

users or projects that might extract knowledge from a given dataset could also factor into the value of a certain dataset.

The value of some data changes over time as well. Inputs for real-time analysis can quickly lose value in some cases, and in fact real-time needs have a further impact on scheduling. Uncertainty Quantification (UQ) techniques can generate data that is valuable for a short period of time, while decisions are made as to what part of a design space should be investigated next.

- Changes to I/O architectures
  - While the workshop members were not largely advocating starting from scratch, it appears clear that there is room for improvement in the organization of the I/O software stack as it exists today. One area that was identified as an issue was in the primitives that are currently provided by the lowest software layer: the parallel file system. The object model from current parallel file systems appears to be one possible underlying building block, with middleware supporting more complex structures sitting atop this layer. In addition to exposing these objects, providing controls over features such as buffering and prefetching would allow middleware to more carefully tune behavior. Finally, smart compression might be one way to provide more effective space for application datasets, but it is not clear where this functionality would be best implemented (e.g., in the object storage layer, in data model middleware).

- Data models for computational science
  - Many computational scientists have embraced higher-level data model support from tools such as HDF5, netCDF and Parallel netCDF, and ADIOS. These tools provide a small subset of the data models in use by the community, and they should be augmented to provide support for all the major data models used, of which there are only a handful. The attendees did not see the need for a single "universal" model. The software supporting these data models would map data into storage containers (e.g., files, objects) so that locality could be exploited during analysis, and this functionality might be best implemented as middleware. As an alternative to explicitly storing data in a particular format, we could capture data as it resides in memory so that it may be restored quickly. In either case, operators on the data model will be needed for analysis purposes.

- Dynamic and composable systems
  - It is unclear at this point how dynamic/composable exascale systems will be. Will we be performing "forklift upgrades," or will we be adding/removing components over time? How does this relate to how we manage unexpected loss of components, which is one aspect of fault tolerance? The attendees were not in a position to answer these questions, so they will remain until the architecture and management of exascale systems become better defined.

- Scheduling and co-scheduling
    - The attendees noted that there is a disconnect between the scheduling of HPC systems and the sort of scheduling that occurs in MapReduce environments or on the analysis and visualization side of current deployments. Current HPC systems could benefit from locality of access for analysis operations, but neither HPC storage nor HPC analysis and visualization tools currently leverage this information. The HPC community can learn valuable lessons from the Google/Hadoop successes in this space.

      There is concern about contention between simulation output (checkpoint) and analysis, whether they are occurring on the same machine or in an active storage environment. HEC FSIO Quality of Service (QoS) work should provide some, perhaps partial, solutions, while additional work at the scheduling layer could alleviate this as well.

While it was not discussed in the breakout due to time constraints, indexing seems like a useful tool for reducing I/O demands during analysis. The role of heterogeneous storage (e.g., a mix of SSD and HDD) in analysis use cases was also not covered. Finally, storing data on compute nodes, such as in local SSD, and then rescheduling data analysis on these same nodes.

2010-2011 Analysis and Update

A panel consisting of Carlos Maltzahn from UCSC, Alok Choudhary from Northwestern, Dries Kimpe from Argonne National Laboratory and Raju Rangaswaami from Florida International University started off the discussion of Next Generation I/O Architecutre. Most panelists reviewed their current projects which generally advocated an evolutionary approach to modifying or replacing a layer of the existing software stack. Some general comments were clear; we seem to be working around or working to fix the existing file system or standards that many application programmers use such as POSIX. If application developers would use some higherlevel abstartction like MPI-IO, then possibly some of the work that has been done or in the works could benefit them. If POSIX is the interface, then change will be very slow if at all. Also, storage hierarchies are here and we must exploit them, but how will they be presented to the user.

In addition, some comments were made that we should look at records based I/O and that many of the issues that we face have been solved, albeit at reduced scale, in data bases.

2010 – 2011 Identified Gaps

No new gaps were identified this year.

Communications and Protocols

HEC FSIO 2005 Problem Definition

One of the most important factors in the resulting performance and functionality of a parallel file system is the communication protocol that ties the system together. Typically file system developers build their protocol from scratch on top of low-level networking protocols such as SCSI or IP and tune their protocol to match their architecture and expected workloads. However, much of the functionality in parallel file systems is similar across implementations. Understanding the key components of parallel file system communication and how new technologies fit into these protocols is critical to effective parallel file system designs in the future.

In the near term, research into most effective integration of networking technologies such as advanced network adapters and alternative low-level protocols will help make best use of upcoming networks. Integrating object-based storage concepts into I/O protocols will make for a more efficient mapping of client I/O operations to device operations. Understanding how to best leverage the new features of NFSv4 and 4.1 (including the pNFS capabilities) will be an important step in the direction of improved usability and lowered development costs.

As parallel file systems continue to evolve, communication between I/O servers begins to play an ever-greater role. Our understanding of this type of communication is very limited and better understanding of how this communication differs from client communication, how we might leverage aggregate communication concepts from message passing or group communication concepts from the peer-to-peer community could revolutionize storage architectures.

2005-2006 Progress

Some of these concepts are now being actively researched as a result of NSF HECURA funding:

- How collaborative caching might be best integrated into parallel I/O systems [Choudary 2006]
- The integration of active networks into the I/O stack, bringing a new communication paradigm into the picture [Chandy 2006]
- How servers might play a more active role in initiation of I/O operations [Sun 2006]
- The impact of network placement and migration [Thottlethodi]
- How active storage devices could be first-class citizens in a parallel file system [Wyckoff]
- Alternatives for server-to-server and client-to-client communication in parallel file systems [Ligon06]

HEC FSIO 2006 Analysis and Update

The projects listed above hit on many of the key problems in parallel file system protocols and the newly-created SciDAC2 Petascale Data Storage Institute is active in NFSv4 and pNFS efforts. However, a few areas were noted as needing additional coverage at the 2006 workshop. Improving the interface between applications and the underlying file system with respect to expected access patterns was seen as particularly critical. While some facilities for these types of hints are available on some systems, often this information is lost in the I/O stack well before it hits the parallel file system.

Along these same lines, providing mechanisms through which the file system can report more information back to the application was also seen as very important. This could be data that could be used to better align high-level data structures to storage, optimize network transfers, or better understand I/O system behavior. Finally, investigations into how to properly incorporate one-sided communication models (e.g. RDMA) into I/O systems was seen as a potential win in the long term.

2005 – 2006 Identified Gaps

- Expanding interface to file system in respect to expected access patterns
- Mechanisms for file system to report information back to the application
- Incorporating one-sided communication models into I/O systems

HEC FSIO 2006-2007 Analysis and Update

The area of communications and protocols was not directly addressed during the 2007 Workshop, but previously identified gaps are still open.

HEC FSIO 2007-2008 Analysis and Update

It was suggested that this roadmap be merged with the Next Generation I/O roadmap. Along with the potential of adding this to the Next Gen I/O roadmap, it was suggested that several new elements to the table or even open up an entire new solicitation. The new areas that were suggested were promote research, collaborate with the data intensive computing community, promote collaboration and exploitation of internet services and explore overlapping and commonality across the entire space. A few more points are listed below:

- Protocol performance at scale will be a major issue, so new protocols will be required.
- What about research in active networks?
- New ideas such as MapReduce, Hadoop and BigTable from Google.

HEC FSIO 2008-2009 Analysis and Update

This is the last funding year for the 2006 projects that fell under this area, and no new projects were funded in the 2009 HECURA Solicitation.

The area was not merged with Next Generation I/O after the 2008 meeting, and this year instead attendees identified a number of new gap areas, listed in order of importance (as measured by votes):

- Scalable abstractions for scientific data
- Scalable replication, relocation, failure detection, and fault tolerance
- Topology aware storage layout
- Wide area storage access protocols

The “Alternative I/O Transfer Schemes” area was considered a candidate for removal. Additionally, it was noted that it is particularly important to encourage development work specifically in protocols because the real impact of work in this area is often not seen until critical mass is achieved on adoption.

The proposed new area “Scalable Abstractions for Scientific Data” might include such things as blurring the line between memory and storage, non-POSIX interfaces, and APIs for small object manipulation (e.g., setattrlist). It was noted that such an area might fit as well under next-generation I/O, but it is not present in either area at this time. For now this topic is captured under next-generation I/O “storage abstractions”.

2009 – 2010 Analysis and Update
Communications and Protocols was not a research area targeted for discussion at this year’s workshop. Yet, the problems identified in the roadmap are still areas that need more investigation.

2010 – 2011 Analysis and Update
The year, the Communications and Protocols break out section started with a talk by Dr. John Chandy from the University of Connecticut reviewing his work on Active Networks and Active Object Storage. Other topics that were brought up for discussion during the talk were cloud storage, scalable abstraction for scientific applications, Inter-stack communication and Memory Hierarchy.

Low latency protocols were discussed as an area that needs attention. It’s recognized that products are out, but the operating system community is not taking advantage of the available hardware like IB.

Many participants felt that topology-aware storage is becoming more important with the latest round of supercomputers at the National labs such as Jaguar at Oak Ridge National Lab and Cielo at Los Alamos National Lab. This area was also brought up in Dr. Chandy’s talk under another name; Inter-stack Communications.

Cloud storage was brought up during the discussion as an area that needs investigation, but it’s not clear that the problems that cloud users/vendors experience have enough in common with the high end computing group. For example, latency is a concern for protocols on many high-end computers, but, in the cloud, there is no expectation that your data is available soon after storing it in the cloud. There are issues in cloud storage, such as protocols/standards for moving data between clouds, but most people felt that the overlap in problem areas between cloud storage and HEC storage are small at this time.

2010 – 2011 Identified Gaps
The area of “Scalable Abstractions for Scientific Data” came up again this year as an area that needs focused attention.

Management and RAS

HEC FSIO 2005 Problem Definition

Management and RAS (Reliability, Availability, and Serviceability) are both obvious areas affected by immense scale. The number of storage devices, associated hardware, and software needed to provide the needed scalable file system service in a demanding and mixed workload environment of the future will be extremely difficult to manage given current technology. Advances must be made in massive scale storage management to enable management survival with future file system deployments. Additionally, RAS at scale is another major issue. Given that future file systems will be based on tens of thousands or more mechanical devices with an extremely complicated software stack deployed at scale, it is likely that failure will be more the norm than the exception.

2005-2006 Progress

Several of the funded NSF HECURA projects are directly addressing this important challenge:

- An analysis of formal failure models for storage systems [Arpaci-Dusseau 2006]
- Improving scalability for HEC parallel file systems [Ligon06]
- Automated problem analysis of large scale storage systems [Narasimhan]

Additionally, there was significant progress in this area in terms of completed work. Bianca Schroeder published an analysis [Schroeder] of machine failures at Los Alamos National Lab using publicly provided data (http://institute.lanl.gov/data/lanldata.shtml) recently made available by Gary Grider's group at Los Alamos National Laboratory. Bianca's analysis makes important contributions and contradicts the previous conventional wisdom in terms of failure predictability. This leads itself into her current work in using this new understanding of failure patterns to improve HEC checkpointing. The availability of the Los Alamos data set should continue to pay similar dividends as other groups attempt different analyses.

HEC FSIO 2006 Analysis and Update

Although the workshop attendees were happy with the funded proposals and the work accomplished thus far, there was unanimous agreement that this problem space is particularly challenging and that large amounts of effort must continue to be expended in this space. Specifically, the group feels that HEC researchers would be well served by delving into the research in the modeling community. By emulating the complexity of those models, useful failures models could be developed to aid the HEC community research management and RAS. Further, the group agreed that more widespread dissemination and analysis of reliability information (such as the Los Alamos failure data) is particularly important.

Identified Gaps

- Dissemination of reliability information
- More formal failure analysis
- Semantics for asynchronous events completion
- Autonomics

HEC FSIO 2006-2007 Analysis and Update

This continues to be an area where much work and research needs to be applied. Management and RAS is comprised of a very difficult set of areas in which to make progress. We will have to look at some existing commercial solutions and those used in some of the "older" systems, where this has always been a major component of the system. There is a slight potential for getting some of the formalism included into POSIX or some other standards. If in fact these things are included in a standard, then the possibility of making them ubiquitous and accepted is much greater. Although the workshop attendees were happy with the funded proposals and the work accomplished thus far, there was unanimous agreement that this problem space is particularly challenging and that large amounts of effort must continue to be expended in this space. Specifically, the group feels that HEC researchers would be well-served by delving into the research in the modeling community. By emulating the complexity of those models, useful failures models could be developed to aid the HEC community research management and RAS. Further, the group agreed that more widespread dissemination and analysis of reliability information (such as the Los Alamos failure data) is particularly important.

In the 2007 workshop, these topics of management and RAS needs were discussed:

- Many of the issues from 2006 are still open
  - We have a very limited amount of reliability information
    - Can we start polling industry/academia/labs for more?
  - Still minimal "formalism" in failure analysis
    - Still ad-hoc in most cases, no formal methodologies
    - Without sufficient information, errors can not be properly analyzed
    - Can we fix the various drivers?

Identified Gaps

- Situational analysis
  - Where and when did the error(s) happen
  - When do humans get involved
  - Can we have "live" feeds from the system?
- Handling errors in virtualized systems
- Model of the system to analyze errors
  - Meaningful error messages
- Handle multiple leveled timeout propagation

HEC FSIO 2007-2008 Analysis and Update

The overall consensus was that the first three elements in the roadmap for Management and RAS might be better served by having their priorities raised. Below are many of the points that seem to be unresolved at this time.

- Reliability is still a serious issue, it has not been solved.

- Failure analysis is being researched by industry and they have started making some progress. (SMART)
- Analysis of power consumption in the exa-scale regime is not a primary focus of industry. They are in fact attacking the problem at levels they can implement.
- Reliability at scale and in other storage regimes is not well understood.
- What about taking reliability to the component parts such as the channel.
- Automated problem analysis also remains to be an unsolved problem.

HEC FSIO 2008-2009 Analysis and Update

Progress in this area is being made, as the HECURA 2006 projects are coming to a close soon, and new HECURA 2009 projects are being funded this year. This area is still in need of research, and large test-beds for the research to be performed on. Workshop participants were still very interested in seeing research done to figure out how we are going to manage and maintain the systems that are coming. The discussion touched on the fact that there has been some progress made, but we are still very far behind where we need to be to effectively manage storage systems in the coming years. Participants also saw a need to study on the Trade-Offs associated with building large systems, such as power, cooling, disk capacity, redundancy levels, and drive technologies.

2008-2009 Identified gaps as ordered by the workshop:

- Understanding the Failures we face: Current ones, and future technologies.
- Manual and Automated problem analysis and action, including creation of a standard API for error reporting.
- Proactive Health Methods: changing reliability over time
- Improved Scalability: managing all parts of the storage stack
- Reliability and degraded performance in HEC systems
- Power consumption and efficiency
- Formal Failure analysis and tools for storage systems

Proactive health methods is a new gap this year.

Proactive Health Methods are methods we can use to improve the data integrity and reduce unplanned outages of storage devices. For example, can we improve disk reliability by powering all disks off at regular intervals, or periodically power cycling redundant controllers. One example used in the industry now is disk scrubbing, or periodically reading data off all disks in a system to determine if data is valid, and drives are working.

The voting was mostly balanced across the first 5 areas, with Power Consumption and Formal Failure analysis receiving very little support from the attendees of the breakout session.

New HECURA 2009 project in the Management and RAS areas:

- Optimization Algorithms for Large-scale, Thermal-aware Storage Systems; [Khuller]

- Performance- and Energy-Aware HEC Storage Stacks; [Zadok 2009]

HEC FSIO 2009-2010 Analysis and Update

Since management and RAS is likely to be required to keep storage systems running at the exascale, the 2010 workshop held two breakout sessions on this topic; “Dealing with End-to-end” and “Leveraging Google Style Copies/Other Protection/Availability Scheme”.

From the end-to-end data protection discussion, “end-to-end” means protecting the data between inception and future use.

From the “thinking out of the box” corner came the idea that uncertainty quantification could be used to limit, or bound, data error or corruption in much the same way it is used in large simulation codes. The idea is that it will be applied in order that the application can gain a level of confidence in the storage and, so, avoid heavy error detection and correction. The very valid point was made that whether we attempt to guarantee data quality and ability to accurately retrieve it or give the application some idea about how confident it may be relative to the quality of the data the goal is the same; We want to “make sure the user is able to get his job done.”

We discussed standards-based protection schemes such as T10DIF, or now T10PI. Various schemes were discussed as solutions such as distributed checksum at the compute nodes with a later sum of the values. A number of qualifications and caveats were raised relative to end-to-end protection schema. For instance, it was noted that there does not seem to be a single solution that fits all cases. Another, particularly relevant, point that was made noted that not all data had the same value and that the value of stored data could change over time. This seemed particularly interesting if one considers that data could increase in value once stored and, so, implies we should consider how one might increase the protection on a data set at rest. The topic could be controversial at times. For instance there was a strong suggestion that the hardware should just be reliable and, nearly simultaneously, that we must allow for and accommodate unreliable hardware. Doing the latter will require error detection and correction at multiple levels in the stack, potentially very high in the stack and, potentially, within the application. It was noted, though, that we should do this judiciously. Having it everywhere adds cost but can be simply redundant. In any case, errors and corruption will still occur and we should cease to attempt to hide them from the application when they do. The application must become an involved, integrated part of the solution. Other thoughts on this topic included an expressed anxiety that we would end up with design targets only slightly in excess of hardware, thinking we won't survive a failure there, which could in turn generate similar thinking from the hardware folks. The end result, of which, could be an unfortunate downward ratcheting effect. Another worthwhile thought, mentioned but not explored, was whether there might be value in having multiple, duplicate (but not copies) paths through the IO stack; Similar to space-based solutions involving components from multiple vendors for the same task and a vote determining the result.

We discussed the cost/value proposition. Namely, can it become more expensive to run the job twice for a UQ-based solution for instance, than to detect/correct? Thoughts on this varied from a feeling that we must detect and correct anyway to a note that some machines are deterministic and might simply reliably repeat error. The question was posed asking whether application would actually leverage end-to-end protection. A very good answer to this question was the note that, ultimately, the researcher's reputation is on the line. They had better use it, then.

Ultimately, the breakout group seemed to agree that end-to-end will become increasingly important and that error detection must remain, to some degree. There did seem to be a consensus that a variety of mechanisms would be available to deal with the problem, at multiple levels. However, an integration of these methods would be a “doomed exercise” as doing so brings insurmountable issues such as layering exceptions within implementations of the software stack. Leveraging the power of these mechanisms will require active, direct participation from the applications. Finally, the uncertainty quantification bombshell cannot be ignored. The point that the goal is always the same motivates us to investigate whether some UQ effort might allow a less costly solution to the end-to-end protection problem.

From the “Leveraging Google Style Copies/Other Protection/Availability Scheme” breakout session, at a high level, we discussed which external tools could be leveraged by the HPC community. The discussion touched on which external tools could be immediately applicable and which are worth monitoring for future exascale systems.

The first candidate tool discussed was the Google File System (GFS) and especially its open source variant, the Hadoop Distributed File System (HDFS), and the programming paradigm of Map-Reduce (MR) which works with these file systems. Whether the contribution of these systems is the ability to expose and exploit data locality. A study was mentioned in which the read performance was not improved by having three copies instead of just one. Some participants argued that the bigger contribution was forcing programmers to decouple their tasks such that Map-Reduce could automatically provide failure handling. Decoupling tasks in this way also lowers the dependence on low latency networks.

This led into a discussion of whether HPC exascale would require new programming models. Recent discussion at other Exascale meetings has raised the possibility that exascale programming models would be more resistant to jitter than they are today. The question of how much jitter is tolerable led the participants into a larger discussion of whether GFS and HDFS have been successful because they have been co-designed specifically to MR and more generally to data-intensive computing. The assumptions behind these external tools must be understood to ensure that they are applicable to HPC workloads. Further study of HPC workloads are also needed to determine typical ratios of writes to reads.

GFS is also dropping triplication in favor of Reed-Solomon encodings. This encoding is directed from the clients. GFS, and HDFS, do not have unique clients and servers; instead, each machine in these systems is both a client and a server which is a model more similar to peer-to-peer systems than to HPC ones. Again, this is more possible for MR workloads which are not jitter sensitive, whereas in HPC, client-directed coordination is more challenging, especially when it is done asynchronously. The question of whether the replication scheme is client-directed or server-directed is influenced by whether there is more bandwidth available between clients and servers or between servers.

Finally, there was much discussion about co-design with the example given of how facebook has three different systems for their mem cache, hadoop, and mySQL workloads. Traditionally, HPC has a single compute-intensive supercomputer design and applications must shoe-horn themselves to it. Perhaps we can leverage some of these co-designs and perhaps we can leverage the example of MR and how it forced users to rewrite their applications such that a new, highly efficient, architectural design could be exploited. Data aware schedulers can be leveraged from external work; there is a lot of money and energy currently being spent in these communities and this large of an investment should be leveraged if possible. Quality of service is needed in both HPC as well as in MR.

Finally, these tools are still emerging and are still being actively developed. This means that now is the right time to get involved. If we donate time and energy now, our changes are much more likely to be accepted whereas if we wait until these tools are stable, then proposals to change them will be much less likely to be well-received. On a practical level, we are also more likely to entice new young people into HPC if we offer them a change to develop these exciting new tools instead of the more long-in-the-tooth set of existing HPC tools and file systems.

2010 – 2011 Analysis and Update

Management and RAS still continues to be a needed area of research, as many of the issues have not been solved as of yet. Being able to do proper management of a large storage system that will statistically be in a degraded state of some sort because of failed file system components is still very important to the HEC community.

Gaining a formal understanding of disk, controller and system failures continues to be a needed area of research. There is some research available for how Magnetic disks fail, but new technologies for magnetic disk’s will be available in the next few years, and these failure types need to be understood. Solid State Storage disks have been widely available for a few years now, and research into the data that is available is important.

One researcher wanted to know how we can properly ‘age’ a peta-byte of SSD’s properly so that we can understand how it will exhibit failures in a production system.

There was a large discussion about file system consistency in a large-scale deployment. Currently it takes many hours to days to check large deployments off line, with research clusters sitting idle while the check is done. The consensus of the group is that any large-scale storage system will be required to have an on-line health check system, or it will not be viable in a production environment.

Research into the power needs of various storage related components is currently in progress. This research can lead to savings in the power budget of large deployments. This power requirement is typically only 7% of the overall needs of a large computing center, and one dissenting remark came forward that pointed at much bigger targets for power savings, such as the memory to CPU power needs of about 50% of the system.

As we expect that Storage Systems will be in a degraded state at all times, vendors and computing centers need to understand clearly what this means to procurements and production benchmarks. Procurements will need to have requirements about speeds and access during degraded operation. Vendors will need to provide failure tolerances before performance degrades and data is lost. Research into how tolerance levels are evaluated will need to be done. It will also be good to have a good understanding of the difference between peak bandwidth and actual bandwidth in a system.

2010 – 2011 Identified Gaps

No new gaps were identified this year.

Archive

HEC FSIO 2005 Problem Definition

The 2005 workshop identified I/O and file system areas of particular concern to high-performance archives supporting HEC systems. The scale and longevity of data in HEC archives adds some particular slants to these research areas:

- The complicating factors of RAS at the very large scale are a foremost concern for deep archives on disk.
- Long-lived archives experience extremes in namespace size, making efficient storage, management, and retrieval of file system metadata imperative and research into new namespace technologies attractive. Content-addressable storage and similar technologies show promise in finding, tracking, and managing large archives over long periods, but more work is needed.
- The longevity of archive files makes more imperative the ability to set and enforce policy to manage the data. Policy is also important in the lifecycle movement of data among layers in the archive hierarchy.
- The X/DSM POSIX file system interface for offline data is more than ten years old; modern, high-performance archives call for a new generation replacement.
- Long-term archives must contend with migration to new generations of hardware; emerging technologies such as object-based storage architectures may be

particularly well suited for optimizing such movement and for enabling larger scale parallelism in the archive system.

2005-2006 Progress

During 2005-2006, the HECURA File System and I/O research awards supported efforts in a number of areas directly pertinent to archives that support high-end computing.

Security requirements for archive are addressed in the research on long term (data and key) management in a high-performance hierarchical archive environment [Du]. Another area focused on flexible object-driven policy for combinations of security and performance needs [Sivasubramaniam].

The more extreme RAS needs of archive systems are addressed by research into more complete failure and problem analysis, both from the viewpoint of more complete thus more insight into failure-handling problems [Arpaci-Dusseau 2006], as well as automating problem analysis [Narasimhan]. In addition, studies supporting monitoring, profiling, and analyzing large storage systems [Zadok 2006] should contribute both to enhanced RAS and improved archive performance.

Research into adaptive I/O stack frameworks, including multi-layer coordination, holds especial promise in improving coordination of resources and layers in hierarchical archives [Ma 2006].

Several awards focused on research with significant foci on metadata and scaling, including:

- Scalable, adaptive metadata management contending with access patterns, load balancing and caching in parallel and distributed file system environments [Jiang 2006]
- Investigating active caching, buffering, communication, and autonomics in support of scalable metadata operations, improvements in handling small unaligned data accesses, and auto-configuration [Ligon06]
- Exploring parallel versions of new methods that show promise for significant performance increases in indexing and scanning (meta)data [Bender06]

Many other research awards included aspects that are central to current and future concerns for archive. Support for scalable archive metadata operations would be encompassed in the exploration of the suitability of object based storage for parallel file systems [Wyckoff]. Offline techniques for deriving metadata [Maccabe] and support for filtering and other active functions “in the data path” [Maccabe, Wycoff, Reddy] could be used for data “scrubbing”, extraction, and transformation or conversion to new archive formats.

In addition, the collaborative and education opportunities provided by the two SCIDAC2 I/O projects [Gibson, Shoshani] and the LANL/UCSC educational institute on scientific data management will advance areas of interest to archive as will the efforts on HEC extensions to POSIX standards.

HEC FSIO 2006 Analysis and Update

At the 2006 workshop, the archive gap discussion focused on the following areas:

- Standard, transparent, interoperable means to deal with archives and archive file systems:
    - Standardizing archive attributes
    - Developing POSIX standards for archive interfaces, including policies on files and directories, as well as the ability to search part of an object
    - Possibility of developing a new standard VFS layer for archives, or some other standard means, that would provide user-transparent, interoperable ways to deal with archives and archive file systems as an alternative to modifying existing POSIX interfaces to accommodate special needs of archives
- Alternate access methods to archives and their components: these ranged from higher-level aspects contending with archive content (e.g., map reduction techniques such as those employed by the Google API) to lower-level functions such as alternate interfaces for accessing and addressing/naming storage hardware and objects
- Reliability is becoming paramount because of proliferation of devices in archives but RAS techniques often conflict with HEC needs; e.g., proactive, prophylactic device "scrubbing" is at odds with the need to minimize power consumption
- Growing interest in parallel archives, especially as they can be aligned with "Information Lifecycle Management" activities, from commercial HEC sites
- Automating the generation of archive attributes and content-metadata
- Management of versions and snapshots in archives including compression and navigation techniques

Identified Gaps

- Standardizing archive attributes and automated generation of archive attributes
- POSIX Standards for archive interfaces
- Standardized VFS layer for archive
- Alternative access methods involving indexing
- Long term disk device reliability
- Commercial parallel archives
- Managing versioning in archives

HEC FSIO 2006-2007 Analysis and Update

In the 2007 workshop, the archive discussion did not yield any new topics or gaps.

HEC FSIO 2007-2008 Analysis and Update

- The attendees felt that the HEC FSIO advisory group should change their views on archive metadata. While file system metadata scaling does cover many of the scaling needs for Archive, the types of metadata and the types of searches as well as the overall size of the archive metadata can be quite different from file systems. There is a place for research in this area. Also, exploiting the natural data movement from device to device in a living archive system should be exploited to help with metadata and indexing. This area is one that is not being capitalized on at HEC sites.

- ILM and related standards was also discussed. The fact that most ILM solutions are being done at user level each with its own API is presenting problems in that a lot of effort is being wasted interfacing to proprietary APIs. It is possible that the file systems standards area could help via extended attributes.

2007 – 2008 Identified Gaps

- Metadata in Archives
- ILM standards assistance

HEC FSIO 2008-2009 Analysis and Update

- There was some frustration that the HECFSIO group has overly limited the Archive area to be an interface and security only agenda, and the lines between file systems and archives are getting more blurry over time
- There was passion displayed for enabling knowledge discovery for usefulness of dedup for our environments. The concept was to add some fingerprinting in the Archive API to help us determine if dedup would be of value
- There was also passion about enabling the implementation of and use of extended attributes for use in search and management tied to file system extended attributes

2008 – 2009 Identified Gaps

- Need for Deduplication exploration
- More emphasis on extended attribute deployment and use
- Less emphasis on long term attribute driven security

2009 – 2010 Analysis and Update

Archive was not a research area targeted for discussion at this year's workshop. Yet, the problems identified in previous years, and identified in the roadmap, are still areas that need more investigation.

2010 – 2011 Analysis and Update

Attendees at the 2011 Archive Breakout did not explicitly re-evaluate 2010 Archive Gap areas, but instead used the opportunity to assess the current state of HEC archives and consider future directions. The discussion indicated that the HEC Archive area is at an inflection point.

The initial discussion focused on the distinguishing characteristics of HEC archives, particularly as compared to POSIX filesystems. For example, HEC archives that serve large Government laboratories and petroleum industries are primarily implemented to accommodate both vast quantities of data for which disk would be too costly and large amounts of infrequently accessed data. Government laboratories typically have longstanding HEC archives implemented via Hierarchical Storage Management Systems (HSMs) that use tape media in the lower tiers. Attendees discussed HSMs' shortcomings (and possible point solution approaches), including APIs, scheduling, policies, service requirements, and workflows.

Attendees recognized that HSMs comprise a small and fragmented niche market with high prices for development and maintenance. Government representatives expressed the desire to facilitate research into HEC-needed features that would be adopted by a broader market, and attendees discussed commercial developments that might be built-upon to make progress in closing HEC Archive gaps.

Attendees speculated that federally mandated compliance requirements for the finance and healthcare industries (e.g., Sarbanes-Oxley and Health Insurance Portability and Accountability Acts) should be leading vendors to develop policy and workflow architectures that might be adapted for HEC archive policy and workflow needs.

Attendees observed that development and acceptance of official standards was extremely slow, and discussed alternatives to formal standards. As one example, use of distributed repositories accessed via Representational state transfer (REST-ful) web services was presented as a possible alternative to POSIX filesystem interfaces in front of tape-based HSM archives.

It was noted that existing large commercial cloud repositories (e.g., Amazon S3) already contend with many of the same extremes that complicate evolving HEC archives, including enormous namespaces and accompanying metadata, requirements to support search at very large scales, and large counts of heterogeneous hardware and software elements. In light of this, attendees raised the question of whether cloud-like architectures could replace traditional HEC archives. Attendees felt that current cloud offerings' performance, RAS, and strategies for security and data longevity were not yet adequate for Government HEC requirements, but many felt that research investments in cloud-like solutions could make them viable to replace traditional HEC archive architectures.

It was suggested that the entire set of current HEC Archive Gap areas be shelved in favor of focusing research in cloud-like architectures so that these commercial offerings could be modified to support HEC needs. Several attendees opined that research efforts should not be restricted to a specific underlying architecture, and that current HEC archive issues should inform recommendations for research that enables future HEC archives. Attendees also expressed concern about how the small HEC community could influence mainstream providers to adopt HEC-related developments in their cloud-like architectures. Time restrictions prevented further discussion, but it was clear that cloud-like architectures merit further investigation in support of HEC archives.

2010 – 2011 Identified Gaps

The 2010 Archive Gaps were not re-evaluated at the 2011 workshop.

Assisting with Standards, Research and Education

Assisting Standards

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 standard which is currently in draft 1, nearing draft 2, specifies the interfaces to object storage devices. The HEC FSIO community has provided input and funding to various parts of this specification development and will continue to do so. One avenue of HEC FSIO involvement in this effort is the SciDAC2 Petascale Data Storage Institute.
- The IETF NFSv4 standard including the new pNFS portion of the NFSv4.1 minor revision of the NFSv4 specification. The HEC FSIO community has provided input and funding to various parts of this specification development. One of the prime sites for NFSv4 activity has been funded partially by DOE. The pNFS, the age old idea of separation of data and control, concept came directly from HEC needs and pressures in the late 1990's. The HEC FSIO team with the HEC FSIO community will continue to provide input and funding of various kinds to this effort. One avenue of HEC FSIO involvement in this effort is the SciDAC2 Petascale Data Storage Institute.
- 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. This effort initially is focusing on extending the POSIX I/O API to accommodate the needs of HEC, in particular for highly parallel/cooperating/concurrent applications. The HEC FSIO team and the HEC I/O community will continue this valuable work. One avenue of HEC FSIO involvement in this effort is the SciDAC2 Petascale Data Storage Institute.

The HEC FSIO group and the HEC I/O community will no doubt continue to make progress on affecting these and other standards efforts. At the both the HEC FSIO 2006 and 2007 workshops, an update was given on NFSv4.1/pNFS and the POSIX I/O API efforts. These presentations are available at the conference web site http://institute.lanl.gov/hec-fsio/workshops/

HEC FSIO 2005-2006 Analysis and Update

- Continued support of POSIX HEC Extensions efforts including prototype implementations
- Continued support for OBSD standards including keeping a reference implementation up to date
- Continued support for NFSv4, especially pNFS prototype work

HEC FSIO 2006-2007 Analysis and Update

- Continued support of POSIX HEC Extensions efforts including prototype implementations,

    - patch management and submission to be done by PDSI/CITI at the University of Michigan
    - revamp of man pages given input from Linux FS kernel developers will be done by PDSI/SDM/HEC
- Continued support for OBSD standards including keeping a reference implementation up to date
- Continued support for NFSv4, especially pNFS prototype work, and testing at scale.

HEC FSIO 2007-2008 Analysis and Update

Little new was discussed at the 2008 workshops except for the need for assistance with the ILM standards from the file systems community.

HEC FSIO 2008-2009 Analysis and Update

Little new was discussed at the 2009 workshops. pNFS seems to be moving forward, albeit a bit slowly from the Linux Kernel server side point of view. Additionally, a new file layout query call has made it into the POSIX file system API.

2009 – 2010 Analysis and Update

An update on progress made towards and time line for the introduction of pNFS was given at this year's workshop. In addition, some of the HEC POSIX extensions that were proposed several years ago have been introduced into various Linux kernels. From the "Futures" panel at this year's workshop, panelists feel that we may have an opportunity for a new api in replacement to or in addition to POSIX.

2010 – 2011 Analysis and Update

No updates on standards were discussed.Assisting Research and Education

In addition to recommended research topics discussed in detail in the previous sections, the HEC FSIO 2005 also called for the HEC FSIO community to provide university I/O center support in the forms of computing and simulation equipment availability, availability of operational data to enable research, and HEC involvement in the educational process were called out as areas needing assistance. The following sections discuss things that have been done in the past year in these areas as well as the input received in these areas at the HEC FSIO 2006 Workshop.

HEC FSIO 2005-2006 Analysis and Update

- Availability of computing equipment
- Availability of simulation tools
- Availability of trace and performance data
- Availability of failure and RAS data

HEC FSIO 2006-2007 Analysis and Update

- Availability of computing equipment, Incite and NSF infrastructure has helped but there is still a serious need for an at-scale FSIO/computer science test facility that would allow for disruptive testing
- Availability of simulation tools
- Availability of trace and performance data

HEC FSIO 2007-2008 Analysis and Update

- Availability of computing resources seems to not be a solved problem yet
- Availability of simulation tools
- Availability of highly documented operational data
- Standards and methods rewarding data release

HEC FSIO 2008-2009 Analysis and Update

- Availability of computing resources seems to not be a solved problem yet, donating equipment to universities came up
- Availability of simulation tools is still a problem but the Clemson simulator may address this to some extent
- Suggestions concerning creating a HEC storage module for graduate education and maybe undergraduate. Creation of a reading list for a graduate storage course was also suggested. Additionally, writing a text book came up but no consensus in support of this was garnered. A suggestion of sites sharing horror stories in HPC storage would be of great use.
- Paying for students, as many as possible, was unanimously supported

2009 – 2010 Analysis and Update

Assisting research and education was not a focus of the 2010 workshop, but remains and important issues for the FSIO community.

2010 – 2011 Analysis and Update

Assisting research and education was not a focus of the 2011 workshop, but remains and important issues for the FSIO community.

Availability of failure and RAS data

HEC FSIO 2005 Problem Definition

At the HEC 2005 workshop, participants identified access to central clearinghouses of HEC data as necessary to understand and make progress on problems seen at HEC sites. Data that would be useful to these researchers include trace data from real HEC applications, synthetic applications that approximate the behavior of real HEC applications, historical data about failure rates of HEC systems, and very basic machine and environment information.

Challenges in providing this data are both political and technical. Politically, it may be impractical to provide data which may be classified to outside entities. Technical challenges in providing traces or synthetic applications involve the level of detail to provide. As the focus of this workshop is on HEC I/O, one simple answer would be to provide data only about the file system activity of the applications. However, this detail may be insufficient as other aspects of the application's behavior may influence their interaction with the file system. For example, a trace of an HEC application run on a machine with a large buffer cache may show less file system activity due to caching than the same application run on a machine with a smaller buffer cache. Additionally, storage and distribution solutions for this clearinghouse data need to be designed and implemented.

2005-2006 Progress

In response to the requests made for data from the HEC 2005 workshop, substantial progress has been made with the public release of data and more is in progress. Los Alamos National Lab has released general machine information, failure, event, and usage data for 22 of their premier supercomputers from the past nine years. In some cases, the records released cover the entire life of the machines from initial use to decommissioning. In addition to releasing over 23,000 failure, event, and usage records, LANL released some papers on computer failures and are publicly available at http://institute.lanl.gov/data/lanldata.shtml.

LANL researchers also made available to the public a synthetic application that mimics the I/O profiles of many of their important, but classified, applications. The synthetic benchmark is available at http://public.lanl.gov/jnunez/ .

HEC FSIO 2006 Analysis and Update

During the 2006 workshop, two presentations were given on two technical reports from Carnegie Mellon derived from the released data and the synthetic application. Bianca Schroeder presented joint work with Garth Gibson on their technical report "A Large-scale Study of Failures in High-performance-computing Systems" [Schroeder] and Mike Mesnier presented "//TRACE: Parallel Trace Replay with Approximate Causal Events" [Mesnier]. During this session, the formation of the Computer Failure Data Repository was announced as a central place to store and make available failure, usage, and event data and contain pointers to sites that already make this data available. In addition, it was announced that there are several other organizations that are planning to release data including HP Labs, the Library of Congress, and Pittsburgh Supercomputing, and that LANL will be releasing synthetic and real application I/O traces.

After the presentations, the Data Availability discussion was broken up into two groups: one focusing on failure, usage, and RAS data and the other on traces and performance data.

In the area of traces and performance data, participants identified the following as gaps in making forward progress:

- lack of analysis tools
    - Instead of writing new analysis tools, we can leverage others efforts by reviewing existing tools and look into extending these existing tools to apply to HEC I/O.
- set of benchmarks that exhibit certain categories of behavior

    - Applications are evolving, which change their I/O patterns. Benchmarks that are general or cover a category of behavior has more of a chance of being relevant than very specific benchmarks.
- specify and use common formats**,** collection programs, obfuscation programs
    - Trace formats must be widely extensible because the number of things to trace will evolve over time with experience and requests from researchers using the traces.
    - There is a tremendous amount of data that should be captured that make the traces and benchmark results more relevant. System-wide profiler-like tools are needed to capture metadata describing traces in terms of application domain, number of procs, etc.
- common clearinghouse for traces, synthetics, benchmarks, and performance data
    - Consolidate the number of sites that provide data or provide pointers to those sites in one location
    - The owner of the repository should be neutral. Both SNIA and USENIX have expressed interest or put effort in this area and Los Alamos National Lab and Carnegie Mellon University have data repositories

In the area of failure, usage and RAS data availability, besides expressing concern over the anonymity of the data and how it will be used by vendors and their marketing, participants identified the following as gaps in making forward progress:

- need for standards, best practices, and guidance on what and how to capture relevant data
- evaluate existing models
    - Common trace formats etc. from the commercial computing world
    - Can we learn from the medical industry on anonymous data
    - Commercial autonomics work
- work with cluster vendors/industry/users to capture any of this data from the thousands of clusters in the world
- Sell this as a pre-competitive research topic
- Engage the vendors on how do we get proactive failure handling
- There may be no way to ever get vendors to give this stuff away

HEC FSIO 2005-2006 Analysis and Update

- Availability of data from HEC sites
- Lack of analysis tools
- Need for standards, best practices, and guidance on what and how to capture relevant data

HEC FSIO 2006-2007 Analysis and Update

- Progress has been made on the availability of data from HEC sites, but more is needed, especially trace data and related information

- Lack of analysis tools to analyze the data
- Need for standards, best practices, and guidance on what and how to capture relevant data

HEC FSIO 2007-2008 Analysis and Update

- Progress has been made on the availability of data from HEC sites, but more is needed, especially trace data and related information, especially well documented data
- Lack of analysis tools to analyze the data
- Need for standards, best practices, and guidance on what and how to capture relevant data
- Need for a way to reward data release via citation etc.

HEC FSIO 2008-2009 Analysis and Update

- New data releases were shared from HEC sites

There was discussion about the need for common trace formats. Many felt this was just not that important, but this was not unanimous

2009 – 2010 Analysis and Update

No new data releases were announced at this year's workshop.

2010 – 2011 Analysis and Update

Both Argonne and Los Alamos National Labs announced that they have or will release new trace data and updated machine failure data.

Education, Community, and Center Support

HEC FSIO 2005 Problem Definition

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.

2005-2006 Progress

There were a few success stories within the HEC FSIO community to help address this important need of education, community, and center support.

- Probably the most notable of the accomplishments was the NSF/HECURA research funding effort in the FSIO area. This is one of the first government funding efforts directed at the FSIO area and sent a signal to universities and vendors that the HEC FSIO effort is serious and capable of producing benefits.

- The SciDAC2 FSIO awards both have outreach to universities and industry in their scope. The PDSI SciDAC Institute calls out directly working with universities to raise awareness and understanding of the FSIO problems and to encourage advancement of curricula and other educational endeavors in the HEC FSIO area.
- An educational institute for scalable scientific data management was put in place by LANL with UCSC. This institute provides funding for shaping curricula and student fellowships in the area of HEC FSIO.
- Of course, there were numerous joint papers and other joint research done between HEC sites, vendors, and universities as well.

HEC FSIO 2005 - 2006 Analysis and Update
At the 2006 workshop, this topic was broken up into two areas: curricula needs and center/community support.

Curricula needs
A number of very good ideas were brought out in this area. The top ideas are as follows:

- HEC site and industry provided lectures in the class room
- Providing access to large scale computing resources for curricula work (this topic is covered in more detail in the next section about availability of computational resources
- Find a way to harness the gaming industry excitement in the curricula program
- More HEC site and industry internships
- Flashy K-12 and undergraduate shows and tours of HEC sites
    - It would be very helpful if decommissioned equipment could be sent to high schools and even middle schools so that youth could see the inside of a disk drive, or see a working tape robot or other interesting show and tell items. Most high school students have never seen the inside of a computer, or a disk drive, or a tape cartridge. This activity might generate interest in pre-college and even undergraduate students.
- Curricula development funding to assist in
    - Developing new storage or storage centric courses
    - Developing materials, problem sets and answers, text book
    - Inject storage topics into existing classes
    - National Lab/Industry input/endorsement of this curricula building activity

Center/Community support
A number of very good ideas were brought out in this area. The top ideas are as follows:

- Data to study (addressed before in data availability section)
- Large testbed environments and simulation environments to be used for research (addressed in availability of computational resources section)
- Promote a separate research storage related program at NSF that is ongoing, not just a one shot HECURA program.
- Provide more industry/HEC site internships.
- Conduct marketing of the importance of the FSIO and HEC FSIO efforts

    - Come up with an appealing name
    - Provide some awards or recognition for achievement in this field
    - Industry luminary speaking on a popular media like NOVA
    - Exploit ACM and other societies, job fairs, and journals like ACM Queue
    - Flashy technology, how it works, shows at K-12 and undergraduate level
    - Come up with a clean message as why storage/IO matters
    - Engage the Other Discipline (MIS, Business, CE, EE, etc) as part of the community to gain broader recognition
    - Use technology roadmaps as motivators
- Fully exploit NSF and other educational funding opportunities.
- Promote a Grand Challenge for information management.
- Engage the ACM K-12 Program, promote programming contests etc.
- Sponsor a top ten open problem in storage
- Lobby industry leaders to lobby government funding sources/agencies for a sustained investment in the storage and FSIO area.

2006 Identified Gaps

- HEC site and industry provided lectures in the class room.
- Find a way to harness the gaming industry excitement in the curricula program.
- More HEC site and industry internships.
- Flashy K-12 and undergraduate program including shows and tours of HEC sites.
- Promote a separate research storage related program at NSF that is ongoing, not just a one shot HECURA program.
- Conduct marketing of the importance of the FSIO and HEC FSIO efforts
- Promote a Grand Challenge for information management.
- Sponsor a top ten open problem in storage
- Lobby industry leaders to lobby government funding sources/agencies for a sustained investment in the storage and FSIO area.

HEC FSIO 2006 - 2007 Analysis and Update

This area was not covered in any depth at HEC FSIO 2007. One site, LANL, pushed forward on their institute concept with two institutes in the FSIO area.

2007 Identified Gaps

- Should approach NSF about an ongoing investment in FSIO.

HEC FSIO 2007-2008 Analysis and Update

The introduction of the HEC FSIO Roadmaps as a way to manage R&D portfolio was introduced. Additionally NSF has committed to a follow on HECURA call in FY09 to keep the R&D pipeline full for the next few years.

HEC FSIO 2008-2009 Analysis and Update

This area was not covered at any depth at HECFSIO 2009. However, it was agreed upon that the HECFSIO Conference is an excellent community building function and perhaps that is one of its most important attributes.

2009 – 2010 Analysis and Update
The SciDAC2 FSIO Petascale Data Storage Institute ended this year and the SciDAC2 Scientific Data Management Center is scheduled to end in FY11. If both these projects are not renewed, there is a gap in helping applications with their I/O issues, with FSIO education and outreach.

2010 – 2011 Analysis and Update
This area was not covered at any depth at the workshop, but the future state of the HEC FSIO Workshop was discussed. Most attendees agreed that the HEC FSIO Workshop is an excellent community building function, but with NSF discontinuing the HECURA program, the workshop will have no purpose to continue.

Availability of Computational Resources

HEC FSIO 2005 Problem Definition
One of the most frequently echoed problems expressed at the 2005 workshop was the difficulty faced by many researchers; lack of access to real HEC applications and computing platforms. The researches felt that government investment is needed to support efforts that allow them access to both the physical and virtual infrastructure that they need in order to participate in and conduct HEC FSIO research.

Many would-be HEC researchers, particularly those in academia, lack access to the large parallel computers typically used in HEC applications. Therefore, the development and maintenance of open testbeds would enable these researchers to contribute in the area of HEC. There are at least four possible ways in which this physical infrastructure could be provided.

- The first is through direct acquisition: by providing large amounts of funding, new testbeds could be directly purchased. However, the cost of purchasing very large parallel systems could be prohibitively expensive.
- A second approach would be to develop sharing and sandboxing mechanisms by which “guests” could use the computational resources at other institutions. The sharing mechanisms would need to ensure appropriate priority schemes such that the guests would not pre-empt the compute time of the presumably more important local users.
- A third approach is using a virtual machine such that one computer architecture is made to appear like another through the use of complex software. One disadvantage of early virtual machine systems was a significant performance cost, but more recent work in this area has reduced this penalty. To date however, there is no virtual machine system for large parallel architectures; as such, this could be a project worthy of government investment. However, creating a parallel architecture from a physical architecture which is not parallel is problematic and it could probably only be done with large performance penalties. However, projects dealing with non-performance related aspects such as fault tolerance, functionality, and correctness would be feasible.

- A forth approach is developing *simulation platforms*. With realistic timing models, these simulation platforms could even provide realistic performance evaluations. However, simulated systems cannot run unmodified applications as can virtual machines. Simulations can provide insight into future system development by exploring trends and simulating systems that do not currently exist.

2005-2006 Progress

There were a few success stories within the HEC FSIO community to help address the availability of computational resources to enable research. Three ways for university researchers to gain access to large scale HEC computational resources were provided.

- The DOE Office of Science 2007 INCITE Program - Now in its fourth year, the Innovative and Novel Computational Impact on Theory and Experiment (INCITE) program using ORNL/LBNL/PNNL/ANL world class resources can be sought by researchers.
- The NSF has a research infrastructure funding program that can be applied for by researchers.
- Also, there is always the possibility of collaboration with HEC sites such as national labs on jointly interesting problems. Frequently, this results in getting university researchers access or exposure to large computational resources. The SciDAC2 teams would be good candidates for such collaboration.

HEC FSIO 2006 Analysis and Update

While this topic was not discussed directly at the 2006 workshop, it did come up in several settings, especially in the area of center/community support and curricula needs breakout sessions.

The following were the most prevalent comments.

- The INCITE program might work pretty well for simulation and modeling activities but given the non mainstream computing resources being offered and the lack of system privileges available, the INCITE program may not be a very complete answer.
- The NSF infrastructure grants could help a small amount
- There is a need to get access both for research and for educational/class activities.
- It would be very helpful if decommissioned equipment could be sent to high schools and even middle schools so that youth could see the inside of a disk drive, or see a working tape robot or other interesting show and tell items. Most high school students have never seen the inside of a computer, or a disk drive, or a tape cartridge. This activity might generate interest in pre-college and even undergraduate students.

2006 Identified Gaps

- Availability of computational resources for research (destructive) and educational (non-destructive) activities
- Disk and system simulators
- Virtual parallel machines

2007 Identified Gaps

- Availability of computational resources for research (destructive) and educational (non-destructive) activities, this the more important recurring need
- Disk and system simulators; the Ligon HECURA work should help with this
- Virtual parallel machines; the PDSI/PNL - Felix/Farber work presented at the workshop should help some with this.

HEC FSIO 2007-2008 Analysis and Update

- Availability of computational resources for research (destructive) and educational (non-destructive) activities, this the more important recurring need
- Disk and system simulators; the Ligon HECURA work should help with this

HEC FSIO 2008-2009 Analysis and Update

- Availability of computational resources for research (destructive) and educational (non-destructive) activities is the important recurring need.
- Donation of equipment to universities was offered up as an idea
- Disk and system simulators; the Ligon HECURA work could help with this but this is a strong recurring theme.

2009 – 2010 Analysis and Update

There were no new compute resources for the community announced.

2010 – 2011 Analysis and Update

The NSF-sponsored Parallel Reconfigurable Observational Environment (PRObE) project was announced with the project aimed at providing a large-scale low-level system research facility. PRObE is a collaborative effort by the New Mexico Consoritum, Los Alamos National Lab, Carnegie Mellon University, the University of Utah and the University of New Mexico.

Research Outcome to Industry

Given that many of the HECURA research projects will be nearing their end in 2008-2009, there must be a way to track the progress of research-proven ideas into industry Thus, this report will have this new section each year from 2007 on discussing research that has made it into industry. Reporting of research ideas into industry will be nontrivial, as industry is not always free to disclose future technical direction.

New items for 2007 include

- In the area of research outcomes to industry
    - One HECURA project is becoming a small business with help from DOE to commercialize partially funded HECURA outcomes.
    - One HECURA project is having great success working with industry to accept HECURA outcomes in product.

Nothing new to mention for 2008.

Nothing new to mention for 2009.

2009 – 2010 Analysis and Update

In 2010, we polled all the past and present HECURA and CPA PIs to collect statistics on number of graduated students supported by HECURA/CPA projects, number of publications, etc. Of the PIs that responded, we found

- published around 250 conference and journal proceedings
- contributed 13 open-source packages
- funded at least some portion of time for over 50 professors and 100 students
- produced 16 PhD and 22 Masters graduates
    - 27 of these graduates have brought their expertise in these problems of national interest to industry
    - 4 to government
    - 11 to academia–ensuring that future generations of students will either continue helping in this area or at least have some familiarity with it.
- Additionally, at least three commercial companies have emerged from these projects, providing valuable and supported products as a return to the community.

2010 – 2011 Analysis and Update

- There were no updates announced for research outcome to industry.

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.
The following is a summary of some of the accomplishments in the HEC FSIO area:

- The publishing of the document, “HPC File Systems and Scalable I/O: Suggested Research and Development Topics for the Fiscal 2005-2009 Time Frame” [Appendix C]and the designation of file systems and I/O as a national focus area beginning in FY06 laid the framework for the HEC FSIO 2005 Workshop. This workshop helped identify, categorize, and prioritize the needed research in this area of HEC. Using the research topics document and the HEC 2005 Workshop document, the HEC organization and the HEC I/O community made major progress in the 2005-2006 timeframe.
- 23 HECURA awards were made from a careful analysis of the proposals and the DOE Office of Science awarded two SciDAC2 FSIO projects.
- The LANL release of failure, event, and usage data began a trend for more sites to provide research data
- The HEC FSIO 2006 workshop in August 20-22 in Washington DC introduced the 23 HECURA and two SciDAC2 FSIO research activities, programs available to get computing resources, and activities to make operational data available to enable research.
- The HEC FSIO 2007 workshop showcased the 23 HECURA projects and the two SciDAC FSIO projects, introduced the five CPA FSIO projects, gave a standards update, described how the program will expand to track the migration of research to production, and collected gaps for future research needs.
- The HEC FSIO 2008 workshop showcased the 23 HECURA projects, the two SciDAC FSIO projects, the seven CPA FSIO projects, described how the program has expanded to track the migration of research to production, and collected gaps for future research needs. A follow on to HECURA with a solicitation in FY09 was announced by NSF.
- The HEC FSIO 2009 workshop introduced HECURA 21 projects and attendees heard from most of the HECURA PIs.
- The HEC FSIO 2010 workshop focused on HECURA projects that potentially will help with the scale and issues that will arise from exascale compute systems. Breakout sessions were held to solicit attendee’s thoughts on new architectures and on leveraging the work being done in data intensive computing.

The HEC FSIO activities have accomplished a great deal, but much work remains.

The HEC FSIO team thanks the workshop coordinators and participants for helping to conduct and participating in each of the HEC FSIO workshops. The workshops have been extremely successful and useful to help the HEC FSIO technical advisory team make recommendations to the HECIWG on coordination of R&D in this important area.

The HEC FSIO community would like to thank the NSF, DOD, DARPA, and DOE for making the HECURA, CPA, and SciDAC2 FSIO awards possible.

The HEC organization would also like extend a special thank you to:

- Almadena Chtchelkanova at NSF,
- Fred Johnson, Thuc Hoang, Robert Meisner at DOE

for their unwavering support of the HEC FSIO activity.

References

[Arpaci-Dusseau06] Arpaci-Dusseau, Remzi H., Andrea C. Arpaci-Dusseau, Benjamin R. Liblit, Miron Livny, and Michael M. Swift. “Formal Failure Analysis for Storage Systems.” High End Computing University Research Activity NSF 06-503 (2006)

- Andrew Krioukov, Lakshmi N. Bairavasundaram, Garth R. Goodson, Kiran Srinivasan, Randy Thelen, Andrea C. Arpaci-Dusseau, Remzi H. Arpaci-Dusseau. "Parity Lost and Parity Regained," FAST 2008, 2008, p. 127.

- Cindy Rubio-Gonzalez, Haryadi S. Gunawi, Ben Liblit, Remzi H. Arpaci-Dusseau, and Andrea C. Arpaci-Dusseau. Error Propagation Analysis for File Systems. In Proceedings of the ACM SIGPLAN 2009 Conference on Programming Language Design and Implementation (PLDI '09), Dublin, Ireland, June 2009

- Cindy Rubio-Gonzalez, Haryadi S. Gunawi, Ben Liblit, Andrea C. Arpaci-Dusseau, Remzi H. Arpaci-Dusseau. "Error Propagation Analysis for File Systems," Proceedings of the 2009 Conference on Programming Language Design and Implementation (PLDI '09), v.1, 2009, p. 1.

- Haryadi Gunawi, Abhishek Rajimwale, Andrea Arpaci-Dusseau, Remzi Arpaci-Dusseau. "SQCK: A Declarative File System Checker," OSDI, 2008.

- Haryadi S. Gunawi, Cindy Rubio-GonzÃ¡lez, Andrea C. Arpaci-Dusseau, Remzi H. Arpaci-Dusseau, Ben Liblit. "EIO: Error Handling is Occasionally Correct," FAST 2008, 2008, p. 207.

- Lakshmi Bairavasundaram, Meenali Rungta, Andrea Arpaci-Dusseau, Remzi Arpaci-Dusseau. "Limiting Trust in the Storage Stack," StorageSS 2006, 2006, p. 37.

- Lakshmi N. Bairavasundaram, Garth R. Goodson, Bianca Schroeder, Andrea C. Arpaci-Dusseau, Remzi H. Arpaci-Dusseau. "An Analysis of Data Corruption in the Storage Stack," FAST 2008, 2008, p. 223.

- Lakshmi N. Bairavasundaram, Swaminathan Sundararaman, Andrea C. Arpaci-Dusseau, Remzi H. Arpaci-Dusseau. "Tolerating File-System Mistakes with EnvyFS," Proceedings of the Usenix Annual Technical Conference (USENIX '09), v.1, 2009, p. 1.

- Lakshmi N. Bairavasundaram, Meenali Rungta, Nitin Agrawal, Andrea C. Arpaci-Dusseau, Remzi H. Arpaci-Dusseau, and Michael M. Swift. Systematically Benchmarking the Effects of Disk Pointer Corruption. In DSN '08, Anchorage, Alaska, June 2008

- Sriram Subramanian, Yupu Zhang, Rajiv Vaidyanathan, Haryadi S. Gunawi, Andrea C. Arpaci-Dusseau, Remzi H. Arpaci-Dusseau, Jeffrey F. Naughton. "Impact of Disk Corruption on Open-Source DBMS," 26th International Conference on Data Engineering (ICDE '10), v.1, 2010, p. 1.

- Swaminathan Sundararaman, Sriram Subramanian, Abhishek Rajimwale, Andrea C. Arpaci-Dusseau, Remzi H. Arpaci-Dusseau, Michael M. Swift. "Why panic()? Improving Reliability with Restartable File Systems," Workshop on Hot Topics in Storage and File Systems (HotStorage '09), v.1, 2009, p. 1.

- Swaminathan Sundararaman, Sriram Subramanian, Abhishek Rajimwale, Andrea C. Arpaci-Dusseau, Remzi H. Arpaci-Dusseau, Michael M. Swift. "Membrane: Operating System Support for Restartable File Systems," Proceedings of the 8th Conference on File and Storage Technologies (FAST '10), v.1, 2010, p. 1.

- Vinod Ganapathy, Arini Balakrishnan, Mike Swift, Somesh Jha. "Microdrivers: A New Architecture for Device Drivers," HotOS 2007, 2007, p. 85.

- Yupu Zhang, Abhishek Rajimwale, Andrea C. Arpaci-Dusseau, Remzi H. Arpaci-Dusseau,. "End-to-end Data Integrity for File Systems: A ZFS Case Study," Proceedings of the 8th Conference on File and Storage Technologies (FAST '10), v.1, 2010, p. 1.

[Arpaci-Dusseau09] Arpaci-Dusseau, Remzi and Andrea Arpaci-Dusseau. “HaRD: The Wisconsin Hierarchically-Redundant, Decoupled Storage Project.” High End Computing University Research Activity NSF 09-530 (2009)

[Bender06] Bender, Michael A. and Martin Farach-Colton. “Techniques for Streaming File Systems and Databases.” High End Computing University Research Activity NSF 06-503 (2006)

- Fernandez Anta, M. A. Mosteiro, and C. Thraves. "Deterministic Communication in the Weak Sensor Model," Proceedings of the 11th International Conference On Principles Of Distributed Systems (OPODIS), 2007, p. 119.

- Arkin, EM; Bender, MA; Demaine, ED; Fekete, SP; Mitchell, JSB; Sethia, S. "Optimal covering tours with turn costs," SIAM JOURNAL ON COMPUTING, v.35, 2006, p. 531-566.

- Arkin, EM; Bender, MA; Fekete, SP; Mitchell, JSB; Skutella, M. "The Freeze-Tag Problem: How to wake up a swarm of robots," ALGORITHMICA, v.46, 2006, p. 193-221.

- Bender, CM; Bender, MA. "Optimal shape of a blob," JOURNAL OF MATHEMATICAL PHYSICS, v.48, 2007.

- Bender, MA; Clifford, R; Tsichlas, K. "Scheduling algorithms for procrastinators," JOURNAL OF SCHEDULING, v.11, 2008, p. 95-104.

- Bender, MA; Farach-Colton, M; Mosteiro, MA. "INSERTION SORT is O(n log n)," THEORY OF COMPUTING SYSTEMS, v.39, 2006, p. 391-397.

- Bender, MA; Fineman, JT; Gilbert, S. "Contention resolution with heterogeneous job sizes," ALGORITHMS - ESA 2006, PROCEEDINGS, v.4168, 2006, p. 112-123.

- Bille, P; Farach-Colton, M. "Fast and compact regular expression matching," THEORETICAL COMPUTER SCIENCE, v.409, 2008, p. 486-496.

- M. Bender and M. A. Bender. "What Is the Optimal Shape of a Blob?," Journal of Mathematical Physics, v.48, 2007, p. 073518.

- Farach-Colton, M; Fernandes, RJ; Mosteiro, MA. "Lower bounds for clear transmissions in radio networks," LATIN 2006: THEORETICAL INFORMATICS, v.3887, 2006, p. 447-454.

- Farach-Colton, M; Fernandes, RJ; Mosteiro, MA. "Bootstrapping a Hop-Optimal Network in the Weak Sensor Model," ACM TRANSACTIONS ON ALGORITHMS, v.5, 2009.

- K. Agrawal, M. A. Bender, and J. T. Fineman. "The Worst Page-Replacement Policy," Theory of Computing Systems, v.44, 2009, p. 175.

- L. Arge, M. A. Bender, E. D. Demaine, B. Holland-Minkley, and J. I. Munro. "Cache-Oblivious Priority Queue and Graph Algorithm Applications," SIAM Journal on Computing, v.36, 2007.

- M. A. Bender and C. A. Phillips. "Scheduling DAGs on Asynchronous Processors," Proceedings of the 19th ACM Symposium on Parallelism in Algorithms and Architectures (SPAA), 2007.

- M. A. Bender and C. A. Phillips. "Scheduling DAGs on Asynchronous Processors," Proceedings of the 19th ACM Symposium on Parallelism in Algorithms and Architectures (SPAA), 2007, p. 35.

- M. A. Bender and H. Hu. "An Adaptive Packed-Memory Array," Proceedings of the 25th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems (Winner: Best Newcomer Award), 2006.

- M. A. Bender and H. Hu. "An Adaptive Packed-Memory Array," Transactions on Database Systems, Special Issue on PODS '06, v.32, 2007.

- M. A. Bender and H. Hu. "An Adaptive Packed-Memory Array," Proceedings of the 25th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems (Winner: Best Newcomer Award), 2006, p. 20.

- M. A. Bender, B. Bradley, G. Jagannathan, and K. Pillaipakkamnatt. "Sum-of-Squares Heuristics for Bin Packing and Memory Allocation," Journal of Experimental Algorithms, v.12, 2007.

- M. A. Bender, D. Ge, S. He, H. Hu, R. Y. Pinter, S. Skiena, and F. Swidan. "Improved Bounds on Sorting with Length-Weighted Reversals," Journal of Computer and Systems Sciences, v.74, 2008, p. 744.

- M. A. Bender, D. P. Bunde, E. D. Demaine, S. P. Fekete, V. J. Leung, H. Meijer, and C. A. Phillips. "Communication-Aware Processor Allocation for Supercomputers: Finding Point Sets of Small Average Distance," gorithmica, Special Issue on WADS '05, v.50, 2008, p. 279.

- M. A. Bender, G. S. Brodal, R. Fagerberg, R. Jacob, and E. Vicari. "Optimal Sparse Matrix Dense Vector Multiplication in the I/O-Model," Proceedings of the 19th ACM Symposium on Parallelism in Algorithms and Architectures (SPAA), 2007.

- M. A. Bender, G. S. Brodal, R. Fagerberg, R. Jacob, and E. Vicari. "Optimal Sparse Matrix Dense Vector Multiplication in the I/O-Model," Proceedings of the 19th ACM Symposium on Parallelism in Algorithms and Architectures (SPAA), 2007, p. 61.

- M. A. Bender, J. T. Fineman, and S. Gilbert. "Contention Resolution with Heterogeneous Job Sizes," Proceedings of the 14th Annual European Symposium on Algorithms (ESA), 2006.

- M. A. Bender, J. T. Fineman, and S. Gilbert. "A New Approach to Incremental Topological Ordering," Proceedings of the 20th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA), 2009, p. 1108.

- M. A. Bender, J. T. Fineman, and S. Gilbert. "Contention Resolution with Heterogeneous Job Sizes," Proceedings of the 14th Annual European Symposium on Algorithms (ESA), 2006, p. 112.

- M. A. Bender, M. Farach-Colton, B. C. Kuszmaul. "Cache-Oblivious String B-Trees," Proceedings of the 25th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems (PODS), 2006.

- M. A. Bender, M. Farach-Colton, J. T. Fineman, Y. Fogel, B. C. Kuszmaul, and J. Nelson. "Cache-Oblivious Streaming B-Trees," Proceedings of the 19th ACM Symposium on Parallelism in Algorithms and Architectures (SPAA), 2007.

- M. A. Bender, M. Farach-Colton, J. T. Fineman, Y. Fogel, B. C. Kuszmaul, and J. Nelson. "Cache-Oblivious Streaming B-Trees," Proceedings of the 19th ACM Symposium on Parallelism in Algorithms and Architectures (SPAA), 2007, p. 81.

- M. A. Bender, R. Clifford, and K. Tsichlas. "Scheduling Algorithms for Procrastinators," Journal of Scheduling, v.11, 2008, p. 95.

- M. A. Bender, S. P. Fekete, T. Kamphans, N. Schweer. "Maintaining Arrays of Contiguous Objects," 17th International Symposium on the Fundamentals of Computation Theory (FCT), LLNCS Volume 6599, 2009, p. 14.

- M. Farach-Colton and M. A. Mosteiro. "Sensor Network Gossiping or How to Break the Broadcast Lower Bound," Proceedings of the 18th International Symposium on Algorithms and Computation (ISAAC), LLNCS, v.4835, 2007, p. 232.

- M. Farach-Colton and M. A. Mosteiro. "Initializing Sensor Networks of Non-uniform Density in The Weak Sensor Model," Proceedings of the 10th International Workshop on Algorithms and Data Structures (WADS), LLNCS, v.4619, 2007, p. 565.

- M. Farach-Colton and M. Mosteiro. "Initializing Sensor Networks of Non-uniform Density in the Weak Sensor Model," Workshop on Data structures and Algorithms (WADS), 2007.

- M. Farach-Colton and Y. Huang. "Lattice Based Clustering of Temporal Gene-Expression Matrices," 7th SIAM International Conference on Data Mining (SDM), 2007.

- M. Farach-Colton and Y. Huang. "A Linear Delay Algorithm for Building Concept Lattices," Symposium on Combinatorial Pattern Matching (CPM), 2008.

- Shah, R; Farach-Colton, M. "On the complexity of ordinal clustering," JOURNAL OF CLASSIFICATION, v.23, 2006, p. 79-102.

- A. Fernandez Anta, C. Georgiou and M. A. Mosteiro. "Designing Mechanisms for Reliable Internet-based Computing," In Proceedings of the 7th IEEE International Symposium on Network Computing and Applications, Trustworthy Network Computing workshop (NCA-TNC), 2008.

- A. Fernandez Anta, M. A. Mosteiro and C. Thraves. "Deterministic Communication in the Weak Sensor Model," In Proceedings of the 11th International Conference On Principles Of Distributed Systems (OPODIS), volume 4878 of Lecture Notes in Computer Science, 2007.

- Arkin, EM; Bender, MA; Fekete, SP; Mitchell, JSB; Skutella, M. "The Freeze-Tag Problem: How to wake up a swarm of robots," ALGORITHMICA, v.46, 2006, p. 193-221.

- Bender, MA; Farach-Colton, M; Mosteiro, MA. "INSERTION SORT is O(n log n)," THEORY OF COMPUTING SYSTEMS, v.39, 2006, p. 391-397.

- Bender, MA; Fineman, JT; Gilbert, S. "Contention resolution with heterogeneous job sizes," ALGORITHMS - ESA 2006, PROCEEDINGS, v.4168, 2006, p. 112-123.

- Farach-Colton, M; Fernandes, RJ; Mosteiro, MA. "Lower bounds for clear transmissions in radio networks," LATIN 2006: THEORETICAL INFORMATICS, v.3887, 2006, p. 447-454.

- Farach-Colton, M; Huang, Y. "A linear delay algorithm for building concept lattices," COMBINATORIAL PATTERN MATCHING, v.5029, 2008, p. 204-216.

- Farach-Colton, M; Mosteiro, MA. "Sensor network gossiping or how to break the broadcast lower bound," ALGORITHMS AND COMPUTATION, v.4835, 2007, p. 232-243.

- K. Agrawal, M. A. Bender, and J. T. Fineman.. "The Worst Page-Replacement Policy.," Proceedings of the 4th International Conference on Fun With Algorithms (FUN), 2007.

- L. Arge, M. A. Bender, E. D. Demaine, B. Holland-Minkley, and J. I. Munro. "Cache-Oblivious Priority Queue and Graph Algorithm Applications.," SIAM Journal on Computing, v.36, 2007.

- M. A. Bender and H. Hu.. "An Adaptive Packed-Memory Array.," Proceedings of the 25th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, 2006.

- M. A. Bender and C. A. Phillips.. "Scheduling DAGs on Asynchronous Processors.," Proceedings of the 19th ACM Symposium on Parallelism in Algorithms and Architectures (SPAA), 2007.

- M. A. Bender, B. Bradley, G. Jagannathan, and K. Pillaipakkamnatt.. "Sum-of-Squares Heuristics for Bin Packing and Memory Allocation," Journal of Experimental Algorithms, v.12, 2007.

- M. A. Bender, G. S. Brodal, R. Fagerberg, R. Jacob, and E. Vicari.. "Optimal Sparse Matrix Dense Vector Multiplication in the {I/O}-Model.," Proceedings of the 19th ACM Symposium on Parallelism in Algorithms and Architectures (SPAA), 2007.

- M. A. Bender, M. Farach-Colton, B. C. Kuszmaul.. "Cache-Oblivious String B-Trees.," Proceedings of the 25th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems (PODS), 2006.

- M. A. Bender, M. Farach-Colton, J. T. Fineman, Y. Fogel, B. C. Kuszmaul, and J. Nelson.. "Cache-Oblivious Streaming B-Trees.," Proceedings of the 19th ACM Symposium on Parallelism in Algorithms and Architectures (SPAA), 2007.

- M. Farach-Colton and M. A. Mosteiro. "Initialiazing Sensor Networks of Non-uniform Density in The Weak Sensor Model," In Proceedings of the 10th International Workshop on Algorithms and Data Structures (WADS), volume 4619 of Lecture Notes in Computer Science, 2007.

- M. Farach-Colton and M. A. Mosteiro. "Sensor Network Gossiping or How to Break the Broadcast Lower Bound," In Proceedings of the 18th International Symposium on Algorithms and Computation (ISAAC), volume 4835 of Lecture Notes in Computer Science, pages 23, 2007.

- M. Farach-Colton and M. Mosteiro. "Initializing Sensor Networks of Non-uniform Density in the Weak Sensor Model," Workshop on Data structures and Algorithms (WADS), 2007.

- M. Farach-Colton and M. Mosteiro.. "Initializing Sensor Networks of Non-uniform Density in the Weak Sensor Model.," Workshop on Data structures and Algorithms, 2007.

- M. Farach-Colton and Y. Huang.. "Lattice based clustering of temporal gene-expression matrices.," 7th SIAM International Conference on Data Mining (SDM?07)., 2007.

- M. Farach-Colton, G. Landau, C. Sahinalp and D. Tsur. "Optimal spaced seeds for Faster Approximate String Matching.," Journal of Computer and System Science, v.73, 2007, p. 1035.

- Shah, R; Farach-Colton, M. "On the complexity of ordinal clustering," JOURNAL OF CLASSIFICATION, v.23, 2006, p. 79-102.

[Bender09] Bender, Michael A., Martin Farach-Colton, Charles E. Leiserson and Bradley C. Kuszmaul. “Multidimensional and String Indexes for Streaming Data.” High End Computing University Research Activity NSF 09-530 (2009)

- Agrawal K.; Bender, M. A.; and Fineman, J. T.. "The Worst Page-Replacement Policy," Theory of Computing Systems, v.44, 2009, p. 175.

- Agrawal, Kunal; Lee, I-Ting Angelina; and Sukha, Jim. "Serial-Parallel Reciprocity in Dynamic Multithreaded Languages (Brief Announcement)," Proceedings of the 22nd ACM Symposium on Parallelism in Algorithms and Architectures (SPAA), 2010.

- Agrawal, Kunal; Leiserson, Charles E.; and Sukha, Jim. "Helper Locks for Fork-Join Parallel Programming," Proceedings of the 15th ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming (PPoPP), 2010.

- Agrawal, Kunal; Leiserson, Charles E.; and Sukha, Jim. "Executing Task Graphs Using Work-Stealing," Proceedings of the 24th IEEE International Parallel and Distributed Processing Symposium (IPDPS), 2010.

- Bender, M. A.; Fekete, S. P.; Kamphans, T., and Schweer, N.. "Maintaining Arrays of Contiguous Objects.," Proceedings of the 17th International Symposium on the Fundamentals of Computation Theory (FCT), v.LLNCS V 6599, 2009, p. 14.

- Kuszmaul, Bradley C.. "What is a Performance Model for SSD?," High Performance Transaction Systems (HPTS) 2009, 2009.

- Kuszmaul, Bradley C. "How TokuDB Fractal Tree Databases Work," MySQL User Conference, Santa Clara, CA, 2009.

- Kuszmaul, Bradley C. "An Open-Source Storage Engine API," OpenSQL Camp 2009, Portland OR, 2009.

- Kuszmaul, Bradley C. "How Fractal Trees Work," OpenSQL Camp 2009, Portland OR, 2009.

- Kuszmaul, Bradley C. "Using Obliviousness to Achieve High Performance," ScalPerf'09, 2009.

- Kuszmaul, Bradley C. "SSD Performance Modeling (Ignite Talk)," MySQL User Conference, Santa Clara, CA, 2010.

- Mitrofanova, A; Farach-Colton, M; Mishra, B. "Efficient and Robust Prediction Algorithms for Protein Complexes using Gomory-Hu Tree," Pacific Symposium on Biocomputing (PSB), 2009, p. 215.

- Agrawal, K; Benoit, A; Dufosse, F; Robert, Y. "Mapping Filtering Streaming Applications With Communication Costs," in 21st ACM Symposium on Parallelism in Algorithms and Architectures., 2009, p. 19-28.

- Buluc, A; Fineman, JT; Frigo, M; Gilbert, JR; Leiserson, CE. "Parallel Sparse Matrix-Vector and Matrix-Transpose-Vector Multiplication Using Compressed

Sparse Blocks," in 21st ACM Symposium on Parallelism in Algorithms and Architectures., 2009, p. 233-244.

- Frigo, M; Halpern, P; Leiserson, CE; Lewin-Berlin, S. "Reducers and Other Cilk plus plus Hyperobjects," in 21st ACM Symposium on Parallelism in Algorithms and Architectures., 2009, p. 79-90.

- Kuszmaul, BC. "Brief Announcement: TeraByte TokuSampleSort Sorts 1TB in 197s," in 21st ACM Symposium on Parallelism in Algorithms and Architectures., 2009, p. 127-129.

- Sukha, J. "Brief Announcement: A Lower Bound for Depth-Restricted Work Stealing," in 21st ACM Symposium on Parallelism in Algorithms and Architectures., 2009, p. 124-126.

- Bender, MA; Hu, H; Kuszmaul, BC. "Performance Guarantees for B-trees with Different-Sized Atomic Keys," 29th ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems (PODS), 2010, p. 305.

- Farach-Colton, M; Fernandes, RJ; Mosteiro, MA. "Bootstrapping a Hop-Optimal Network in the Weak Sensor Model," ACM TRANSACTIONS ON ALGORITHMS, v.5, 2009.

- Montes, P; Memelli, H; Ward, C; Kim, J; Mitchell, JSB; Skiena, S. "Optimizing Restriction Site Placement for Synthetic Genomes," 21st Annual Symposium on Combinatorial Pattern Matching (CPM 2010), v.LLNCS 6, 2010, p. 323.

- Sukha, J. "Brief Announcement: A Lower Bound for Depth-Restricted Work Stealing," 21st ACM Symposium on Parallelism in Algorithms and Architectures, 2009, p. 124.

[Brandt06] Brandt, Scott A., Darrell D. E. Long, and Carl Maltzahn. “End-to-End Performance Management for Large Distributed Storage.” High End Computing University Research Activity NSF 06-503 (2006)

- Anna Povzner, Darren Sawyer, Scott Brandt. "Horizon: Efficient Deadline-Driven Disk I/O Management for Distributed Storage Systems," The ACM International Symposium on High Performance Distributed Computing (HPDC 2010), 2010.

- Anna Povzner, Scott Brandt, Richard Golding, Theodore Wong, Carlos Maltzahn. "Virtualizing Disk Performance with Fahrrad," Work-in-Progress session of the USENIX Conference on File and Storage Technology (FAST 2008), 2008.

- Anna Povzner, Tim Kaldewey, Scott A. Brandt, Richard Golding, Theodore Wong, and Carlos Maltzahn. "Efficient Guaranteed Disk Request Scheduling with Fahrrad," Eurosys 2008, 2008.

- Anna Povzner, Tim Kaldewey, Scott A. Brandt, Richard Golding, Theodore Wong, and Carlos Maltzahn. "Efficient Guaranteed Disk Request Scheduling with Fahrrad," ACM SIGOPS Operating Systems Review, v.42, 2008, p. 13.

- Changkyu Kim Jatin Chhugani, Nadathur Satish, Eric Sedlar, Anthony Nguyen,Tim Kaldewey, Victor Lee, Scott Brandt, Pradeep Dubey. "FAST: Fast Architecture Sensitive Tree Search onModern CPUs and GPUs," ACM SIGMOD, 2010.

- David Bigelow, Scott Brandt, John Bent, H.B. Chen. "Mahanaxar: Quality of Service Guarantees in High-Bandwidth, Real-Time Streaming Data Storage," IEEE / NASA Conference on Mass Storage Systems and Technologies, 2010.

- David O. Bigelow, Suresh Iyer, Tim Kaldewey, Roberto C. Pineiro, Anna Povzner, Scott A. Brandt, Richard Golding, Theodore Wong, and Carlos Maltzahn. "End-to-End Performance Management for Scalable Distributed Storage," Petascale Data Storage Workshop at SC07, 2007.

- Esteban Molina-Estolano, Carlos Maltzahn, Scott Brandt, and John Bent. "Comparing the Performance of Different Parallel Filesystem Placement Strategies," Work-in-Progress session of the USENIX Conference on File and Storage Technology (FAST 2009), 2009.

- Greg Levin, Shelby Funk, Caitlin Sadowski, Ian Pye, Scott Brandt. "DP-FAIR: A Simple Model for Understanding Optimal Multiprocessor Scheduling," Euromicro Conference on Real-Time System (ECRTS 2010), 2010.

- Ian Pye, Scott Brandt, Carlos Maltzahn. "Ringer: A Global-Scale Lightweight P2P File Service," Work-in-Progress session of the USENIX Conference on File and Storage Technology (FAST 2008), 2008.

- Joel Wu, Scott A. Brandt. "Providing Quality of Service Support in Object-based File Systems," IEEE / NASA Goddard Conference on Mass Storage Systems and Technologies, 2007.

- Roberto Pineiro and Scott Brandt. "Predictable and Guaranteeable Performance with Throughput, Latency, and Firmness Controls in Buffer-Cache," Work-in-Progress session of the USENIX Conference on File and Storage Technology (FAST 2009), 2009.

- Scott A. Brandt, Carlos Maltzahn, Anna Povzner, Roberto Pineiro, Andrew Shewmaker, Tim Kaldewey. "An Integrated Model for Performance Management in a Distributed System," Workshop on Operating Systems Platforms for Embedded Real-Time Applications (OSPERT 2008), 2008.

- Tim Kaldewey, Andrew Shewmaker, Richard Golding, Carlos Maltzahn, Theodore Wong, Scott Brandt. "RADoN: QoS in Storage Networks," Work-in-Progress session of the USENIX Conference on File and Storage Technology (FAST 2008), 2008.

- Tim Kaldewey, Anna Povzner, Theodore Wong, Richard Golding, Scott Brandt, Carlos Maltzahn. "Virtualizing Disk Performance," IEEE Real-Time and Embedded Technology and Applications Symposium (RTAS 2008), 2008.

- Tim Kaldewey, Jeff Hagen, Eric Sedlar, Andrea Di Blas, and Scott Brandt. "Memory Matters," Work-in-Progrees session of the IEEE Real-Time Systems Symposium (RTSS 2008), 2008.

[Brandt10] Brandt, Scott and Maya Gokhale. “Damasc: Adding Data Management Services to Parallel File Systems.” Scientific Data Management and Analysis at Extreme Scale ASCR FOA-10-0000256 (2010)

[Burns09] Burns, Randal and John Griffin. “CRAM: A Congestion-Aware Resource and Allocation Manager for Data-Intensive High Performance Computing.” High End Computing University Research Activity NSF 09-530 (2009)

- X. Wang, E. Perlman, R. Burns, T. Malik, T. Budava ri, C. Meneveau, and A. Szalay. JAWS: Job-aware workload scheduling for the exploration of turbulence simulations. In Supercomputing (SC), 2010

[Chandy06]Chandy, John A. “Active Storage Networks for High End Computing.” High End Computing University Research Activity NSF 06-503 (2006)

- Ajithkumar Thamarakuzhi and John A. Chandy. "Scaling the NetFPGA switch using Aurora over SATA," Proceedings of NetFPGA Developers Workshop, 2010.

- Ajithkumar Thamarakuzhi and John A. Chandy. "2-Dilated Flattened Butterfly: A Nonblocking Switching Network," Proceedings of International Conference on High Performance Switching and Routing, 2010, p. 134.

- Chandy, John A; Singaraju, Janardhan. "Hardware parallelism vs. software parallelism," USENIX Workshop on Hot Topics in Parallelism, 2009.

- Janardhan Singaraju and John A. Chandy. "FPGA Based String Matching for Network Processing Applications," Microprocessors and Microsystems, v.32, 2008, p. 210.

- John A. Chandy. "A Case for Active Storage Networks in High Performance Computing," Proceedings of Boston Area Architecture Workshop, 2007, p. 23.

- John A. Chandy. "A generalized replica placement strategy to optimize latency in a wide area distributed storage system," Proceedings of HPDC International Workshop on Data Aware Distributed Computing, 2008, p. 49.

- John A. Chandy and Sumit Narayan. "Reliability Tradeoffs in Personal Storage Systems," ACM Operating Systems Review, v.41, 2007, p. 37.

- Narayan, S; Chandy, JA. "ATTEST: ATTributes-based Extendable STorage," JOURNAL OF SYSTEMS AND SOFTWARE, v.83, 2010, p. 548-556.

- Sumit Narayan and John A. Chandy. "I/O Characterization on a Parallel File System," Proceedings of International Symposium on Performance Evaluation of Computer and Telecommunication Systems, 2010.

- Sumit Narayan and John A. Chandy. "Parity Redundancy in a Clustered Storage System," Proceedings of International Workshop on Storage Network Architecture and Parallel I/Os, 2007, p. 15.

- Sumit Narayan, John A. Chandy, Samuel Lang, Philip Carns, and Robert Ross. "Uncovering errors: the cost of detecting silent data corruption," Proceedings of the 4th Annual Workshop on Petascale Data Storage, 2009, p. 37.

- Sumit Narayan, Rohit K. Mehta, and John A. Chandy. "User space storage system stack modules with file level control," Proceedings of Annual Linux Symposium, 2010, p. 125.

- Tina Miriam John. Active storage for database applications. M.S. Thesis, University of Connecticut, December 2008. Department of Electrical and Computer Engineering

- Tina Miriam John, Anuradharthi Thiruvenkata Ramani, and John A. Chandy. Active storage using object-based devices. In Proceedings of CLUSTER Workshop on High Performance I/O Systems and Data Intensive Computing, pages 472-478, October 2008

- Vamsi Kundeti and John A. Chandy. "FEARLESS: Flash Enabled Active Replication of Low End Survivable Storage," Workshop on Integrating Solid-state Memory into the Storage Hierarchy, 2009.

CONFERENCE PROCEEDINGS

- John, TM; Ramani, AT; Chandy, JA. "Active Storage using Object-Based Devices," in IEEE International Conference on Cluster Computing., 2008, p. 472-478.

- Narayan, S; Chandy, JA. "Parity redundancy in a clustered storage system," in 4th International Workshop on Storage Network Architecture and Parallel I/Os., 2007, p. 17-24.

[Chandy09]Chandy, John. “Active Object Storage to Enable Scalable and Reliable Parallel File System.” High End Computing University Research Activity NSF 09-530 (2009)

[Chiueh06] Chiueh, Tzi-Cker. “Quality of Service Guarantee for Scalable Parallel Storage Systems.” High End Computing University Research Activity NSF 06-503 (2006)

- Bautin, M; Dwarakinath, A; Chiueh, TC. "Graphic engine resource management - art. no. 68180O," in 15th Conference on Multimedia Computing and Networking 2008., v.6818, 2008, p. O8180-O8180.

- Chiang, JH; Chiueh, TC. "Accurate Clock Synchronization for IEEE 802.11-Based Multi-Hop Wireless Networks," in 17th IEEE International Conference on Network Protocols., 2009, p. 11-20.

- Chiueh, TC. "Fast bounds checking using debug register," in 3rd International Conference on High Performance Embedded Architectures and Compilers., v.4917, 2008, p. 99-113.

- Chiueh, TC. "Fast bounds checking using debug register," in 3rd International Conference on High Performance Embedded Architectures and Compilers., v.4917, 2008, p. 99-113.

- Chiueh, TC; Bajpai, S. "Accurate and efficient inter-transaction dependency tracking," in 24th IEEE International Conference on Data Engineering., 2008, p. 1209-1218.

- Chiueh, TC; Bajpai, S. "Accurate and efficient inter-transaction dependency tracking," in 24th IEEE International Conference on Data Engineering., 2008, p. 1209-1218.

- Gang, P; Chiueh, TC. "Availability and Fairness Support for Storage QoS Guarantee," in 28th International Conference on Distributed Computing Systems., 2008, p. 589-596.

- Guo, FL; Chiueh, TC. "Comparison of QoS guarantee techniques for VoIP over IEEE802.11 wireless LAN - art. no. 68180J," in 15th Conference on Multimedia Computing and Networking 2008., v.6818, 2008, p. J8180-J8180.

- Guo, FL; Chiueh, TC. "Comparison of QoS guarantee techniques for VoIP over IEEE802.11 wireless LAN - art. no. 68180J," in 15th Conference on Multimedia Computing and Networking 2008., v.6818, 2008, p. J8180-J8180.

- Guo, FL; Chiueh, TC. "DAFT: Disk Geometry-Aware File System Traversal," in 17th IEEE International Symposium on Modeling, Analysis and Simulation of Computer and Telecommunication Systems., 2009, p. 484-493.

- Guo, FL; Chiueh, TC. "Device-Transparent Network-Layer Handoff for Micro-Mobility," in 17th IEEE International Symposium on Modeling, Analysis and Simulation of Computer and Telecommunication Systems., 2009, p. 319-328.

- Krishnan, R; Raniwala, A; Chiueh, TC. "Design of a channel characteristics-aware routing protocol," in 27th IEEE Conference on Computer Communications (INFOCOM 2008)., 2008, p. 346-350.

- Krishnan, R; Raniwala, A; Chiueh, TC. "Design of a channel characteristics-aware routing protocol," in 27th IEEE Conference on Computer Communications (INFOCOM 2008)., 2008, p. 346-350.

- Li, W; Chiueh, TC. "Automated format string attack prevention for Win32/X86 binaries," in 23rd Annual Computer Security Applications Conference., 2007, p. 398-407.

- Li, W; Chiueh, TC. "Automated format string attack prevention for Win32/X86 binaries," in 23rd Annual Computer Security Applications Conference., 2007, p. 398-407.

- Lu, MH; Chiueh, TC. "File Versioning for Block-Level Continuous Data Protection," in 29th IEEE International Conference on Distributed Computing Systems., 2009, p. 327-334.

- Lu, MH; Chiueh, TC. "Fast Memory State Synchronization for Virtualization-based Fault Tolerance," in 39th Annual IEEE/IFIP International Conference on Dependable Systems and Networks., 2009, p. 534-543.

- Lu, MH; Chiueh, TC; Lin, SB. "An Incremental File System Consistency Checker for Block-Level CDP Systems," in 27th IEEE International Symposium on Reliable Distributed Systems., 2008, p. 157-162.

- Nanda, S; Chiueh, TC. "Execution Trace-Driven Automated Attack Signature Generation," in 24th Annual Computer Security Applications Conference., 2008, p. 195-204.

- Nanda, S; Lam, LC; Chiueh, TC. "Web Application Attack Prevention for Tiered Internet Services," in 4th International Symposium on Information Assurance and Security., 2008, p. 186-191.

- Nanda, S; Lam, LC; Chiueh, TC. "Web Application Attack Prevention for Tiered Internet Services," in 4th International Symposium on Information Assurance and Security., 2008, p. 186-191.

- Wu, G; Chiueh, TC. "Passive and Accurate Traffic Load Estimation for Infrastructure-Mode Wireless LAN," in 10th ACM/IEEE International Symposium on Modeling Analysis and Simulation of Wireless and Mobile Systems., 2007, p. 109-116.

- Yu, Y; Kolam-govindarajan, H; Lam, LC; Chiueh, TC. "Applications of a Feather-weight Virtual Machine," in 4th International Conference on Virtual Execution Environments., 2008, p. 171-180.

[Choudhary06] Choudhary, Alok N., Mahmut T. Kandemir and Rajeev S. Thankur "Scalable I/O Middleware and File System Optimizations for High-Performance Computing." High End Computing University Research Activity NSF 06-503 (2006)

- Alok Choudhary, Wei-keng Liao, Kui Gao, Arifa Nisar, Robert Ross, Rajeev Thakur, and Robert Latham.. "Scalable I/O and Analytics," the Journal of Physics: Conference Series, v.180, 2009.

- Florin Isaila, Francisco Javier Garcia Blas, Jesus Carretero, Wei-keng Liao, and Alok Choudhary. "A Scalable Message Passing Interface Implementation of an Ad-Hoc Parallel I/O System.," the International Journal of High Performance Computing Applications, v.24, 2010, p. 164.

- C. Patrick, R. Garg, R. Prabhakar, S.-W. Son, M. Kandemir. "Shared Storage Cache Management for I/O-Intensive Workloads," Technical Report, Department of Computer Science of Engineering, 2008.

- Arifa Nisar, Waseem Ahmad, Wei-keng Liao, and Alok Choudhary. High Performance Parallel/Distributed Biclustering Using Barycenter Heuristic. In SIAM International Conference on Data Mining, April 2009

- Florin Isaila, Francisco Javier Garcia Blas, Jesus Carretero, Wei-keng Liao, and Alok Choudhary. AHPIOS: An MPI-based Ad-hoc Parallel I/O System. In the 14th International Conference on Parallel and Distributed Systems, December 2008

- Wei-keng Liao and Alok Choudhary. Dynamically Adapting File Domain Partitioning Methods for Collective I/O Based on Underlying Parallel File System Locking Protocols. In International Conference for High Performance Computing, Networking, Storage and Analysis, November 2008

- Arifa Nisar, Wei-keng Liao, and Alok Choudhary. Scaling Parallel I/O Performance through I/O Delegate and Caching System. In the International

Conference for High Performance Computing, Networking, Storage and Analysis, November 2008

- Wei-keng Liao, Avery Ching, Kenin Coloma, Arifa Nisar, Alok Choudhary, Jackie Chen, Ramanan Sankaran, and Scott Klansky. Using MPI File Caching to Improve Parallel Write Performance for Large-Scale Scienti_c Applications. In the International Conference for High Performance Computing, Networking, Storage and Analysis, November 2007

- Avery Ching, Robert Ross, Wei-keng Liao, Lee Ward, and Alok Choudhary. Noncontiguous Locking Techniques for Parallel File Systems. In the International Conference for High Performance Computing, Networking, Storage and Analysis, November 2007

- Wei-keng Liao, Kenin Coloma, Alok Choudhary, and Lee Ward. Cooperative Clientside File Caching for MPI Applications. International Journal of High Performance Computing Applications, 21(2):144-154, May 2007

- Wei-keng Liao, Avery Ching, Kenin Coloma, Alok Choudhary, and Lee Ward. An Implementation and Evaluation of Client-Side File Caching for MPI-IO. In the International Parallel and Distributed Processing Symposium, March 2007

- Wei-keng Liao, Avery Ching, Kenin Coloma, Alok Choudhary, and Mahmut Kandemir. Improving MPI Independent Write Performance Using A Two-Stage Write-Behind Buffering Method. In the Next Generation Software Workshop, held in conjunction with the 21st International Parallel and Distributed Processing Symposium, March 2007

[Choudhary09] Choudhary, Alok. “An Application Driven I/O Optimization Approach for PetaScale Systems and Scientific Discoveries.” High End Computing University Research Activity NSF 09-530 (2009)

- Sanchit Misra, Ankit Agrawal, and Wei-keng Liao Alok Choudhary. Anatomy of a Hash-based Long Read Sequence Mapping Algorithm for Next Generation DNA Sequencing. Bioinformatics, November 2010

- Seong Jo Kim, Yuanrui Zhang, SeungWoo Son, Ramya Prabhakar, Mahmut Kandemir, Christina Patrick, Wei-keng Liao, and Alok Choudhary. Automated Tracing of I/O Stack. In the 17th European MPI Users Group conference (EuroMPI), September 2010

- Seung Woo Son, Samuel Lang, Philip Carns, Robert Ross, Rajeev Thakur, Berkin Ozisikyilmaz, Prabhat Kumar, Wei-keng Liao, and Alok Choudhary. Enabling Active Storage on Parallel I/O Software Stacks. In the 26th IEEE Symposium on Massive Storage Systems and Technologies, May 2010

- Sanchit Misra, Ramanathan Narayanan, Wei-keng Liao, Alok Choudhary, and Simon Lin. pFANGS: Parallel High Speed Sequence Mapping for Next Generation 454-Roche Sequencing Reads. In Ninth IEEE International Workshop on High Performance Computational Biology, held in conjunction with the International Parallel and Distributed Parallel Processing Symposium, April 2010

- Florin Isaila, Francisco Javier Garcia Blas, Jesus Carretero, Wei-keng Liao, and Alok Choudhary. A Scalable Message Passing Interface Implementation of an Ad-Hoc Parallel I/O System. International Journal of High Performance Computing Applications, 24(2):164-184, 2010

- Nithin Nakka, Alok Choudhary, Wei-keng Liao, Lee Ward, Ruth Klundt, and Marlow Weston. Detailed Analysis of I/O Traces for Large Scale Applications. In the International Conference on High Performance Computing, December 2009

- Kui Gao, Wei-keng Liao, Arifa Nisar, Alok Choudhary, Robert Ross, and Robert Latham. Using Subfiling to Improve Programming Flexibility and Performance of Parallel Shared-file I/O. In the International Conference on Parallel Processing, September 2009

- Kui Gao, Wei-keng Liao, Alok Choudhary, Robert Ross, and Robert Latham. Combining I/O Operations for Multiple Array Variables in Parallel NetCDF. In the Workshop on Interfaces and Architectures for Scienti_c Data Storage, held in conjunction with the IEEE Cluster Conference, September 2009

[Choudhary10] Choudhary, Alok, Rob Latham, Nagiza Smatova and Quincey Koziol. “The Damsel Project: A Data Model Storage Library for Exascale Science.” X-Stack Software Research ASCR FOA-10-0000257 (2010)

[Dennis09] Dennis, Jack, Guang Gao and Vivek Sarkar. “Programming Models and Storage System for High Performance Computation with Many-Core Processors.” High End Computing University Research Activity NSF 09-530 (2009)

- Peter J. Denning and Jack B. Dennis. "The Profession of IT: The Resurgence of Parallelism," Communications of the ACM, v.53, No., 2010, p. 30.

- V. Sarkar, W. Harrod, A.E. Snavely. "Software Challenges in Extreme Scale Systems," SciDAC Review Special Issue on Advanced Computing: The Roadmap to Exascale, 2010, p. 60.

- Mihailo Kaplarevic, Alison E. Murray, Stephen C Cary, and Guang R. Gao. Engenius - environmental genome informational utility system. J Bioinform Comput Biol, 6(6):1193-211, 2008

- Rishi L. Khan, Rajanikanth Vadigepalli, Mary K. McDonald, Robert F. Rogers, Guang R. Gao, and James S. Schwaber. Dynamic transcriptomic response to acute

hypertension in the nucleus tractus solitarius. American Journal of Physiology - Regulatory, Integrative and Comparative Physiology, 295(1):R15-R27, 2008

- Hongbo Rong, Alban Douillet, and Guang R. Gao. Register allocation for software pipelined multidimensional loops. ACM Trans. Program. Lang. Syst., 30:23:1-23:68, August 2008

- Guangming Tan, Vugranam Sreedhar, and Guang Gao. Analysis and performance results of computing betweenness centrality on ibm cyclops64. The Journal of Supercomputing, 53:1-24, 2009. 10.1007/s11227-009-0339-9

- Guangming Tan, Ninghui Sun, and G.R. Gao. Improving performance of dynamic programming via parallelism and locality on multicore architectures. Parallel and Distributed Systems, IEEE Transactions on, 20(2):261-274, February 2009

- Chen Chen, Joseph B. Manzano, Ge Gan, Guang R. Gao, and Vivek Sarkar. A study of a software cache implementation of the openmp memory model for multicore and manycore architectures. In Proceedings of the 16th international Euro-Par conference on Parallel processing: Part II, Euro-Par'10, pages 341{352, Berlin, Heidelberg, 2010. Springer-Verlag

- Long Chen, O. Villa, S. Krishnamoorthy, and G.R. Gao. Dynamic load balancing on single- and multi-gpu systems. In Parallel Distributed Processing (IPDPS), 2010 IEEE International Symposium on, pages 1-12, April 2010

- Ge Gan, Xu Wang, Joseph Manzano, and Guang R. Gao. Tile percolation: An openmp tile aware parallelization technique for the cyclops-64 multicore processor. In Proceedings of the 15th International Euro-Par Conference on Parallel Processing, Euro-Par '09, pages 839{850, Berlin, Heidelberg, 2009. Springer-Verlag

- Ge Gan, Xu Wang, Joseph Manzano, and Guang R. Gao. Tile reduction: The first step towards tile aware parallelization in openmp. In Proceedings of the 5th International Workshop on OpenMP: Evolving OpenMP in an Age of Extreme Parallelism, IWOMP '09, pages 140{153, Berlin, Heidelberg, 2009. Springer-Verlag

- Elkin Garcia, Ioannis E. Venetis, Rishi Khan, and Guang R. Gao. Optimized dense matrix multiplication on a many-core architecture. In Proceedings of the 16th international Euro-Par conference on Parallel processing: Part II, Euro-Par'10, pages 316-327, Berlin, Heidelberg, 2010. Springer-Verlag

- Joseph J. Grzymski, Alison E. Murray, Barbara J. Campbell, Mihailo Kaplarevic, Guang R. Gao, Charles Lee, Roy M. Daniel, Amir Ghadiri, Robert A. Feldman, and Stephen C. Cary. Metagenome analysis of an extreme microbial symbiosis reveals eurythermal adaptation and metabolic exibility. In Proceedings of the

National Academy of Sciences of the United States of America - PNAS, Volume. 105, no. 45, 2008

- Joseph B. Manzano, Andres Marquez, and Guang G. Gao. Moda: A memory centric performance analysis tool. In LCI International Conference on High-Performance Clustered Computing, 2010

- Daniel A. Orozco and Guang R. Gao. Mapping the fdtd application to many-core chip architectures. In Proceedings of the 2009 International Conference on Parallel Processing, ICPP '09, pages 309{316, Washington, DC, USA, 2009. IEEE Computer Society

- Daniel Orozco, Elkin Garcia, and Guang G. Gao. Locality optimization of stencil applications using data dependency graphs. In The 23rd International Workshop on Languages and Compilers for Parallel Computing (LCPC2010), 2010

- Guangming Tan, Dongrui Fan, Junchao Zhang, Andrew Russo, and Guang R. Gao. Experience on optimizing irregular computation for memory hierarchy in manycore architecture. In Proceedings of the 13th ACM SIGPLAN Symposium on Principles and practice of parallel programming, PPoPP '08, pages 279{280, New York, NY, USA, 2008. ACM

- Guangming Tan, Vugranam C. Sreedhar, and Guang R. Gao. Just-in-time locality and percolation for optimizing irregular applications on a manycore architecture. In Jose Nelson Amaral, editor, Languages and Compilers for Parallel Computing, chapter Just-In-Time Locality and Percolation for Optimizing Irregular Applications on a Manycore Architecture, pages 331{342. Springer-Verlag, Berlin, Heidelberg, 2008

- Ioannis E. Venetis and Guang R. Gao. Mapping the lu decomposition on a many-core architecture: challenges and solutions. In Proceedings of the 6th ACM conference on Computing frontiers, CF '09, pages 71{80, New York, NY, USA, 2009. ACM

- Liping Xue, Long Chen, Ziang Hu, and Guang R. Gao. Performance tuning of the fast fourier transform on a multi-core architecture. In the First Workshop on Programmability Issues for Multi-Core Computers (MULTIPROG), 2008

- Handong Ye, R. Pavel, A. Landwehr, and G.R. Gao. Tiny threads on bluegene/p: Exploring many-core parallelisms beyond the traditional os. In Parallel Distributed Processing, Workshops and Phd Forum (IPDPSW), 2010 IEEE International Symposium on, pages 1-8, April 2010

- Yuan Zhang, Evelyn Duesterwald, and Guang R. Gao. Concurrency analysis for shared memory programs with textually unaligned barriers. In Vikram Adve, Mar__a Jes_us Garzar_an, and Paul Petersen, editors, Languages and Compilers for Parallel Com- puting, chapter Concurrency Analysis for Shared Memory Programs with Textually Unaligned Barriers, pages 95{109. Springer-Verlag, Berlin, Heidelberg, 2008

- Rajkishore Barik, Jisheng Zhao, David Grove, Igor Peshansky, Zoran Budimli, and Vivek Sarkar. Communication optimizations for distributed-memory x10 programs. In IPDPS '11: Proceedings of the 2011 IEEE International Symposium on Parallel and Distributed Processing. (to appear)

- Raghavan Raman, Jisheng Zhao, Vivek Sarkar, Martin Vechev, and Eran Yahav. Effcient data race detection for async-_nish parallelism. In RV'10, Proceedings of the 1th International Conference on Runtime Veri_cation. Springer, Nov 2010.

- Yi Guo. A Scalable Locality-aware Adaptive Work-stealing Scheduler for Multi-core Task Parallelism. PhD thesis, Department of Computer Science, Rice University, August 2010

- Jisheng Zhao, Jun Shirako, V. Krishna Nandivada, and Vivek Sarkar. Reducing task creation and termination overhead in explicitly parallel program. In PACT'10, Proceedings of the 19th International Conference on Parallel Architectures and Compilation Techniques. IEEE Computer Society, Sep 2010

- Martin Vechev, Eran Yahav, Raghavan Raman, and Vivek Sarkar. Verifying determinism of structured parallel programs. In SAS'10, Proceedings of the 17th International Statical Analysis Symposium. Springer, Sep 2010

- Chen Chen, Joseph B. Manzano, Ge Gan, Guang R. Gao, and Vivek Sarkar. A study of a software cache implementation of the openmp memory model for multicore and manycore architectures. In Euro-Par '10: Proceedings of the 15th International Euro-Par Conference on Parallel Processing, Berlin, Heidelberg, 2010. Springer-Verlag

- Yi Guo, Jisheng Zhao, Vincent Cave, and Vivek Sarkar. Slaw: a scalable locality-aware adaptive work-stealing scheduler. In IPDPS '10: Proceedings of the 2010 IEEE International Symposium on Parallel&Distributed Processing, pages 1-12, Washington, DC, USA, Apr 2010. IEEE Computer Society

- Jun Shirako and Vivek Sarkar. Hierarchical phasers for scalable synchronization and reduction. In IPDPS '10: Proceedings of the 2010 IEEE International Symposium on Parallel&Distributed Processing, pages 1-12, Washington, DC, USA, 2010. IEEE Computer Society

- Yonghong Yan, Jisheng Zhao, Yi Guo, and Vivek Sarkar. Hierarchical place trees: A portable abstraction for task parallelism and data movement. In Languages and Compilers for Parallel Computing, 22nd International Workshop, LCPC 2009, volume 5898 of Lecture Notes in Computer Science. Springer, 2009

- Rajkishore Barik. Effcient Optimization of Memory Accesses in Parallel Programs. PhD thesis, Rice University, Oct 2009

- Rajkishore Barik and Vivek Sarkar. Interprocedural load elimination for dynamic optimization of parallel programs. In PACT'09, Proceedings of the 18th International Conference on Parallel Architectures and Compilation Techniques, pages 41-52, Washington, DC, USA, Sep 2009. IEEE Computer Society

- Jun Shirako, Jisheng Zhao, V. Krishna Nandivada, and Vivek Sarkar. Chunking parallel loops in the presence of synchronization. In ICS '09: Proceedings of the 23rd international conference on Supercomputing, pages 181-192, New York, NY, USA, 2009. ACM

- Yi Guo, Rajkishore Barik, Raghavan Raman, and Vivek Sarkar. Work-first and help-first scheduling policies for async-_nish task parallelism. In IPDPS '09: Proceedings of the 2009 IEEE International Symposium on Parallel&Distributed Processing, pages 1-12, Washington, DC, USA, May 2009. IEEE Computer Society

- Jun Shirako, David M. Peixotto, Vivek Sarkar, and William N. Scherer. Phaser accumulators: A new reduction construct for dynamic parallelism. In IPDPS '09: Proceedings of the 2009 IEEE International Symposium on Parallel&Distributed Processing, pages 1-12, Washington, DC, USA, 2009. IEEE Computer Society

- Raghavan Raman. Master thesis: Compiler support for work-stealing parallel runtime systems, May 2009

- Yuan Zhang, Vugranam C. Sreedhar, Weirong Zhu, Vivek Sarkar, and Guang R. Gao. Minimum lock assignment: A method for exploiting concurrency among critical sections. pages 141-155, 2008

[Dickens07] Dickens, Phillip. “Object Based Caching for MPI-IO.” Foundations of Computing Processes and Artifacts NSF 06-585 (2007)

- Dickens, PM; Logan, J. "A high performance implementation of MPI-IO for a Lustre file system environment," CONCURRENCY AND COMPUTATION-PRACTICE & EXPERIENCE, v.22, 2010, p. 1433-1449.

- Logan, J. and Dickens, P.. "Using Object Based Files for High Performance Parallel I/O," IEEE International Workshop on Intelligent Data Acquisition and Advanced Computing Sysytems: Technology and Applications, 2007.

- Dickens, P; Logan, J. "Towards a High Performance Implementation of MPI-IO on the Lustre File System," in On the Move Confederated International Conference and Workshops., v.5331, 2008, p. 870-885.

- Dickens, PM; Logan, J. "Y-Lib: A User Level Library to Increase the Performance of MPI-IO in a Lustre File System Environment," in 18th ACM International Symposium on High Performance Distributed Computing (HPDC 2009)., 2009, p. 31-37.

- Logan, J; Dickens, P. "Towards an Understanding of the Performance of MPI-IO in Lustre File Systems," in IEEE International Conference on Cluster Computing., 2008, p. 330-335.

- Logan, J; Dickens, P. "Improving I/O Performance through the Dynamic Remapping of Object Sets," in 5th IEEE International Workshop on Intelligent Data Acquisition and Advanced Computing Systems., 2009, p. 259-265.

- Logan, J; Dickens, PM. "Using object based files for high performance parallel I/O," in 4th IEEE International Workshop on Intelligent Data Acquisition and Advanced Computing Systems., 2007, p. 149-154.

[Dhall09] Dhall, Sudarshan. “EAGER: Autonomous Data Partitioning Using Data Mining for High End Computing.” High End Computing University Research Activity NSF 09-530 (2009)

[Felix06] Felix, Evan, Gary Grider, Rob Hill, Bill Loewe, Rob Ross, Lee Ward. “HPC File Systems and Scalable I/O: Suggested Research and Development Topics for the fiscal 2005-2009 time frame.” 10 Oct. 2006 ftp://ftp.lanl.gov/public/ggrider/HEC-IWG-FS-IO-Workshop-08-15-2005/FileSystems-DTS-SIO-FY05-FY09-R&D-topics-final.pdf

[Ganger06] Ganger, Greg. “Performance Insulation and Predictability for Shared Cluster Storage.” Foundations of Computing Processes and Artifacts

- Amar Phanishayee, Elie Krevat, Vijay Vasudevan, David G. Andersen, Gregory R. Ganger, Garth A. Gibson, Srinivasan Seshan. "Measurement and analysis of TCP throughput collapse in cluster-based storage systems," USENIX Conference on File and Storage Technologies (FAST), 2007.

- Elie Krevat, Vijay Vasudevan, Amar Phanishayee, David G. Andersen, Gregory R. Ganger, Garth A. Gibson, Srinivasan Seshan. "On application-level approaches to avoiding TCP throughput collapse in cluster-based storage systems," Petascale Data Storage Workshop, 2007.

- Matthew Wachs, Gregory R. Ganger. "Co-scheduling disk head time in cluster-based storage," Proceedings of Symposium on Reliable Distributed Systems, 2009.

- Matthew Wachs, Michael Abd-El-Malek, Eno Thereska, Gregory R. Ganger. "Argon: Performance Insulation for Shared Storage Servers," Proceedings of the 5th USENIX Conference on File and Storage Technologies (FAST '07), 2007.

- Michael P. Mesnier, Matthew Wachs, Raja R. Sambasivan, Alice X. Zheng, Gregory R. Ganger. "Modeling the Relative Fitness of Storage," ACM SIGMETRICS, 2007.

- Michael P. Mesnier, Matthew Wachs, Raja R. Sambasivan, Alice X. Zheng, Gregory R. Ganger. "Relative Fitness Modeling," Communications of the ACM, 2009.

- Vijay Vasudevan, Amar Phanishayee, Hiral Shah, Elie Krevat, David G. Andersen, Gregory R. Ganger, Garth A. Gibson, Brian Mueller. "Safe and Effective Fine-grainder TCP Retransmissions for Datacenter Communication," ACM SIGCOMM, 2009.

[Gao07] Gao, Guang. “A High Throughput Massive I/O Storage Hierarchy for PETA-scale High-end Architectures.” Foundations of Computing Processes and Artifacts NSF 06-585 (2007)

[Gibson06] Gibson, Garth, Evan Felix, Gary Grider, Peter Honeyman, William Kramer, Darrell Long, Philip Roth, Lee Ward. PetaScale Data Storage Institute, SciDAC2 10 Oct. 2006 http://www.scidac.gov/compsci/PDSI.html

[He09] He, Xubin, Tao Li and Tong Zhang. “Cross-Layer Exploration of Non-Volatile Solid-State Memories to Achieve Effective I/O Stack for High-Performance Computing Systems.” High End Computing University Research Activity NSF 09-530 (2009)

[HeScott09] He, Xubin and Stephen L. Scott. “RUI: Automatic Identification of I/O Bottleneck and Run-time Optimization for Cluster Virtualization.” High End Computing University Research Activity NSF 09-530 (2009)

[Huang09] Huang, H. Howie and Alexander Szalay. “Balanced Scalable Architectures for Data-Intensive Supercomputing.” High End Computing University Research Activity NSF 09-530 (2009)

- Carliles, S., Budavari, T., Heinis, S., Priebe,C., Szalay, A.S.. "Random Forests for Photometric Redshifts," The Astrophysical Journal, v.712, 2010, p. 511.

- Szalay, A., Bell, G., Huang, H. H., Terzis, A., White, A.. "Low-Power Amdahl-Balanced Blades for Data Intensive Computing," ACM SIGOPS Operating Systems Review, v.44, 2010, p. 71.

- H. Howie Huang. "A control-theoretic approach to automated local policy enforcement in computational grids," Future Generation Computer Systems, v.26, 2010, p. 787.

[Iskra10] Iskra, Kamil and Maya Gokhale. “NoLoSS: Investigating the Roles of Node Local Storage in Exascale Systems.” Advanced Architectures and Critical Technologies for Exascale Computing ASCR FOA-10-0000255 (2010)

[Jiang06] Jiang, Hong, Yifeng Zhu, David R. Swanson, and Jun Wang. “SAM^2 Toolkit: Scalable and Adaptive Metadata Management for High-End Computing.” High End Computing University Research Activity NSF 06-503 (2006)

- Bo Mao, Dan Feng, Hong Jiang, Suzhen Wu, Jianxi Chen, Lingfang Zeng. "GRAID: A Green RAID Storage Architecture with Improved Energy Efficiency and Reliability," Proceedings of the 16th Annual Meeting of the IEEE International Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunication Systems (MASCOTS 08), 2008.

- Bo Mao, Hong Jiang, Dan Feng, Suzhen Wu. "HPDA: A Hybrid Parity-based Disk Array for Enhanced Performance and Reliability," Proceedings of the 24th IEEE International Parallel & Distributed Processing Symposium (IPDPS 2010), 2010.

- Chao Jin, Hong Jiang, Dan Feng, Lei Tian. "P-Code: A New RAID-6 Code with Optimal Properties," the Proceedings of the 23rd ACM International Conference on Supercomputing, 2009.

- D. Feng, Q. Zou, H. Jiang, and Y. Zhu. "A Novel Model for Synthesizing Parallel I/O Workloads in Scientific Applications," Proceedings of IEEE International Conference on Cluster Computing, 2008, p. 252.

- Dongyuan Zhan, Hong Jiang, and Sharad Seth. "Exploiting Set-Level Non-Uniformity of Capacity Demand to Enhance CMP Cooperative Caching," Proceedings of the 24th IEEE International Parallel & Distributed Processing Symposium (IPDPS 2010), 2010.

- Hailong Cai, Ping Ge, Jun Wang. "Applications of Bloom Filters in Peer-to-peer Systems: Issues and Questions," Proceedings of International Conference on Networking, Architecture, and Storage, 2008.

- Hui Tian, Ke Zhou, Hong Jiang, Dan Feng. "Digital Logic Based Encoding Strategies for Voice-over-IP Steganography," Proceedings of the ACM Multimedia 2009 conference, 2009.

- J. Yue, Y. Zhu and Z. Cai. "An Energy-Oriented Evaluation of Buffer Cache Algorithms Under Parallel I/O Workloads," IEEE Transaction on Distributed and Parallel Computing, v.19, 2008, p. 1565.

- J. Yue, Y. Zhu and Z. Cai. "Evaluating Memory Energy Efficiency in Parallel I/O Workloads," Proceedings of IEEE International Conference on Cluster Computing, 2007.

- J. Yue, Y. Zhu, and Z. Cai. "Impacts of Indirect Blocks on Buffer Cache Energy Efficiency," Proceedings of the 37th International Conference on Parallel Processing (ICPP 08), 2008.

- J. Yue, Y. Zhu, Z. Cai. "Energy Efficient Buffer Cache Replacement," Proceedings of 16th Annual Meeting of the IEEE International Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunication Systems, 2008.

- J. Yue, Y. Zhu, Z. Cai, L. Lin. "Energy and Thermal Aware Buffer Cache Replacement Algorithm," Proceedings of 26th IEEE Symposium on Massive Storage Systems and Technologies (MSST), 2010.

- Jiansheng Wei, Hong Jiang, Ke Zhou, Dan Feng. "AD2: A Scalable High-Throughput Exact Deduplication Approach for Network Backup Services," Proceedings of The 26th IEEE Symposium on Massive Storage Systems and Technologies (MSST2010), 2010.

- Jun Wang, Peng Gu, Hailong Cai. "An Advertisement-based Peer-to-Peer Search Scheme," Journal of Parallel and Distributed Computing, v.69, 2009, p. 638.

- Jun Wang, Xiaoyu Yao, Christopher Mitchell, and Peng Gu. "A hierarchical data cache architecture for iSCSI storage server," IEEE Transactions on Computers, v.58, 2009, p. 1.

- Jun Wang, Xiaoyu Yao, Christopher Mitchell, and Peng Gu. "A hierarchical data cache architecture for iSCSI storage server," IEEE Transactions on Computers, v.58, 2009, p. 1.

- L. Lin, M. Li, H. Jang, and Y. Zhu. "AMP: An Affinity-based Metadata Prefetching Scheme in Large-Scale Distributed Storage Systems," Proceedings of the 8th IEEE International Symposium on Cluster Computing and the Grid (CCGrid'08), 2008.

- Lei Tian, Dan Feng, Hong Jiang, Ke Zhou, Lingfang Zeng, Jianxi Chen, Zhikun Wang, and Zhenlei Song. "PRO: A Popularity-based Multi-threaded Reconstruction Optimization for RAID-Structured Storage Systems," Proceedings of the 5th USENIX Conference on File and Storage Technologies (FAST '07), 2007, p. 277.

- Lei Tian, Hong Jiang, Dan Feng, Qin Xin, and Xing Shu. "Implementation and Evaluation of a Popularity-Based Reconstruction Optimization Algorithm in Availability-Oriented Disk Arrays," Proceedings of the 24th IEEE Conference on Mass Storage Systems and Technologies (MSST?07) (acceptance rate: 37%), 2007, p. 233.

- P. Gu, Y. Zhu, H. Jiang, and J. Wang. "Nexus: A Novel Weighted-Graph-Based Prefetching Algorithm for Metadata Servers in Petabyte-Scale Storage Systems," Proceedings of International Symposium on Cluster Computing and the Grid (CCGrid, 2006), 2006, p. 409.

- Peng Gu, Jun Wang, Hailong Cai. "ASAP: An Advertisement-based Search Scheme for Unstructured Peer-to-Peer Systems," International Conference on Parallel Processing (ICPP 2007), 2007.

- Peng Gu, Jun Wang, Robert Ross. "Bridging the Gap Between Parallel File Systems and Local File Systems: A Case Study with PVFS," Proceedings of ICPP, 2008.

- Peng Gu, Jun Wang, Yifeng Zhu, Hong Jiang. "Improving Metadata Server Performance in Petabyte-Scale File Systems," Journal of Parallel and Distributed Computing, 2008.

- Peng Gu, Jun Wang, Yifeng Zhu, Hong Jiang, Pengju Shang. "A Novel Weighted-Graph-Based Grouping Algorithm for Metadata Prefetching," IEEE Transactions on Computers, v.59, 2010, p. 1.

- Peng Xia, Dan Feng, Hong Jiang, Lei Tian, and Fang Wang. "FARMER: A Novel Approach to File Access Correlation Mining And Evaluating Reference Model for Optimizing Peta-Scale File Systems Performance," Proceedings of the 17th ACM/IEEE International Symposium on High Performance Distributed Computing (HPDC 2008) (Acceptance rate: 17%), 2008.

- Q. Zou, D. Feng, Y. Zhu, H. Jiang, X. Ge, and Z. Zhou. "A Novel and Generic Model for Synthesizing Disk I/O Traffic Based on The Alpha-stable Process," Proceedings of 16th Annual Meeting of the IEEE International Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunication Systems (MASCOTS 08), 2008.

- Q. Zou, Y. Zhu and D. Feng. "A study of Self-similarity in Parallel I/O Workloads," Proceedings of 26th IEEE Symposium on Massive Storage Systems and Technologies (MSST), 2010.

- Saba Sehrish, and Jun Wang. "Smart Read/Write for MPI-IO," In the 14th International Workshop on High-Level Parallel Programming Models and Supportive Environments, in conjunction with the 23rd IEEE International Parallel and Distributed Processing Symposium, 2009.

- Saba Sehrish, and Jun Wang. "Smart Read/Write for MPI-IO," the 14th International Workshop on High-Level Parallel Programming Models and Supportive Environments, in conjunction with the 23rd IEEE International Parallel and Distributed Processing Symposium, 2009.

- Saba Sehrish, Jun Wang, and Rajeev Thakur. "Self-detecting Locks to Support MPI-IO Atomicity," EuroPVM/MPI, 2009.

- Suzhen Wu, Dan Feng, Hong Jiang, Bo Mao, Lingfang Zeng, and Jianxi Chen. "JOR: A Journal-guided Reconstruction Optimization for RAID-Structured Storage Systems," Proceedings of the Fifteenth International Conference on Parallel and Distributed Systems (ICPADS'09), 2009.

- Suzhen Wu, Hong Jiang, Dan Feng, Lei Tian, and Bo Mao. "WorkOut: I/O Workload Outsourcing for Boosting the RAID Reconstruction Performance," Proceedings of the 7th USENIX Conference on File and Storage Technologies (FAST '09), 2009.

- Tianming Yang, Hong Jiang, Dan Feng, Zhongying Niu,Ke Zhou, andYaping Wan. "DEBAR: A Scalable High-Performance De-duplication Storage System for Backup and Archiving," Proceedings of the 24th IEEE International Parallel & Distributed Processing Symposium (IPDPS 2010), 2010.

- Xiao Qin and Hong Jiang. "Dynamic Load Balancing for I/O-Intensive Applications on Clusters," ACM Transactions on Storage, v.5, 2009, p. 9.

- Xiao Qin, Hong Jiang, Adam Manzanares, Xiaojun Ruan, Shu Yin. "Communication-Aware Load Balancing for Parallel Applications on Clusters," IEEE Transactions on Computers, v.59, 2010.

- Y. Hua, H. Jiang, Y. Zhu, D. Feng, L. Tian. "SmartStore: A New Metadata Organization Paradigm with Semantic-Awareness," 7th USENIX Conference on File and Storage Technologies, Work-In-Progress, 2009.

- Y. Hua, Y. Zhu, H. Jiang, D. Feng, and L. Tian. "Scalable and Adaptive Metadata Management in Ultra Large-Scale File Systems," Proceedings of the 28th International Conference on Distributed Computing Systems (ICDCS'08), 2008.

- Y. Zhu and H. Jiang. "On the Analysis and Impact of False Rates of Bloom Filters in Distributed Systems," Proceedings of the 35th International Conference on Parallel Processing (ICPP), 2006, p. 255.

- Y. Zhu, and H. Jiang. "RACE: A Robust Adaptive Caching Strategy for Buffer Cache," IEEE Transaction on Computers, v.57, 2008, p. 25-40.

- Y. Zhu, H. Jiang, J. Wang and F. Xian. "HBA: Distributed Metadata Management System for Large Cluster-based Storage," IEEE Transaction on Distributed and Parallel Systems, v.19, 2008, p. 750-763.

- Yang Hu, Hong Jiang, Dan Feng, Lei Tian, Sshuping Zhang, Jingning Liu, Wei Tong. "Achieving Page-Mapping FTL Performance at Block-Mapping FTL Cost by Hiding Address Translation," Proceedings of The 26th IEEE Symposium on Massive Storage Systems and Technologies (MSST2010), 2010.

- Yinliang Yue, Hong Jiang, Lei Tian, Fang Wang, Dan Fang, and Quan Zhang. "RoLo: A Rotated Logging Storage Architecture for Enterprise Data Centers," Proceedings of The 30th International Conference on Distributed Computing Systems (ICDCS 2010), 2010.

- Yu Hua, Dan Feng, Hong Jiang, and Lei Tian. "RBF: A New Storage Structure for Space-Efficient Queries for Multidimensional Metadata in OSS," the 5th USENIX Conference on File and Storage Technologies (FAST '07) Work-in-Progress (WiP) Report, 2007, p. 1.

- Yu Hua, Hong Jiang, Yifeng Zhu, Dan Feng, and Lei Tian. "SmartStore: A New Metadata Organization Paradigm with Metadata Semantic-Awareness for Next-Generation File Systems," Proceedings of The 22nd International Conference on High Performance Computing, Networking, Storage and Analysis (The 22nd Annual Supercomputing Conference -- SC'09), 2009.

- Zhongying Niu, Ke Zhou, Dan Feng, Hong Jiang, Frank Wang, Hua Chai, Wei Xiao,and Chunhua Li. "Implementing and Evaluating Security Controls for an Object-Based Storage System," Proceedings of the 24th IEEE Conference on Mass Storage Systems and Technologies (MSST07), 2007.

CONFERENCE PROCEEDINGS

- Cai, HL; Ge, P; Wang, J. "Applications of Bloom filters in peer-to-peer systems: Issues and questions," in IEEE International Conference on Networking, Architecture, and Storage., 2008, p. 97-103.

- Cai, HL; Ge, P; Wang, J. "Applications of Bloom filters in peer-to-peer systems: Issues and questions," in IEEE International Conference on Networking, Architecture, and Storage., 2008, p. 97-103.

- Feng, D; Zou, Q; Jiang, H; Zhu, YF. "A Novel Model for Synthesizing Parallel I/O Workloads in Scientific Applications," in IEEE International Conference on Cluster Computing., 2008, p. 252-261.

- Feng, D; Zou, Q; Jiang, H; Zhu, YF. "A Novel Model for Synthesizing Parallel I/O Workloads in Scientific Applications," in IEEE International Conference on Cluster Computing., 2008, p. 252-261.

- Hua, Y; Zhu, YF; Jiang, H; Feng, D; Tian, L. "Scalable and Adaptive Metadata Management in Ultra Large-scale File Systems," in 28th International Conference on Distributed Computing Systems., 2008, p. 403-410.

- Hua, Y; Zhu, YF; Jiang, H; Feng, D; Tian, L. "Scalable and Adaptive Metadata Management in Ultra Large-scale File Systems," in 28th International Conference on Distributed Computing Systems., 2008, p. 403-410.

- Yue, JH; Zhu, YF; Cai, Z. "Evaluating Memory Energy Efficiency in Parallel I/O Workloads," in IEEE International Conference on Cluster Computing., 2007, p. 21-30.

- Yue, JH; Zhu, YF; Cai, Z. "Evaluating Memory Energy Efficiency in Parallel I/O Workloads," in IEEE International Conference on Cluster Computing., 2007, p. 21-30.

[Jiang09] Jiang, Hong and Yifeng Zhu. “A New Semantic-Aware Metadata Organization for Improved File-System Performance and Functionality in High-End Computing.” High End Computing University Research Activity NSF 09-530 (2009)

- Bo Mao, Hong Jiang, Dan Feng, Suzhen Wu. "HPDA: A Hybrid Parity-based Disk Array for Enhanced Performance and Reliability," Proceedings of the 24th IEEE International Parallel & Distributed Processing Symposium (IPDPS 2010), 2010.

- Dongyuan Zhan, Hong Jiang, and Sharad Seth. "Exploiting Set-Level Non-Uniformity of Capacity Demand to Enhance CMP Cooperative Caching," Proceedings of the 24th IEEE International Parallel & Distributed Processing Symposium (IPDPS 2010), 2010.

- J. Yue, Y. Zhu, Z. Cai, L. Lin. "Energy and Thermal Aware Buffer Cache Replacement Algorithm," Proceedings of 26th IEEE Symposium on Massive Storage Systems and Technologies (MSST), 2010.

- Jiansheng Wei, Hong Jiang, Ke Zhou, Dan Fen. "MAD2: A Scalable High-Throughput Exact Deduplication Approach for Network Backup Services," Proceedings of The 26th IEEE Symposium on Massive Storage Systems and Technologies (MSST2010), 2010.

- Q. Zou, Y. Zhu and D. Feng. "A study of Self-similarity in Parallel I/O Workloads," Proceedings of 26th IEEE Symposium on Massive Storage Systems and Technologies (MSST), 2010.

- Tianming Yang, Hong Jiang, Dan Feng, Zhongying Niu, Ke Zhou, and Yaping Wan. "DEBAR: A Scalable High-Performance De-duplication Storage System for Backup and Archiving," Proceedings of the 24th IEEE International Parallel & Distributed Processing Symposium (IPDPS 2010), 2010.

- Yang Hu, Hong Jiang, Dan Feng, Lei Tian, Sshuping Zhang, Jingning Liu, Wei Tong. "Achieving Page-Mapping FTL Performance at Block-Mapping FTL Cost by Hiding Address Translation," Proceedings of The 26th IEEE Symposium on Massive Storage Systems and Technologies (MSST2010), 2010.

- Yinliang Yue, Hong Jiang, Lei Tian, Fang Wang, Dan Fang, and Quan Zhang. "RoLo: A Rotated Logging Storage Architecture for Enterprise Data Centers," Proceedings of The 30th International Conference on Distributed Computing Systems (ICDCS 2010), 2010.

- Yu Hua, Hong Jiang, Yifeng Zhu, Dan Feng, and Lei Tian. "SmartStore: A New Metadata Organization Paradigm with Metadata Semantic-Awareness for Next-Generation File Systems," Proceedings of The 22nd International Conference on High Performance Computing, Networking, Storage and Analysis (The 22nd Annual Supercomputing Conference -- SC'09), 2009.

- Yujuan Tan, Hong Jiang, Dan Feng, Lei Tian, and Zhichao Yan. SAM: A Semantic-Aware Multi-Tiered Source De-duplication Framework for Cloud Backup. Proceedings of International Conference on Parallel Processing (ICPP), 2010

- Yu Hua, Yifeng Zhu, Hong Jiang, Dan Feng, and Lei Tian. Supporting scalable and adaptive metadata management in ultra large-scale _le systems. IEEE Transactions on Parallel and Distributed Systems [Accepted, to appear], 2010

- Yu Hua, Hong Jiang, Yifeng Zhu, and Dan Feng. Rapport: Semantic-sensitive Namespace Management in Large-scale File Systems. Technical Report TR-UNL-CSE. Department of Computer Science and Engineering University of Nebraska-Lincoln, 2010

- Suzhen Wu, Hong Jiang, Dan Feng, Lei Tian, and Bo Mao. Improving availability of raid-structured storage systems by workload outsourcing. Special Issue on Dependable Computer Architecture, IEEE Transactions on Computers, 60(1), January 2011

- Yujuan Tan, Hong Jiang, Dan Feng, Lei Tian, Zhichao Yan, and Guohui Zhou. Cabdedup: A causality-based de-duplication performance booster for cloud backup services. Proceedings of the 25th IEEE International Parallel & Distributed Processing Symposium (IPDPS11), May 2011

- Chao Jin, Dan Feng, Hong Jiang, Lei Tian, and Jingning Liu. Trip: Temporal redundancy integrated performance booster for parity-based raid storage systems. Proceedings of the 16th International Conference on Parallel and Distributed Systems (ICPADS 2010), (Best-Paper Award Nominee), December 2010

- Zhongying Niu, Hong Jiang, Ke Zhou, and Dan Feng. Dsfs: Decentralized security for large parallel file systems. Proceedings of the 11th ACM/IEEE International Conference on Grid Computing (Grid 2010), October 2010

- Jian Hu, Hong Jiang, Lei Tian, and Lei Xu. Pud-lru: An erase-effcient write buffer management algorithm for ash memory ssd. Proceedings of the 18th Annual Meeting of the IEEE/ACM International Symposium on Modeling, Analysis and Simulation of Computer and Telecommunication Systems (MASCOTS10), August 2010

- Yujuan Tan, Hong Jiang, Dan Feng, Lei Tian, and Zhichao Yan. Sam: A semantic-aware multi-tiered source de-duplication framework for cloud backup. Proceedings of the 39th International Conference on Parallel Processing (ICPP 2010), September 2010

[Kandemir09] Kandemir, Mahmut, John M. Dennis, Padma Raghavan and Qian Wang. “Adaptive Techniques for Achieving End-to-End QoS in the I/O Stack on Petascale Multiprocessors.” High End Computing University Research Activity NSF 09-530 (2009)

- Christina M Patrick, Nicholas Voshell, and Mahmut Kandemir. Appmap: Minimizing interference through application mapping in multi-level bu_er caches. In ISPASS, 2011

- M. Kandemir, S.P. Muralidhara, M. Karakoy, and S. W. Son. Computation mapping for multi-level storage cache hierarchies. Proc. of HPDC, 2010

- C.M. Patrick, M. Kandemir, M. Karakoy, S. W. Son, and A. Choudhary. Cashing in on hints for better prefetching and caching in pvfs and mpi-io. Proc. of HPDC, 2010

- S.J. Kim, Y. Zhang, S. W. Son, Ramya Prabhakar, M. Kandemir, Christina Patrick, W. K. Liao, and A. Choudhary. Automated tracing of I/O stack. Proc. of EuroMPI, 2010

- R. Prabhakar, S. Srikantaiah, M. Kandemir, and C. Patrick. Adaptive multi-level cache allocation in distributed storage architectures. Proc. of ICS, 2010

- R. Prabhakar, C.M. Patrick, and M. Kandemir. MPISec I/O: Providing Data Confidentiality in MPI-I/O. Proc. of CCGRID, 2009

- R. Garg, S. W. Son, M. Kandemir, P. Raghavan, and R. Prabhakar. Markov model based disk power management for data intensive workloads. Proc. of CCGRID, 2009

- C.M. Patrick, R. Garg, S. W. Son, and M. Kandemir. Improving I/O performance using soft-QoS-based dynamic storage cache partitioning. Proc. of CLUSTER, 2009

- S. W. Son, M. Kandemir, Y. Zhang, and R. Garg. Topology-aware I/O caching for shared storage systems. Proc. of ISCA PDCCS, 2009

- R. Garg, C.M. Patrick, and M. Kandemir. Dynamic storage cache partitioning using feedback control theory. Proc. of ISCA PDCCS, 2009

- R. Garg, R. Prabhakar, and M. Kandemir. Power aware disk allocation. Proc. of ISCA PDCCS, 2009

- R. Prabhakar, S. Srikantaiah, and C. Patrick andM. Kandemir. Dynamic storage cache allocation in multi-server architectures. Proc. of SC, 2009

- M. Kandemir, S.W. Son, and M. Karakoy. Improving I/O performance of applications through compiler-directed code restructuring. Proc. of FAST, 2008

- S.W. Son, M. Kandemir, and M. Karakoy. Improving I/O performance through compiler-directed code restructuring and adaptive prefetching. Proc. of IPDPS, 2008

- S.W. Son, S.P. Muralidhara, O. Ozturk, M. Kandemir, I. Kolcu, and M. Karakoy. Profiler and compiler assisted adaptive I/O prefetching for shared storage caches. Proc. of PACT, 2008

- C.M. Patrick, S.W. Son, and M. Kandemir. Enhancing the performance of MPI-IO applications by overlapping I/O, computation and communication. Proc. of PPoPP, 2008

- C.M. Patrick, S.W. Son, and M. Kandemir. Comparative evaluation of overlap strategies with study of I/O overlap in MPI-IO. Proc. of Operating Systems Review, 2008

- S.W. Son and M. Kandemir. Integrated data reorganization and disk mapping for reducing disk energy consumption. Proc. of CCGRID, 2007

- S.W. Son and M. Kandemir. Runtime system support for software-guided disk power management. Proc. of CLUSTER, 2007

- W. Liao et al. Improving MPI independent write performance using a two-stage write-behind buffering method. Proc. of IPDPS, 2007

- M. Kandemir, S.W. Son, and M. Karakoy. Improving disk reuse for reducing power consumption. Proc. of ISLPED, 2007

- S.W. Son, G. Chen, O. Ozturk, M. Kandemir, and A. Choudhary. Compiler-directed energy optimization for parallel disk based systems. IEEE Trans. Parallel Distrib. Syst., 2007

- S.W. Son and M. Kandemir. A prefetching algorithm for multi-speed disks. Proc. Of HiPEAC, 2007

[Khuller09] Khuller, Samir and Amol V Deshpande. “Optimization Algorithms for Large-scale, Thermal-aware Storage Systems.” High End Computing University Research Activity NSF 09-530 (2009)

[Klasky10] Klasky, Scott, Arie Shoshani, Karsten Schwan. “Runtime System for I/O Staging in Support of In-Situ Processing of Extreme Scale Data.” Scientific Data Management and Analysis at Extreme Scale ASCR FOA-10-0000256 (2010)

[Lang10] Lang, Sam and Chris Carothers. “CODES: Enabling Co-Design of Multi-Layer Exascale Storage Architectures.” Advanced Architectures and Critical Technologies for Exascale Computing ASCR FOA-10-0000255 (2010)

[Leiserson] Leiserson, Charles. “Microdata Storage Systems for High-End Computing.” Foundations of Computing Processes and Artifacts

- Aydin Bulu, Jeremy T. Fineman, Matteo Frigo, John R. Gilbert, Charles E. Leiserson. "Parallel sparse matrix-vector and matrix-transpose-vector multiplication using compressed sparse blocks," Proceedings of the 21st ACM Symposium on Parallelism in Algorithms and Architectures (SPAA), 2009, p. 233.

- Bradley C. Kuszmaul. "Covering Indexes: Orders-of-Magnitude Improvements," MySQL Users Conference, 2009.

- Bradley C. Kuszmaul. "Obliviousness in Scalable Computing Systems," ScalPerf09, 2009.

- Bradley C. Kuszmaul. "What is a Performance Model of SSDs?," Workshop on High Performance Transaction Systems, 2009.

- Bradley C. Kuszmaul and Jim Sukha. "Cache-Oblivious Workloads for Transactional Memory," The Workshop on Transactional Memory Workloads held at The ACM SIGPLAN 2006 Conference on Programming Language Design and Implementation (PLDI), 2006, p. 33.

- Bradley C. Kuszmaul.. "Cilk Provides the ``Best Overall Productivity" for High Performance Computing (and Won the HPC Challenge Award to Prove It).," Symposium on Parallelism in Algorithms and Architectures (SPAA 2007)., 2007, p. 299.

- Bradley C. Kuszmaul:. "Brief announcement: TeraByte TokuSampleSort sorts 1TB in 197s," Proceedings of the 21st ACM Symposium on Parallelism in Algorithms and Architectures (SPAA), 2009, p. 127.

- C. Scott Ananian, Krste Asanovic, Bradley C. Kuszmaul, Charles E. Leiserson, and Sean Lie.. "Unbounded Transactional Memory," The IEEE MICRO Special Issue: Top Picks from Computer Architecture Conferences, 2006, p. 59.

- Charles E. Leiserson. "The Cilk++ concurrency platform (invited paper)," 46th Design Automation Conference, Special Session on Multicore Computing and EDA, 2009.

- Charles E. Leiserson and Ilya Mirman. "How to survive the multicore software revolution (or at least survive the hype)," Journal of Advancing Technology, v.9, 2009, p. 42.

- Christopher Y. Crutchfield, Zoran Dzunic, Jeremy T. Fineman, David R. Karger, and Jacob H. Scott. "Improved Approximations for Multiprocessor Scheduling Under Uncertainty," Proceedings of the Twentieth ACM Symposium on Parallelism in Algorithms and Architectures (SPAA), 2008, p. 246.

- Edya Ladan-Mozes, Charles E. Leiserson. "A consistency architecture for hierarchical shared caches," Proceedings of the twentieth annual symposium on Parallelism in algorithms and architectures, 2008, p. 11.

- Jim Sukha. "Brief Announcement: A Lower Bound for Depth-Restricted Work Stealing," Proceedings of the 21st ACM Symposium on Parallelism in Algorithms and Architectures (SPAA), 2009, p. 124.

- Kunal Agrawal, Anne Benoit, Fanny DufossÃƒÂ©, Yves Robert. "Mapping filtering streaming applications with communication costs," Proceedings of the 21st ACM Symposium on Parallelism in Algorithms and Architectures (SPAA), 2009, p. 19.

- Kunal Agrawal, Charles E. Leiserson, and Jim Sukha. "Memory Models for Open-Nested Transactions," Proceedings of the ACM SIGPLAN Workshop on Memory Systems Performance and Correctness (MSPC), 2006, p. 70.

- Kunal Agrawal, I-Ting A. Lee, Jim Sukha. "Safer open-nested transactions through ownership," Proceedings of the 13th ACM SIGPLAN Symposium on Principles and practice of parallel programming, 2008, p. 291.

- Kunal Agrawal, I-Ting Angelina Lee, Jim Sukha. "Safe open-nested transactions through ownership," Proceedings of the Twentieth Annual Symposium on Parallelism in Algorithms and Architectures, 2008, p. 110.

- Kunal Agrawal, Jeremy T. Fineman, and Jim Sukha. "Nested Parallelism in Transactional Memory," Proceedings of the Second ACM SIGPLAN Workshop on Transactional Computing, 2007.

- Kunal Agrawal, Jeremy T. Fineman, Jim Sukha. "Nested parallelism in transactional memory," Proceedings of the 13th ACM SIGPLAN Symposium on Principles and practice of parallel programming, 2008, p. 163.

- Kunal Agrawal, Michael A. Bender, and Jeremy T. Fineman. "The Worst Page-Replacement Policy," Proceedings of the Fourth International Conference on Fun With Algorithms, 2007, p. 135.

- Kunal Agrawal, Yuxiong He, Charles E. Leiserson. "Adaptive work stealing with parallelism feedback," Proceedings of the 12th ACM SIGPLAN symposium on Principles and practice of parallel programming, 2007, p. 112.

- Kunal Agrawal, Yuxiong He, Wen Jing Hsu, and Charles E. Leiserson. "Adaptive scheduling with parallelism feedback," ACM Transactions on Computing Systems, v.16, 2008, p. 7:1.

- Matteo Frigo, Pablo Halpern, Charles E. Leiserson, Stephen Lewin-Berlin. "Reducers and other Cilk++ hyperobjects," Proceedings of the 21st ACM Symposium on Parallelism in Algorithms and Architectures (SPAA), 2009, p. 79.

- Michael A Bender, H Hu, and Bradley C. Kuszmaul. "Performance Guarantees for B-Trees with Different-Sized Atomic Keys," Proceedings of teh 29th ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems (PODS), 2010, p. 305.

- Michael A. Bender, Jeremy T. Fineman, and Seth Gilbert. "Contention Resolution with Heterogeneous Job Sizes," Proceedings of the 14th Annual European Symposium on Algorithms (ESA), 2006, p. 112.

- Michael A. Bender, Martin Farach-Colton, and Bradley C. Kuszmaul. "Cache-Oblivious String B-Trees," The 25th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems (PODS 2006, 2006, p. 233.

- Michael A. Bender, Martin Farach-Colton, Jeremy T. Fineman, Yonatan Fogel, Bradley Kuszmaul, Jelani Nelson, and Chris Wright. "Cache-Oblivious Streaming B-trees," Symposium on Parallelism in Algorithms and Architectures (SPAA 2007)., 2007, p. 8.

- Yuxiong He, Wen Jing Hsu, and Charles E. Leiserson. "Provably efficient online nonclairvoyant adaptive scheduling," IEEE International Parallel and Distributed Processing Symposium, 2007, p. 1.

- Yuxiong He, Wen Jing Hsu, and Charles E. Leiserson. "Provably efficient online nonclairvoyant adaptive scheduling," IEEE Transactions on Parallel and Distributed Systems, v.19, 2008, p. 1263.

[Ligon06] Ligon, Walter B. “Improving Scalability in Parallel File Systems for High End Computing.” High End Computing University Research Activity NSF 06-503 (2006)

[Long10] Long, Darrell. “Dynamic Non-Hierarchical File Systems for Exascale Storage.” Scientific Data Management and Analysis at Extreme Scale ASCR FOA-10-0000256 (2010)

[KMa09] Ma, Kwan-Liu, Peter H. Beckman and Kamil A. Iskra. “Visual Characterization of I/O System Behavior for High-End Computing.” High End Computing University Research Activity NSF 09-530 (2009)

- Chris Muelder, Francois Gygi, and Kwan-Liu Ma. "Visual Analysis of Inter-Process Communication for Large-Scale Parallel Computing," IEEE Transactions on Visualization and Computer Graphics, v.15, 2009, p. 1129.

[XMa06]Ma, Xiaosong, Anand Sivasubramaniam, Yuanyuan Zhou, John M. Blondin and Vincent W. Freeh. “Application-adaptive I/O Stack for Data-intensive Scientific Computing.” High End Computing University Research Activity NSF 06-503 (2006)

- Zhe Zhang, Chao Wang, Sudharshan S. Vazhkudai, Xiaosong Ma, Gregory G. Pike, John W. Cobb, Frank Mueller. "Optimizing center performance through coordinated data staging, scheduling and recovery," Proceedings of SC2007, 2007.

- Chao Wang, Zhe Zhang, Xiaosong Ma, Sudharshan Vazhkudai, and Frank Mueller. "Improving the Availability of Supercomputer Job Input Data Using Temporal Replication," The International Supercomputing Conference, 2009.

- Sudharshan Vazhkudai and Xiaosong Ma. "Recovering Transient Data: Automated On-demand Data Reconstruction and Offloading for Supercomputers," Operating Systems Review (OSR), Special Issue on File and Storage Systems, 2007.

- Zhe Zhang, Amit Kulkarni, Xiaosong Ma, Yuanyuan Zhou. "Memory Resource Allocation for File System Prefetching -- From a Supply Chain Management Perspective," The European Conference on Computer Systems (EuroSys '09), 2009.

- Zhe Zhang, Kyuhyung Lee, Xiaosong Ma, Yuanyuan Zhou. "PFC: Transparent Optimization of Existing Prefetching Strategies for Multi-level Storage Systems," 2008 International Conference on Distributed Computing Systems, 2008.

- Kim, J. Choi, S. Gurumurthi, A. Sivasubramaniam. "Graceful Operation of Disk Drives Under Thermal Emergencies," Proceedings of the International

Conference on Thermal Issues in Emerging Technologies Theory and Applications, 2007.

- H. Li, W-C Lee, A. Sivasubramaniam, C. Lee Giles. "A Hybrid Cache and Prefetch Mechanism for Scientific Literature Search Engines," Proceedings of the International Conference on Web Engineering, 2007, p. 121.

- Huajing Li, Wang-Chien Lee, Anand Sivasubramaniam, C. Lee Giles:. "Workload analysis for scientific literature digital libraries," Int. J. on Digital Libraries, v.9, 2008, p. 139.

- Shiva Chaitanya, Bhuvan Urgaonkar, Anand Sivasubramaniam. "QDSL: a queuing model for systems with differential service levels," ACM SIGMETRICS, 2008, p. 289.

[XMa09]Ma, Xiaosong, Frank Mueller, Kai Shen and Marianne Winslett. “Automatic Extraction of Parallel I/O Benchmarks from HEC Applications.” High End Computing University Research Activity NSF 09-530 (2009)

- Karthik Vijayakumar, Frank Mueller, Xiaosong Ma, Philip C. Roth. "Scalable Multi-Level I/O Tracing and Analysis," the 4th Petascale Data Storage Workshop (PDSW), in conjunction with Supercomputing 2009 (SC|09), 2009.

- Park, S; Shen, K. "A Performance Evaluation of Scientific I/O Workloads on Flash-Based SSDs," in IEEE International Conference on Cluster Computing (Cluster 2009)., 2009, p. 501-505.

- Min Li, Sudharshan Vazhkudai, Ali Butt, Fei Meng, Xiaosong Ma, Youngjae Kim, Christian Engelmann, and Galen Shipman. Functional partitioning to optimize end- to-end performance on many-core architectures. In SC, 2010

- Karthik Vijayakumar, Frank Mueller, Xiaosong Ma, and Philip C. Roth. Scalable multi- level i/o tracing and analysis. In the 4th Petascale Data Storage Workshop (PDSW), in conjunction with Supercomputing 2009, 2009

- Zhe Zhang, Amit Kulkarni, Xiaosong Ma, and Yuanyuan Zhou. Memory resource allocation for _le system prefetching: from a supply chain management perspective. In Proceedings of the EuroSys Conference, pages 75{88, 2009

- Xiaosong Ma, Sudharshan S. Vazhkudai, and Zhe Zhang. Improving data availability for better access performance: A study on caching scientific data on distributed desktop workstations. J. Grid Comput., 7(4):419{438, 2009

- Chao Wang, Zhe Zhang, Sudharshan S. Vazhkudai, Xiaosong Ma, and Frank Mueller. On-the-y recovery of job input data in supercomputers. In ICPP, pages 620-627, 2008

- Heshan Lin, Pavan Balaji, Ruth Poole, Carlos P. Sosa, Xiaosong Ma, and Wu chun Feng. Massively parallel genomic sequence search on the blue gene/p architecture. In SC, page 33, 2008

- Zhe Zhang, Kyuhyung Lee, Xiaosong Ma, and Yuanyuan Zhou. Pfc: Transparent optimization of existing prefetching strategies for multi-level storage systems. In ICDCS, pages 740-751, 2008

- Sudharshan Vazhkudai and Xiaosong Ma. Recovering transient data: automated on-demand data reconstruction and o_oading for supercomputers. Operating Systems Review, 41(1):14-18, 2007

- Z. Zhang, C.Wang, S. Vazhkudai, X. Ma, G. Pike, J. Cobb, and F. Mueller. Optimizing center performance through coordinated data staging, scheduling and recovery. In Proceedings of Supercomputing 2007 (SC07): Int'l Conference on High Performance Computing, Networking, Storage and Analysis, November 2007

[Maccabe06] Maccabe, Arthur B., Karsten Schwan, Patrick G. Bridges, Greg S. Eisenhauer, Ada Gavrilovska, Patrick A. Widener and Matthew Wolf. “Petascale Storage for High End Computing.” High End Computing University Research Activity NSF 06-503 (2006)

- Kumar, V; Cooper, BF; Cai, ZT; Eisenhauer, G; Schwan, K. "Middleware for enterprise scale data stream management using utility-driven self-adaptive information flows," CLUSTER COMPUTING-THE JOURNAL OF NETWORKS SOFTWARE TOOLS AND APPLICATIONS, v.10, 2007, p. 443-455.

[McDaniel09] McDaniel, Patrick, Radu Sion, Marianne Winslett and Erez Zadok. “Secure Provenance in High-End Computing Systems.” High End Computing University Research Activity NSF 09-530 (2009)

- Kevin Butler, Stephen McLaughlin, and Patrick McDaniel. "Disk-Enabled Authenticated Encryption," Proceedings of the 26th IEEE Symposium on Massive Storage Systems and Technologies, 2010.

- Patrick McDaniel, Kevin Butler, Stephen McLaughlin, Radu Sion, Erez Zadok, and Marianne Winslett. "Towards a Secure and Efficient System for End-to-End Provenance," The Proceedings 2nd USENIX Workshop on the Theory and Practice of Provenance, 2010.

- Miroslava Sotakova, Heike Busch, Stefan Katzenbeisser, Radu Sion. "The PUF Promise," Trust and Trustworthy Computing Conference (TRUST), 2010.

- P. Sehgal. "Optimizing Energy and Performance for Server-Class File System Workloads, Technical Report FSL-10-01," Master's Thesis, Stony Brook University, 2010.

- P. Sehgal, V. Tarasov, and E. Zadok. "Evaluating Performance and Energy in File System Server Workloads," Proceedings of the Eighth USENIX Conference on File and Storage Technologies (FAST '10), 2010.

- Patrick McDaniel, Kevin Butler, Stephen McLaughlin, Radu Sion, Erez Zadok, Marianne Winslett. "Towards a Secure and Efficient System for End-to-End Provenance," USENIX Workshop on the Theory and Practice of Provenance TAPP, 2010.

- Radu Sion, Junichi Tatemura. "Runtime Web-Service Workflow Optimization," Information and Software as Services, Springer-Verlag LNBIP, 2010.

- Radu Sion, Marianne Winslett. "A Road-Map to Regulatory Compliance for Information Systems," The Handbook of Financial Cryptography, CRC Press, 2010.

- Ragib Hasan, Radu Sion, Marianne Winslett. "Secure Provenance: Protecting the Genealogy of Bits," ;login: The USENIX Magazine, 2009.

- Ragib Hasan, Radu Sion, Marianne Winslett. "Preventing History Forgery with Secure Provenance," ACM Transactions on Storage TOS, 2009.

- Ragib Hasan, Radu Sion, Marianne Winslett. "Remembrance: The Unbearable Sentience of Being Digital," Conference on Innovative Data Systems Research CIDR, 2009.

- Yao Chen, Radu Sion, Bogdan Carbunar. "XPay: Practical anonymous payments for Tor routing and other networked services," Workshop on Privacy in the Electronic Society WPES, 2009.

- Kevin Butler and Petros Efstathopoulos. U Can't Touch This: Block-Level Protection for Portable Storage. In Proceedings of the International Workshop on Software Support for Portable Storage (IWSSPS'09), Grenoble, France, October 2009

- Thomas Moyer, Kevin Butler, Joshua Schi_man, Patrick McDaniel, and Trent Jaeger. Scalable Web Content Attestation. In ACSAC '09: Proceedings of the 2009 Annual Computer Security Applications Conference, 2009

- Kevin Butler, Stephen McLaughlin, and Patrick McDaniel. Kells: A Protection Framework for Portable Data. In Proceedings of the 26th Annual Computer Security Applications Conference (ACSAC), Austin, TX, USA, December 2010

- P. McDaniel, K. Butler, S. McLaughlin, R. Sion, E. Zadok, and M. Winslett. Towards a secure and effcient system for end-to-end provenance. In Proceedings

of the 2nd conference on Theory and practice of provenance. USENIX Association, 2010

- Kevin Butler, Stephen McLaughlin, Thomas Moyer, and Patrick McDaniel. New Security Architectures Based on Emerging Disk Functionality. IEEE Security and Privacy, 8(5):34-41, Sep./Oct. 2010

- Kevin Butler, Stephen McLaughlin, and Patrick McDaniel. Disk-Enabled Authenticated Encryption. In Proceedings of the 26th IEEE Symposium on Massive Storage Systems and Technologies (MSST), May 2010

- P. McDaniel, K. Butler, S. Mclaughlin, R. Sion, E. Zadok, and M. Winslett. Towards a Secure and E_cient System for End-to-End Provenance. In Proceedings of the second USENIX workshop on the Theory and Practice of Provenance (TAPP '10), San Jose, CA, February 2010. USENIX Association

- R. Sion et. al. An mls labeling infrastructure for clouds. Technical report, Stony Brook University, 2010

- S. J. Nadgowda and R. Sion. Cloud performance benchmark series: Amazon ebs, s3, and ec2 instance local storage. Technical report, 2010

- Md. B. Uddin, B. He, and R. Sion. Cloud performance benchmark series: Amazon rds tpc-c benchmark. Technical report, 2010

- Md. B. Uddin, B. He, and R. Sion. Cloud performance benchmark series: Amazon ec2 web serving. Technical report, 2010

[Mesnier06] Mesnier, Mike, Gregory R. Ganger, James Hendrix, Julio Lopez, Raja R. Sambasivan, Matthew Wachs. "//TRACE: Parallel Trace Replay with Approximate Causal Events" Carnegie Mellon University Parallel Data Lab Technical Report CMU-PDL-06-108, September 2006.

[Miller09] Miller, Ethan, Margo I. Seltzer and Darrell E. Long. "Scalable Data Management Using Metadata and Provenance." High End Computing University Research Activity NSF 09-530 (2009)

- Aleatha Parker-Wood, Christina Strong, Ethan L. Miller, and Darrell D. E. Long. Security aware partitioning for e_cient _le system search. May 2010

- Andrew W. Leung. Organizing, Indexing, and Searching Large-Scale File Systems. PhD thesis, University of California, Santa Cruz, December 2009

- Andrew Leung, Ian Adams, and Ethan L. Miller. Magellan: A searchable metadata architecture for large-scale _le systems. Technical Report UCSC-SSRC-09-07, University of California, Santa Cruz, November 2009

- Kiran-Kumar Muniswamy-Reddy and Margo Seltzer. Provenance as first class cloud data. ACM SIGOPS Operating Systems Review, 43, 2010

- Kiran-Kumar Muniswamy-Reddy and Margo Seltzer. Provenance as first-class cloud data. In 3rd ACM SIGOPS International Workshop on Large Scale Distributed Systems and Middleware (LADIS'09), 2009

- Kiran-Kumar Muniswamy-Reddy, Peter Macko, and Margo Seltzer. Provenance for the Cloud. In Proceedings of the 8th USENIX Conference on File and Storage Technologies, Feb 2010

- James Cheney, Stephen Chong, Nate Foster, Margo Seltzer, and Stijn Vansummeren. Provenance: a future history. In Proceedings of the 24th ACM SIGPLAN Conference Companion on Object Oriented Programming Systems Languages and Applications (OOPSLA 2009), October 2009

- Uri Braun, Margo Seltzer, Adriane Chapman, Barbara Blaustein, M. David Allen, and Len Seligman. Towards query interoperability: Passing plus. In Proceedings of the 2nd Workshop on the Theory and Practice of Provenance (TaPP '10), February 2010

- Kiran-Kumar Muniswamy Reddy. Foundations for Provenance-Aware Systems. PhD thesis, Harvard School of Engineering and Applied Sciences, March 2010

[Narasimhan06] Narasimhan, Priya, Chuck Cranor and Gregory R. Ganger. “Toward Automated Problem Analysis of Large Scale Storage Systems.” High End Computing University Research Activity NSF 06-503 (2006)

[Odlyzko06] Odlyzko, Andrew, David H. Du, Yongdae Kim and David J. Lilja. “Integrated Infrastructure for Secure and Efficient Long-Term Data Management.” High End Computing University Research Activity NSF 06-503 (2006)

- Biplob Debnath, Mohamed F. Mokbel, David J. Lilja, David Du. "Deferred Updates for Flash Based Storage," The 26th IEEE Symposium on Massive Storage Systems and Technologies (MSST2010)., 2010.

- Biplob Debnath, Mohamed F. Mokbel, David Lilja. "SARD: A Statistical Approach for Ranking Database Tuning Parameters," Proceedings of the Third International Workshop on Self-Managing Database Systems, SMDB 2008, 2008.

- Biplob Debnath, Srinivasan Krishnan, Weijun Xiao, David Lilja, David Du. "Sampling-based Metadata Management for Flash Storage," Laboratory for Advanced Research in Computing Technology and Compilers Technical Report No. ARCTiC 10-01, 2010.

- Biplob K. Debnath, Mohamed F. Mokbel and David J. Lilja. "Exploiting the Impact of Database System Configuration Parameters: A Design of Experiments Approach," IEEE Data Engineering Bulletin, 2008.

- David Du, Dingshan He, Changjin Hong, Jaehoon Jeong, Vishal Kher, Yongdae Kim, Yingping Lu, Aravindan Raghuveer, Sarah Sharafkandi. "Experiences in Building an Object-Based Storage System based on the OSD T-10 Standard," 14th NASA Goddard, 23rd IEEE Conference on Mass Storage Systems and Technologies, 2006.

- ingshan He, Xianbo Zhang, David H.C. Du, Gary Grider. "Coordinating Parallel Hierarchical Storage Management in Object-based Cluster File Systems," 14th NASA Goddard, 23rd IEEE Conference on Mass Storage Systems and Technologies, 2006.

- James Skarie, Biplob K. Debnath, David J. Lilja, and Mohamed F. Mokbel. "SCRAP: A Statistical Approach for Creating a Database Query Workload Based on Performance Bottlenecks," IEEE International Symposium on Workload Characterization, 2007.

- James Skarie, Biplob K. Debnath, David J. Lilja, and Mohamed Mokbel. "SCRAP: A Statistical Approach for Creating a Compact Representational Query Workload Based on Performance Bottlenecks," IEEE International Symposium on Workload Characterization (IISWC), 2007.

- Vishal Kher, Eric Seppanen, Cory Leach, Yongdae Kim. "SGFS: Secure, Efficient and Policy-based Global File Sharing," 14th NASA Goddard, 23rd IEEE Conference on Mass Storage Systems and Technologies, 2006.

- Xianbo Zhang, David Du, Jim Hughes, Ravi Kavuri. "HPTFS: A High Performance Tape File System," 14th NASA Goddard, 23rd IEEE Conference on Mass Storage Systems and Technologies, 2006.

[Panda06] Panda, D.K. and Pete Wyckoff. “Applicability of Object-Based Storage Devices in Parallel File Systems.” High End Computing University Research Activity NSF 06-503 (2006)

[Prabhat10] Prabhat, Quincy Koziol and Palmer. “ExaHDF5: An I/O Platform for Exascale Data Models, Analysis and Performance.” Scientific Data Management and Analysis at Extreme Scale ASCR FOA-10-0000256 (2010)

[Qin07] Qin, Xiao. “BUD: A Buffered-Disk Architecture for Energy Conservation in Parallel Disk Systems.” Foundations of Computing Processes and Artifacts NSF 06-585 (2007)

- Manzanares, X.-J. Ruan, S. Yin, M. Nijim, X. Qin, and W. Luo. "Energy-Aware Prefetching for Parallel Disk Systems: Algorithms, Models, and Evaluation,"

Proc. the 8th IEEE International Symposium on Network Computing and Applications, 2009.

- Manzanares, A. Roth, X.-J Ruan, S. Yin, M. Nijim, and X. Qin. "Conserving Energy in Real- Time Storage Systems with I/O Burstiness," ACM Transactions on Embedded Computing Systems, v.9, 2010.

- Manzanares, D. Hamilton, and X. Qin. "The Relationship Between Software Architecture and Visual Programming Languages," Proc. Grand Challenges in Modeling & Simulation, Edinburgh, 2008.

- Manzanares, K. Bellam, and X. Qin. "A Prefetching Scheme for Energy Conservation in Parallel Disk Systems," Proc. NSF Next Generation Software Program Workshop, 2008.

- Roth, A. Manzanares, K. Bellam, M. Nijim, and X. Qin. "Energy Conservation for Real-Time Disk Systems with I/O Burstiness," Proc. IEEE Int'l Workshop Next Generation Autonomous Storage and High Performance Computing, 2008.

- Liu, X. Qin, and S. Li,. "PASS: Power-Aware Scheduling of Mixed Applications with Deadline Constraints on Clusters," Proc. the 17th Int'l Conf. Computer Communications and Networks (ICCCN), 2008.

- Liu, X. Qin, S. Kulkarni, C.-J. Wang, S. Li, A. Manzanares, and S. Baskiyar. "Distributed Energy-Efficient Scheduling for Data-Intensive Applications with Deadline Constraints on Data Grids," Proc. 27th IEEE International Performance Computing and Communications Conference (IPCCC), 2008.

- J. Tjioe, R. Widjaja, A. Lee, and T. Xie. "DORA: A Dynamic File Assignment Strategy with Replication," The 38th International Conference on Parallel Processing, 2009.

- K. Bellam, A. Manzanares, and X. Qin. "Improving Reliability and Energy Efficiency of Disk Systems," Proc. 46th ACM Southeast Conference, 2008.

- K. Bellam, A. Manzanares, X. Ruan, X. Qin, and Y.-M. Yang. "Improving Reliability and Energy Efficiency of Disk Systems via Utilization Control," Proc. IEEE Symposium on Computers and Communications, 2008.

- K. Bellam, R.K. Vudata, X. Qin, Z.-L. Zong, M. Nijim, and X.-J. Ruan. "Interplay of Security and Reliability using Non-Uniform Checkpoints," Proc. 16th IEEE Int'l Conference on Computer Communications and Networks (ICCCN), 2007.

- M. Nijim, A. Manzanares, X.-J. Ruan, and X. Qin. "HYBUD: An Energy-Efficient Architecture for Hybrid Parallel Disk Systems," Proc. the 18th Int'l Conf. on Computer Communications and Networks, 2009.

- M. Nijim, A. Manzanares, and X. Qin. "An Adaptive Energy-Conserving Strategy for Parallel Disk Systems," Proc. the 12th IEEE Int'l Symp. Distributed Simulation and Real Time Applications (DS-RT), 2008.

- M. Nijim, Z.-L. Zong, K. Bellam, X.-J. Ruan and X. Qin. "Security-Aware Cache Management for Cluster Storage Systems," Proc. the 17th Int'l Conf. Computer Communications and Networks (ICCCN), 2008.

- T. Xie and A. Sharma. "Collaboration- Oriented Data Recovery for Mobile Disk Arrays," The 29th International Conference on Distributed Computing Systems (ICDCS 2009), 2009.

- T. Xie and Y. Sun. "A File Assignment Strategy Independent of Workload Characteristic Assumptions," ACM Transactions on Storage, v.5, 2009.

- T. Xie and H. Wang. "MICRO: A Multi-level Caching-based Reconstruction Optimization for Mobile Storage Systems," IEEE Transactions on Computers, v.57, 2008, p. 1386-1398.

- T. Xie and Y. Sun. "PEARL: Performance, Energy, and Reliability Balanced Dynamic Data Redistribution for Next Generation Disk Arrays," Proc. the 16th Annual Meeting of the IEEE International Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunication Systems (MASCOTS), 2008.

- Tao Xie and Xiao Qin. "Availability-Aware Stochastic Scheduling for Heterogeneous Clusters," Cluster Computing: The Journal of Networks, Software Tools and Applications, 2008.

- Tao Xie and Xiao Qin. "Allocation of Tasks with Availability Constraints in Heterogeneous Systems," IEEE Transactions on Computers, v.57, 2008, p. 188.

- Tao Xie and Xiao Qin. "Allocation of Tasks with Availability Constraints in Heterogeneous Systems," IEEE Transactions on Computers, v.57, 2008, p. 188-199.

- W. Luo, F.-M. Yang, L.-P. Pang, G. Tu, and X. Qin. "TERCOS: A Novel Approach to Exploiting Redundancies in Fault-Tolerant and Real-Time Distributed Systems," Proc. 13th IEEE International Conference on Embedded and Real-Time Computing Systems and Applications, 2007.

- W. Luo, X. Qin, X.-C. Tan, K. Qin, and A. Manzanares. "Exploiting Redundancies to Enhance Schedulability in Fault-Tolerant and Real-Time Distributed Systems," IEEE Transactions on Systems Man & Cybernetics, Part A: Systems and Humans, v.39, 2009, p. 626.

- X. Qin. "Design and Analysis of a Load Balancing Strategy in Data Grids," Future Generation Computer Systems: The Int'l Journal of Grid Computing, v.23, 2007, p. 131-137.

- X. Qin. "Performance Comparisons of Load Balancing Algorithms for I/O-Intensive Workloads on Clusters," Journal of Network and Computer Applications, v.31, 2008, p. 32-46.

- X. Qin, H. Jiang, A. Manzanares, X.-J Ruan, and S. Yin. "Dynamic Load Balancing for I/O-Intensive Applications on Clusters," ACM Transactions on Storage, v.5, 2009.

- X. Qin, H. Jiang, A. Manzanares, X.-J Ruan, and S. Yin. "Communication-Aware Load Balancing for Parallel Applications on Clusters," IEEE Transactions on Computers, v.59, 2010, p. 42.

- X. Qin, M. Alghamdi, M. Nijim, and Z.-L. Zong. "Scheduling of Periodic Packets in Energy-Aware Wireless Networks," Proc. the 26th IEEE Int'l Performance Computing and Communications Conf., 2007.

- X. Qin, M. Alghamdi, M. Nijim, Z.-L. Zong, X.-J. Ruan, K. Bellam, and A. A. Manzanares,. "Improving Security of Real-Time Wireless Networks Through Packet Scheduling," IEEE Transactions on Wireless Communications, v.7, 2008, p. 3273-3279.

- X.-J. Ruan, S. Yin, A. Manzanares, J. Xie, Z.-Y. Ding, J. Majors, and X. Qin. "ECOS: An Energy- Efficient Cluster Storage System," Proc. the 28th International Performance Computing and Communications Conference, 2009.

- X.-J. Ruan*, A. Manzanares, S. Yin, Z. -L. Zong, and X. Qin. "Performance Evaluation of Energy-Efficient Parallel I/O Systems with Write Buffer Disks," Proc. the 38th Int'l Conf. on Parallel Processing, 2009.

- X.-J. Ruan, A. Manzanares, K. Bellam, X. Qin. "DARAW: A New Write Buffer to Improve Parallel I/O Energy-Efficiency," Proc. the 24th Annual ACM Symposium on Applied Computing, 2009.

- X.-J. Ruan, A. Manzanares, K. Bellam, X. Qin. "DARAW: A New Write Buffer to Improve Parallel I/O Energy-Efficiency," the 24th Annual ACM Symposium on Applied Computing, 2009.

- X.-J. Ruan, S. Yin, A. Manzanares, M. Alghamdi, and X. Qin. "A Message Scheduling Scheme for Energy Conservation in Multimedia Wireless Systems," IEEE Trans. on Systems Man & Cybernetics, v.41, 2011, p. 272.

- X.-J. Ruan, X. Qin, M. Nijim, Z.-L. Zong, and K. Bellam. "An Energy-Efficient Scheduling Algorithm Using Dynamic Voltage Scaling for Parallel Applications on Clusters," Proc. 16th IEEE Int'l Conference on Computer Communications and Networks (ICCCN), 2007.

- Xiao Qin. "Performance Comparisons of Load Balancing Algorithms for I/O-Intensive Workloads on Clusters," Journal of Network and Computer Applications, v.31, 2008, p. 32.

- Z.-L. Zong, K. Bellam, X.-J. Ruan, A. Manzanares, X. Qin, and Y.-M Yang. "A Simulation Framework for Energy-efficient Data Grids," Proc. Winter Simulation Conference, 2008.

- Z.-L. Zong, M. Nijim, and X. Qin. "Energy-Efficient Scheduling for Parallel Applications on Mobile Clusters," Cluster Computing: The Journal of Networks, Software Tools and Applications, v.11, 2008, p. 91-113.

- Z.-L. Zong, M.E. Briggs, N.W. O'Connor, X. Qin, M. Alghamdi, and Y.-M. Yang. "Design and Performance Analysis of Energy-Efficient Parallel Storage Systems," Commodity Cluster Symposium, 2007.

- Z.-L. Zong, X. Qinï€ª , M. Nijim, X.-J. Ruan, K. Bellam, and M. Alghamdi. "Energy-Efficient Scheduling for Parallel Applications Running on Heterogeneous Clusters," Proc. 36th International Conference on Parallel Processing (ICPP), 2007.

- Z.-L. Zong, X.-J. Ruan, A. Manzanares, and X. Qin. "Energy-Aware Duplication Strategies for Scheduling Precedence Constrained Parallel Tasks on Clusters," IEEE Transactions on Computers, v.60, 2011, p. 360.

- Z.-L.Zong, M.E. Briggs, N.W. O'Connor, and X. Qin. "An Energy-Efficient Framework for Large-Scale Parallel Storage Systems," Proc. 21st Int'l Parallel and Distributed Processing Symp. (IPDPS), 2007.

[Rangaswami09] Rangaswami, Raju and Ming Zhao. “Streamlining High-End Computing with Software Persistent Memory.” High End Computing University Research Activity NSF 09-530 (2009)

- Y. Xu, L. Wang, D. Clavijo, Y. Liu, R. Figueiredo, and M. Zhao. "Virtualization-based Bandwidth Management for Parallel Storage Systems," 5th Petascale Data Storage Workshop, 2010.

- Jorge Guerra, Leonardo Marmol, Daniel Galano, Raju Rangaswami, and Jinpeng Wei. Software persistent memory. FIU SCIS Technical Report TR-2010-12-01, 2010

[Reddy06] Reddy, A. L. Narasimha. “Active Data Systems.” High End Computing University Research Activity NSF 06-503 (2006)

- S. Kang and A. L. Narasimha Reddy. "An approach to virtual allocation in storage systems," ACM Transactions on Storage, v.2, 2006, p. 371.

[Riska09] Riska, Alma and Mary Evgenia Smirni. “Interleaving Workloads with Performance Guarantees on Storage Cluster.” High End Computing University Research Activity NSF 09-530 (2009)

- Riska and E. Riedel. "Evaluation of disk-level workloads at different time-scales," 2009 IEEE International Symposium on Workload Characterization, IISWC 2009, 2009, p. 158.

- Riska, N. Mi, G. Casale, and E. Smirni. "Feasibility regions: exploiting tradeoffs between power and performance in disk drives," ACM SIGMETRICS Performance Evaluation Review, v.37(3), 2009, p. 43.

- Alma Riska, Ningfang Mi, Evgenia Smirni, and Giuliano Casale. Feasibility regions: exploiting tradeoffs between power and performance in disk drives. ACM SIGMETRICS Performance Evaluation Review, 37(3), 2009

- Xenia Mountrouidou, Alma Riska, and Evgenia Smirni. Adaptive workload shaping for power savings on disk drives. In International Conference in Performance Engineering (ICPE), 2011

- Xenia Mountrouidou, Alma Riska, and Evgenia Smirni. PREFiguRE: a performance, power, and reliability framework for disk drives. Technical report, 2010

- Alma Riska and Evgenia Smirni. Autonomic exploration of trade-o_s between power and performance in disk drives. In International Conference on Autonomic Computing and Communications (ICAC), pages 131-140, 2010

- Lei Lu, Ludmila Cherkasova, Vittoria de Nitto Persone, Ningfang Mi, and Evgenia Smirni. Await: E_cient overload management for busy multi-tier web services under bursty workloads. In International Conference on Web Engineering (ICWE), pages 81-97, 2010

[Rodrigues10] Rodrigues, Arun, John Shalf, Keren Bergman and Bruce Jacob. “Data Movement Dominates: Adding Data Management Services to Parallel File Systems.” Advanced Architectures and Critical Technologies for Exascale Computing ASCR FOA-10-0000255 (2010)

[Schroeder06] Schroeder, Bianca and Garth Gibson. “A Large-scale Study of Failures in High-performance-computing Systems.” Proceedings of the International Conference on Dependable Systems and Networks (DSN2006), Philadelphia, PA, USA, June 25-28, 2006.

[Shen06] Shen, Kai. “Concurrent I/O Management for Cluster-based Parallel Storages.” High End Computing University Research Activity NSF 06-503 (2006)

- Chuanpeng Li, Kai Shen, and Athanasios E. Papathanasiou. "Competitive Prefetching for Concurrent Sequential I/O," Proc. of the Second EuroSys Conference, 2007, p. 189.

- Kai Shen. "Parallel Sparse LU Factorization on Different Message Passing Platforms," Journal of Parallel and Distributed Computing (JPDC), v.66, 2006, p. 1387.

- Kai Shen, Christopher Stewart, Chuanpeng Li, and Xin Li. "Reference-Driven Performance Anomaly Identification," Proc. of ACM SIGMETRICS, 2009, p. 85.

- Kai Shen, Ming Zhong, Sandhya Dwarkadas, Chuanpeng Li, Christopher Stewart, and Xiao Zhang. "Hardware Counter Driven On-the-Fly Request Signatures," Proc. of the 13th International Conference on Architectural Support for Programming Languages and Operating Systems, 2008, p. 189.

- Ming Zhong, Kai Shen, and Joel Seiferas. "Correlation-Aware Object Placement for Multi-Object Operations," Proc. of the 28th International Conference on Distributed Computing Systems, 2008.

- Ming Zhong, Kai Shen, and Joel Seiferas. "Replication Degree Customization for High Availability," Proc. of the Third EuroSys Conference, 2008, p. 55.

- Pin Lu and Kai Shen. "Multi-Layer Event Trace Analysis for Parallel I/O Performance Tuning," Proc. of the 36th International Conference on Parallel Processing, 2007.

- Pin Lu and Kai Shen. "Virtual Machine Memory Access Tracing With Hypervisor Exclusive Cache," Proc. of the USENIX Annual Technical Conference, 2007, p. 29.

- Shen, K. "Parallel sparse LU factorization on different message passing platforms," JOURNAL OF PARALLEL AND DISTRIBUTED COMPUTING, v.66, 2006, p. 1387-1403.

- Zhong, M; Shen, K; Seiferas, J. "The convergence-guaranteed random walk and its applications in peer-to-peer networks," IEEE TRANSACTIONS ON COMPUTERS, v.57, 2008, p. 619-633.

CONFERENCE PROCEEDINGS

- Lu, P; Shen, K. "Virtual machine memory access tracing with hypervisor exclusive cache," in 2007 USENIX Annual Technical Conference., 2007, p. 29-43.

- Shen, K; Zhong, M; Dwarkadas, S; Li, CP; Stewart, C; Zhang, X. "Hardware counter driven on-the-fly request signatures," in 13th International Conference on Architectural Support for Programming Languages and Operating Systems., v.43, 2008, p. 189-200.

- Zhong, M; Lu, P; Shen, K; Seiferas, J. "Optimizing Data Popularity Conscious Bloom Filters," in 27th Annual ACM Symposium on Principles of Distributed Computing., 2008, p. 355-364.

- Zhong, M; Shen, K; Seiferas, J. "Correlation-Aware Object Placement for Multi-Object Operations," in 28th International Conference on Distributed Computing Systems., 2008, p. 512-521.

- Zhong, M; Shen, K; Seiferas, J. "Object Replication Degree Customization for High Availability," in 16th Annual ACM Symposium on Principles of Distributed Computing., 2007, p. 344-345.

[Shoshani06] Shoshani, Arie, Ilkay Altinas, Alok Choudhary, Terence Critchlow, Bill Gropp, Chandrika Kamath, Wei-Keng Liao, Bertram Ludaescher, Jarek Nieplocha, Steve Parker, Rob Ross, Doron Rotem, Nagiza Samatova, Rajeev Thakur, Jeff Vetter, Mladen Vouk. Scientific Data Management Center for Enabling Technologies, SciDAC2 10 Oct. 2006 http://www.scidac.gov/compsci/SDM.html

[Sivasubramaniam06] Sivasubramaniam, Anand and Patrick D. McDaniel. “Exploiting Asymmetry in Performance and Security Requirements for I/O in High-end Computing.” High End Computing University Research Activity NSF 06-503 (2006)

- Kevin Butler, Stephen McLaughlin, and Patrick McDaniel. "Non-Volatile Memory and Disks: Avenues for Policy Architectures," Proceedings of the 1st ACM Computer Security Architectures Workshop, 2007.

- Kevin Butler, Stephen McLaughlin, and Patrick McDaniel. "Disk-Enabled Authenticated Encryption," Proceedings of the 26th IEEE Symposium on Massive Storage Systems and Technologies: Research Track, May 2010, 2010.

- Kevin Butler, Stephen McLaughlin, Thomas Moyer, Joshua Schiffman, Patrick McDaniel, and Trent Jaeger. "Firma: Disk-Based Foundations for Trusted Operating Systems," Technical Report NAS-TR-0114-2009, Networking and Security Research Center, Department of Computer Science and Engineering, 2009.

- S. Chaitanya, K. Butler, A. Sivasubramaniam, P. McDaniel, M. Vilayannur. "Design, implementation and evaluation of security in iSCSI-based network storage systems. ," Proceedings of International Workshop on Storage Security and Survivability (StorageSS '06), 2006, p. 1.

[Smirni08] Smirni, Evgenia. “Effective Resource Allocation under Temporal Dependence Architectures.” Foundations of Computing Processes and Artifacts (2008)

- Riska, N. Mi, G. Casale, and E. Smirni. "Feasibility Regions: Exploiting Trade-offs between Power and Performance in Disk Drives," HotMetrics'09, Performance Evaluation Review, v.37(3), 2009, p. 43.

- G. Casale, E. Smirni. "MPA-AMVA: Approximate Mean Value Analysis of Bursty Systems," Proceedings of DSN/PDS (Dependable Systems and Networks), 2009, p. 1.

- G. Casale, E. Z. Zhang, E. Smirni. "Trace data characterization and fitting for Markov modeling," Performance Evaluation, v.59(1), 2010, p. 66.

- L. Cherkasova, K. Ozonat, N. Mi, J. Simmons, and E. Smirni. "Automatic Anomaly Detection and Performance Modeling," ACM Transactions on Computer Systems, v.27(3), 2009, p. 1.

- N. Mi , A. Riska, Q. Zhang, E. Smirni, and E. Riedel. "Efficient Management of Idleness in Systems," ACM Transactions on Storage (TOS), 2009.

- N. Mi, A. Riska, X. Li, E. Smirni, E. Riedel. "Restraint Utilization of Idleness for Transparent Scheduling of Background Tasks," Proceedings of ACM SIGMETRICS 2009, 2009, p. 205.

- N. Mi, G. Casale, A. Riska, Q. Zhang, and E. Smirni. "Autocorrelation-Driven Load Control in Distributed Systems," Proceedings of MASCOTS'09, 2009, p. 269.

- N. Mi, G. Casale, L. Cherkasova, and E. Smirnii,. "Burstiness in Multi-Tier Applications: Symptoms, Causes, and New Models," Proceedings of MIddleware 2008, Lecture Notes in Computer Science,, v.5346, 2008, p. 265.

- Ningfang Mi, Giuliano Casale, Ludmila Cherkasova, and Evgenia Smirni. "Injecting Realistic Burstiness to a Traditional Client-Server Benchmark," Proceedings of the 6th International Conference on Autonomic Computing (ICAC'09), 2009, p. 149.

- Q. Zhang, A. Heindl, A. Stathopoulos, and E. Smirni. "Comparison of two output models for the BMAP/MAP/1 departure process," Proceedings of QEST 2009, IEEE Press, 2009, p. 1.

[Sun06] Sun, Xian-He. William D. Gropp and Rajeev S. Thakur. “The Server-Push I/O Architecture for High End Computing.” High End Computing University Research Activity NSF 06-503 (2006)

- S. Byna, Y. Chen, X.-H. Sun. "Taxonomy of data prefetching for multicore processors," Journal of Computer Science and Technology, v.24, 2009, p. 405.

- X-H. Sun, S. Byna and Y. Chen. ""Server-based Data Push Architecture for Multi-processor Environments"," Journal of Computer Science and Technology, v.Vol. 22, 2007, p. 641.

- X.-H. Sun, Y. Chen. "Reevaluating Amdahl's Law in the Multicore Era," Journal of Parallel and Distributed Computing, 2009.

- Y. Chen, X.-H. Sun, and M. Wu. "Algorithm-System Scalability of Heterogeneous Computing," Journal of Parallel and Distributed Computing, v.68, 2008, p. 1403.

[Sun09] Sun, Xian-He, Surendra Byna, William D. Gropp and Rajeev S. Thakur. “A Dynamic Application-specific I/O Architecture for High End Computing.” High End Computing University Research Activity NSF 09-530 (2009)

- Yong Chen, Huaiyu Zhu, Philip C. Roth, Hui Jin, and Xian-He Sun. Global-aware and multi-order context-based prefetching for high-performance processors. the International Journal of High Performance Computing Applications, 2011

- Yong Chen, Xian-He Sun, Rajeev Thakur, Philip C. Roth, and William Gropp. Lacio: A new layout-aware collective i/o strategy for parallel i/o systems. In Parallel Distributed Processing (IPDPS), 2011 IEEE International Symposium on, May 2011

- Zhiling Lan, Jiexing Gu, Ziming Zheng, Rajeev Thakur, and Susan Coghlan. A study of dynamic meta-learning for failure prediction in large-scale systems. J. Parallel Distrib. Comput., 70:630-643, June 2010

- Xian-He Sun and Yong Chen. Reevaluating amdahl's law in the multicore era. J. Parallel Distrib. Comput., 70:183-188, February 2010

- Hui Jin, Xian-He Sun, Yong Chen, and Tao Ke. Remem: Remote memory as checkpointing storage. In Proceedings of the 2nd International Conference on Cloud Computing, CloudCom '10, 2010

- Rong Ge, X. Feng, J. Hu, and Xian-He Sun. Assessing energy effciency of parallel i/o systems(poster presentation). In Proceedings of the ACM/IEEE SuperComputing Conference, SC '10, 2010

- Huaiming Song, Xian-He Sun, Hui Jin, and Yong Chen. Trace-based adaptive data layout optimization for parallel _le systems (poster presentation). In 5th

Petascale Data Storage Workshop, in conjunction with the Supercomputing '10, PDSW '10, 2010

- Yong Chen, Xian-He Sun, Rajeev Thakur, Huaiming Song, and Hui Jin. Improving parallel i/o performance with data layout awareness. Cluster Computing, IEEE International Conference on, 0:302-311, 2010

- Hui Jin, Yong Chen, Huaiyu Zhu, and Xian-He Sun. Optimizing hpc fault-tolerant environment: An analytical approach. Parallel Processing, International Conference on, 0:525-534, 2010 33

- Yong Chen, Huaiyu Zhu, Hui Jin, and Xian-He Sun. Improving the e_ectiveness of context-based prefetching with multi-order analysis. Parallel Processing Workshops, International Conference on, 0:428{435, 2010

- Yong Chen, Huaiming Song, Rajeev Thakur, and Xian-He Sun. A layout-aware optimization strategy for collective i/o. In Proceedings of the 19th ACM International Symposium on High Performance Distributed Computing, HPDC '10, pages 360{363, New York, NY, USA, 2010. ACM

- Huaiyu Zhu, Yong Chen, and Xian-He Sun. Timing local streams: improving timeliness in data prefetching. In Proceedings of the 24th ACM International Conference on Supercomputing, ICS '10, pages 169{178, New York, NY, USA, 2010. ACM

- Yong Chen, Huaiyu Zhu, and Xian-He Sun. An adaptive data prefetcher for high-performance processors. In Proceedings of the 2010 10th IEEE/ACM International Conference on Cluster, Cloud and Grid Computing, CCGRID '10, pages 155-164, Washington, DC, USA, 2010. IEEE Computer Society

- Rong Ge, Xizhou Feng, S. Subramanya, and Xian he Sun. Characterizing energy effciency of i/o intensive parallel applications on power-aware clusters. In Parallel Distributed Processing, Workshops and Phd Forum (IPDPSW), 2010 IEEE International Symposium on, pages 1-8, April 2010

- Ming Wu and Xian-He Sun. IGI Global, May 2009

- Cong Du, P.Shukla, and Xian-He Sun. CRC Press, January 2009

- Surendra Byna, Yong Chen, and Xian-He Sun. A taxonomy of data prefetching mechanisms. In Proceedings of the The International Symposium on Parallel Architectures, Algorithms, and Networks, pages 19-24, Washington, DC, USA, 2008. IEEE Computer Society

- Yawei Li, Zhiling Lan, Prashasta Gujrati, and Xian-He Sun. Fault-aware runtime strategies for high-performance computing. IEEE Trans. Parallel Distrib. Syst., 20:460-473, April 2009

- Xian-He Sun, S. Byna, and D. Holmgren. Modeling data access contention in multicore architectures. In Parallel and Distributed Systems (ICPADS), 2009 15th International Conference on, pages 213-219, December 2009

- Bing Xie, Yong Chen, Xian-He Sun, and Hui Jin. Performance under failure of multitier web services. In Proceedings of the 2009 15th International Conference on Parallel and Distributed Systems, ICPADS '09, pages 776{781, Washington, DC, USA, 2009. IEEE Computer Society

- Xian-He Sun, Yong Chen, and Yanlong Yin. Data layout optimization for petascale file systems. In Proceedings of the 4th Annual Workshop on Petascale Data Storage, PDSW '09, pages 11{15, New York, NY, USA, 2009. ACM

- Xian-He Sun, Cong Du, Hongbo Zou, Yong Chen, and P. Shukla. V-mcs: A configuration system for virtual machines. In Cluster Computing and Workshops, 2009. CLUSTER '09. IEEE International Conference on, pages 1-7, September 2009

- Hui Jin, Xian-He Sun, Ziming Zheng, Zhiling Lan, and Bing Xie. Performance under failures of dag-based parallel computing. In Proceedings of the 2009 9th IEEE/ACM International Symposium on Cluster Computing and the Grid, CCGRID '09, pages 236-243, Washington, DC, USA, 2009. IEEE Computer Society

- Zhibin Fang, Xian-He Sun, Yong Chen, and Surendra Byna. Core-aware memory access scheduling schemes. Parallel and Distributed Processing Symposium, International, 0:1-12, 2009

- Yong Chen, Xian-He Sun, and MingWu. Algorithm-system scalability of heterogeneous computing. J. Parallel Distrib. Comput., 68:1403{1412, November 2008

- Luciano Piccoli, James B. Kowalkowski, James N. Simone, Xian-He Sun, Hui Jin, Donald J. Holmgren, Nirmal Seenu, and Amitoj G. Singh. Lattice qcd workows: A case study. In Proceedings of the 2008 Fourth IEEE International Conference on eScience, pages 620-625, Washington, DC, USA, 2008. IEEE Computer Society

- Yong Chen, Surendra Byna, Xian-He Sun, Rajeev Thakur, and William Gropp. Hiding i/o latency with pre-execution prefetching for parallel applications. In Proceedings of the 2008 ACM/IEEE conference on Supercomputing, SC '08, pages 40:1-40:10, Piscataway, NJ, USA, 2008. IEEE Press

- Surendra Byna, Yong Chen, Xian-He Sun, Rajeev Thakur, and William Gropp. Parallel i/o prefetching using mpi file caching and i/o signatures. In Proceedings of the 2008 ACM/IEEE conference on Supercomputing, SC '08, pages 44:1{44:12, Piscataway, NJ, USA, 2008. IEEE Press

- Xian-He Sun, Yong Chen, and Surendra Byna. Scalable computing in multicore era. In Proceedings of International Symposium on Parallel Algorithms, Architectures and Programming 2008, PAAP '08, 2008

- Surendra Byna, Yong Chen, Xian-He Sun, Rajeev Thakur, and William Gropp. Parallel i/o prefetching using mpi _le caching and i/o signatures. In Proceedings of the 2008 ACM/IEEE conference on Supercomputing, SC '08, pages 44:1-44:12, Piscataway, NJ, USA, 2008. IEEE Press

- Jiexing Gu, Ziming Zheng, Zhiling Lan, John White, Eva Hocks, and Byung-Hoon Park. Dynamic meta-learning for failure prediction in large-scale systems: A case study. In Proceedings of the 2008 37th International Conference on Parallel Processing, ICPP '08, pages 157-164, Washington, DC, USA, 2008. IEEE Computer Society

- Luciano Piccoli, J. Simone, and J. Kowalkowski. Tracking lqcd workflows(poster presentation). In Lattice 2008, 2008

- Surendra Byna, Yong Chen, and Xian-He Sun. A taxonomy of data prefetching mechanisms. In Proceedings of the The International Symposium on Parallel Architectures, Algorithms, and Networks, pages 19-24, Washington, DC, USA, 2008. IEEE Computer Society

- L. Piccoli, Xian-He Sun, and J. Simone. The lqcd workow experience: What we have learned (poster presentation). In Supercomputing, 2007. SC '07. Proceedings of the 2007 ACM/IEEE Conference on, November 2007

- Kirk W. Cameron, Rong Ge, and Xian-He Sun. lognp and log3p: Accurate analytical models of point-to-point communication in distributed systems. IEEE Trans. Computers, 56(3):314-327, 2007

- Xian-He Sun, Surendra Byna, and Yong Chen. Server-based data push architecture for multi-processor environments. J. Comput. Sci. Technol., 22:641-652, September 2007

- Xian-He Sun, Zhiling Lan, Yawei Li, Hui Jin, and Zhiming Zheng. Towards a fault aware computing environment. In Proc. of the High Availability and Performance Computing Workshop (HAPCW), March 2007

- Ming Wu, Xian-He Sun, and Hui Jin. Performance under failures of high-end computing. In Supercomputing, 2007. SC '07. Proceedings of the 2007 ACM/IEEE Conference on, pages 1 {11, November 2007

- Yong Chen, Surendra Byna, and Xian-He Sun. Data access history cache and associated data prefetching mechanisms. In Proceedings of the 2007 ACM/IEEE conference on Supercomputing, SC '07, pages 21:1-21:12, New York, NY, USA, 2007. ACM

- Prashasta Gujrati, Yawei Li, Zhiling Lan, Rajeev Thakur, and John White. A meta-learning failure predictor for blue gene/l systems. In Proceedings of the 2007 International Conference on Parallel Processing, ICPP '07, pages 40{, Washington, DC, USA, 2007. IEEE Computer Society

- Yawei Li, Prashasta Gujrati, Zhiling Lan, and Xian-he Sun. Fault-driven re-scheduling for improving system-level fault resilience. In Proceedings of the 2007 International Conference on Parallel Processing, ICPP '07, pages 39{, Washington, DC, USA, 2007. IEEE Computer Society

- Xian-He Sun and Ming Wu. Quality of service of grid computing: Resource sharing. In Proceedings of the Sixth International Conference on Grid and Cooperative Computing, GCC '07, pages 395{402, Washington, DC, USA, 2007. IEEE Computer Society

- Zhiling Lan, Yawei Li, Prashasta Gujrati, Ziming Zheng, Rajeev Thakur, and John White. A fault diagnosis and prognosis service for teragrid clusters. In Proceedings of conference of TeraGrid'07, TeraGrid '07, 2007

- Kun Xiao, Nianen Chen, Shangping Ren, Limin Shen, Xianhe Sun, Kevin Kwiat, and Michael Macalik. A workow-based non-intrusive approach for enhancing the survivability of critical infrastructures in cyber environment. In Proceedings of the Third International Workshop on Software Engineering for Secure Systems, SESS '07, pages 4-, Washington, DC, USA, 2007. IEEE Computer Society

- Cong Du, Xian-He Sun, and Ming Wu. Dynamic scheduling with process migration. In Proceedings of the Seventh IEEE International Symposium on Cluster Computing and the Grid, CCGRID '07, pages 92{99, Washington, DC, USA, 2007. IEEE Computer Society

- Xian-He Sun, S. Byna, and Yong Chen. Improving data access performance with server push architecture. In Parallel and Distributed Processing Symposium, 2007. IPDPS 2007. IEEE International, pages 1-6, March 2007

[Thain] Thain, Douglas. “Deconstructing Clusters for high end biometrics.” Software and Hardware Foundation/Compilers/ Advanced Computer Research Program

- Bui, H; Kelly, M; Lyon, C; Pasquier, M; Thomas, D; Flynn, P; Thain, D. "Experience with BXGrid: a data repository and computing grid for biometrics research," CLUSTER COMPUTING-THE JOURNAL OF NETWORKS SOFTWARE TOOLS AND APPLICATIONS, v.12, 2009, p. 373-386.

- Christopher Moretti, Jared Bulosan, Douglas Thain, and Patrick Flynn. "All-Pairs: An Abstraction for Data Intensive Cloud Computing," Proceedings of IEEE International Parallel and Distributed Processing Symposium, 2008.

- Douglas Thain and Christopher Moretti. "Efficient Access to Many Small Files in a Filesystem for Grid Computing," IEEE Grid Computing, 2007.

- Douglas Thain and Christopher Moretti. "Efficient Access to Many Small Files in a Filesystem for Grid Computing," Proceedings of IEEE Grid Computing, 2007.

- Hoang Bui, Deborah Thomas, Michael Kelly, Christopher Lyon, Douglas Thain, and Patrick J. Flynn. "Poster: BXGrid: A Data Repository and Workflow Abstraction for Biometrics Research," Proceedings of the IEEE International Conference on e-Science, 2008.

- Li Yu, Christopher Moretti, Andrew Thrasher, Scott Emrich, Kenneth Judd, and Douglas Thain. "Harnessing Parallelism in Multicore Clusters with the All-Pairs, Wavefront, and Makeflow Abstractions," Journal of Cluster Computing, v.13, 2010, p. 243.

- Li Yu, Christopher Moretti, Scott Emrich, Kenneth Judd, and Douglas Thain. "Harnessing Parallelism in Multicore Clusters with the All-Pairs and Wavefront Abstractions," Proceedings of IEEE High Performance Distributed Computing, 2009.

- Moretti, C; Bui, H; Hollingsworth, K; Rich, B; Flynn, P; Thain, D. "All-Pairs: An Abstraction for Data-Intensive Computing on Campus Grids," IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS, v.21, 2010, p. 33-46.

- Thain, D; Moretti, C; Hemmes, J. "Chirp: a practical global filesystem for cluster and Grid computing," JOURNAL OF GRID COMPUTING, v.7, 2009, p. 51-72.

CONFERENCE PROCEEDINGS

- Moretti, C; Bulosan, J; Thain, D; Flynn, PJ. "All-pairs: An abstraction for data-intensive cloud computing," in 22nd IEEE International Parallel and Distributed Processing Symposium (IPDPS 2008)., 2008, p. 1853-1863.

- Raicu, I; Foster, IT; Zhao, Y; Little, P; Moretti, CM; Chaudhary, A; Thain, D. "The Quest for Scalable Support of Data-Intensive Workloads in Distributed

Systems," in 18th ACM International Symposium on High Performance Distributed Computing (HPDC 2009)., 2009, p. 207-216.

- Thain, D; Moretti, C. "Efficient access to many small files in a filesystem for grid computing," in 8th IEEE/ACM International Conference on Grid Computing., 2007, p. 74-81.

- Yu, L; Moretti, C; Emrich, S; Judd, K; Thain, D. "Harnessing Parallelism in Multicore Clusters with the All-Pairs and Wavefront Abstractions," in 18th ACM International Symposium on High Performance Distributed Computing (HPDC 2009)., 2009, p. 1-10.

[Thottethodi06] Thottethodi, Mithuna S., Vijay S. Pai, Rahul T. Shah, T. N. Vijaykumar and Jeffrey S. Vitter. “Performance Models and Systems Optimization for Disk-Bound Applications.” High End Computing University Research Activity NSF 06-503 (2006)

- J. Dyaberi and K. Kannan and V. Pai. "Storage Optimization for a Peer-to-Peer Video-on-Demand Network," Proceedings of the ACM Multimedia Systems Conference, 2010.

- Mohamed Y. Eltabakh, Wing-Kai Hon, Rahul Shah, Walid G. Aref, Jeffrey Scott Vitter. "The SBC-tree: an index for run-length compressed sequences," In international conference on Extending Database Technology, 2008.

- Paolo Ferragina, Roberto Grossi, Ankur Gupta, Rahul Shah, Jeffrey Scott Vitter. "On searching compressed string collections cache-obliviously," In ACM Symposium on Principles of Database Systems, 2008.

- Sheng-Yuan Chiu, Wing-Kai Hon, Rahul Shah, Jeffrey Scott Vitter. "I/O-Efficient Compressed Text Indexes: From Theory to Practice," IEEE Data Compression Conference, 2010, p. 426.

- W. Hon, T. Lam, R. Shah, S. Tam, J. S. Vitter. "Cache Oblivious Index for approximate string matching," In Combinatorial Pattern Matching CPM 2007, 2007.

- Wing-Kai Hon, Manish Patil, Rahul Shah, Shih-Bin Wu. "Efficient index for retrieving top-k most frequent documents," Journal of Discrete Algorithms, v.8(4), 2010, p. 402.

- Wing-Kai Hon, Rahul Shah, Peter J Varman, Jeffrey Scott Vitter. "Tight Competitive Ratios for Parallel Disks Prefetching and Caching," In ACM symposium on Parallelism in Algorithms and Architectures (SPAA) 2008, 2008, p. 352.

- Wing-Kai Hon, Rahul Shah, Sharma V. Thankachan, Jeffrey Scott Vitter. "On Entropy-Compressed Text Indexing in External Memory," In international conference on String Processing and Information Retrieval, 2009.

- Yu-Feng Chien, Wing-Kai Hon, Rahul Shah, Jeffrey Scott Vitter. "Geometric Burrows-Wheeler Transform: Linking Range Searching and Text Indexing," In IEEE Data Compression Conference, 2008.

- Yung Ryn Choe and Vijay S. Pai. "Achieving Reliable Parallel Performance in a VoD Storage Server Using Randomization and Replication," Proceedings of the 21st International Parallel and Distributed Processing Symposium, 2007.

- Yung Ryn Choe, Chase Douglas, and Vijay S. Pai. "A Model and Prototype of a Resource-Efficient Storage Server for High-Bitrate Video-on-Demand," IPDPS Workshop on Performance Modeling, Evaluation, and Optimization of Parallel and Distributed Systems, 2007.

[Tosun07] Tosun, Ali Saman. “High Throughput I/O for Large Scale Data Repositories.” Foundations of Computing Processes and Artifacts NSF 06-585(2007)

- Ali Saman Tosun. "Divide-and-Conquer Scheme for Strictly Optimal Retrieval of Range Queries," ACM Transactions on Storage, v.5, 2009, p. 1.

- Tosun, AS. "Analysis and comparison of replicated declustering schemes," IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS, v.18, 2007, p. 1578-1591.

[Urgaonkar08] Urgaonkar, Bhuvan. “HybridStore: An Enterprise-scale Storage System Employing Solid-State Memory and Hard Disk Drives.” Foundations of Computing Processes and Artifacts (2008)

[Vetter10] Vetter, Jeffrey, Robert Schreiber, Trevor Mudge, Yuan Xie. “Blackcomb: Hardware-Software Co-design for Non-Volatile Memory in Exascale Systems.” Advanced Architectures and Critical Technologies for Exascale Computing ASCR FOA-10-0000255 (2010)

[Zadok06] Zadok, Erez, Ethan L. Miller and Klaus Mueller. “File System Tracing, Replaying, Profiling, and Analysis on HEC Systems.” High End Computing University Research Activity NSF 06-503 (2006)

- Leung, E. L. Miller, and S. Jones. "Scalable Security for Petascale Parallel File Systems," Proceedings of the 2007 International Conference for High Performance Computing, Networking, Storage and Analysis, v.1, 2008, p. 1.

- Leung, E. Lalonde, J. Telleen, C. Lee, and J. Davis. "Using Comprehensive Analysis for Performance Debugging in Distributed Storage Systems," In Proceedings of the 24th IEEE Conference on Mass Storage Systems and Technologies (MSST 2007), 2007, p. 281.

- Leung, S. Pasupathy, G. Goodson, and E. L. Miller.. "Measurement and Analysis of Large-Scale Network File System Workloads," Proceedings of the 2008 Annual USENIX Technical Conference, v.1, 2008.

- Traeger and E. Zadok and E. L. Miller and D. D. E. Long. "Findings from the First Annual Storage and File Systems Benchmarking Workshop," ;login: The USENIX Magazine, v.33(5), 2008, p. 113.

- Traeger, I. Deras, and E. Zadok.. "DARC: Dynamic Analysis of Root Causes of Latency Distributions," Proceedings of the 2008 International Conference on Measurement and Modeling of Computer Systems (SIGMETRICS 2008), v.1, 2008, p. 277.

- Traeger, K. Thangavelu, and E. Zadok. "Round-Trip Privacy with NFSv4," In Proceedings of the Third ACM Workshop on Storage Security and Survivability (StorageSS 2007), 2007, p. 1.

- Traeger, N. Joukov, C. P. Wright, and E. Zadok. "A Nine Year Study of File System and Storage Benchmarking," ACM Transactions on Storage (TOS), v.4, 2008, p. 25.

- Energy and Performance Evaluation of Lossless File Data Compression on Server Systems. "R. Kothiyal and V. Tarasov and P. Sehgal and E. Zadok," Proceedings of the Israeli Experimental Systems Conference (ACM SYSTOR '09), 2009.

- J. Nam, M. Maurer, and K. Mueller. "High-Dimensional Feature Descriptors to Characterize Volumetric Data," 2nd Workshop on Knowledge-Assisted Visualization (KAV), 2008.

- J. Nam, M. Maurer, K. Mueller. "A High-Dimensional Feature Clustering Approach to Support Knowledge-Assisted Visualization," Computers & Graphics, 33(5), 2009, p. 607.

- J. Nam, M. Maurer, K. Mueller.. "Semantic Visualization Facilitated By Cluster Analysis," First Workshop of Knowledge-Assisted Visualization (KAV 2007)., v.1, 2007.

- K. McDonnell, K. Mueller.. "Illustrative Parallel Coordinates," Computer Graphics Forum, v.27(3), 2008, p. 1031.

- L. Wang, A. Li, J. Giesen, P. Zolliker, K. Mueller.. "Color Design for Illustrative Visualization," Submitted to IEEE Transactions on Visualization and Computer Graphics, 2008.

- L. Wang, J. Giesen, K. McDonnell, P. Zolliker, and K. Mueller. "Color Design for Illustrative Visualization," IEEE Transactions on Visualization and Computer Graphics, v.14(6), 2008, p. 1739.

- M. W. Storer and K. M. Greenan and I. F. Adams and E. L. Miller and D. D. E. Long and K. Voruganti. "Logan: Automatic Management for Evolvable, Large-Scale, Archival Storage," 3rd International Petascale Data Storage Workshop archival storage (PDSW 08), 2008.

- N. Joukov, A. Traeger, R. Iyer, C. P. Wright, and E. Zadok. . "Operating System Profiling via Latency Analysis ," In Proceedings of the 7th Symposium on Operating Systems Design and Implementation (OSDI 2006). , 2006, p. 89.

- P. Sehgal and V. Tarasov and E. Zadok. "Evaluating Performance and Energy in File System Server Workloads extensions," Proceedings of the Eighth USENIX Conference on File and Storage Technologies (FAST '10), 2010, p. 253.

- R. Kothiyal and V. Tarasov and P. Sehgal and E. Zadok. "Energy and Performance Evaluation of Lossless File Data Compression on Server Systems," Proceedings of the Israeli Experimental Systems Conference (ACM SYSTOR '09), 2009.

- S. Garg, J. Nam, IV Ramakrishnan, K. Mueller.. "Model-Driven Visual Analytics," Proceedings of Visual Analytics Science and Technology Symposium (VAST 2008), v.1, 2008.

- S. Weil, S. Brandt, E. Miller, D. Long, C. Maltzahn. . "Ceph: A Scalable, High-Performance Distributed File System ," In Proceedings of the 7th Symposium on Operating Systems Design and Implementation (OSDI 2006). , 2006, p. 307.

- W. Xu and K. Mueller. "Parameter Space Visualizer: an Interactive Parameter Selection Interface for Iterative CT Reconstruction Algorithms," Proc. SPIE, vol. 7625 (SPIE Medical Imaging), 2010.

- W. Xu, K. Mueller. "Learning Effective Parameter Settings for Iterative CT Reconstruction Algorithms," Fully 3D Image Reconstruction in Radiology and Nuclear Medicine, 2009.

[Zadok09] Zadok, Erez and Geoffrey H Kuenning . “Performance- and Energy-Aware HEC Storage Stacks.” High End Computing University Research Activity NSF 09-530 (2009)

- P. Sehgal, V. Tarasov, and E. Zadok. "Evaluating Performance and Energy in File System Server Workloads," Proceedings of the Eighth USENIX Conference on File and Storage Technologies (FAST '10), 2010, p. 253.

- R. Spillane, S. Dixit, S. Archak, S. Bhanage, and E. Zadok. "Exporting Kernel Page Caching for Efficient User-Level I/O," Proceedings of the 26th International IEEE Symposium on Mass Storage Systems and Technologies, 2010, p. 1.

[Zhang] Zhang, Xiaodong. “Memory Caching and Prefetching to Improve I/O Performance in High-End Systems.” Foundations of Computing Processes and Artifacts/Advanced Computer Research Program

- Feng Chen, Song Jiang, and Xiaodong Zhang . "SmartSaver: turning flash drive into a disk energy saver for mobile computers," Proceedings of the 11th International Symposium on Low Power Electronics and Design (ISLPED'06), v.1, 2006, p. 412.

- Shuang Liang, Song Jiang, and Xiaodong Zhang. "STEP: Sequentiality and Thrashing Detection based Prefetching to improve performance of networked storage servers," Proceedings of the 27th International Conference on Distributed Computing Systems (ICDCS'07), v.1, 2007, p. 1.

- Xiaoning Ding, Song Jiang, Feng Chen, Kei Davis, and Xiaodong Zhang. "DiskSeen: exploiting disk layout and access history to enhance I/O prefetch," Proceedings of the 2007 USENIX Annual Technical Conference, (USENIX'07), v.1, 2007, p. 1.

[Zhang07] Zhang, Xiaodong and Song Jiang. “Algorithms Design and Systems Implementation to Improve Buffer Management for Fast I/O Data Access.” Foundations of Computing Processes and Artifacts NSF 06-585 (2007)

- Feng Chen, David Koufaty, and Xiaodong Zhang. "Understanding intrinsic characteristics and system implications of flash memory based solid state drives," Proceedings of 2009 ACM SIGMETRICS Conference on Measurement and Modeling of Computer Systems}, (SIGMETRICS/Performance 2009), 2009, p. 181.

- Jiang Lin, Qingda Lu, Xiaoning Ding, Zhao Zhang, Xiaodong Zhang, and P. Sadayappan. "Enabling software management for multicore caches with a lightweight hardware support," Proceedings of 22nd ACM/IEEE Annual Conference on Supercomputing (SC09), 2009, p. 1.

- Qingda Lu, Jiang Lin, Xiaoning Ding, Zhao Zhang, Xiaodong Zhang, and P. Sadayappan. "Soft-OLP: improving hardware cache performance through software-controlled object-level partitioning," Proceedings of 18th International Conference on Parallel Architectures and Compilation Techniques (PACT 2009), 2009, p. 1.

- Rubao Lee, Xiaoning Ding, Feng Chen, Qingda Lu, and Xiaodong Zhang. "MCC-DB: minimizing cache conflicts in muli-core processors for databases," Proceedings of 35th International Conference on Very Large Data Bases, 2009, p. 1.

- Shuang Liang, Song Jiang, Xiaodong Zhang. "STEP: Sequentiality and Thrashing Detection Based Prefetching to Improve Performance of Networked Storage Servers," The 27th IEEE International Conference on Distributed Computing Systems (ICDCS'07), 2007.

- Xiaoning Ding, Song Jiang, and Xiaodong Zhang. "BP-Wrapper: A Framework Making Any Replacement Algorithms (Almost) Lock Contention Free," Proceedings of the 25th IEEE Int'l Conference on Data Engineering (ICDE'09), 2009, p. 369.

- Xiaoning Ding, Song Jiang, Feng Chen, Kei Davis, and Xiaodong Zhang. "DiskSeen: Exploiting Disk Layout and Access History to Enhance I/O Prefetch," Proceedings of 2007 USENIX Annual Technical Conference (USENIX'07), 2007.

[Zhao09] Zhao, Ming and Renato J. Figueiredo. “QoS-driven Storage Management for High-end Computing Systems.” High End Computing University Research Activity NSF 09-530 (2009)

- Y. Xu, L. Wang, D. Clavijo, Y. Liu, R. Figueiredo, and M. Zhao. Virtualization-based bandwidth management for parallel storage systems. In 5th Petascale Data Storage Workshop(co-located with Supercomputing), 2010

APPENDIX A: HECURA, CPA, SciDAC2 and X-Stack FSIO Projects

High End Computing University Research Activity (HECURA) I/O 2006 Projects

- Collaborative Research: Petascale I/O for High End Computing;
  - Maccabe, Arthur B/ Schwann, Karsten; UNM/ Georgia Tech Research Corporation – GA Inst of Tech
  - HEC topics: Metadata, Next generation I/O architectures
  - Keywords
    - Higher level I/O abstractions via I/O graphs
    - Flexible metadata management by metabots
    - Rich metadata
    - Lightweight file systems

Motivation

Data-intensive HPC applications are becoming increasingly important, adding substantial challenges to the already daunting input/output requirements of MPP codes. A well-known example is the interpretation of data from seismic exploration. In these applications, I/O problems occur both from the large data volumes produced by seismic sensing and from the fact that this data must be manipulated to fit simulation requirements for translating the time series data from multiple sensor locations into a format ready for 3-D subsurface reconstruction. Similarly, in online collaboration systems, visualizations require conversion and/or filtering to meet client needs.

Problem Statement and Solution Approach

The difficulties faced by scientists and engineers in attaining high performance I/O for data-intensive MPP applications are exacerbated by the low level of abstraction presented by current I/O systems. This research will create higher level I/O abstractions for developers. Specifically, the SSDS framework we propose models I/O as I/O Graphs that `connect' application components with input or output mechanisms like file systems based on metadata constructed offline by autonomous metabots. SSDS enhances the I/O functionality available to end users in several ways. I/O Graphs can be programmed to realize application-specific I/O functionality, such as data filtering and conversion, data remeshing, and similar tasks. Their management is automated, including the mapping of their logical graph nodes to underlying physical MPP and distributed machine resources. I/O performance in SSDS will be improved by integrating the computational I/O actions of I/O Graphs with the backend file systems that store high volume data and with the I/O actions already taken by applications, and by moving metadata management offline into metabots.

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.

The implementation of the SSDS system and its I/O Graph model will impact a substantial HPC user community, due to its planned integration with the Lightweight File System (LWFS) currently under development at Sandia National Laboratories (SNL). This file system and its SSDS extensions will be deployed on large-scale machines at Sandia to demonstrate scalability and application utility. SSDS will be integrated with the file formats and file systems used by other groups at Sandia and at Oak Ridge National Laboratories (ORNL) (with whom we are also collaborating). In fact, Georgia Tech and UNM have a long history of collaboration with SNL and ORNL. Finally, our team has been working with the vendors of future processor technology: with IBM as part of the PERCS project, and with Intel, to better understand the implications for the high performance domain of future processor virtualization techniques

- Collaborative Research: Techniques for Streaming File Systems and Databases;
  - Bender, Michael A/Farach-Colton, Martin; SUNY at Stony Brook/Rutgers University New Brunswick
  - HEC topics: Metadata, Next generation I/O architectures, and File System and related Communication Protocols
  - Keywords
    - Streaming B-trees and variants for efficient data layout on disk, databases

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.

- Applicability of Object-Based Storage Devices in Parallel File Systems;
  - Wyckoff, Pete; Ohio State University Research Foundation
  - HEC topics: Metadata, File System and related Communication Protocols, and Next generation IO architectures
  - Keywords
    - Objects trade-offs and attributes
    - Applicability of OSDs in parallel file systems
    - Metadata

While continued improvements in processing speeds and disk densities improve computing over time, the most fundamental advances come from changing the ways in which components interact. Delegating responsibility for some operations from the host processor to intelligent peripherals can improve application performance. Traditional storage technology is based on simple fixed-size accesses with little assistance from disk drives, but an emerging standard for object-based storage devices (OSDs) is being adopted. These devices will offer improvements in performance, scalability and management, and are expected to be available as commodity items soon.

When assembled as a parallel file system, for use in high-performance computing, object-based storage devices offer the potential to improve scalability and throughput by permitting clients to securely and directly access storage. However, while the feature set offered by OSD is richer than that of traditional block-based devices, it does not provide all the functionality needed by a parallel file system.

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.

- Collaborative Research: SAM^2 Toolkit: Scalable and Adaptive Metadata Management for High-End Computing;
  - Jiang, Hong/Zhu, Yifeng; University of Nebraska-Lincoln/University of Maine
  - HEC topics: Metadata
  - Keywords
    - Scalable adaptive Metadata Management (SAM^2) tools
    - Predictive metadata access patterns
    - Bloom filters for load balance and scalability
    - Adaptive cache coherence protocol for metadata caching
    - Decentralized metadata group schemes

The increasing demand for Exa-byte-scale storage capacity by high end computing applications requires a higher level of scalability and dependability than that provided by current file and storage systems. The proposal deals with file systems research for metadata management of scalable cluster-based parallel and distributed file storage systems in the HEC environment. It aims to develop a scalable and adaptive metadata management (SAM2) toolkit to extend features of and fully leverage the peak performance promised by state-of-the-art cluster-based parallel and distributed file storage systems used by the high performance computing community.

The project involves the following components: 1. Develop multi-variable forecasting models to analyze and predict file metadata access patterns. 2. Develop scalable and adaptive file name mapping schemes using the duplicative Bloom filter array technique to enforce load balance and increase scalability 3. Develop decentralized, locality-aware metadata grouping schemes to facilitate the bulk metadata operations such as prefetching. 4. Develop an adaptive cache coherence protocol using a distributed shared object model for client-side and server-side metadata caching. 5. Prototype the SAM2 components into the state-of-the-art parallel virtual file system PVFS2 and a distributed storage data caching system, set up an experimental framework for a DOE CMS Tier 2 site at University of Nebraska-Lincoln and conduct benchmark, evaluation and validation studies.

- Improving Scalability in Parallel File Systems for High End Computing;
  - Ligon, Walter B; Clemson University
  - HEC topics: Metadata, Management and RAS, and Next generation I/O architectures
  - Keywords
    - Active caching and buffering
    - Server to server and client to client communication
    - Autonomics

        - Scalable metadata
        - Small unaligned data access
        - Reliability through redundancy

As high end computing systems (HECs) grow to several tens of thousands of nodes, file I/O is becoming a critical performance issue. Current parallel file systems such as PVFS2 and others, can reasonably stripe data across a hundred nodes and achieve good performance for bulk transfers involving large aligned accesses. Serious performance limits exist, however, for small unaligned accesses, metadata operations, and accesses impacted by the consistency semantics (any time one process writes data that is read by another).

The proposed research would address a few of these most critical issues through a straightforward application of engineering and research. The approach would build heavily on what is already known about similar problems in other distributed systems, especially distributed shared memory systems. These existing techniques would be studied in the new context of a parallel file system, adjusted, adapted, and where prudent, rejected for a novel approach. *The fundamental approach is to build quantitative evidence in support of each technique using analytical and simulation techniques, and to finally develop prototypes for PVFS2.* It is expected that the same techniques could be applied to any other parallel file system as well. A major focus would be on scalability as the key unit of evaluation. It is unclear if we would have the opportunity to test on a very large HEC, but we intend to simulate such machines and use machines we do have access to for validation of those simulations.

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 autonomics.* 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.

- The Server-Push I/O Architecture for High End Computing;
    - Sun, Xian-He; Illinois Institute of Technology
    - HEC topics: File systems and related Communication Protocols and Next generation I/O architectures
    - Keywords
        - Server side push
        - Collective I/O aware access patter prediction

Unlike traditional I/O designs where data is stored and retrieved by request, a new I/O architecture for High End Computing (HEC) is proposed based on a novel "Server-Push" model where a data access server proactively pushes data from a file server to the compute node's memory. The objective of this research is two fold: 1) increasing fundamental understanding of data access delay, 2) producing an effective I/O architecture that minimizes I/O latency. The PIs plan to increase the fundamental understanding through the study of data access pattern identification, prefetching algorithms, data replacement strategy, and extensive experimental testing. The PIs will verify the performance improvement with their file server design for various critical I/O intensive applications by using a combination of simulation and actual implementation in the PVFS2 file system.

- Collaborative research: Scalable I/O Middleware and File System Optimizations for High-Performance Computing;
  - Choudhary, Alok N/Kandemir, Mahmut T; Northwestern University/Pennsylvania State University University Park
  - File Systems and related Communication Protocols, Next generation I/O architectures, and Measurement and Understanding
    - Middleware cache
    - Small I/O
    - Collective
    - New APIs
    - New benchmarks

This project entails research and development to address several parallel I/O problems in the HECURA initiative. In particular, the main goals of this project are to design and implement novel I/O middleware techniques and optimizations, parallel file system techniques that scale to ultra-scale systems, design and development of techniques that efficiently enable newer APIs and flexible I/O benchmarks that mimic real and dynamic I/O behavior of science and engineering applications. The fundamental premise is that, to achieve extreme scalability, incremental changes or adaptation of traditional techniques for scaling data accesses and I/O will not succeed because they are based on pessimistic and conservative assumptions of parallelism and interactions. We will develop techniques to optimize data accesses that utilize the understanding of high-level access patterns ("intent"), and use that information through middleware and file systems to enable optimizations. Specifically, the objectives are to (1) design and develop middleware I/O optimizations and cache system that are able to capture small, unaligned, irregular I/O accesses from large number of processors and uses access pattern information to optimize for I/O; (2) incorporate these optimizations in MPICH2's MPI-IO implementation to make them available to a large number of users; (3) design and evaluate enhanced APIs for file system scalability, and (4) develop flexible, execution oriented and scalable I/O benchmarks that mimic the I/O behavior of real science, engineering and bioinformatics applications.

- Collaborative Research: Application-adaptive I/O Stack for Data-intensive Scientific Computing;
  - Ma, Xiaosong/Sivasubramaniam, Anand/Zhou, Yuanyuan; North Carolina State University/Pennsylvania State University University Park/University of Illinois at Urbana-Champaign
  - HEC topics: Next Generation I/O Architectures and QOS
  - Keywords
    - Parallel Adaptive I/O (PATIO)
    - Multilevel cache/vertical layer caching and pre-fetching

        - Access pattern recognition
        - Tunable consistency semantics
        - Content addressable storage
        - Cache partitioning between multiple workloads
        - Storage QoS

Advances in computational sciences have been greatly accelerated by the rapid growth of high-end computing (HEC) facilities. However, the continuous speedup of end-to-end scientific discovery cycles relies on the ability to store, share, and analyze the terabytes and petabytes of data generated by today's supercomputers. With the growing performance gap between I/O systems and processor/memory units, data storage and accesses are inevitably becoming more bottleneck-prone.

In this proposal, we address the I/O stack performance problem with adaptive optimizations at multiple layers of the HEC I/O stack (from high-level scientific data libraries to secondary storage devices and archiving systems), and propose effective communication schemes to integrate such optimizations across layers. In particular, our proposed PATIO (Parallel AdapTive I/O) framework explores multi-layer caching/prefetching that coordinates storage resources ranging from processors to tape archiving systems. This novel approach will bridge existing disjoint optimization efforts at each individual layer and responds to the critical call of improving the overall I/O system performance with increasingly deep HEC I/O stacks.

- Active Storage Networks for High End Computing;
    - Chandy, John A; Univ of Connecticut
    - HEC topics: Next Generation I/O Architectures
    - Keywords
        - Active storage networks-computation at the networks such as reductions and transformations

Recent developments in object-based storage systems and other parallel I/O systems with separate data and control paths have demonstrated an ability to scale aggregate throughput very well for large data transfers. However, there are I/O patterns that do not exhibit strictly parallel characteristics. For example, HPC applications typically use reduction operations that funnel multiple data streams from many storage nodes to a single compute node. In addition, many applications, particularly non-scientific applications, use small data transfers that can not take advantage of existing parallel I/O systems. In this project, we suggest a new approach called active storage networks (ASN) - namely putting intelligence in the network along with smart storage devices to enhance storage network performance. These active storage networks can potentially improve not only storage capabilities but also computational performance for certain classes of operations. The main goals of this project will include investigation of ASN topologies and architectures, creation of ASN switch from reconfigurable components, studying HEC applications for ASNs, protocols to support programmable active storage network functions, and storage system optimizations for ASNs.

- Active Data Systems;
    - Reddy, A.L. Narasimha; Texas A & M University

  - HEC topics: Next generation I/O architectures and QoS
  - Keywords
    - Broadening active disk applicability by examining running multiple applications at disk concurrently
    - Scheduling
    - Security
    - Sharing

This project plans to address several issues related to broadening the practicality of active storage. More specifically, this project plans to study and investigate:
(1) The impact of mixed workloads (both active and normal requests) at the active devices. (2) The impact of multiple active applications at the active devices. (3) The resource scheduling and QOS policies for a diverse set of workloads. (4) The impact of intelligent allocation in active storage systems.
In order to address these issues, the project plans to develop (a) an "active data" model to allow flexible processing of data, either at devices or at the requester. (b) QOS algorithms and security mechanisms for mixed workloads. (c) Algorithms and prototypes for exploiting the nature of data to develop content-based active storage.

- Quality of Service Guarantee for Scalable Parallel Storage Systems;
  - Chiueh, Tzi-Cker; SUNY at Stony Brook
  - HEC topics: Next generation I/O architectures, Measurement and Understanding, and QoS
  - Keywords
    - Platypus – storage system
    - QoS trace replay
    - Bandwidth guarantees
    - Prefetching using decoupled architecture by extracting a prefetch thread from the computation thread

The Platypus project will develop a parallel I/O system that supports guaranteed storage QoS for concurrently running parallel applications while maximizing the parallel storage system's utilization efficiency. In addition, it will implement a timing-accurate parallel trace play-back tool to evaluate the effectiveness and efficiency of the proposed parallel I/O system

- Concurrent I/O Management for Cluster-based Parallel Storages;
  - Shen, Kai; University of Rochester
  - HEC topics: Next generation I/O architectures
  - Keywords
    - Concurrent I/O workload
    - Disk seek/spin reduction by prefetching and anticipatory I/O scheduling
    - Server level coscheduling
    - Load adaptive parallel data aggregation

High-end parallel applications that store and analyze large scientific datasets demand scalable I/O capacity. One recent trend is to support high-performance parallel I/O using clusters of commodity servers, storage devices, and communication networks. When many processes in a parallel program initiate I/O operations simultaneously, the resulted concurrent I/O workloads present challenges to the storage system. At each individual storage server, concurrent I/O may induce frequent disk seek/rotation and thus degrade the I/O efficiency. Across the whole storage cluster, concurrent I/O may incur synchronization delay across multiple server-level actions that belong to one parallel I/O operation.

This project investigates system-level techniques to efficiently support concurrent I/O workloads on cluster-based parallel storages. Our research will study the effectiveness of I/O prefetching and scheduling techniques at the server operating system level. We will also investigate storage cluster level techniques (particularly co-scheduling techniques) to support better synchronization of parallel I/O operations. In parallel to developing new techniques, we plan to develop an understanding on the performance behavior of complex parallel I/O systems and explore automatic ways to help identify causes of performance anomalies in these systems.

- Performance Models and Systems Optimization for Disk-Bound Applications;
  - Thottethodi, Mithuna S; Purdue University
  - HEC topics: Next generation I/O architectures, Measurement and Understanding, and File System and related Communications Protocols
  - Keywords
    - Disk array modeling/algorithms
    - Network aware placement and migration
    - Power and thermal optimization via entropy-aware disk caching

Despite many recent breakthroughs in the understanding and optimization of data-intensive applications and disk-array-based systems, significant challenges remain in system modeling, algorithm design, and performance optimization. Existing analytical models do not incorporate application characteristics, internal disk behavior, and I/O interconnection network contention; these shortcomings cause two key problems.
First, optimization opportunities are lost since designers are compelled to design for the worst case rather than for specific application characteristics that may be significantly more benign. We propose an application characterization-driven approach wherein the behavior of the application (e.g., entropy, locality) shapes the optimization decisions.
Second, inaccurate models may lead to wasted design effort because of differences between model-predicted performance and actual disk-array performance. We propose a unified and flexible disk-array access model that improves accuracy by accounting for (a) the contention on the interconnection network between disks and memory and (b) internal disk behavior. We propose to develop and distribute an integrated execution-driven simulation environment that incorporates all the individual components described above.
We envision that the insights from our models and simulator will lead to a range of optimizations such as network-contention-aware data placement and migration policies, improved caching and pre-fetching policies and techniques to ameliorate power and thermal problems in large disk arrays.

- Exploiting Asymmetry in Performance and Security Requirements for I/O in High-end Computing;
    - Sivasubramaniam, Anand; Pennsylvania State University University Park
    - HEC topics: Security
    - Keywords
        - Data Vault – security
        - Tunable tradeoff between security and performance for site-specific policies
        - Visualization dashboard

Application sciences are more collaborative, with sharing of data sets becoming prevalent not just between users/applications of a single organization, but across organizations as well placing even higher performance requirements on the storage system. Given the sensitive nature of many of these applications, in addition to the performance demands, there is an impending need to secure such data from adversarial attacks. The consequences of security breaches can have far reaching consequences, over and beyond the costs of detecting and investigating such breaches. At the same time, one cannot fully confine the data physically since these need to be shared by collaborative applications from different administrative domains. Regulations are also mandating the maintenance of audit records and provenance of data.

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.

- Integrated Infrastructure for Secure and Efficient Long-Term Data Management;
    - Odlyzko, Andrew; University of Minnesota-Twin Cities
    - HEC topics: Security, Archive
    - Keywords
        - Security
        - Hierarchical cluster-based archive
        - Long-term key management

To achieve the level of security and privacy for enterprise data that is increasingly required by laws or industry standards, data should be encrypted both at rest and in transit. Yet, numerous recent privacy breaches through loss or theft of archival tapes or notebook computers show that today most data, even of extremely sensitive nature, is not encrypted. The main reason is that we do not have a flexible system for key management. Loss of the encryption key (through lapses of memory, death of staff members, or destruction of stored copies) would mean that the owner of the data would effectively lose it completely, with potentially catastrophic consequences.

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.

- Formal Failure Analysis for Storage Systems;
  - Arpaci-Dusseau, Remzi H; Univ of Wisconsin-Madison
  - HEC topics: Management and RAS and Measurement and Understanding
  - Keywords
    - Formal analysis of failures with Wisconson's Program Analysis of Storage Systems (PASS) program

Building scalable storage systems requires robust tolerance of the many faults that can arise from modern devices and software systems. Unfortunately, many important storage systems handle failure in a laissez-faire manner. In this proposal, we describe the Wisconsin Program Analysis of Storage Systems project (PASS), wherein we seek to develop the techniques needed to build the high-end, scalable, robust storage systems of tomorrow. Our focus in PASS is to bring a more formal approach to the problem, utilizing programming language tools to build, analyze, test, and monitor these storage systems. By applying these techniques, we will raise the level of trust in the failure-handling capabilities of high-end storage systems by an order of magnitude.

The PASS project will change the landscape of storage systems in three fundamental ways. First, by developing more formal failure analysis techniques, we will be able to uncover a much broader range of storage system failure-handling problems. Second, within PASS we will develop more robust and scalable testing infrastructure; such a framework will be of general use to the development of any future storage system. Finally, through run-time instrumentation of a large Condor cluster, we plan to gather information as to what types of faults occur in practice as well as how they manifest themselves as failures. Such data will be invaluable to future designs and implementations of robust, scalable storage systems.

- Toward Automated Problem Analysis of Large Scale Storage Systems;
  - Narasimhan, Priya; Carnegie-Mellon University
  - HEC topics: Measurement and Understanding and Management and RAS
  - Keywords
    - Continuous performance and anomaly tracing
    - Auto blame assignment and performance diagnosis
    - Automated analysis of failure and performance degradation

This research explores methodologies and algorithms for automating analysis of failures and performance degradations in large-scale storage systems. Problem analysis includes such crucial tasks as identifying which component(s) misbehaved, likely root causes, and supporting evidence for any conclusions. Automating problem analysis is crucial to achieving cost-effective storage at the scales needed for tomorrow's high-end computing systems, whose scale will make problems common rather than anomalous. Moreover, the distributed software complexity of such systems make by-hand analysis increasingly untenable.

Combining statistical tools with appropriate instrumentation, the investigators hope to significantly reduce the difficulty of analyzing performance and reliability problems in deployed storage systems. Such tools, integrated with automated reaction logic, also provide an essential building block for the longer-term goal of self-healing. The research involves understanding which statistical tools work and how well in this context for the problems of problem detection/prediction, identifying which components need attention, finding root causes, and diagnosing performance problems. It will also involve quantifying the impact of instrumentation detail on the effectiveness of those tools so as to guide justification for associated instrumentation costs. Explorations will be done primarily in the context of the Ursa Minor/Major cluster-based storage systems via fault injection and analysis of case studies observed in its deployment.

- File System Tracing, Replaying, Profiling, and Analysis on HEC Systems;
  - Zadok, Erez; SUNY at Stony Brook
  - HEC topics: Measurement and Understanding, Next generation I/O architectures, and File System and related Communication Protocols
  - Keywords
    - Visualization
    - Tracing and replaying file system activity

File systems are difficult to analyze, as they are affected by OS internals, hardware used, device drivers, disk firmware, networking, and applications. Traditional profiling systems have focused on CPU usage, not on I/O latencies. Worse, existing tools for profiling, analysis, and visualization are too simplistic, cannot cope with massive and complex data streams, and do not scale to large clusters. We have expertise in single-host file system tracing, replaying, profiling, and benchmarking---as well as having developed over 20 file systems; large data analysis and visualization; and designing and implementing petabyte-size storage clusters.

In this project we are developing tools and techniques that will work on large clusters and scale well. We are conducting large scale tracing and replaying, collecting vital information useful to analyze the cluster's performance given a specific application. We use automated and user-driven feedback to raise or lower the level of tracing on individual cluster nodes to (1) ``zoom in'' on hot-spots and (2) trade off information accuracy vs. overheads. We use advanced data analysis techniques to identify performance bottlenecks, and we will visualize them for cluster users for ease of analysis. The end goal is to help identify I/O bottlenecks in running distributed applications, so as to improve their performance significantly---resulting in more effective use of these expensive clustering resources by scientists worldwide.

- End-to-End Performance Management for Large Distributed Storage;
  - Brandt, Scott A; University of California, Santa Cruz
  - HEC topics: QoS
  - Keywords
    - QoS server side
    - Server I/O scheduling
    - Server and client cache management
    - Client-to-server network flow control
    - Client-to-server connection management

End-to-end Performance Management for Large Distributed Storage Scott Brandt, Darrell Long, and Carlos Maltzahn, UC Santa Cruz Richard Golding and Theodore Wong, IBM Almaden Research Center.

Storage systems for large and distributed clusters of compute servers are themselves large and distributed. Their complexity and scale make it hard to manage these systems and, in particular, to ensure that applications using them get good, predictable performance. At the same time, shared access to the system from multiple applications, users, and internal system activities leads to a need for predictable performance.

This project investigates mechanisms for improving storage system performance in large distributed storage systems through mechanisms that integrate the performance aspects of the path that I/O operations take through the system, from the application interface on the compute server, through the network, to the storage servers. We focus on five parts of the I/O path in a distributed storage system: I/O scheduling at the storage server, storage server cache management, client-to-server network flow control, client-to-server connection management, and client cache management.

- Performance Insulation and Predictability for Shared Cluster Storage
  - Ganger, Greg R.; Carnegie-Mellon University
  - HECIWG topics: QoS
  - Keywords:
    - Performance Insulation

This research explores design and implementation strategies for insulating the performance of high-end computing applications sharing a cluster storage system. In particular, such sharing should not cause unexpected inefficiency. While each application may see lower performance, due to only getting a fraction of the total attention of the I/O system, none should see less work accomplished than the fraction it receives. Ideally, no I/O resources should be wasted due to interference between applications, and the I/O performance achieved by a set of applications should be predictable fractions of their non-sharing performance. Unfortunately, neither is true of most storage systems, complicating administration and penalizing those that share storage infrastructures.

Accomplishing the desired insulation and predictability requires cache management, disk layout, disk scheduling, and storage-node selection policies that explicitly avoid interference. This research combines and builds on techniques from database systems (e.g., access pattern shaping and query-specific cache management) and storage/file systems (e.g., disk scheduling and storage-node selection). Two specific techniques are: (1) Using prefetching and write-back that is aware of the applications associated with data and requests, efficiency-reducing interleaving can be avoided; (2) Partitioning the cache space based on per-workload benefits, determined by recognizing each workload's access pattern, one application's data cannot get an unbounded footprint in the storage server cache.

- Microdata Storage Systems for High-End Computing
  - Leiserson, Charles; MIT
  - HECIWG Topics: Metadata / Next generation I/O architectures, File System and related Communication Protocols
  - Keywords:
    - Cache Oblivious Data Structures
    - Buffered Repository B-trees
    - Virtual-memory-based transactional memory

This research project is aimed at understanding and developing microdata storage systems, a technology which is needed for many application ares, including genome processing and radar knowledge formation. Microdata storage systems are designed to perform well for small files (microfiles), as well as for large files (macrofiles). Today's filesystems are optimized for reading and writing data in large blocks, but they perform poorly when dealing with large volumes of microdata.

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.

The investigators employ benchmarks, such as the DARPA HPC SSCA#3 benchmark (an I/O-only version of which they developed), to evaluate the impact of microdata storage systems on high-end computing. The investigators are also developing course materials on microdata storage systems which will be made freely available under the MIT OpenCourseware initiative http://ocw.mit.edu.

- Memory Caching and Prefetching to Improve I/O Performance in High-End Systems
  - Zhang, Xiaodong; Ohio State University
  - HECIWG Topic: Measurement and Understanding
  - Keywords:

        - Buffer Cache Management

This research project will focus on a buffer caching topic: to develop and test a general clock-based system framework for caching management in a large scope of storage hierarchy for core, distributed and Internet systems. The PI will design and implement a clock-based and unified memory buffer management framework with following unique merits: (1) it does not require any global synchronization, and it is system independent; (2) it will be easily used by any types of buffer management at any level of the storage hierarchy, such as buffer caches for I/O data, data buffer for large scientific data bases, memory buffers for large data streams, and others; and (3) it will be designed to flexibly adopt and test different types of novel ideas of exploiting data access localities.

- Deconstructing Clusters for high end biometrics
    - Thain, Douglas; Notre Dame
    - HECIWG Topic: Next Generation I/O Architectures
    - Keywords:
        - Deconstructing Clusters

Today's user of scientific computing facilities has easy access to thousands of processors. However, this bounty of processing power has led to a data crisis. A conventional computing system often dispatches hundreds or thousands of jobs that simultaneously access a centralized server, which inevitably becomes a bottleneck. To support large data intensive applications, clusters must expose control of their internal storage and computing resources to an external scheduler that can make more informed placement decisions. This technique is called deconstructing clusters.

This project attacks a particular data-intensive problem in high-end biometric research: the pair-wise comparison of hundreds of thousands of face images. The technique of deconstructing clusters will be used to parallelize the workload across large computing clusters. If successful, this project will reduce the time to develop and analyze a new biometric matching algorithm from years to days, thus improving the productivity of biometric researchers. The broader impact upon society will be an improvement in the accuracy and efficiency of biometric identification for commercial and national security. The software will be published in open source form in order to benefit other scientific computations with a similar pair-wise computation model.

Foundations of Computing Processes and Artifacts (CPA) 2007 I/O Projects

- BUD: A Buffered-Disk Architecture for Energy Conservation in Parallel Disk Systems
    - Qin, Xiao; New Mexico Institute of Mining and Technology
    - HECIWG Topic: Management and RAS
    - Keywords:
        - Power management and Energy Aware

Parallel disks consisting of multiple disks with high-speed switched interconnect are ideal for data-intensive applications running in high-performance computing systems.

Improving the energy efficiency of parallel disks is an intrinsic requirement of next generation high-performance computing systems, because a storage subsystem can represent 27% of the energy consumed in a data center. However, it is a major challenge to conserve energy for parallel disks and energy efficiently coordinate I/Os of hundreds or thousands of concurrent disk devices to meet high-performance and energy-saving requirements. This research investigates novel energy conservation techniques to provide significant energy savings while achieving low-cost and high-performance for parallel disks. In this research project, the investigators take an organized approach to implementing energy-saving techniques for parallel disks, simulating energy-efficient parallel disk systems, and conducting a physical demonstration. This research involves four tasks: (1) design and develop a buffer-disk (BUD) architecture to reduce energy dissipation in parallel disk systems; (2) develop innovative energy-saving techniques, including an energy-related reliability model, energy-aware data partitioning, disk request processing, data movement, data placement, prefetching strategies, and power management for buffer disks; (3) implement a simulation toolkit (BUDSIM) used to develop a variety of energy-saving techniques and their integration in the BUD architecture; and (4) validate the BUD architecture along with our innovative energy-conservation techniques using real data-intensive applications running on high-performance clusters. This research can benefit society by developing economically attractive and environmentally friendly parallel disk systems, which are able to lower electricity bills and reduce emissions of air pollutants. Furthermore, the BUD architecture and the energy-conservation techniques can be transferable to embedded disk systems, where power constraints are more severe than conventional disk systems.

- Algorithms Design and Systems Implementation to Improve Buffer Management for Fast I/O Data Access
  - Zhang, Xiaodong/ Jiang, Song; Ohio State University Research Foundation/ Wayne State University
  - HECIWG Topic: Measurement and Understanding
  - Keywords:
    - Using disk layout to improve buffer cache

Although processor cycles, memory size, and disk capacity all become increasingly abundant, there is still a serious deficiency in the system support for handling data-intensive applications, which is the long latency of hard disk accesses, measured by the time to get the first byte of requested data. This latency improvement has significantly lagged behind other system component improvement, including disk peak bandwidth. To address this critical issue, the investigators will develop new and efficient buffer cache management systems that adapt to the dramatic technology changes and the high demand of data-intensive applications with complicated access patterns. Aiming at making the memory buffer as a truly effective agent between the requests from applications and services provided by disks, the investigators will leverage the cache and prefetch mechanisms in the memory buffer to improve effective I/O system performance, perceived by applications, by minimizing the cost (both energy and time) of expensive disk accesses. A unique approach to be adopted in the research is to put the disk layout information directly on the map of buffer management and effectively integrate both temporal and spatial localities. The investigators will design and implement a system infrastructure that analyzes and exploits data layout information on disks. With this critical system support, the investigators will further design and implement dual-side-aware memory buffer management algorithms that adapt to characteristics exhibited at both programs' side and disks' side.

- High Throughput I/O for Large Scale Data Repositories
  - Tosun, Ali Saman; University of Texas at San Antonio
  - HECIWG Topic: Metadata
  - Keywords:
    - Declustering, high dimensional data

Declustering 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

- Object Based Caching for MPI-IO
  - Dickens, Phillip M; University of Maine
  - HECIWG Topic: Next Generation I/O Architectures
  - Keywords:
    - Small unaligned I/O, Next generation middleware

As the size of large-scale computing clusters increases from thousands to tens of thousands of nodes, the challenge of providing high-performance parallel I/O to MPI applications executing in such environments becomes increasingly important and difficult. There are many factors that make this problem so challenging. The most often cited difficulties include the I/O access patterns exhibited by scientific applications (e.g., non-contiguous I/O), poor file system support for parallel I/O optimizations, strict file consistency semantics, and the latency of accessing I/O devices across a network. However, we believe that a more fundamental problem, whose solution would help alleviate all of these challenges, is the legacy view of a file as a linear sequence of bytes. The problem is that application processes rarely access data in a way that matches this file model, and thus a large component of the scalability problem is the cost of dynamically translating between the process data model and the file data model. In fact, the data model used by applications is more accurately defined as an object model, where each process maintains a collection of (perhaps) unrelated objects. We believe that aligning these different data models will significantly enhance the performance of parallel I/O for large-scale, data-intensive applications. This research is developing the infrastructure to merge the power and flexibility of the MPI-IO parallel I/O interface with a more powerful object-based file model. Toward this end, we are developing an object-based caching system that serves as an interface between MPI applications and object-based files. The object-based cache is based on MPI file views, or, more precisely, the intersections of such views. These intersections, which we term objects, identify all of the file regions within which conflicting accesses are possible and (by extension) those regions for which there can be no conflicts (termed shared-objects and private-objects respectively). This information will be leveraged by the runtime system to maximize the parallelism of file accesses and minimize the cost of enforcing strict file consistency semantics and global cache coherence. In this way, the performance and scalability characteristics of large-scale, data-intensive MPI applications will be significantly enhanced.

- A High Throughput Massive I/O Storage Hierarchy for PETA-scale High-end Architectures
  - Gao, Guang R.; University of Delaware
  - HECIWG Topic: Next Generation I/O Architectures
  - Keywords:
    - Small unaligned I/O, Next generation middleware

There has been significant progress in the research and development of modern high-end computer (HEC) architecture that is comprised of tens-of-thousands of processors or more. This has widened the performance gap between the computing power and the storage and I/O performance that can support and sustain such calculations. This gap presents a great challenge for the scalability of future parallel I/O architecture models and I/O middleware support. To address these challenges, we propose a new I/O architecture model in which each node has a dedicated high-bandwidth connection to its own local solid state storage (FLASH memory). We will propose and develop an I/O middleware model and software support that will exploit the features of the proposed I/O architecture model. We will also develop new management and RAS (reliability, accessibility and serviceability) capabilities that can scale to the new peta-scale architecture. Two flash memories will be visible to each node, the local flash memory and a neighboring node’s flash memory that will keep a backup copy. Dedicated service agents make this dual connection configuration transparent to nodes, by managing the traffic according to priority, current usage, and availability. We will implement the proposed solutions by leveraging the extension of an experimental HEC system software testbed to simulate the proposed I/O architecture and middleware models as well as the RAS support. We also plan to demonstrate the effectiveness of our proposal for the most common set of third party I/O benchmarks.

Foundations of Computing Processes and Artifacts (CPA) 2008 I/O Projects

- Effective Resource Allocation under Temporal Dependencend Architectures
  - Smirni, Evgenia; William and Mary
  - HECIWG Topic: Next Generation I/O Architectures
  - Keywords:
    - Next generation middleware

Temporal dependence within the workload of any computing or networking system has been widely recognized as a significant factor affecting performance. More specifically, autocorrelation in flows, is catastrophic for performance. In a simple single server system, autocorrelation in the arrival intensities or service demands may result in user response times that are slower by several orders of magnitude. In homogeneous clusters where size-based load balancing policies have been proved optimal for performance, autocorrelation in the arrival intensities of jobs obliterates any performance benefit of traditional load balancing policies. In multi-tiered systems, if a service process of any of the tiers is autocorrelated, then user response times are very high, in spite of the fact that the bottleneck resource in the system is far from saturation and that the measured throughput and utilizations in all other tiers are also modest, falsely indicating that the system can sustain higher capacities. In storage systems, autocorrelation in the arrival or service processes at the disk level may result in significant user-perceived performance degradation. This project aims at providing a practical way to characterize and quantify the performance impacts of autocorrelated flows in systems. The main focus is on the development of new technologies for resource allocation that consider autocorrelation as an important characteristic of any stochastic process. On-line monitoring of autocorrelation provides the necessary information for scheduling parameterization, making an important step toward the development of autonomic systems.

- HybridStore: An Enterprise-scale Storage System Employing Solid-State Memory and Hard Disk Drives
  - Urgaonkar, Bhuvan; Penn State University
  - HECIWG Topic: Next Generation I/O Architectures
  - Keywords:
    - Novel storage devices

The mechanical movement inherent in the operation of the hard disk poses access speed limits for many workloads and storage systems are consuming increasing amounts of power. Flash memory overcomes some key limitations of the hard disk including faster access to non-sequential data and significantly lower power usage. Encouraged by these advantages offered by flash memory and the recent emergence of high-capacity flash drives, this research will design and evaluate a hybrid system. Named HybridStore, this system will exploit complementary properties of these two media to provide improved performance, service differentiation, and thermal/power behavior in enterprise-scale storage. HybridStore will comprise a dynamic data management solution that will adapt the use of available flash to workload conditions. Techniques for improving performance (e.g., moving non-sequential content to flash, use of flash as a write buffer) will be investigated. The investigators will explore how flash can facilitate improved service differentiation by reducing the variance of access times inherent in the operation of disks. Finally, the the feasibility of selected replication of popular content on disk and flash and diverting more IO traffic to flash during periods of thermal emergencies will be investigated. Power savings resulting from opportunities to slow down disks, without compromising performance, will also be explored. The investigators will implement a Linux-based prototype Direct-Attached Storage HybridStore system that will manage a hard disk drive and a SATA-enabled flash drive attached to the shared IO bus. To explore other hybrid configurations (such as flash on disk or RAID controller), a comprehensive simulator called HybridSim will be implemented. The PIs will enhance the graduate and undergraduate curricula at Penn State with topics related to this research.

High End Computing University Research Activity (HECURA) 2009 I/O Projects

- HaRD: The Wisconsin Hierarchically-Redundant, Decoupled Storage Project
    - Arpaci-Dusseau, Remzi; University of Wisconsin-Madison
    - HECIWG Topic: Next Generation I/O Architecture
    - Keywords:
        - xxxx

The Wisconsin Hierarchically-Redundant, Decoupled storage project (HaRD) investigates the next generation of storage software for hybrid Flash/disk storage clusters. The main objective of the project is to improve the performance of storage in a variety of diverse scenarios, including new application environments such as photo storage as found in Facebook and Flickr, high-end scientific processing as found in government labs, and large-scale data processing such as that found in Google and Microsoft. The HaRD project focuses on three key issues in order to improve performance of these important applications: client-side Flash-based RAID and file-system integration, server-side memory reduction and multicore scheduling of file-system tasks, and scheduled network transfers. HaRD pulls together these technologies into a synthesized whole through three targeted storage systems: a scalable photo server, a high-performance checkpoint subsystem, and an improved file system for MapReduce workloads. The impact of this project is significant, as HaRD helps to shape the storage software architecture of the next generation of cloud computing services, which are of increasing relevance to both industry and society at large.

- CRAM: A Congestion-Aware Resource and Allocation Manager for Data-Intensive High-Performance Computing

  - Burns, Randal; Johns Hopkins University
  - HECIWG Topic: Quality of Service
  - Keywords:
    - xxxx

This project will develop a job scheduling and resource allocation system for data-intensive high-performance computing (HPC) based on the congestion pricing of a systems' heterogeneous resources. This extends the concept of resource management beyond processing: it allocates memory, disk I/O, and the network among jobs. The research will overcome the critical shortcomings of processor-centric resource management, which wastes huge portions of cluster and supercomputer resources for data-intensive workloads, e.g. I/O bandwidth governs the performance of many modern HPC applications but, at present, it is neither allocated nor managed. The research will develop techniques that (1) reconfigure the degree of parallelism of HPC jobs to avoid congestion and wastage, (2) support lower-priority, allocation elastic jobs that can be scheduled on arbitrary numbers of nodes to consume unallocated resource fragments, and (3) co-schedule batch-processing workloads that use system resources that are unoccupied due to asymmetric utilization and temporal shifts in the foreground jobs. These techniques will be implemented and supported for free public use as extensions to an open-source resource-management framework. If used broadly, the software has the potential to provide much better utilization of the national investment in HPC facilities.

- Active Object Storage to Enable Scalable and Reliable Parallel File System
  - Chandy, John; University of Connecticut
  - HECIWG Topic: Next Generation I/O Architecture
  - Keywords:

The increasing performance and decreasing cost of processors has enabled increased system intelligence at peripherals such as disk drives. This computational capability at the disk has led to the development of object-based storage whereby some of the file system functionality is moved to the disk. The computation capability can also enable computation at the storage node in what has been called active disks or active storage. This active storage computation serves as a mechanism to enable parallel computation using distributed storage nodes.

This research focuses on the use of these active disks for parallel file system and storage management. A functional active storage system architecture built on the standardized object-storage device specification is being developed. The architecture supports a variety of execution engines allowing multiple programming languages and models. Using this active object storage architecture, mechanisms to improve overall scalability and large-scale system reliability are being investigated. In addition, active and object storage are used to enable customizable and extensible file systems including autonomic (self-configuring and self-managing) storage as well as application aware storage such that the storage can be optimized for application and user needs.

- An Application Driven I/O Optimization Approach for PetaScale Systems and Scientific Discoveries
  - Choudhary, Alok and Wei-keng Liao; Northwestern University
  - HECIWG Topic: Next Generation I/O Architecture
  - Keywords:
    - Small unaligned I/O, Next generation middleware

This research focuses on developing scalable parallel file access methods for multi-scale problem domain decompositions, such as the one presented in Adaptive Mesh Refinement (AMR) based algorithms. Existing parallel I/O methods concentrate on optimizing the process collaboration under a fairly evenly-distributed request pattern. However, they are not suitable for data structures in AMR, because the underlying data distribution is highly irregular and dynamic. Process synchronization in the existing parallel I/O methods can penalize the I/O parallelism if the process collaboration is not carefully coordinated. This research addresses such synchronization issue by developing scalable solutions in the Parallel netCDF library (PnetCDF), particularly to address AMR structured data and its I/O patterns. PnetCDF is a popular I/O library used by many computer simulation communities. A scalable solution for storing and accessing AMR data in parallel is considered a challenging task. This research will design a process-group based parallel I/O approach to eliminate unrelated processes and thus avoid possible I/O serialization. In addition, a new metadata representation will also be developed in pnetCDF for conserving tree-structured AMR data relationship in a portable form.

- EAGER: Autonomous Data Partitioning Using Data Mining for High End Computing
  - Dhall, Sudarshan; University of Oklahoma
  - HECIWG Topic:
  - Keywords:
    - xxxx

Query response time and system throughput are the most important metrics when it comes to database and file access performance. Because of data proliferation, efficient access methods and data storage techniques have become increasingly critical to maintain an acceptable query response time and system throughput. One of the common ways to reduce disk I/Os and therefore improve query response time is database clustering, which is a process that partitions the database/file vertically (attribute clustering) and/or horizontally (record clustering). To take advantage of parallelism to improve system throughput, clusters can be placed on different nodes in a cluster machine.

This project develops a novel algorithm, AutoClust, for database/file clustering that dynamically and automatically generates attribute and record clusters based on closed item sets mined from the attributes and records sets found in the queries running against the database/files. The algorithm is capable of re-clustering the database/file in order to continue achieving good system performance despite changes in the data and/or query sets. The project then develops innovative ways to implement AutoClust using the cluster computing paradigm to reduce query response time and system throughput even further through parallelism and data redundancy. The algorithms are prototyped on a Dell Linux Cluster computer with 486 compute nodes available at the University of Oklahoma. For broader impacts, performance studies are conducted using not only the decision support system database benchmark (TPC-H) but also real data recorded in database and file formats collected from science and healthcare applications in collaboration with domain experts, including scientists at the Center for Analysis and Prediction of Storms (CAPS) at the University of Oklahoma. The project also makes important impacts on education as it provides training for graduate and undergraduate students working on this project in the areas of national critical needs: database and file management systems, and high-end computing and applications. The developed algorithm and prototype, real datasets and performance evaluation results are made available to the public at the Website: http://www.cs.ou.edu/~database/AutoClust.html.

- RUI: Automatic Identification of I/O Bottleneck and Run-time Optimization for Cluster Virtualization
  - He, Xubin, Tennessee Technological University
  - HECIWG Topic: Measurement and Understanding
  - Keywords:
    - xxxx

Extending virtualization technology into high-performance, cluster platforms generates exciting new possibilities. However, I/O efficiency in virtualized environments, specifically with respect to disk I/O, remains little understood and hardly tested.

The objective of this research is to investigate fundamental techniques for virtual clusters that not only facilitate rigorous performance studies, but also identify places where performance is suffering and then optimize the system to lessen the impact of such bottlenecks. To accomplish this objective, the following research tasks will be conducted: 1) An in-depth analysis of I/O efficiency in virtualized environments and investigation of intelligent and automated I/O bottleneck identification schemes; 2) Design and development of techniques to optimize I/O to address the detected I/O bottlenecks; 3) Development of an extensible framework for characterizing I/O workloads across virtualized clusters.

This research will greatly contribute to understanding virtualized I/O, identifying I/O bottlenecks and optimizing I/O, and thus facilitate the cluster systems to most effectively utilize virtualization technology. This project will also contribute to the society through promoting research and engaging under-represented groups that leads students to advancing their careers in science and engineering.

- Adaptive Techniques for Achieving End-to-End QoS in the I/O Stack on Petascale Multiprocessors
  - Kandemir, Mahmut/ John Dennis; Pennsylvania State Univ University Park/ National Center For Atmospheric Research
  - HECIWG Topic: Quality of Service
  - Keywords:

        - xxxx

Emerging high-end computing platforms, such as leadership-class machines at the petascale, provide new horizons for complex modeling and large-scale simulations. These machines are used to execute data intensive applications of national interest such as climate modeling, cosmic microwave background radiation, and astrophysical thermonuclear flashes. While these systems have unprecedented levels of peak computational power and storage capacity, a critical challenge concerns the design and implementation of scalable I/O (input-output) system software (also called I/O stack) that makes it possible to harness the power of these systems for scientific discovery and engineering design. Unfortunately, currently, there are no available mechanisms that accommodate I/O stack-wide, application-level QoS (quality-of-service) specification, monitoring, and management.

This project investigates a revolutionary approach to the QoS-aware management of the I/O stack using feedback control theory, machine learning, and optimization. The goal is to maximize I/O performance and thus improve overall performance of large scale applications of national interest. The project uses (1) machine learning and optimization to determine the best decomposition of application-level QoS to sub-QoSs targeting individual resources, and (2) feedback control theory to allocate shared resources managed by the I/O stack such that the specified QoSs are satisfied throughout the execution. The project tests the developed I/O stack enhancements using the workloads at NCAR, LBNL and ANL systems. It also involves two efforts in broadening participation: CISE Visit in Engineering Weekends (VIEW) and NASA-Aerospace Education Services Project (NASA-AESP) at the Center for Science and the Schools (CSATS).

- Optimization Algorithms for Large-scale, Thermal-aware Storage Systems
    - Khuller, Samir; University of Maryland College Park
    - HECIWG Topic: Management an RAS
    - Keywords:
        - xxxx

This project investigates optimization problems that arise while performing thermal management in very large data storage centers. To satisfy the growing data management needs, such storage centers contain possibly hundreds of thousands of hard disks and other components, and typically are consistently active. These generate a lot of heat, and hence the storage system must be cooled to maintain reliability, resulting in significant cooling costs. The cooling mechanism and the workload assignments in a storage center are intricately tied together.

This project is developing a general science of thermal management for large scale storage systems, by focusing on thermal modeling and management at different levels of the system hierarchy. Thermal aware techniques for allocating data access tasks to specific disks on which data is located, for controlling the schedules and speeds of thousands of tasks and disks to optimize quality of service, and for reorganizing data layouts on disks are being developed. This project will enable better thermal management in data storage centers, which can potentially result in significant reductions in the carbon footprint caused by those. The project will train several Ph.D. students in conducting research both at the University, and through internships at Industrial Research Labs.

- Multidimensional and String Indexes for Streaming Data
  - Leiserson, Charles E. /Bradley C. Kuszmaul (MIT)/ Michael A. Bender (Stony Brook) / Martin Farach-Colton (Rutgers)
  - HECIWG Topic: Metadata
  - Keywords:
    - XXXXX

This research project aims to understand and develop systems for maintaining superlinear indexes for streaming data. A superlinear index describes over an abstract space that cannot easily be linearized. In contrast, a linear index, typified by a B-tree, supports point and range queries on totally ordered data.

Examples of superlinear indexes include (1) multidimensional indexes, which can be over a geometric domain, such as geographic data, or which can be over multiple linear indexes; and (2) full text queries, which can include searching for a particular word or substring.

The superlinear indexes found in today's databases cannot support high rates of insertion. On traditional mechanical disk drives, the existing superlinear indexes can only support about one hundred insertions per second in the worst case. For many important applications, that is too slow, and so database users often avoid superlinear indexing. Even traditional linear indexes based on B-trees cannot support the high insertion rates demanded by many databases.

This research investigates streaming superlinear indexes, that is, indexes that efficiently support full text or multidimensional queries, and can be updated at speeds that are related to disk bandwidth rather than seeks per second.

Among the significant research issues are the following: (1) design efficient files structures for streaming superlinear indexes; (2) investigate how streaming superlinear indexes might pave the way to improved file systems; (3) determine whether cache-oblivious algorithms technology can enhance streaming superlinear indexes; and (4) program complex data structures for transactions and recovery.

If successful, this research will show how to build filesystems that achieve dramatically better performance than today's B-tree-based filesystems, how to maintain rich geometrical data and multidimensional nongeographical databases in real time, and how to maintain full-text searchable databases in real time. For example, some of today's file systems try to maintain an full-text index to find strings in files quickly, but these systems often fall behind at high data write rates. A streaming superlinear index would allow such a file system to keep up, and would improve the usability of both high-end storage systems and relatively small consumer storage systems that are nonetheless too large to index with today's indexes.

The researchers are developing course materials on streaming indexing technology which will be made freely available under the MIT OpenCourseWare initiative (http://ocw.mit.edu).

- Cross-Layer Exploration of Non-Volatile Solid-State Memories to Achieve Effective I/O Stack for High-Performance Computing Systems
  - Li, Tao/ He, Xubin/ Zhang, Tong,; University of Florida/Tennessee Technological University/ Rensselaer Polytechnic Institute
  - HECIWG Topic: Next Generation I/O Architecture
  - Keywords:
    - XXXX

The objective of this research is to develop techniques that utilize solid-state memory technologies from device, circuit, architecture, and system perspectives across I/O hierarchy in order to exploit their true potential for improving I/O stack performance in high-performance computing systems.

I/O friendly memory system architectures will be developed to enable hybrid processor-memory 3D integrations with largely reduced off-chip I/O traffic. Adaptive cache management and hotspot prediction methods will be developed to address the low random write performance of solid-state drives, and data processing techniques will be developed to enable run-time configurable trade-offs among solid-state drive performance characteristics. A comprehensive full-system simulation infrastructure will be developed to evaluate and demonstrate the research under diverse high-performance computing workloads.

The research will facilitate the high-performance computing systems to most effectively utilize existing/emerging memory and processing technologies to tackle the grand I/O stack design challenge. It can greatly contribute to enabling high-performance computing systems to stay on track of their historic scaling, and hence benefit numerous real-life applications such as biology, chemistry, earth science, health care, etc. This project will also contribute to the society through engaging under-represented groups, research infrastructure dissemination for education and training, and outreach to high school students.

- Visual Characterization of I/O System Behavior for High-End Computing
  - Ma, Kwan-Liu/ Kamil A Iskr; University of California-Davis/ University of Chicago
  - HECIWG Topic: Measurement and Understanding
  - Keywords:
    - XXXX

Modern supercomputers are complex, hierarchical systems consisting of huge number of cores, systems for disk storage, and nodes for I/O forwarding. These numbers will continue to grow and the need for tools to understand the behavior of the system software becomes paramount: without these tools it will be impossible to effectively tune system software, and high degrees of efficiency will be unattainable by applications. This project addresses the challenge of understanding behavior of complex system software on very

large-scale compute platforms like the current petascale computers. In particular, this project will develop software infrastructure to provide end-to-end analysis and visualization of I/O system software. Specifically, the objectives of this project are to develop, improve, and deploy
(1) end-to-end, scalable tracing integrated into the I/O system (MPI-IO, I/O forwarding, and file system); (2) information visualization tools for inspecting traces and extracting knowledge; (3) testing components that drive this system to generate example patterns, including an "I/O system failure" component to generate anomalies; and (4) tutorials and tools for helping other system software developers incorporate this analysis and visualization system into their production software. It is clear that the software and techniques developed in this project will be directly applicable to and useful in other system software libraries, such as communication libraries, which perform complex interactions on large systems.

- Automatic Extraction of Parallel I/O Benchmarks from HEC Applications
  - Ma, Xiaosong/ Shen, Kai/ Winslett, Marianne; North Carolina State University; University of Rochester; University of Illinois at Urbana-Champaign
  - HECIWG Topic: Measurement and Understanding
  - Keywords:
    - Automatic Benchmark extraction

I/O performance is often an issue for high-end computing (HEC) codes, due to their increasingly data-intensive nature and the ever-growing CPU-I/O performance gap. Portable parallel I/O benchmarks can help

(1) application owners to improve their codes' performance,
(2) HEC storage systems architects to improve their designs, and
(3) future and current owners of HEC platforms to reduce hardware cost and improve application performance through better system provisioning and configuration.

To keep up with the growing scale and complexity of HEC applications, this project develops automated generation of parallel I/O benchmarks, analogous to the SPEC and NAS benchmarks for computation. Our approach will be embedded in BenchMaker, a prototype tool that takes a real-world, large-scale parallel application and automatically distills it into a compact, human-intelligible, I/O-intensive, and parameterized benchmark. Such a benchmark accurately reflects the original application's I/O characteristics and I/O performance, yet with shorter execution time, reduced need for libraries, better portability, and easy scalability.

This research will produce benchmarks and tools that benefit the computational science community at large. Our benchmark prototypes will be used for parallel computing course projects and student research contests.

- Secure Provenance in High-End Computing Systems

  - Patrick McDaniel/ Radu Sion/ Marianne Winslett; Pennsylvania State Univ University Park/ SUNY at Stony Brook/ University of Illinois at Urbana-Champaign
  - HECIWG Topic: Security
  - Keywords:
    - Data Provenance
    - Security Protocols

Data provenance documents the inputs, entities, systems, and processes that influence data of interest---in effect providing a historical record of the data and its origins. The generated evidence supports essential forensic activities such as data-dependency analysis, error/compromise detection and recovery, and auditing and compliance analysis.

This collaborative project is focused on theory and systems supporting practical end-to-end provenance in high-end computing systems. Here, systems are investigated where provenance authorities accept host-level provenance data from validated provenance monitors, to assemble a trustworthy provenance record. Provenance monitors externally observe systems or applications and securely record the evolution of data they manipulate. The provenance record is shared across the distributed environment.

In support of this vision, tools and systems are explored that identify policy (what provenance data to record), trusted authorities (which entities may assert provenance information), and infrastructure (where to record provenance data). Moreover, the provenance has the potential to hurt system performance: collecting too much provenance information or doing so in an inefficient or invasive way can introduce unacceptable overheads. In response, the project is further focused on ways to understand and reduce the costs of provenance collection.

- Scalable Data Management Using Metadata and Provenance
  - Miller, Ethan/ Seltzer, Margo I; University of California, Santa Cruz/ Harvard University
  - HECIWG Topic: Metadata
  - Keywords:
    - xxxx

This project is developing new techniques for identifying and managing files, replacing tree-structured file names with content- and metadata- based search access. By leveraging existing work in search and recognizing the explosion in the volume of data stored, this project enables users to find and access their data in natural and intuitive ways, based on the files' contents, tags the user has assigned, system metadata, and provenance (information about the file's origins). This research targets high-end computing (HEC) users, who manage billions of files generated by measurement devices, experimentation, or scientific workflows. The techniques and system developed are also applicable to general-purpose computing.

Realizing this goal requires advances in several areas. First, the project is designing and developing fast, scalable mechanisms to gather, maintain and index the large volume of metadata and provenance that HEC applications and users generate. This project is also exploring search algorithms that operate on graph structures, enabling users to find files "near" their current workspace. To enable users to access this functionality, the project is developing a new "language" that facilitates the kind of searches that users need.

- Streamlining High-End Computing with Software Persistent
  - Rangaswami, Raju/ Jason X Liu/ Ming Zhao; Florida International University/ Florida International University/ Florida International University
  - HECIWG Topic: Next Generation I/O Architecture
  - Keywords:
    - Software Persistent Memory (SoftPM)

Current high-end computing (HEC) applications explicitly manage persistent data, including both application state and application output. This practice not only increases development time and cost, but also requires an application developer to be intimately aware of the underlying platform-dependent storage mechanisms to achieve good application I/O performance. Such vertical development also makes the application software less portable.

The Software Persistent Memory (SoftPM) project builds a lightweight abstraction and practical infrastructure for streamlining data management in next generation HEC applications. SoftPM eliminates the duality of data management in HEC applications by allowing applications to allocate persistent memory in much the same way volatile memory is allocated and easily restore, browse, and interact with past versions of persistent memory state. This simplifies the implementation of three broad capabilities required in HEC applications -- recoverability (e.g., checkpoint-restart), record-replay (e.g., data-visualization), and execution branching (e.g., simulation model-space exploration).

The SoftPM project is organized in three modules. The first module builds an evolvable SoftPM API and addresses memory management issues. The second module addresses high-performance I/O and the atomicity of persistence points for local storage and parallel file systems. The final module builds several HEC application case-studies to illustrate the different capabilities supported by SoftPM in HEC environments.

- Interleaving Workloads with Performance Guarantees on Storage Cluster
  - Riska, Alma, College of William and Mary
  - HECIWG Topic: Quality of Service
  - Keywords:
    - xxxx

This research focuses on the design and implementation of a lightweight, yet, versatile middleware framework that provides effective and scalable solutions to the problem of interleaving storage workloads with a wide spectrum of demands. The framework uses simple and non-intrusive collection of workload statistics such as workload histograms and measures of temporal dependence to provide accurate forecasting of system workload characteristics and their impact on system metrics.

The framework maps accurately and swiftly complex processes that exist and interact in storage clusters into robust allocation decisions. Central to the framework is its ability to estimate beforehand the effect of resource allocation policies on system metrics, which enables navigating through multiple possible allocations of system resources and selecting the on that best meets system targets.

This research has the potential to revolutionize autonomic resource management in storage systems and provide methodologies to meet conflicting targets such as discovering trade-offs and dependencies between performance and other metrics including cost, energy consumption, reliability, and availability.

This project will enable enhancement of graduate courses on parallel and distributed systems with aspects of emerging paradigms such as data intensive, cloud, and green computing, as well as will advance the education of the multiple students directly involved.

- Programming Models and Storage System for High Performance Computation with Many-Core Processors
    - Sarkar, Vivek/ Dennis, Jack/ Gao, Guang; William Marsh Rice University/ MIT/ University of Delaware
    - HECIWG Topic: Next Generation I/O Architecture
    - Keywords:
        - xxxx

A major challenge for future High End Computing (HEC) systems built using many-core chips is the storage system since the available memory and bandwidth per processor core is starting to decline at an alarming rate, with the rapid increase in the number of cores per chip. Data-intensive applications that require large data sets and/or high input-output bandwidth will be especially vulnerable to these trends. Historically, the storage architecture of an HEC system has been constrained to a large degree by the filesystem interfaces in the underlying Operating System (OS). The specific focus of this research is on exploring a new storage model based on write-once tree structures. This research will explore three programming models for users of the storage system, all of which can inter-operate through shared persistent data: 1) a declarative programming model in which any data structure can be directly made persistent in the storage system, with no programmer intervention, 2) a strongly-typed imperative programming model in which a type system extension will be used to enforce a separation between data structures that can be directly made persistent and those that cannot, and 3) a weakly-typed runtime interface that enables low-level C programs to access the storage system.

- A Dynamic Application-specific I/O Architecture for High End Computing
    - Sun, Xian-He; Illinois Institute of Technology
    - HECIWG Topic: Next Generation I/O Architecture
    - Keywords:
        - xxxx

Disk I/O on high-end computing machines continues to be a significant performance bottleneck. Parallel file systems have been developed to improve parallel I/O performance. However, most of these methods are application dependent and their performance varies largely from application to application. The performance of parallel I/O can be improved with better understanding of I/O access characteristics at both client and file-server side. There is a great need for research into next-generation intelligent and application-specific I/O architectures to meet the demand of highend computing.

We propose a dynamic application-specific I/O architecture that tailors various parallel I/O optimizations based on I/O characteristics of applications. This architecture is dynamic in the sense that its underlying optimization strategies are able to adapt to the variations in different applications for best performance. The proposed research is twofold: 1) understanding I/O behavior, 2) developing application-specific optimizations for data layout, prefetching, and caching to form an integrated application-specific I/O architecture. Several technical hurdles have been identified, which include I/O access signature, compiler analysis, global-aware coordinated caching, collective prefetching, data layout optimization and distribution strategies. Solutions are proposed and detailed plans are provided to test these newly proposed solutions and techniques under the PVFS2 parallel file system.

- Balanced Scalable Architectures for Data-Intensive Supercomputing
    - Szalay, Alexander/ Huang, H. Howie; Johns Hopkins University/ George Washington University
    - HECIWG Topic: Next Generation I/O Architecture
    - Keywords: XXXX

The nature of scientific computing is changing – it is becoming increasingly data-centric. We use Amdahl's Laws to quantify what is (i) a data-intensive computational problem, (ii) and what is a data-intensive computational architecture. Based on these objective metrics we propose several different architectural approaches, including some next-generation, low-power processors and storage devices, e.g., Solid State Disks (SSDs), and consider how these architectures might offer substantial benefits over the existing ones. As data volumes grow, transferring data to where our computational resources are is becoming increasingly difficult. As a result we need to bring our computing as close to the data storage as possible. In this research, we plan to explore approaches where the first steps of the scientific data processing are performed on the backplane of the database servers – the closest we can get to low level scientific data. There are several challenges in using databases for large scale scientific computations; not all data structure map equally well to relational database tables. We will also explore how we can use trees and arrays, representing very large scientific data sets, in both relational databases and on a MapReduce/Hadoop like environment. We are in a fortunate situation that the necessary hardware components of our research have been provided by funds from the Gordon and Betty Moore Foundation and Microsoft Research. We seek a moderate support for graduate students, beyond fractional salaries of the senior personnel involved.

Intellectual merits: Our premise is that in the near future the only feasible scalable data-intensive environment will consist of a massively scaled-down and -out system, where data partitioning, fault tolerance, and massive parallelism will play a much larger role than today. While SSDs can provide an excellent performance for both sequential and random I/Os, we will still need traditional hard disks for the bulk volume. How this additional tier in our storage system can be used in the most effective way, both with databases and without, is yet to be explored. We propose to use real workloads from our existing experiments to evaluate these compromises, both on our 1.1PB GrayWulf system, and on small-scale hybrid systems built out of low-power motherboards, SSDs, and hard disks. The proposed research will focus on the challenge of designing and developing massively distributed query environments, that is, providing data storage and access support for scientific applications, to name a few, SkyQuery, Turbulence, and Viz. This research will rethink and redesign our existing database clusters and data management systems in the presence of solid-state drives and low-power processors, as the emergence of these new devices has undoubtedly introduced exciting opportunities for improving not only performance but also energy efficiency.

Broader impacts: We feel that the research here is *transformative*, and has a *very broad impact* to much of the next generation of scientific computing architectures, in all scientific disciplines. It is clear that current systems (BeoWulf) will not scale well beyond the next two years, mostly due to power requirements, and the difference in the scaling laws of multi-core CPUs and hard-disk based IO subsystems. Our approach, using Amdahl's Laws as an objective criterion, enable us to design systems from the first principles and match architectures to applications. The development of a balanced data-intensive supercomputing system that can effectively utilize multi-core processors and SSDs will lead not only to a reduced cost of ownership, but also performance improvement for many scientific applications. The education and outreach plan in this proposal consists of three tasks: developing educational materials, mentoring underrepresented students, and developing collaborations in the industry. The PI plans to interleave them to maximize the broader impact on multiple fronts.

- Performance- and Energy-Aware HEC Storage Stacks
    - Zadok, Erez/Geoff Kuenning; SUNY at Stony Brook/ Harvey Mudd College
    - HECIWG Topic: Management and RAS/Measurement and Understanding
    - Keywords:
        - Adaptive Cluster Reconfiguration
        - Energy tracing/profiling

High-End Computing (HEC) systems are designed for performance, not energy efficiency. In recent years, HEC users have found that as energy costs increase, network and disks have become significant bottlenecks; worse, scientific workloads vary wildly, exercising different parts of HEC clusters, making it impossible to understand where the bottlenecks are and where energy is being wasted.

This project explores the impact of storage-stack configurations on power and performance, using actual cluster configurations and realistic scientific workloads. The research follows three thrusts: tracing and analysis, adaptive cluster reconfiguration, and new storage software stacks.

(1) Traces are collected and analyzed which combine both performance and energy data

on a large set of scientific workloads: I/O-, network-, memory-, and CPU-intensive. Three popular scientific cluster configurations are investigated, varying many configuration parameters. (2) Tools are being developed to dynamically adapt a cluster's configurations to a given workload, so as to optimize power and performance prior to running long-term scientific experiments or simulations. (3) New operating systems software is developed specifically to optimize power and performance for scientific workloads: a new lightweight file system and disk I/O scheduler.

The long-term results of this project help society save energy in computing without unduly hurting performance.

- QoS-driven Storage Management for High-end Computing Systems
  - Zhao, Ming/Figueiredo, Renato J; Florida International University/ University of Florida
  - HECIWG Topic: Quality of Service
  - Keywords:
    - XXXX

In today’s high-end computing (HEC) systems, the parallel file system (PFS) is at the core of the storage infrastructure. PFS deployments are shared by many users and applications, but currently there are no provisions for differentiation of service - data access is provided in a best-effort manner. As systems scale, this limitation can prevent applications from efficiently utilizing the HEC resources while achieving their desired performance and it presents a hurdle to support a large number of data-intensive applications concurrently. This NSF HECURA project tackles the challenges in quality of service (QoS) driven HEC storage management, aiming to support I/O bandwidth guarantees in PFSs by addressing the following four research aspects: 1. Per-application I/O bandwidth allocation based on PFS virtualization, where each application gets its specific I/O bandwidth share through its dynamically created virtual PFS. 2. PFS management services that control the lifecycle and configuration of per-application virtual PFSs as well as support application I/O monitoring and storage resource reservation. 3. Efficient I/O bandwidth allocation through autonomic, fine-grained resource scheduling across applications that incorporate coordinated scheduling and optimizations based on profiling and prediction. 4. Scalable application checkpointing based on performance isolation and optimization on virtual PFSs customized for checkpointing I/Os.

- A New Semantic-Aware Metadata Organization for Improved File-System Performance and Functionality in High-End Computing
  - Zhu, Yifeng/ Hong Jiang; University of Maine/University of Nebraska-Lincoln
  - HECIWG Topic: Metadata
  - Keywords:
    - XXXX

Existing data storage systems based on the hierarchical directory-tree organization do not meet the scalability and functionality requirements for exponentially growing datasets and increasingly complex metadata queries in large-scale Exabyte-level file systems with billions of files. This project focuses on a new decentralized semantic-aware metadata organization that exploits semantics of file metadata to improve system scalability, reduce query latency for complex data queries, and enhance file system functionality.

The research has four major components: 1) exploit metadata semantic-correlation to organize metadata in a scalable way, 2) exploit the semantic and scalable nature of the new metadata organization to significantly speed up complex queries and improve file system functionality, 3) fully leverage the semantic-awareness of the new metadata organization to optimize storage system designs, such as caching, prefetching, and data de-duplication, and 4) implement the new metadata organization, complex query functions, and system design optimizations in large-scale storage systems. This project has broader impact to data-intensive scientific and engineering applications, graduate and undergraduate education, and K-12 education through its contributions to storage system research and its integration with an existing NSF-REU site award and an NSF-ITEST award.

- An Application Driven I/O Optimization Approach for PetaScale Systems and Scientific Discoveries.
    - Klasky, Scott; Oak Ridge National Lab and Beck, Micah;
    - HECIWG Topic: Metadata
    - Keywords:
        - I/O Middleware

Application workflows typically involve large-scale simulations, applications, and subsequent analysis, verification and validation. One of the most important requirements shared by various applications running on petascale systems is fast, portable, scalable I/O which is componentized, metadata rich and easy to use. The Adaptable I/O System (ADIOS) serves as a high-level I/O interface for an application to select which I/O libraries and formats to use without changing the application program. Codes such as GTC generate hundreds of TBs of data on hundreds of thousands of cores in a twenty four hour period. One must therefore optimize the I/O both for fast output in the generation phase and for fast input in the analysis phase. Both the writing and reading efficiency of I/O are critical for knowledge discovery. Development of a high level software infrastructure to allow optimization of I/O for entire workflows (including High-Performance I/O when reading data with different patterns) would greatly improve end-to-end performance in the knowledge discovery cycle.

This project plans to develop efficient I/O methods which will enable application scientists to optimize data for writing, and which will be able to re-organize the data to obtain optimal performance for common reading patterns used by scientists. This project directly impacts the I/O performance of many petascale applications, including the GTC, GTS, XGC-1, Chimera, and S3D codes, and work directly with these teams to optimize the I/O in all stages of their scientific workflow.

Scientific Discovery through Advanced Computing (SciDAC2) I/O Projects

PetaScale Data Storage Institute

Garth Gibson (Lead PI) - Carnegie Mellon University
Evan Felix - Pacific Northwest National Laboratory
Gary Grider – Los Alamos National Lab
Peter Honeyman - University of Michigan at Ann Arbor
William Kramer - Lawrence Berkeley National Laboratory/NERSC
Darrell Long - University of California at Santa Cruz
Philip Roth - Oak Ridge National Laboratory
Lee Ward – Sandia National Lab

- Leveraging experience in applications and diverse file and storage systems expertise of its members, the institute will enable a group of researchers to collaborate extensively on developing requirements, standards, algorithms, and development and performance tools.

  Petascale computing infrastructures for scientific discovery make petascale demands on information storage capacity, performance, concurrency, reliability, availability, and manageability. The last decade has shown that parallel file systems can barely keep pace with high performance computing along these dimensions; this poses a critical challenge when petascale requirements are considered. This proposal describes a Petascale Data Storage Institute that focuses on the data storage problems found in petascale scientific computing environments, with special attention to community issues such as interoperability, community buy-in, and shared tools. Leveraging experience in applications and diverse file and storage systems expertise of its members, the institute allows a group of researchers to collaborate extensively on developing requirements, standards, algorithms, and development and performance tools. Mechanisms for petascale storage and results are made available to the petascale computing community. The institute holds periodic workshops and develops educational materials on petascale data storage for science.

  The Petascale Data Storage Institute is a collaboration between researchers at Carnegie Mellon University, National Energy Research Scientific Computing Center, Pacific Northwest National Laboratory, Oak Ridge National Laboratory, Sandia National Laboratory, Los Alamos National Laboratory, University of Michigan, and the University of California at Santa Cruz.

  The Institute’s work will be organized into six projects:

- Petascale Data Storage Outreach: (Type: Dissemination) Development and deployment of training materials, both tutorials for scientists and course materials for graduate students; support and advise other SciDAC projects and institutes; and development of frequent workshops drawing together experts in the field and petascale science users.

- Protocol/API Extensions for Petascale Science Requirements: (Type: Dissemination) Drive deployment of best practices for petascale data storage systems through development and standardization of application programmer interfaces and protocols, with specific emphasis on Linux APIs. Validate and demonstrate these APIs in large scale scientific computing systems.

- Petascale Storage Application Performance Characterization: (Type: Data Collection) Capture, characterize, model and distribute workload, access trace, benchmark and usage data on terascale and projected petascale scientific applications, and develop and distribute related tools.

- Petascale Storage System Dependability Characterization: (Type: Data Collection) Capture, characterize, model and distribute failure, error log and usage data on terascale and projected petascale scientific systems, and develop and distribute related tools.

- Exploration of Novel Mechanisms for Emerging Petascale Science Requirements: (Type: Exploration) In anticipation of petascale challenges for data storage, explore novel mechanisms such as global/ WAN high performance file systems based on NFS; security aspects for federated systems, collective operations, and ever higher performance systems; predictable sharing of high performance storage by heavy storage load applications; new namespace/search and attribute definition mechanisms for ever large namespaces; and integration and specialization of storage systems for server virtualization systems.

- Exploration of Automation for Petascale Storage System Administration: (Type: Exploration) In anticipation of petascale challenges for data storage, explore and develop more powerful instrumentation, visualization and diagnosis methodologies; data layout planning and access scheduling algorithms; and automation for tuning and healing configurations.

Scientific Data Management Center for Enabling Technologies

Arie Shoshani (PI), Doron Rotem, Lawrence Berkeley National Laboratory
Rob Ross, Bill Gropp, Rajeev Thakur, Argonne National Laboratory
Terence Critchlow, Chandrika Kamath, Lawrence Livermore National Laboratory
Nagiza Samatova, Jeff Vetter, Oak Ridge National Laboratory
Jarek Nieplocha, Pacific Northwest National Laboratory

Alok Choudhary, Wei-Keng Liao, Northwestern University
Mladen Vouk, North Carolina State University
Steve Parker, University of Utah
Bertram Ludaescher, University of California at Davis
Ilkay Altinas, San Diego Supercomputer Center

Managing scientific data has been identified as one of the most important emerging needs by the scientific community because of the sheer volume and increasing complexity of data being collected. Effectively generating, managing, and analyzing this information requires a comprehensive, end-to-end approach to data management that encompasses all of the stages from the initial data acquisition to the final analysis of the data.

Based on the community input, we have identified three significant requirements. First, more efficient interactions with disks and the resulting files are needed. In particular, parallel file system improvements are needed to write and read large volumes of data without slowing a simulation, analysis, or visualization engine. Second, scientists require improved access to their data, in particular the ability to effectively perform complex data analysis and searches over large data sets. Specialized feature discovery and statistical analysis techniques are needed before the data can be understood or visualized. Finally, generating the data, collecting and storing the results, data post-processing, and analysis of results is a tedious, fragmented process. Tools for automation of these workflows this process in a robust, tractable, and recoverable fashion are required to enhance scientific exploration.

We have organized our activities in three layers abstracting the end-to-end data flow described above: the Storage Efficient Access (SEA), Data Mining and Analysis (DMA), and Scientific Process Automation (SPA) layers. The SEA layer is immediately on top of hardware, operating systems, file systems, and mass storage systems, and provides parallel data access technology. On top of the SEA layer exists the DMA layer, consisting of indexing, feature selection, and parallel statistical analysis. The SPA layer, which is on top of the DMA layer, provides the ability to compose scientific workflows from the components in the DMA layer as well as application specific modules. Together these layers provide an integrated system for data management in computational science.

X-Stack Software Research 2010 I/O Projects

- The Damsel Project: A Data Model Storage Library for Exascale Science
    - Choudhary, Alok/Latham, Rob/Smatova, Nagiza/Koziol, Quincey; Northwestern University/Argonne National Lab/North Carolina State University/The HDF Group
    - HECIWG Topic: Next Generation I/O Architectures
    - Keywords:
        - I/O Middleware

Computational science applications are steadily increasing the complexity of grids, solution methods on those grids, and data that link the two together on modern petascale computers. Several common motifs can be described, having distinct grid types and computation/communication patterns; examples include structured AMR, unstructured finite element/volume, and spectral elements. From a storage and I/O perspective, these applications exhibit distinct data organization and access patterns during simulation, analysis, and visualization. However, these codes continue to interact with I/O data libraries much the same as they have since the 1990s, in terms of relatively low-level data structures like vertex positions, connectivity arrays, and multi-dimensional solution data arrays. Although these high-level I/O libraries have had a beneficial impact on parallel applications in terms of read/write performance, the impedance mismatch with application data models is growing as applications go to more complex data models and interactions. This mismatch is making it more difficult to achieve I/O performance close to system peak capabilities, while also not supporting the full range capabilities in today's computational science data models.

A different approach is needed, that leverages performance improvements and best practices learned from previous approaches, while raising the level of interaction with application data models and access patterns. As the International Exascale Software Project (IESP) report observes, "[...] The purpose of I/O by an application can be a very important source of information that can help scalable I/O performance when hundreds of thousands (to millions) of cores simultaneously access the I/O system [...]" In other words, the high-level view of the *data model* is overlooked rather than exploited. Also, the *data layout* used in these codes and how that layout interacts with I/O software used to save the data to or read the data from storage systems are highly relevant. Increasing the complexity with which applications can interact with I/O libraries will reduce the impedance mismatch between the two, while also streamlining the I/O process by reducing unnecessary data copying between applications and I/O libraries. This process will be simplified further by developing customized interfaces and formats, or "verticals", for the data model motifs used in today's petascale applications. This model represents the underlying approach of our project.

The goal for Damsel is to enable Exascale computational science applications to interact conveniently and efficiently with storage through abstractions that match their data models. We are pursuing four major activities: (1) constructing a unified, high-level data model that maps naturally onto a set of data model motifs used in a representative set of high-performing computational science applications; (2) developing a data model storage library, called *Damsel*, that supports the unified data model, provides efficient storage data layouts, incorporates optimizations to enable exascale operation, and is tolerant to failures; (3) assessing the performance of this approach through the construction of new I/O benchmarks or the use of existing I/O benchmarks for each of the data model motifs; and (4) productizing Damsel and working with computational scientists to encourage adoption of this library by the scientific community.

Scientific Data Management and Analysis at Extreme Scale 2010 I/O Projects

- Damasc: Adding Data Management Services to Parallel File Systems
  - Brandt, Scott and Gokhale, Maya; University of California, Santa Cruz and Lawrence Livermore National Lab
  - HECIWG Topic: Next Generation I/O Architectures
  - Keywords:
    - XXXX

Scientific applications need high-level data management services. Scientists use file systems to store and manipulate large volumes of data, yet unstructured file systems provide poor support for highly structured scientific data and scientific codes and data often outlast the systems upon which they are processed and stored. To facilitate software development and minimize platform-specific impact on scientific codes, the high-end computing community has adopted multiple layers of abstractions and specialized file formats such as parallel NetCDF, NetCDF-4, HDF5, and MPI-IO. These APIs are limited in terms of the performance and functionality they can provide. Associated libraries providing access to the highly structured contents of scientific data files stored in the (unstructured) file systems can only optimize to the extent that the file system interfaces permit. As intermediate layers such as MPI-IO evolve, high-level layers such as HDF5 do not always keep up.

Motivated by the need to manage and analyze large-scale scientific data effectively, we are developing Damasc, an enhanced file system with rich data management services for scientific computing provided as a native part of the file system. Damasc will make it easy to pose queries and updates over files stored in their native byte-stream format, without losing the inherent performance benefits of file system data storage by allowing scientists to write *declarative* queries and updates over *views* of underlying files. Views provide a flexible mechanism on how the underlying files should be interpreted and may remain virtual, with little or no overhead. To achieve this goal, we are adding a configurable thin layer to the file system that exposes the contents of files in a logical data model through which views can be defined and used for queries and updates. Queries and updates so posed can be optimized to efficient execution plans evaluated inside the file system. This enables a declarative querying interface that facilitates data analysis and the development of applications over efficiently stored files. The logic of data management is offloaded to the distributed file system and the application can focus on manipulating useful information extracted via appropriate queries and updates. We are also developing two important services on top of our enhanced file system: *self-organizable storage* and *provenance capture*.

The Damasc design includes six components: 1) A thin layer of software that can process files of different formats to extract structured views in our logical data model and translate accesses to the underlying native byte-stream file system format; 2) A declarative query and update language for accessing files via the structured interface; 3) An optimization module that does cost-based rewriting to efficiently evaluate declarative queries; 4) A self-organizing storage service that automatically indexes files based on query patterns to improve performance; 5) A provenance capture service that captures critical provenance information for subsequent understanding and analysis; and 6) An integration of the extended data management services into a petabyte-scale parallel file system without sacrificing scalability and performance. We will prototype these components inside the Ceph file system. Damasc has four key benefits for the development of data-intensive scientific code: 1) applications can employ high-level services, such as declarative queries, views, and provenance tracking, that are currently available only within a database system; 2) the use of the new services becomes easier, as they are provided within a familiar file-based ecosystem; 3) common optimizations, such as indexing and caching, are readily supported across several file formats, thus avoiding effort duplication; and, 4) the potential for significant performance benefits as data processing is integrated more tightly with data storage.

- An Information-Theoretic Framework for Enabling Extreme-Scale Science Discovery
    - Shen, Han-Wei/Ross, Rob/Chiang, Yi-Jen; Ohio State University/Argonne National Lab/Polytechnic Institute of New York University;
    - HECIWG Topic: Measurement and Understanding
    - Keywords:
        - Visualization
        - Information Entropy
        - Data Saliency

As scientists eagerly anticipate the benefits of extremescale computing, roadblocks to science discovery at scale threaten to impede their progress. The disparity between computing and storing information, and the gap between stored information and the understanding derived from it, are two of the main barriers to success. This project addresses two difficulties faced by computational scientists. The first is deciding what data are the most essential for analysis, given that only a small fraction can be retained. The second is transforming these data into visual representations that rapidly convey the most insight to the viewer. We will quantify the amount of information in data using information-theoretic approaches. Computing the information entropy of data allows decisions to be made as to how data should be stored and subsequently analyzed. Data saliency will further be used to inform and steer visualization algorithms automatically, including temporal analyses, and it can enable new types of analyses to be performed. We will construct a data analysis and visualization framework based on information theory that allows us to evaluate the information content of simulation output, and we will test our approaches in applications that represent the next generation of extreme-scale science. We will work together with scientists to evaluate the results of our information-theoretic algorithms. With these tools, scientists will be able to preserve important and discard irrelevant data, enabling them to see results sooner. Informed visualization algorithms will generate more meaningful displays. This will result in more knowledge, faster, and will impact decisions critical to the mission of the Department of Energy.

- Dynamic Non-Hierarchical File Systems for Exascale Storage
    - Long, Darrell; University of California, Santa Cruz
    - HECIWG Topic: Next Generation I/O Architectures
    - Keywords:
        - Provenance
        - Dynamic Non-Hierarchical File System
        - Scalable Indexing
        - Metadata Clustering

Modern high-end computing (HEC) systems must manage petabytes of data stored in billions of files, yet current techniques for naming and managing files were developed 40 years ago for collections of thousands of files. HEC users are therefore forced to adapt their usage to fit an outdated file system model and interface, unsuitable for exascale systems. Attempts to enrich the interface, such as augmentation or replacement with databases, or the layering of additional interfaces and semantic extensions atop existing file systems result in performance-limited systems that do not adequately scale.

Parallels exist between HEC systems and the web, where locating and browsing data sets has rapidly become dominated by search. The strengths and weaknesses of the web provide several useful lessons from which we have learned: 1) Although the web implements a hierarchical namespace, search has become the dominant navigation tool in the face of the massive volume of data that is accessible; 2) While finding *some* information is easy, finding the *right* or *good* information is not; 3) The easier it is for people to contribute information to a repository, the more critical it becomes to determine the veracity of that data; 4) The links that relate documents provide valuable insight into the importance of documents. From these observations we can see that simply modifying existing high performance filesystems to support search, and the requisite storage of additional semantic metadata, would be woefully inadequate.

We propose to develop a radically different filesystem structure that addresses these problems directly, and which will leverage provenance (a record of the data and processes that contributed to its creation), file content, and rich semantic metadata to provide a scalable and searchable file namespace. Such a namespace would allow the tracking of data as it moves through the scientific workflow. This allows scientists to better find and utilize the data *they* need, using both content and data history to identify and manage stored information. We take advantage of the familiar search-based metaphor to provide an initial easy to-use interface that enables users to find the files they need and evaluate the authenticity and quality of those files. Realizing this vision requires research success in dynamic, nonhierarchical file systems design and implementation, large-scale metadata management, efficient scalable indexing, and automatic provenance capture.

- ExaHDF5: An I/O Platform for Exascale Data Models, Analysis and Performance
    - Prabhat/Koziol/Palmer; Lawrence Berkeley National Lab/The HDF Group/Pacific Northwest National Lab;
    - HECIWG Topic: XXXX

    - Keywords:
        - XXXX

It is reasonably well accepted that one of the primary bottlenecks in modern computational and experimental sciences is coping with the sheer volume and complexity of data. Storing, reading, finding, analyzing, and sharing data are tasks common across virtually all areas of science, yet advances in data management infrastructure, particularly I/O, have not kept pace with our ability to collect and produce scientific data. This "impedance mismatch" between our ability to produce and store/analyze data continues to grow and could, if not addressed, lead to situations where science experiments are simply not conducted or scientific data not analyzed for want of the ability to overcome data-related challenges.

Our project consists of three thrust areas that address the challenges of data size and complexity on current and future computational platforms:

We are extending the scalability of I/O middleware to make effective use of current and future computational platforms.
We are incorporating advanced index/query technology to accelerate operations common to scientific data analysis.
We are building upon our existing work on data model APIs that simplify simulation and analysis code development by encapsulating the complexity of parallel I/O.

We are conducting these activities in close collaboration with specific DOE science code teams to ensure the new capabilities are responsive to scientists' needs and are usable in production environments. Our approach includes a clear path for maintainability and production release.

- Runtime System for I/O Staging in Support of In-Situ Processing of Extreme Scale Data
    - Klasky, Scott/Shoshani, Arie/Schwan, Karsten; Oak Ridge National Lab/Lawrence Berkeley National Lab/Georgia Tech
    - HECIWG Topic: XXXX
    - Keywords:
        - ADIOS
        - In-situ pipeline

As we approach the extreme scale in computing, we must realize new strategies to deal with the daunting challenge of managing and exploiting the massive quantities of complex data produced by scientific simulations. The challenge is exacerbated by the fact that I/O and memory systems have not seen increases in performance at the same rate as those observed for computational elements. This not only leads to unfavorable tradeoffs concerning machine power consumption for I/O and memory vs. computation, but it also means that the time scientists will spend on analyzing and visualizing the results produced by their simulations will greatly slow down the knowledge discovery process.

Our research will create and evaluate an I/O infrastructure and tools for extreme-scale applications and facilities so that they can reduce the time to discovery at small cost in machine resources and consequent power consumption.
New tools must be highly scalable, portable, and easy-to-use, so that scientists can gain control of their science and concentrate on producing important scientific discovery in their own domain. Accelerating the rate of insight and scientific productivity, therefore, demands new solutions to managing the avalanche of data expected at extreme scale. Partnering with many application teams and working on petascale machines, our team has developed an approach and delivered proven technology that accelerates I/O and the knowledge discovery process by reducing, analyzing, and indexing the data produced by a simulation while it is still in memory (referred to as "in-situ" processing of data). These technologies include the Adaptable I/O system (ADIOS), FastBit indexing, and Parallel R. For the proposed project, we will leverage those technologies and integrate them to create a runtime system that will allow scientists to create easy-to-use scientific workflows that will run in situ on select nodes of the extreme scale machine. This will not only accelerate simulation and I/O, but it will also provide scientists with immediate and valuable insights into data with online methods that pre-analyze, index, visualize, and reduce the overall amount of data produced by their simulations.

Advanced Architectures and Critical Technologies for Exascale Computing 2010 I/O Projects

- Blackcomb: Hardware-Software Co-design for Non-Volatile Memory in Exascale Systems
  - Vetter, Jeffrey/Schreiber, Robert/Mudge, Trevor/Xie, Yuan; Oak Ridge National Lab/HP Labs/University of Michigan/Penn State University
  - HECIWG Topic: Next Generation I/O Architectures
  - Keywords:
    - Non-Volatile Memory

Memory, not processing, is the crux of the exascale co-design problem. Exascale machines will push the limits of memory capacity, power, and performance. DRAM, the universal memory technology of today, may not scale to meet the needs of exascale applications. Disk storage, critical for checkpointing and for archiving computational inputs and results, may also fail to provide adequate performance, reliability, and power efficiency by the end of this decade. We confront a memory/storage crisis.
The Blackcomb effort seeks to create and understand new memory technologies, develop their roles in exascale systems, adapt applications to them, and assess their relative merits. We focus on emerging nonvolative memory (NVM) technologies, including spin-torque-transfer RAM (STT-RAM), phase-change RAM (PC-RAM), and memristor (resistive RAM, or R-RAM).

- NoLoSS: Investigating the Roles of Node Local Storage in Exascale Systems

  - Iskra, Kamil and Gokhale, Maya; Argonne National Lab/Lawrence Livermore National Lab;
  - HECIWG Topic: Next Generation I/O Architectures
  - Keywords:
    - In-system storage solutions
    - SSD

The international computational science community is on a path to build exaFLOP-capable systems by the year 2018. These exascale systems will enable transformative science discoveries in a number of areas, including climate, combustion, nuclear energy, and national security. A key exascale barrier is the need for scalable storage of persistent state: one that provides the necessary I/O bandwidth and capacity without overwhelming the power, cooling, and cost budgets of an exascale system. Traditional global storage system approaches simply cannot scale to meet these requirements.

With the development of inexpensive, nonvolatile memory technologies such as flash memory and phase change memory, it is feasible to include solid state persistent memory on every node in a future exascale system – enabling in-system storage (also referred to as node local storage). In-system storage augments the memory hierarchy, potentially reducing DRAM requirements and thus the node's power requirements. It streamlines and simplifies checkpointing, increasing system reliability. In-system storage reduces the peak bandwidth requirements of a global exascale storage system, offering a scalable checkpoint/restart solution. However, there remain considerable research challenges to realizing these potential benefits, especially if one wants to hide the complexity introduced by another layer in the storage hierarchy from the user.

The goal of the Node Local Storage Systems (NoLoSS) project is to conduct a detailed assessment of the potential roles and benefits of in-system storage in exascale computational science. We are exploring existing hardware options for NLS and assess the software mechanisms that best exploit them based on a detailed analysis of existing Office of Science applications. We are implementing important examples of those mechanisms and determining how modifications to the existing hardware mechanisms could better support them. We will continue this three-pronged, iterative process throughout the project's lifetime, including anticipating how our successes will alter I/O usage patterns of emerging exascale applications.

- CODES: Enabling Co-Design of Multi-Layer Exascale Storage Architectures
  - Lang, Sam and Carothers, Chris; Argonne National Lab and Rensselaer Polytechnic Institute
  - HECIWG Topic: Measurement and Understanding
  - Keywords:
    - Resilience

The data demands of science and the limited rates of data access place a daunting challenge on the designers of exascale storage architectures. *Co-design* of these systems will be necessary to find the best possible design points for exascale systems. Designers must consider performance, reliability, and power consumption in the context of the I/O patterns and requirements of applications and analysis tools at exascale. Meeting these constraints will require the development of a multi-layer hardware and software architecture incorporating devices that do not yet exist. The most promising approach for codesign of such systems is simulation.

The goal of this project is to enable the exploration and co-design of exascale storage systems by providing a detailed, accurate, and highly parallel simulation toolkit for exascale storage. We will develop models to realistically represent application checkpoint and analysis workloads. These models will be joined together using the Rensselaer Optimistic Simulation System (ROSS), a discrete event simulation framework that allows simulations to be run in parallel, decreasing the simulation run time of massive simulations to hours. Building on our prior work in highly parallel simulation and using our new high-resolution models, our system will capture the complexity, scale, and multi-layer nature of exascale storage hardware and software, and it will execute in a time frame that enables "what if" exploration of design concepts.

With this new toolkit we will investigate design options and trade-offs related to improving the reliability, performance at scale, and power consumption of potential exascale storage architectures. We will work with industry, DOE computing facilities, and the computer and computational science communities to refine our models and to encourage the use of this powerful tool in the design of future extreme-scale storage systems.

- Data Movement Dominates: Adding Data Management Services to Parallel File Systems
  - Rodrigues, Arun/Shalf, John/Bergman, Keren/Jacob, Bruce; Sandia National Lab/Lawrence Berkeley National Lab/Columbia University/University of Maryland
  - HECIWG Topic: Next Generation I/O Architectures /Measurement and Understanding
  - Keywords:
    - 3D memory Stacking
    - Optical chip-to-chip communication

Energy is the fundamental barrier to exascale computing, and is dominated by the cost of moving data, not computation. Further, data movement, not computation, dominates the performance of real applications in HPC environments. This project will addresses the problems of data movement by examining three critical technologies: 3D integration, optical chip-to-chip communication and hardware support for logic operations in the memory system.

Simulation of the proposed systems will be accomplished by merging and improving several existing simulation models: the PhoenixSim optical interconnect simulator; the DRAMsim advanced memory simulator; and the Structural Simulation Toolkit (SST), which will provide processor and I/O models as well as a parallel simulation and power analysis infrastructure. This unified simulation infrastructure will provide accurate physical layer device models as well as more abstract designs for architectural exploration.

APPENDIX B: HEC FSIO 2011 Attendees

Mohammad Abbasi	Oak Ridge National Laboratory
Michael Albrecht	Ultra Systems Research Center
Dave Anderson	Seagate Technology
Remzi Arpaci-Dusseau	University of Wisconsin, Madison
Michael Bender	Stony Brook University and Tokutek
John Bent	Los Alamos National Laboratory
Medha Bhadkamkar	New Mexico Consortium
Scott Brandt	University of California, Santa Cruz
Joe Buck	Los Alamos National Laboratory
Randal Burns	Johns Hopkins University
Phil Carns	Argonne National Laboratory
John Carrier	Cray, Inc.
Robert Chadduck	The National Archives of the U.S.
John Chandy	University of Connecticut
Yong Chen	Texas Tech University
Ron Chiang	George Washington University
Alok Choudhary	Northwestern University
Carolyn Connor	Los Alamos National Laboratory
Jason Cope	Argonne National Laboratory
Chuck Cranor	Carnegie Mellon University
Matthew Curry	Sandia National Laboratories
Nikita Danilov	Xyratex International, Inc.
Frederica Darema	Air Force Office of Scientific Research
Jack Dennis	Massachusetts Institute of Technology
Amol Deshpande	University of Maryland
Phillip Dickens	University of Maine
David Du	University of Minnesota
Sorin Faibish	EMC Corporation
Evan Felix	Pacific Northwest National Laboratory
Jim Finlayson	Department of Defense
Greg Ganger	Carnegie Mellon University
Garth Gibson	Carnegie Mellon University & Panasas
Bradford Glade	EMC Corporation
Gary Grider	Los Alamos National Laboratory
Thuc Hoang	Department of Energy
Stephen Hodson	Oak Ridge National Laboratory
Peter Honeyman	University of Michigan
Howie Huang	George Washington University
Frank Indiviglio	National Oceanic and Atmospheric Administration
Mahmut Kandemir	Pennsylvania State University
Suzanne Kelly	Sandia National Laboratories
Samir Khuller	University of Maryland

Dai Kim	Office of the Assistant Secretary of Defense for Research and Engineering
Ben Kobler	NASA Goddard Space Flight Center
Olga Kornievskaia	University of Michigan
Quincey Koziol	The HDF Group
Geoff Kuenning	Harvey Mudd College
Bradley Kuszmaul	MIT & Tokutek
Mark Laurri	OASD (R&E)
Wei-keng Liao	Northwestern University
Walter Ligon	Clemson University
Jay Lofstead	Sandia National Laboratories
Julio Lopez	Carnegie Mellon University
Xiaosong Ma	North Carolina State University/Oak Ridge National Laboratory
Tina Macaluso	SAIC
Adnan Majeed	State University of New York, Binghamton
Carlos Maltzahn	University of California, Santa Cruz
Adam Manzanares	Los Alamos National Laboratory
Xiaoxuan Meng	University of Delaware
Linden Mercer	Naval Research Laboratory
Ethan Miller	University of California, Santa Cruz
Christopher Mitchell	Los Alamos National Laboratory
Christopher Morrone	Lawrence Livermore National Laboratory
Frank Mueller	North Carolina State University
Jose Munoz	National Science Foundation
Joe Naps	Los Alamos National Laboratory
Thomas Ndousse-Fetter	Department of Energy
Henry Newman	Instrumental, Inc.
Mark Nossokoff	NetApp
James Nunez	Los Alamos National Laboratory
Matthew O'Keefe	University of Minnesota
Aleatha Parker-Wood	University of California, Santa Cruz
Milo Polte	Panasas
Stephen Poole	Oak Ridge National Laboratory
Branislav Radovawovic	NetApp
Raju Rangaswami	Florida International University
Narasimha Reddy	Texas A&M University
Erik Riedel	EMC Corporation
Scott Rife	Library of Congress
Robert Ross	Argonne National Laboratory
Ellen Salmon	NASA
Saba Sehrish	Northwestern University
Yukiko Sekine	Department of Energy
Brad Settlemyer	Oak Ridge National Laboratory
Kai Shen	University of Rochester
Galen Shipman	Oak Ridge National Laboratory

Stan Skelton	NetApp
Ankur Srivastava	University of Maryland
Tom St. John	University of Delaware
Xian-He Sun	Illinois Institute of Technology
Sai Susarla	NetApp
Nisha Talagala	Fusion-io
William Terrell	NetApp
Oercy Tzelnic	EMC Corporation
Ahsen Uppal	George Washington University
Sudharshan Vazhkudai	Oak Ridge National Laboratory
Lee Ward	Sandia National Laboratories
Noah Watkins	Los Alamos National Laboratory
Geoffrey Wehrman	SGI
Brent Welch	Panasas
Kenneth Wilson	Clemson University
Marianne Winslett	University of Illinois at Urbana-Champaign
Edward Wobber	Microsoft Research Silicon Valley
Weikuan Yu	Auburn University
Erez Zadok	Stony Brook University
Xiaodong Zhang	Ohio State University
Ming Zhao	Florida International University

APPENDIX C: Roadmaps

Roadmaps 2010

Metadata

2010 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/ANL – 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	■	■	■	■					
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
	UCSD – NVM Characterization			■	■	■				

This gap area is best represented in Archive. Thus, this gap sub area will be removed from the Metadata Road Map.

2010 Metadata Gap Area

Area	Researcher	Fiscal Year								Rankings
		07	08	09	10	11	12	13	14	
Data Transparency and Access Methods	***None***									⊜ ○ research focused on tadata

This gap area was merged with “Extensibility and name Spaces” in the Metadata gap area. Thus, this gap sub area will be removed from the Metadata Road Map.

● 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

2010 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	■	■	■	▒					🔴 ⊜ ○ A comprehensive tool is nowhere in sight; problem is complex. This gap area includes monitoring.
	Reddy	■	■	■	▒					
	Smirni		▒	■	■	▒				
	Zadok	■	■	■	▒					
	Narashimhan	■	■	▒						
	Riska			▒	■	■	▒			
	He			▒	■	■	▒			
	Zadok (2009 HECURA)			▒	■	■	▒			
	SciDAC - PDSI	■	■	■	■					
	SciDAC - SDM	■	■	■	■	■				
Standards and common practices for HEC I/O benchmarks	Zadok/Miller		🟨	🟨	🟨					⊜ ○ ○ Danger of over simplifying problem and could drive vendors to incorrect solutions.
	High Productivity Computing Systems (HPCS) Benchmarks				■	■				
	Ma/Shen/Winslett			▒	■	■	▒			
Modeling, simulation and test environments.	Clemson - Ligon	■	■	■	▒					🔴 ⊜ ○ Simulators are being developed. PROBE's testbeds for use are retired clusters. No real testbeds being built. This problem will only get worse over time, i.e. as systems get bigger.
	CODES – ANL/RPI					■	■	■		
	PROBE – LANL/CMU					■	■	■		
	Thottethodi	■	■	■	■					
	UCSC - Maltzahn			🟨	🟨					
	DMD – Sandia /LBNL /UMD /Columbia					■				
Applying cutting edge analysis tools to large scale I/O	Reddy	■	■	■	▒					🔴 🔵 ○ Data are becoming available from Labs including I/O traces. Many opportunities to evaluate this research. This includes applying analysis and visualization tools to I/O traces
	Zadok	■	■	■	▒					
	LANL/CMU – Trace replay and Visualizer		🟨	🟨	🟨					
	Ma/Iskra			▒	▒					
	Shen, Ross, Chiang - Ohio State/ANL/PolyTech NY					■	■	■		

🔴 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

2010 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			▨	▨					
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

2010 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	🟨	🟨	🟨	🟨					
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. Traditional block-based solutions ready for commercialization. Alternative interfaces not yet well explored.
	Urgaonkar		▒	■	■	▒				
	Szalay/ Huang			▒	■	■	▒			
	He			▒	■	■	▒			
	Rangaswami				▒	▒				
	Arpaci-Dusseau (2009 HECURA)			▒	▒					
	UCSD (Swanson/Gupta) - NVTM				■	■				
	NoLoSS - ANL/LLNL				■	■	■			
	Blackcomb – ORNL/ HP/ UM/ Penn State				■	■	■			
	PNNL	🟨	🟨	🟨	🟨					
HEC systems with multi-	Choudhary/Kandemir	■	■	■	▒					▤ ▤ ▤
	Dickens	▒	■	■	▒					

2010 Next Generation I/O Architectures Gap Area

Area	Researcher	Fiscal Year								Rankings
		07	08	09	10	11	12	13	14	
million way parallelism doing small I/O operations	Gao	▨	■	■	▨					Good initial research; needs to be moved into testing. More fundamental solutions being pondered including non-volatile solid state storage.
	Sun	■	■	■	▨					
	Zhang/ Jiang	▨	■	■	▨					
	Sun			▨	■	■	▨			
	FASTOS – I/O Forwarding		▨	■	■					
	LANL/CMU – PLFS			■	■					
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

2010 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

2010 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 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

2010 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	■	■	■						⊜ ⊜ ○ 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					🟨				
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

2010 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
	Clemson - OrangFS				■	🟨				
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				■	■				
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

Past years are status, future years are identified needs or desires

2010 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 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 Updated Disksim including MEMS simulation SNL releasing kernel I/O tracing tool	PFS Sim		
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			

Roadmaps 2009

Metadata

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 - MDFS		🟨					
	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

2009 Metadata Gap Area

Area	Researcher	FY 07	FY 08	FY 09	FY 10	FY 11	FY 12	Rankings
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

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	■	■	■	▒			🔴 ⊝ ○ A comprehensive tool is nowhere in sight; problem is complex. This gap area includes monitoring.
	Narasimhan	■	■	▒				
	Reddy	■	■	■	▒			
	Smirni		▒	■	■	▒		
	Zadok	■	■	■	▒			
	Riska			▒	■	■	▒	
	He			▒	■	■	▒	
	Zadok (2009 HECURA)			▒	■	■	▒	
	SciDAC - PDSI	■	■	■	■			
	SciDAC - SDM	■	■	■	■	■		
Standards and common practices for HEC I/O benchmarks	Zadok/Miller		🟨	🟨	🟨			⊜ ◯ ○ Danger of over simplifying problem and could drive vendors to incorrect solutions.
	Ma/Shen/Winslett			▒	■	■	▒	
Modeling, simulation and test environments.	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	■	■	■	■			
	Maltzahn			🟨	🟨	🟨		
Applying cutting edge analysis tools to large scale I/O	Reddy	■	■	■	▒			🔴 🔵 ○ Data are becoming available from Labs including I/O traces. Many opportunities to evaluate this research. This includes applying analysis and visualization tools to I/O traces
	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

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

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. Traditional block-based solutions ready for commercialization. Alternative interfaces not yet well explored.
	Urgaonkar		▨	■	■	▨		
	Szalay/ Huang			▨	■	■	▨	
	He			▨	■	■	▨	
	Rangaswami				▨	▨		
	Arpaci-Dusseau (2009 HECURA)			▨	▨			
	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		▨	■	■			
	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

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							▤ ○ ○ 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*							▤ ▤ ○

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.

●	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

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

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

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

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 ANSI OBSD IETF pNFS	PDSI UM CITI patch pushing/maintenance Revamp of manual pages V2 nearing publication V 4.1 nearing pub Assistance in testing may be needed	First Linux full patch set Some file system pilot test Initial products	Layout Query going into POSIX V2 ratified 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 station file system statistic data		

Roadmaps 2008

Metadata

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 – LiFS/facets	🟨	🟨	🟨				
	CMU/ANL - MDFS		🟨					
	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	***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

2008 Measurement and Understanding Gap Area

Area	Researcher	FY 07	FY 08	FY 09	FY 10	FY 11	FY 12	Rankings
Understand system workload in HEC environment	Arpaci-Dusseau	■	■	▨				● ⊜ ○ A comprehensive tool is nowhere in sight; problem is complex.
	Narasimhan	■	■	▨				
	Reddy	■	■	▨				
	Smirni		▨	■	■	▨		
	Zadok	■	■	▨				
	SciDAC - PDSI	■	■	■	■	■		
	SciDAC - SDM	■	■	■	■	■		
Standards and common practices for HEC I/O benchmarks and trace formats	Zadok/Miller		🟨	🟨	🟨			⊜ ○ ○ Danger of over simplifying problem and could drive vendors to incorrect solutions.
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 analysis tools to large scale I/O	Reddy	■	■	▨				● ● ○ Data are becoming available from Labs including I/O traces. Many opportunities to evaluate this research.
	Zadok	■	■	▨				
	LANL/CMU – Trace replay and Visualizer		🟨	🟨	🟨			

● 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

2008 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.
	Chiueh	■	■	▨				
	Ganger	■	■					
Standard Interfaces 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	🟨	🟨	🟨	🟨	🟨	🟨	

🔴 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

2008 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 - Scalable file system architectures	Choudhary	■	■	▨				🔴 🔵 ⊜ 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	▨	■	▨				
	Wyckoff	■	■	▨				
	SciDAC – SDM	■	■	■	■	■		
	SciDAC – PDSI	■	■	■	■	■		
	PNNL	🟨	🟨	🟨	🟨			
Understanding file system abstractions - naming and organization	Bender/Farach-Colt	This Gap Area has been integrated into "Understanding file system abstractions – Scalable file system architectures" in Next Generation I/O Arch. and/or Metadata "Extensibility and Name Spaces"						🔴 🔵 ⊜ 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	This Gap Area has been integrated into "Understanding file system abstractions – Scalable file system architectures" in Next Generation I/O Arch.						🔴 🔵 ○ 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 HEC area. Includes power consumption. Traditional block-based solutions ready for commercialization. Alternative interfaces not yet well explored.
	Urgaonkar		▨	▨				
	PNNL	🟨	🟨	🟨	🟨			
HEC systems with multi-million way parallelism doing small I/O	Choudhary	■	■	▨				⊜ ⊜ ⊜ Good initial research; needs to be moved into testing. More fundamental solutions
	Dickens	▨	■	■	▨			
	Gao	▨	■	■	▨			

2008 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.
	Zhang/ Jiang	▨	■	■	▨			
	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

2008 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	■	■	▨				⊜ ○ ○ 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 ⊜ 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

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
	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
Metadata scaling	Bender/Farach-Colton Jiang/Zhu Leiserson Panasas Lustre ANL/CMU							⊜ ○ ● urrent research is in terms of le systems, not archive, but his work can be applied to rchive. ile system research will be more than fast enough for archive.
Policy driven management	*None*							○ ○ ⊜ Sarbanes-Oxley Act is solving this problem

This gap area has been integrated into Metadata "scaling" since work in file system scaling is related and applicable to archive.

● 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

2008 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.
	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.
Reliability, and degraded performance in HEC systems	*None*							⊜ ⊜ ⊜ Industry is working on 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

Security

2008 Security Gap Area

Area	Researchers	FY 06	FY 07	FY 08	FY 09	FY 10	FY 11	Rankings
Long term key management	Odlyzko							⊜ ⊜ ○ Current researcher need data to validate designs This is not a file system issue or HEC FSIO, but a problem everyone has. We are hampered by this problem
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.	*None*							● ⊜ ○ Industry will help some, but not in HEC context.
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. Note: NSF should incorporate this into a call for security research; this topic is larger than FSIO.

This gap area is recognized as not a file system specific problem, but a more general problem. Thus, the gap sub area is being removed from the Security Roadmaps. It must be noted that this problem is NOT

● 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

Past years are status, future years are identified needs or desires

2008 Assisting with Standards, Research and Education

Area	FY07	FY 08	FY 09	FY 10	FY 11
Standards: POSIX HEC ANSI OBSD IETF pNFS	PDSI UM CITI patch pushing/maintenance Revamp of manual pages V2 nearing publication V 4.1 nearing pub Assistance in testing may be needed	First Linux full patch set Some file system pilot test Initial products	V2 ratified NFS v4.1 final voting ("last call")		
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 station file system statistic data		

Roadmaps 2007

Metadata

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	***None***							● ● ○ Research is being done, but no research focused on metadata
Data Transparency and Access Methods	***None***							⊜ ⊜ ○ 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

Measurement and Understanding

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	***None***							○ ○ ○ 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

Quality of Service

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

Communication and Protocols

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

Archive

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	■	■	▨				⊜ ⊜ ○ Current research is in terms of file systems, not archive. API merging with POSIX and API for searching lacking
	Tosun	▨	■	■	▨			
	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

Management and RAS

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	*None*							⊜ ○ ○ 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

Security

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.	*None*							● ⊜ ○ 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

Assisting with Standards, Research and Education

Past years are status, future years are identified needs or desires

2007 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			
ANSI OBSD	V2 nearing publication	Some file system pilot test			
IETF pNFS	V 4.1 nearing pub Assistance in testing may be needed	Initial products			
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		
Education	LANL Institutes as one example PDSI	Other Institute like activites			
Research Data	Failure, usage, event data	Many more traces, FSSTATS, more disk failure data			

APPENDIX D: Inter-Agency HPC FSIO R&D Needs Document

HPC File Systems and Scalable I/O: Suggested Research and Development Topics for the fiscal 2005-2009 time frame

DOE Office of Science

Rob Ross ANL
Evan Felix PNL

DOE NNSA

Bill Loewe LLNL
Lee Ward SNL
Gary Grider LANL

DOD

Rob Hill, NSA

Executive Summary

The need for immense and rapidly increasing scale in scientific computation drives the need for rapidly increasing scale in storage for scientific processing. Individual storage devices are rapidly getting denser while bandwidth is not growing at the same pace.
In the past several years, Research and Development (R&D) into highly scalable file systems, high level I/O libraries, and I/O middleware was done to provide some solutions to the problems that arise from massively parallel storage. This document primarily concentrates on file systems and I/O middleware since high level I/O libraries have been addressed in many Data Management discussions and calls for R&D. The purpose of this document is to present areas of needed research and development in the HPC file systems and scalable I/O area which should be pursued by the government.

In the last five years, supercomputers with thousands of nodes and over ten thousand processors have been deployed. Additionally, thousands of small clusters have been deployed worldwide. File systems and I/O has come a long way since the simple cross mounted NFS that was used with early clusters. File systems and I/O middleware have been developed and deployed for these supercomputers and clusters systems which have enabled bandwidths beyond ten gigabytes/sec and metadata performance beyond one thousand file inserts/sec. It is now possible to scale bandwidth, and there is competition in the scalable global parallel file systems market space through products that span the gamut from completely proprietary to open source. I/O Middleware is maturing which is enabling many applications to get good performance from the underlying file systems.

Recently several new I/O R&D efforts have begun to try to address future needs in HPC file systems and I/O middlware. Efforts in the areas of:

- Relaxation of POSIX Semantics for parallelism
- Scalable metadata operations in a single directory
- NFSv4 security and the pNFS effort to allow NFS to get more native file system performance through separation of data and control which enables parallelism
- I/O Middleware enhancements to enable dealing with small, overlapped, and unaligned I/O
- Tighter integration between high level I/O libraries and I/O middlware
- Initial autonomic storage management
- Sharing of global parallel file systems between multiple clusters
- Initial active storage research

In the near future, sites will deploy supercomputers with tens of thousands or even hundreds of thousands of processors. Immense bandwidth, metadata, security, and management needs are emerging. Work flows for efficient complex science will begin to approach the exabyte range, and the ability to handle a more varied I/O workload including small to extremely large I/O operations, extremely high metadata activities, and multiple simultaneous workloads will be required. File systems will be so large that complete checks or rebuilds, entire tree walks, and other large operations will not be able to be contemplated. Management of these large storage systems will become increasingly difficult.

To meet the demands posed by the future HPC environments, investment in R&D and standards work need to be undertaken. The following are key areas to consider investment in:

- Scaling of metadata , security, reliability, availability, and management to deal with the enormity of future systems
- Enhancements in the POSIX I/O API and I/O middleware in the areas of ordering, coherence, alternate metadata operations, shared file descriptors, locking schemes, and portability of hinting for layouts and other information.
- Support for machine architectures which do not have full operating systems on each node and exploitation of hierarchies of node types
- Additional work in active storage concepts, including use in the LAN/SAN, WAN, application integration, and object archives
- Continued support for development and standards work on NFSv4 and pNFS
- Tracing, Simulation, and benchmarking, including application realistic benchmarking and simulation
- New metadata layouts other than tree based directories

More focused and complete government investment needs to be made in the file systems and I/O middlware area of HPC, given its importance and its lack of sufficient funding levels in the past, compared to other elements of HPC. Scalable I/O is perhaps the most overlooked area of HPC R&D, and given the information generating capabilities being installed and contemplated, it is a mistake to continue to neglect this area of HPC. Many areas in need of new and continued investment in R&D and standardization in this crucial HPC I/O area have been summarized in this document.

Background

The need for immense and rapidly increasing scale in scientific computation is well understood and documented. To keep up with the need for immense scaling, Moore's Law, which maps how individual processors get faster, and the idea of ever increasing parallelism are combined. In providing storage for scientific processing, similar and well documented concepts apply. It is well known that individual storage devices are getting denser at an amazing rate, keeping up with the ever faster speeds of processors. It is also well known that bandwidth to/from individual storage devices is getting faster but at an alarmingly slower rate than density of the devices. To deal with the ever increasing need for bandwidth to storage, massive parallelism is required.

In the past several years, Research and Development (R&D) into highly scalable file systems and I/O, was done to provide some solutions to the problems that arise from massively parallel storage. The I/O software stack is primarily composed of 4 basic layers: higher level I/O libraries, I/O Middleware, file systems, and storage devices. Higher level I/O libraries, for the most part, provide the needed high level abstraction for application programmers who are familiar with the science involved in the application. I/O Middleware, for the most part, provide abstractions that are computer science based, and deal with distributed/parallel nature of the memories involved and the distributed/parallel access to the storage devices. File systems, of course, provide the standard interfaces for organization, storage and retrieval of data and deal with coordination of the storage devices. The goal of the I/O software stack is to maintain performance and correctness from the high level I/O interface all the way to the storage devices and to fully leverage innovations in storage devices and the connectivity to those devices.

We define HPC file systems and scalable I/O as a subset of the overall scientific data management function, that subset of the I/O stack dealing with the storage devices, networks to reach those storage devices, data movement and metadata protocols, file systems, and I/O Middleware functions.

Purpose

The purpose of this document is to present areas of needed research and development in the HPC file systems and scalable I/O area which should be pursued. This document does not heavily delve into R&D needs in the high level I/O libraries area, because that topic is better addressed by Scientific Data Management (SDM) experts and scientific application designers and developers, although it is understood that coordination with high level I/O libraries is necessary and good. Nor does this document address R&D needs for advancement of physical storage devices. The R&D at the storage device level is largely driven by the needs of the multi-billion dollar storage industry and is done at a scale that is well beyond the proposed R&D investments in this document. The focus area of this document is the I/O Middleware and file system layers of the I/O software stack, primarily, with some extension upwards into high level I/O libraries and down into storage devices.

Additionally, this document does not address breakthrough storage technologies that would fundamentally change I/O paradigms. This document assumes an evolution of storage devices with no extremely disruptive technologies. While there are new storage technologies nearly ready to enter the market place like Micro-Electro-Mechanical Systems (MEMS) (http://www.memsnet.org/mems/what-is.html) and Magnetic Random Access Memory (MRAM) (http://computer.howstuffworks.com/mram.htm), given the market drivers behind these technologies, for the HPC market, it is very unlikely these technologies will be disruptive enough, in the next several years, to cause radical or wholesale changes in the I/O stack. While it is important to remain vigilant in exploiting new evolutionary technologies and watching for disruptive technologies, this document only addresses evolutionary technologies and ideas.

This document is also based on the idea that high end supercomputing will continue to be based on physically distributed memories. Much of the I/O software stack assumes and deals with this current distributed memory reality.

Frequently used terms

I/O – input/output
File system – A combination of hardware and software that provides applications access to persistent storage through an application programming interface (API), normally the Portable Operating System Interface (POSIX) for I/O.
POSIX - Portable Operating System Interface (POSIX), the standard user interfaces in the UNIX based and other operating systems (http://web.access.net.au/felixadv/files/output/book/x1164.html)
Global – refers to accessible globally (by all), often implies all who access see the same view
Parallel – multiple coordinated instances, such as streams of data, computational elements
Scalable – decomposition of a set of work into an arbitrary number of elements, the ability to subdivide work into any number of parts from 1 to infinity
Metadata – information that describes stored data, examples are location, creation/access dates/times, sizes, security information, etc.
Higher level I/O library – software libraries that provide applications with high level abstractions of storage systems, higher level abstractions than parallelism, examples are the Hierarchical Data Formats version 5 library (HDF5) (http://www-rcd.cc.purdue.edu/~aai/HDF5/html/RM_H5Front.html) and parallel Network Common Data Formats library (PnetCDF) (http://www-unix.mcs.anl.gov/parallel-netcdf/sc03_present.pdf)
I/O Middleware – software that provide applications with higher level abstractions than simple strings of bytes, an example is the Message Passing Interface – I/O library (MPI-IO) (http:www-unix.mcs.anl.gov/romio)
WAN – Wide Area Network, refers to connection over a great distance, tens to thousands of miles
SAN – Storage Area Network, network for connecting computers to storage devices

QoS – Quality of Service

Historical perspective

To put into context the rationale for the proposed set of research and development topics, a high level summary of more than the last half decade of HPC file systems and Scalable I/O related research and development is presented here.

The early 1990's

In the early 1990's, most cluster based supercomputers were relatively small by today's standards, typically with tens of nodes, and for the most part used Network File System (NFS) servers for both home (program/source code/etc.) and scratch space. Sometimes only one NFS server was used, but frequently multiple cross mounted NFS servers were used. Given the small nature of these clusters, this approach provided sufficient performance and acceptable access. Given the primitive and limited scale of the applications in this time frame, parallel access methods to I/O typically were not needed.

The mid 1990's

In the mid 1990's, large scale cluster based supercomputers, hundreds to a few thousand nodes, began to show up. In order to serve the storage bandwidth needs of these clusters, hundreds of redundant array of independent disks (RAID) controllers with over a thousand individual disks used in a coordinated manner were required. Given that the Linux cluster phenomenon had not occurred yet, these clusters were for the most part proprietary operating system based clusters. NFS or cross mounted NFS was simply not an option on these machines due to their need for scalable bandwidth and manageability at scale. The file systems used on these clusters were first generation client/server based proprietary cluster file systems. Examples include IBM's General Purpose File System (GPFS) (http://www.almaden.ibm.com/StorageSystems/file_systems/GPFS/), Meiko's Parallel File System (Meiko PFS), and Intel's PFS (Intel PFS). Additionally, due to the large scale parallelism in the clusters of this era, middleware libraries like the MPI-IO library to enhance parallel access were begun. Projects to develop Storage Area Network (SAN) based file systems, which relied, at the time, on every file system client having direct access to the storage devices, typically via a fibre channel SAN, were also begun. Examples of SAN approaches at that time are the University of Minnesota/Sistina Global File System (GFS) (http://www.redhat.com/software/rha/gfs/), the Silicon Graphic Incorporated (SGI) Clustered-XFS (C-XFS) (http://www.sgi.com/products/storage/cxfs.html), the Advanced Digital Information Corporation (ADIC) Clustered Virtual File System (CVFS) (http://www2.adic.com/stornext/). These SAN file systems were primarily intended for use in smaller node count clusters (tens to a few hundred) due to the difficulty and expense in building large secure Fibre Channel SAN's. Fibre Channel SAN's lack the ability to share data at a granularity smaller than a volume (disk or virtual disk) and all nodes/machines that have direct access to the SAN that are sharing a volume have root style access to that entire volume, so there is no concept of sharing at the file or object level with authentication based security. Another concept that took shape in this timeframe was the use of an old idea, the separation of metadata activity from data movement activity. Some file systems deployed this concept though the use of separate metadata servers and others shared the metadata responsibility among the file systems clients through the use of lock mechanisms.

The late 1990's to present

In the late 1990's to the present, cluster supercomputers are now routinely in the thousands of nodes with tens of thousands being contemplated. Given the slow improvement of individual disk device bandwidth improvement compared to processor speed increases, the number of individual disk devices needed to be coordinated to achieve the needed bandwidth for these super clusters is currently several thousand. Open source Linux clusters are common place, implying that other, more open and Linux based, solutions for global parallel file systems are needed and past proprietary solutions for global parallel file systems are diminished in value.

Both SAN based and client server file systems have grown in popularity as clustered computing has become more prevalent in thousands of computing centers worldwide. Pure SAN file systems, again, where all file system clients have access to the disk devices, still suffer from scalability for very large clusters due to the difficulty in building scalable Fibre Channel SANs, also have security issues due to clients having direct access to storage. Most SAN based file systems have begun to offer ways to extend access to beyond clients that have direct access the disk devices through a variety of mechanisms like SCSI protocol over IP (ISCSI), gateway functions, and others. In part, due to government sponsored R&D, new and less proprietary client server file systems have become popular. Examples include: the DOE Office of Science/Argonne National Laboratory/Clemson Parallel Virtual File System (PVFS) (http://www.parl.clemson.edu/pvfs/desc.html), the DOE/NNSA/tri-labs/HP/CFS/Intel Lustre file system (http://www.lustre.org), and the Panasas PanFS (http://www.panasas.com) file system. These new and less proprietary client server file systems leverage heavily separation of data and control, abstracting the storage device functions like allocation and date-write operations for less than full block updates away from the file system proper.

Recently the ANSI T10/1355-D Object based Storage Devices (OSD) (http://www.t10.org/ftp/t10/drafts/osd/osd-r10.pdf) standard was introduced, putting the industry on a course to standardize the interface to storage devices that provide allocation, read-update-write on partial block writes, and transactional security for the storage devices. Additionally, using separate metadata servers for metadata operations has become even more popular. The MPI-IO library and other parallel access methods became popular in parallel applications during this time.

Additionally, in this time frame several important HPC file systems and I/O related R&D efforts were begun including:

Relaxation of Posix Semantics

A realization that the POSIX semantics had to be broken to enable scaling occurred. Ordering semantics to ensure last writer wins on overlapped I/O operations and lazy updates of last update times/dates and file sizes for files being written to by many clients concurrently are both examples of where relaxations have been made. File systems like PVFS, Lustre, and Panasas have all implemented options for relaxed POSIX semantics where absolutely necessary for scaling. These three file systems and others have implemented special (non POSIX standard) I/O controls (IOCTL's) which allow applications to control widths, depths, and other striping oriented layout information for files. The PVFS is probably the best example of extending POSIX semantics for scaling. This is accomplished by providing a user space library interface to the file system which is well integrated with the MPI-IO parallel I/O library. To assist with situations where overlapped I/O is necessary and control over last writer wins semantics is important for the application, Northwestern University has done important work in enabling detecting and handling these overlapped I/O operations within the MPI-IO library, where multiple clients can coordinate these activities in an intelligent way.

Scaling

Extremely scalable parallel data movement bandwidth has been achieved by client server based file systems. Both Lustre and Panasas have shown coordinated bandwidths in excess of 10 GigaBytes/sec and both have plans of exceeding 30-50 GigaBytes/sec in fy05. For the most part, the data movement bandwidth scaling problem has been solved at least for non-overlapped, large buffer, parallel I/O operations to/from file systems. Some initial research has begun at the University of California Santa Cruz (UCSC) in the area of scalable metadata. The problem here is handling tens to hundreds of thousands of inserts, deletes, and queries per second in file systems that will manage billions of files. Some file systems like the IBM SAN file system and others have introduced scalable metadata systems, but only for scalability across multiple directories. Scalability within a single directory is a very hard problem especially when posed with the possibility that mass operations like tens of thousands of inserts could be requested into the same directory or subdirectory all within a few milliseconds. UCSC has come up with some novel ideas in trying to address this problem. Additionally, the Lustre and Panasas file systems are building first generation engineering approaches to the scalable metadata problem for scalability across multiple directories as well as within one directory.

NFS version 4

The Internet Engineering Task Force (IETF) NFS version 4 (NFSv4) (http://nfsv4.org) effort has seen progress in the last two years. The NFS has been riddled with security problems since its inception. The NFSv4 effort is addressing this issue through the use of the General Security Services (GSS) infrastructure. Many government sites see this as a very useful move for the NFS. Additionally, the NFSv4 effort has new compound operations capability which allows for multiple operations to be coalesced to allow for efficiencies never before possible with the NFS. Additionally, there is a new parallel NFS effort (pNFS) (http://www.ietf.org/proceedings/04mar/slides/nfsv4-1.pdf) which is a part of the overall IETF NFSv4 project. This pNFS effort promises to allow NFSv4 clients to perform metadata operations on file systems via the NFSv4 server and then bypass the NFSv4 server for data movement operations. There have been many non-standard modifications to the NFS over the years to bypass the NFS server for data movement operations, going directly to the storage devices. This pNFS effort legitimizes these approaches via the IETF standardization process. This pNFS effort promises to make NFS clients first class clients to parallel file systems for performance. This allows for a large variety of heterogeneous clients to have high performance clients to parallel file systems. Important work by the University of Michigan is enabling a Linux implementation of these NFSv4 features. Many vendors are participating in building NFSv4 releases as well as working on the NFSv4 and pNFS IETF standards effort. Garth Gibson of Panasas and Carnegie Mellon University (CMU) has been instrumental in working on the pNFS standards effort.

Higher level I/O libraries and integration with I/O Middleware

The I/O software has to be able to extract the available performance from the underlying storage devices. The high level I/O library must be well integrated into the overall I/O stack so that performance can be attained. It is vital to recognize important performance work done by assisting with tighter integration between I/O Middleware and higher level I/O libraries like the Hierarchical Data Formats version 5 (HDF5) from the National Center for Supercomputer Applications (NCSA) and parallel Network Common Data Format (PnetCDF) Argonne National Laboratory (ANL) and Northwestern University. It is important to tune these higher level libraries and tune how applications utilize these libraries. Work by ANL to utilize the PnetCDF library integrated with the MPI-IO library and the PVFS achieved promising performance results. The joint work by Los Alamos National Laboratory and NCSA using the Unified Data Models (UDM) library in conjunction with the HDF5 library which uses the MPI-IO I/O Middleware has also achieved promising performance for applications.

Enterprise Class Global Parallel File System (GPFS – not to be confused with the IBM GPFS – General Purpose File system)

Supercomputer sites have been deploying more than one computing cluster for many years. Sometimes sites will have a very large cluster for simulation with smaller clusters for pre/post processing of data, reduction, analysis, and visualization. In the last few years this has become common for larger supercomputer sites to have more than one very large cluster for simulation. This often arises from sites installing a new large cluster every year or two but keeping the past generation or two before decommissioning. Sites with more than one cluster increasingly want to share access to the same data between clusters. In the past it was very common to have a separate parallel or high bandwidth file system on each cluster with some means to move the data between clusters for sharing, either through direct data movement or via a common archive capability. Many supercomputer sites are now expressing the desire to have and deploy parallel and scalable file systems that are shared between all the supercomputers at the site. We refer to these deployments as Enterprise Class Global Parallel File Systems (GPFS). Often sites desiring this Enterprise Class GPFS capability even want to provide access not only to just the clusters in the enterprise, but for workstations and other servers in the enterprise. Giving access to multiple clusters and workstations to a common file system gives rise to new issues like Quality of Service (QoS) guarantees to more important clients, differing security between different kinds of clients, and heterogeneous access. Space allocation on a filesytem according to various policies, which relate to which system, project, or user is creating the data, will also become important, as multiple funding sources will be dictating how the resources are used. Relevant work in this area includes the NFSv4 and pNFS work already mentioned, as well as some early design work for QoS in the Lustre file system and in the ANSI T10/1355-D Object based Storage Device standardization effort.

Utilize processing power near the storage devices – Active Storage

With the success of client server oriented file systems, the opportunity to utilize server side processing power near the disk storage device was recognized. Many organizations have contemplated using this power near the disk. Early research at the Intelligent Storage Consortia (ISC) at the University of Minnesota has looked at ways to utilize the power near the storage device to provide functions like hierarchical storage management (HSM), indexing, and mining. Other universities have also done preliminary research into how the power near the storage device could be used. Proof of Active Storage concepts has been done at PNNL(http://www.emsl.pnl.gov/), by augmenting the Lustre filesystem. Additionally, industry has begun to deploy storage systems with processing capabilities, probably the most noteworthy are the Content Addressable Storage released recently by the EMC^2 Corporation (http://www.emc.com/products/systems/centera/) which allows access to data objects by their content, and the Data Appliance also released recently by the Netteza Company (http://www.netezza.com) provides database query operations. This area of R&D represents great promise, but only initial work has been done. The advent of the ANSI T10/1355-D OSD standard which provides a standard and secure way to request actions from an intelligent storage device is an important step to being able to utilize processing power near the storage device.

Summary of current state

It would be appropriate to call the late 1990's to the present the era spent enabling extremely scalable data movement bandwidth. Multiple client/server based file systems have achieved greater than 10 GigaBytes/sec and have plans to exceed 30-50 GigaBytes/sec in fy05. Supercomputing sites have deployed scalable global parallel file systems for individual extremely large clusters as well as for multiple clusters in an enterprise. It is now possible to scale bandwidth, and there is competition in the scalable global parallel file systems market space through products that span the gamut from completely proprietary to open source. Standards are emerging, like the ISCSI standard, the ANSI T10/1355-D OSD standard, and others, which will lead to even more competition. Some work has been done to deal with POSIX limitations; the NFSv4/pNFS efforts appear to be headed in a good direction for security, heterogeneous access, and WAN access; tighter integration of the I/O software stack has yielded good results; and some promising initial work on metadata scaling and how to utilize the power near the disk has been started. Much progress has been made in the HPC file systems and scalable I/O area in the last decade.

Demands of the next 5 years

In the near future, sites will deploy supercomputers with tens of thousands of processors routinely, perhaps even hundreds of thousands. Bandwidth needs to storage will go from tens of GigaBytes/sec to TeraBytes/sec. Online storage needs 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 including small to extremely large I/O operations, extremely high metadata activities, and multiple simultaneous workloads will be required. Additionally, new access methods, such as access methods to files/objects in arrangements other than tree based directories, will be required. Global or virtual enterprise and wide 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, making the need for integrity and reliability schemes to be far more scalable than available today. Management for enterprise class global parallel file/storage systems will become increasingly difficult due to the number of elements involved and the extreme varied workloads. The challenges of the future are formidable.

Key areas of possible future research, development, and standards work

As indicated by the challenges for the future, the needed R&D activities need to shift from scaling of data movement bandwidth for large I/O operations to scaling performance for high volumes of small I/O operations, high volumes of metadata operations, and extreme mixing of a variety of workloads. Improving reliability, integrity, availability, and manageability by orders of magnitude must be addressed. Additionally, R&D into new access methods, issues for enterprise wide sharing, WAN and heterogeneous access, security, as well as extreme scaling issues for metadata operations, and security will be required. Novel approaches to these issues such as leveraging the power near the storage devices, tighter integration of the I/O software stack, as well as other approaches should be studied.

It is important that the natural evolution from ideas to prototypes, from prototypes to user level library implementations, from user level library implementations to standards, standard implementations and even products, be supported and managed. Thus, all efforts throughout this evolution should be undertaken where promising results are indicated.

Below, categories of future R&D and standards work are discussed.

The POSIX I/O API

The POSIX API is "unnatural" for high-end computing applications. Opportunity abounds to make the POSIX I/O API more friendly to HPC and parallelism. The entire set of operations should be combed over and carefully, consistently, altered for high-end computing needs. Then, the result must be re-integrated in such a way that it is enabled by all applications in a positive fashion. The resulting semantics, after all, are less than useful for legacy applications as well as single platform applications.

Ordering

One of the prime reasons the POSIX I/O API is awkward for HPC stems from the "stream of bytes" paradigm in which POSIX I/O is based. POSIX I/O was developed to provide an interface from a single machine with a single memory space to a streaming device with some simple random access capabilities. HPC/parallel I/O applications are based on distributed memories being mapped to many storage devices. A re-interpretation towards the concept of a "vector of bytes" would be more appropriate for HPC applications versus a "stream of bytes" model. This would entail a careful reexamination of the POSIX I/O-related interfaces to eliminate or relax stream-oriented semantics.

Coherence

Probably the worst offense for really high bandwidth I/O is the fundamental read and write calls. They have two glaring problems. The first is coherency. The last-writer-wins semantic of the write call without cache-snooping is difficult, perhaps impossible, to achieve in a scalable fashion. Similarly, again without cache-snooping, the overhead of cache invalidation is enormous for reads. Especially if attempted for regions for which an application might never have interest. Additionally, block boundaries can present coherence issues for the application. A standard way for applications to assume all responsibility for coherency is needed, which implies that application control of cache flushing/invalidation at some level is also needed. Additionally dealing with block boundaries and alignment needs to be dealt with in the API in a consistent manner which takes alignment/block boundary issues away from the application. This problem is particularly bad in the implementation of O_Direct today.

Extension issues

The POSIX I/O API is organized as a set of mandatory functions and sets of extensions to that set of mandatory functions. Current extensions of the API are awkward for use in HPC. For instance, the real-time extensions for list-based I/O (listio) are useful, however they are awkward in that they have the restriction that the memory regions of interest must coincide exactly with a region in the file. For many applications, relationship of memory regions to regions in the file is often not possible. Instead, two separate vectors, one of memory regions and one of file regions, could be passed and the two lists reconciled by the implementation. Such a concept allows scatter/gather-gather/scatter semantics.

Missing capabilities

The POSIX I/O API is also not as complete as one would like. For instance, there is a call to position and write a buffer of data (pwrite) but no call to position and write a vector of described memory regions, like a pwritev.

Metadata

The classic "wish" in this area is for the support of "lazy" attributes. These are results to the stat call, where some values may not be maintained at the same granularity normally expected. The most obvious fields are those that record timestamps for last update and modify. Many file systems implement these but no two in the same way. A standard, portable set of semantics would be useful. Explorations into a more descriptive API for metadata management and query to allow applications to deal with the needed information could be helpful in this area. For years, the backup/archive industry has needed a portable bulk metadata interface to the metadata of file systems. There are many opportunities for R&D in the area of an overhaul of how metadata is handled and the API's with which it is accessed given the extremely limited implementations in today's file systems.

Locking schemes

Locking schemes that support cooperating groups of processes is also necessary. The current POSIX semantics assume only a single process would ever want exclusive write access. In the HPC/parallel world, groups of processes may want to gain exclusive access to a file. Perhaps a mandatory version of the fcntl and flock is needed. It is necessary that more options for how locks can be requested and revoked be provided. Legacy codes must continue to work as expected, so current locking semantics must be maintained.

Shared file descriptors

Shared file descriptors between nodes in a cluster would be of great value, not just between processes on a node. The component count for supercomputers is going up in the next generation of supercomputers. Full lookups in the name space are going to have to become a thing of the past for parallel codes. Some mechanism decreasing the need for mass name-space traversal is desperately needed, even if it requires new semantics to accomplish it. It is possible to implement this via higher level function in the I/O stack. Implementation of shared file descriptors at the file system level might be difficult but none the less would be quite useful, if achievable with a reasonable amount of R&D investment. As mentioned in the metadata section above, an alternate API for name space traversal as well as alternate file organization (something other than a tree) might also be a way to assist in this area.

Portability of hinting for layouts and other information

There is a need for proper hinting calls. Things like stripe, stride, depth, raid layout options, etc. need to be accomplished in some portable way. Additionally, there needs to be mechanisms for adding standard hinting without major new standardization efforts. Perhaps the MPI-IO Info approach for hinting can serve as a prototype, particularly in terms of the semantics, like ignoring of unknown hints and the mechanism for getting the values to use. For users to understand and use these hints effectively, they need to be as easy to use as things like umasks, shell variables, or file permissions.

Necessary determinism

Additionally, all operations done on the basis of time are awkward to deal with on supercomputers with light weight operating systems due to the inability to respond via asynchronous signaling to call back mechanisms. Supercomputing applications need more deterministic behavior and more control over the hardware throughout the entire computation and I/O hardware stacks. Operations with the ability to be driven from clients that can't listen for call backs is vital. It is quite likely that some variants of supercomputers with hundreds of thousands of processors simply won't be able to be bothered with call back mechanisms at all.

Active storage concepts

Active storage concepts are those ideas where CPU power near the storage device is utilized for better overall application performance or machine throughput. The work on active storage and associated concepts is interesting, although it is important that the value proposition be examined closely. Just because processing power near the storage device can be used to participate in the problem solution does not mean there is value in doing the processing near the disk as opposed to on other hardware. Many examples of particular applications and classes of applications enjoying significant benefits from this technology are available. However, to date, no pursuit of a generally useful, secure interface has been made. Research into a library, or hardware/firmware, interface supporting "sandboxes" and rich programming support could go a long way toward motivating applications to make use of the scheme. Without proper interfaces, this proposed new function will always be a "one-off" for any application, generally useful and portable for that application, but extremely hard to reuse, especially in an environment where more than one application might like to leverage active storage paradigms simultaneously. This R&D area is in its infancy, still mostly in prototypes. More work needs to be done to understand better the value proposition this idea brings and how it might be used in a more standard way.

Application of active storage concepts – Network Topology

In the local, machine, or system area network setting, the processing power near the storage device has no real advantage in bandwidth or latency to the storage device over any other processor, it is unclear that applications actually benefit from the processor near the disk versus just adding another general purpose processor to the application pool or a co-processor near the disk used only for application specific code. There is risk, availability/reliability risk, in putting application oriented function directly in the path of a highly shared item like a storage device. The processing power near the storage device does enjoy one advantage in this setting, that being location. All accesses to the storage device go through this processor. It is possible that this advantage could be exploited in some manner.

In the wide area network setting, the processing power near the storage device offers at least a latency advantage and possible also a bandwidth advantage over a general purpose processor. In this setting, there are many advantages to exploit, including smart batching of requests to hide latency, retrieving only the data needed for transmission over the WAN, etc.

Application of active storage concepts – enhancing the I/O stack function

Another important aspect is the ability to utilize the processing power near the storage device to simplify the higher layers in the I/O stack. Pushing more function nearer the storage device could have the benefit of allowing more innovation to occur for file systems, I/O Middle Ware, and high level I/O libraries. Exposing data layout information to the processor near the storage device could help that processor better map I/O operations to the geometry of the underlying storage and open up new possibilities for I/O stack exploitation of this concept. Database systems live in this world, so it is likely that many ideas can be formulated by studying database technology. R&D in this area could pay big dividends for some applications.

Leveraging active storage based file system technology, Archive/HSM, Other approaches

One example of utilizing active storage to allow for enhancement of the I/O stack is the possibility of integration of Archive/HSM function. Disk-based file systems for clusters are increasingly using multiple software "mover" components to accomplish parallel data transfers. These movers, most often, function by exporting access to unique, independent data stores.

Classic hierarchical storage management (HSM) methodologies also employ multiple movers, but curiously, usually to support more connections, not parallel connections. Lessons learned from recent file systems work could be used to simplify the back-end data path in HSMs by using a metadata service to maintain tape and location layout information. Perhaps a realistic core set of requirements for archive products for science use might be a stepping-off point to an acceptable interface to HSM software with a usable lifetime greater than a decade. The marriage of modern file system designs with a subset of classic HSM software could yield a seamless infinite global parallel file system solution which could eliminate the need for a separate parallel or serial archive capability.

Further, there are doubtless additional approaches not yet considered for using active storage. There needs to be an effort to pursue other ways one might design a scalable file system with computational power near the storage devices to contribute to providing solutions for data intensive related problems?

Application integration with Active Storage

As has been mentioned earlier, tighter integration in the I/O stack is becoming important for applications to effectively tap the performance of the I/O Middleware, file systems, and storage devices. Extending this integration from the application all the way to the storage device through the use of processing power near the storage device, which has been shown in the past to have promising performance, warrants a closer look. As mentioned above, if the processing power near the storage device had information about data layout, much could be done to exploit that fact higher in the I/O stack. A generic capability for applications to securely and effectively utilize processing power near the disk should be explored, prototypes of this environment need to be developed and tested to determine if the performance value proposition is worth the risks of destabilizing a highly shared device like a storage device. More work in understanding the performance payoff for applications, the API(s) needed to accomplish this capability, and the risks in providing this capability needs to be done.

To date, though, all research in this area has only been applicable to a single application at any given moment. To be really useful, sandboxes or other technology must be applied to allow independent applications access simultaneously.

NFSv4

As has been mentioned earlier in this document, the NFSv4 effort has yielded a much more secure NFS capability. There is still good work in the pipeline from the NFSv4 effort which must continue to be supported. Additions of directory leasing capabilities for WAN access performance, load balanced NFS serving, and the important pNFS effort which promises to allow heterogeneous NFSv4 clients to access file systems more natively by bypassing the NFS server for data movement operations, are all vital parts of the NFSv4 effort that need to be accomplished. In order for these efforts to be successful, development and standards work need to continue. The IETF requires two interoperating implementations, so this requires persistence in funding and oversight to see these projects through to completion and insertion officially into the IETF.

Enterprise Class Global Parallel File System (GPFS – not to be confused with the IBM GPFS – General Purpose File system)

The use of a global parallel file system by multiple clusters within an enterprise and extending access to the desktop workstations of an enterprise causes a set of issues to arise. These issues have mostly to do with treating one set of clients differently than others. There may be a need for security or other services to behave differently based on file system client or sets of clients. Additionally, this idea applied to performance implies a QoS solution is needed to enable one set of applications/clients to be treated differently than others applications/clients. R&D and standards work need to occur to enable these capabilities to support this enterprise class sharing concept in a portable way. Further, when connecting multiple clusters of different technologies and workstations to an Enterprise Class GPFS, scalable backbone technologies that allow heterogeneous interconnection at extremely scalable bandwidths with high reliability and availability are needed. Normally, single cluster interconnects are designed to scale to very high cross sectional bandwidth, but intra-cluster networks are not designed to scale that broadly. This multi-heterogeneous-cluster to common GPFS scalable network is a new development and needs to be studied. It is possible that Internet Protocol version 6 (IPv6) may be of assistance with this issue.

Scaling

As mentioned before, clusters of unprecedented scale are on the drawing boards with tens to hundreds of thousands of processors. Given that data movement scaling has been accomplished, R&D to address scaling other attributes of file systems and I/O is desperately needed.

Metadata

Clustered file systems seem to be converging on an architecture that employs a centralized metadata service to maintain layout and allocation information among multiple, distinct movers. While this has had significant, positive impact on the scalability in the data path, it has been at the expense of the metadata service. Due to scale up, the transaction rates against the metadata service have increased. As well, the amount of information communicated between the metadata service component and the clients has increased. There is some belief that such a file system design is problematic. The reason for this is seek-latency in disk media. Additionally, alternate metadata access methods, like bulk metadata access and perhaps alternative to tree shaped access of the metadata might provide both new needed function and relieve some of the metadata scaling issues.   It is vital that continued R&D investments be made in this vital scaling of metadata performance.

Data movement bandwidth with small and unaligned I/O operations

With the incredible success in scaling data movement bandwidth using large I/O operations, it is now time to concentrate on dealing well with the scaling of small I/O operations. Many applications have not been able to take advantage of the enormous improvements in file system scalability in the last several years due to the small I/O operation sizes used by these applications. It is vital that all applications be able to have scalable I/O available to them. Often, it is inconvenient, or impossible, for the applications to be altered. They are expensive, proven codes and, in some cases, the host machines do not have the memory it would require to efficiently rearrange working sets for efficient data transfer. For non dusty deck applications, R&D in areas of more aggressive caching in high level I/O libraries, I/O middleware, and alternate file system consistency close-to-open semantics could pay off, particularly in applications with lots of small I/O operations to independent files. It is also possible that active storage pursuits could assist in this area significantly by providing data layout information to the processor near the storage devices to allow for better mapping of the workload to the underlying storage geometry. For the more dusty deck oriented applications, this is a very difficult and perhaps nearly intractable problem, however, if R&D could assist these applications in their ability to use global parallel file systems more effectively and efficiently, there is a win for users.

High Level library exploitation of scalability enhancements

As has been mentioned several times in this document, integration up and down the entire I/O software stack has yielded good performance benefits. If file systems become better at scaling of metadata and small I/O operations as mentioned above, further re-integration of higher level I/O libraries to take advantage of these file system improvements may be possible and should be explored. As an example, it is possible that formatting libraries, which currently put all data and metadata for a single application run into a single file, might leverage scalable metadata operations by keeping a family of related files associated with a single application run. It is important to not resort to thousands of files per application or one file per process in this endeavor, but some modest number of files based on access patterns could be exploited. In the storage management field, the term "collections" is used to describe this concept. All advancements in the file system layer and below should be exploited if possible by higher layers in the I/O software stack.

Security

Another dimension for file systems that must be addressed in a scalable fashion is security. Allowing file system clients to access storage devices in a scalable way may require transactional oriented security so storage devices can trust that clients are authorized to perform the request. Additionally, with metadata services becoming more scalable, security related workload for authentication and authorization, which the metadata server must do, is increasing. The necessity for security even as scaling increases means security services must be scalable too. For this reason, R&D investments in security scaling must be undertaken.

Reliability and availability

As has been discussed, clusters of unprecedented scale are being planned. To provide the needed file system bandwidth to these clusters, unprecedented numbers of storage devices will need to be used in a coordinated way. Striping data from a single application over enormous numbers of disks will eventually lead to difficulties in protecting against data unavailability or loss. Current RAID protection and availability technologies are not designed to provide sufficient protection for such immense scale. One concept that could be pursued is lazy redundancy, producing redundant data at specific points in time, perhaps associated with checkpoints or snapshots. This concept allows for variable redundancy on a per file basis which allows for trading off reliability for performance based on expected usage. Another concept that could be pursued is the ideas related to raiding memories in compute nodes. This concept works quite well for pure defensive I/O and dovetails nicely with MPI-2 features of dynamic process/communicator growth. There are no doubt other redundancy concepts that could be pursued as well.

Management

Yet another area affected by immense scale is management. The number of devices needed to provide the needed scalable file system service in a demanding and mixed workload environment of the future will be extremely difficult to manage given current technology. Advances must be made in massive scale storage management to enable management survival with future file system deployments.

Autonomic Storage Management concepts

The storage industry is currently working on management solutions that are fully automated. Ideas like, storage that self configures, self heals, self migrates, and self tunes are all being pursued. These ideas are all good ideas and the related projects need to be at least followed by the HPC I/O community. Additionally if these features are to be useful in the HPC environment, where things like determinism in parallel are important, it would be very useful for the HPC community to be involved in this Autonomic Storage Management R&D to ensure that these good ideas are implemented in a way that the HPC community can benefit. As an addition to this thinking, automated mining of data is also being pursued. These features also could be useful in the HPC environment but must be developed with consideration for the HPC I/O environment.

Hierarchical I/O architectures

Many supercomputers are arranged with compute nodes, I/O/routing nodes, and storage. I/O. All I/O operations on behalf of the compute nodes are routed in some manner through the I/O/routing nodes. Some of the newer architectures emerging, like the Red Storm and Blue Gene/Light architectures, require this type of I/O arrangement. An excellent research questions is: “Can part of the I/O function be placed at these I/O/router nodes to assist in performance for the user, especially in the areas of caching and aggregation”?

Tools for scalable I/O tracing and file system simulation

In parallel application building and tuning, there are a multitude of correctness and performance tools available to applications. In the area of scalable I/O and file systems, there are few generally applicable tools available. Tools and benchmarks for use by application programmers, library developers, and file system managers would be of enormous use for the future.

Tracing

Tools to quickly get tracing information from parallel applications, analyzing these traces to characterize the applications I/O footprint, and even being able to replay the traces in parallel against real or simulated parallel file systems would be of great use.

Simulation

Tools for simulation of portions or entire global parallel file systems would also be of great use to assist in understanding design trade-offs for I/O performance for applications, libraries, and file systems. Most other areas of the computer industry rely heavily on simulation tools. While asking for a parallel file system simulator is a tall order, the value it would have could be enormous.

Benchmarking

As in other areas of computing, benchmarking is a vital part of the I/O professional’s toolkit. Benchmarks in the areas of interactive benchmarks (simulating user experience items like ls, cp, rm, tar, etc.), throughput benchmarks (including both peak performance and I/O kernels from real applications), and metadata benchmarks (collective opens, creates, resizes, etc) are needed. Some of these benchmarks exist and others do not. R&D in the benchmarking area is needed to collect and index current benchmarks and design and build new benchmarks to fill any gaps. At the very least, a clearing house for all I/O benchmarks and their usage could be of benefit. It currently is quite difficult to determine if a benchmark exists for a particular function or workload.

New metadata layouts other than tree based directories

In order to assist applications with managing enormous amounts of data, application programmers and data management specialists are calling for the ability to store and retrieve data in organizations other than the age old file system tree based directory structure. The data formats libraries currently provide some of this function, but it is not at all well mated to the underlying file system capabilities. Databases are often called upon to provide this capability but they are not designed for petabyte or exabyte scale stores with immense numbers of clients. Exploratory work in providing new metadata layouts is vital to address this identified need.

Kernel/user space data movement and control passing

There are many benefits to providing I/O related services completely in user space. Due to the kernel based file system layer paradigm on which most all operating system function rests, it is necessary to continue providing file system services through the Unix kernel. If a zero-copy data path from user space to kernel and back were standardized in Unix/Linux, it would be possible to implement more file system services in user space without a penalty in bandwidth or latency performance. This development could open up new and innovative I/O and file system services never before possible due to the abundance of user space developers. As well, user-space implementations tend to be more highly portable and cheaper to implement.

IPv6

If machines continue to grow in power at the same rate – Moore's law again – then the number of components in the I/O system must increase dramatically. Most of these components are uniquely addressable. While the internet protocol (IP) is ubiquitous, its address space is partitioned into only very small remaining chunks. The advent of IPv6 presents an opportunity to cleanly craft addressable sub-units in the I/O system. Unfortunately, little attention has been paid to this relevant streamlining.

Support of Storage Centers of Excellence

As part of the overall investment strategy, consideration should be given to supporting the approximately 3 university storage centers of excellence within the US at some base level. These centers provide ongoing file systems, storage, and scalable I/O research funded by industry partners. Supporting and being involved in these centers leads to more access to industry planners, more leverage over research topics, and more access to students and faculty. Additionally, supporting these centers produces the next generation of researchers in the field.

Conclusion

More focused and complete government investment needs to be made in this area of HPC, given its importance and its lack of sufficient funding levels in the past, compared to other elements of HPC. Scalable I/O is perhaps the most overlooked area of HPC R&D, and given the information generating capabilities being installed and contemplated, it is a mistake to continue to neglect this area of HPC. Many areas in need of new and continued investment in R&D and standardization in this crucial HPC I/O area have been summarized in this document.