HEC FSIO Session 2: Next Generation I/O Architectures Talks and Roadmap

Steve Poole, Oak Ridge National Lab

August 2009

OAK RIDGE NATIONAL LABORATORY

Current R&D Gaps

- Understanding file system abstractions – file system architectures
- Understanding file system abstractions – naming and organization
  - 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”
- Self-assembling, Self-reconfiguration, Self-healing storage components
- Architectures using 10^6 storage components
  - This Gap Area has been integrated into “Understanding file system abstractions – Scalable file system architectures” in Next Generation I/O Arch.
- Hybrid architectures leveraging emerging storage technologies
- HEC systems with multimillion way parallelism doing small I/O operations

2006 HECURA/CPA Projects

- Petascale I/O for High End Computing; Maccabe, Arthur, UNM and Schwann, Karsten, Georgia Tech Research Corporation – GA Inst of Tech
- Application-adaptive I/O Stack for Data-intensive Scientific Computing; Xiaosong Ma, North Carolina State University, Anand Sivasubramaniam, Pennsylvania State University University Park, and Yuanyuan Zhou, University of Illinois at Urbana-Champaign
- Active Storage Networks for High End Computing; John Chandy, University of Connecticut
- Active Data Systems; A.L. Narasimha Reddy, Texas A & M University
- Concurrent I/O Management for Cluster-based Parallel Storages; Kai Shen, University of Rochester
- Deconstructing Clusters for high end biometrics; Douglas Thain, Notre Dame
- Object Based Caching for MPI-IO; Philip Dickens, University of Maine
- A High Throughput Massive I/O Storage Hierarchy for PETA-scale High-end Architectures; Guang R. Gao, University of Delaware
- Effective Resource Allocation under Temporal Dependencend Architectures; Evgenia Smirni, William and Mary
- HybridStoe: An Enterprise-scale Storage System Employing Solid-State Memory and Hard Disk Drives; Bhuvan Urgaonkar, Penn State University

2009 HECURA Projects and Presentations

- HaRD: The Wisconsin Hierarchically- Redundant, Decoupled Storage Project; Remzi Arpaci-Dusseau, Wisconsin
- A Dynamic Application-specific I/O Architecture for High End Computing; Xian-He Sun, Illinois Institute of Technology
- Cross-Layer Exploration of Non-Volatile Solid-State Memories to Achieve Effective I/O Stack for High-Performance Computing Systems; Tong Zhang, Rensselaer Polytech
- Streamlining High-End Computing with Software Persistent Memory; Raju Rangaswami, Florida International University
- Programming Models and Storage System for High Performance Computation with Many-Core Processors; Jack Dennis, MIT
- RUI: Automatic Identification of I/O Bottleneck and Run-time Optimization for Cluster Virtualization; Xubin He, Tennessee Technological University
- Balanced Scalable Architecture for Data-Intensive Supercomputing; Alexander Szalay, John Hopkins (Tuesday)

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							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 HEC area. Includes power consumption. Traditional block-based solutions ready for commercialization. Alternative interfaces not yet well explored.
	Urgaonkar							
	PNNL							

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"

This Gap Area has been integrated into "Understanding file system abstractions – Scalable file system architectures" in Next Generation I/O Arch.

2008 Next Generation I/O Architectures Gap Area

Area	Researcher	FY 07	FY 08	FY 09	FY 10	FY 11	FY 12	Rankings
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 storage.
	Dickens							
	Gao							
	Sun							
	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

Here is the URL to the abstracts of the previous HECURA projects: http://institutes.lanl.gov/hec-fsio/rande/

Current R&D Gaps (Notes)

- Understanding file system abstractions – file system architectures
  - Living with POSIX, replacement ? What might be new ?
- Understanding file system abstractions – naming and organization
  - 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”
- Self-assembling, Self-reconfiguration, Self-healing storage components
  - Power awareness (UBIQ)(efficient)
- Architectures using 10^6 storage components
  - This Gap Area has been integrated into “Understanding file system abstractions – Scalable file system architectures” in Next Generation I/O Arch.
- Hybrid architectures leveraging emerging storage technologies
  - Hetero/hybrid computing impact on I/O
  - Inverse

Current R&D Gaps (Notes)

- HEC systems with multimillion way parallelism doing small I/O operations
  - Perf requirements, customer reqs , emerging reqs, not know what data req are needed, Can HEC help DB folks or the inverse ? How to hide ? New technology (Google) , KDD, I/O Book ?, Optimize I/O techniques ? Optimize whole data path.
  - Get rid of storage. (Karsten) in situ analysis, better support of Viz data
  - Why are we still stuck on “blocks” smarter protocols (OBSD)++
  - Richer vocab for storage seperation
  - GP + Special Purpose
  - Comm I/O breakout in with this
  - Stream processing

Next Gen I/O Breakout Results

1. Understanding file system abstractions – file system architectures
   1. 3 3 3 3 3 3 3 3 3 2 2 2 1 1 1 = 36 (Rank #2)
2. Understanding file system abstractions – naming and organization

   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”

   1. 3 3 3 3 3 3 2 2 2 2 2 2 1 1 1 1 = 34 (Rank #4)
3. Self-assembling, Self-reconfiguration, Self-healing storage components
   1. 3 3 3 3 3 2 2 2 2 2 1 1 1 1 1 1 1 1 1 1 = 35 (Rank #3)
4. Architectures using 10^6 storage components
   1. 3 2 2 2 2 2 1 1 1 1 1 1 1 = 20 (Rank #6)

   – This Gap Area has been integrated into “Understanding file system abstractions – Scalable file system architectures” in Next Generation I/O Arch.

Next Gen I/O Breakout Results

1. Hybrid architectures leveraging emerging storage technologies
    1. 3 3 3 3 3 3 3 3 3 3 3 3 3 3 2 2 2 2 2 2 1 1 1 1 1 1 1 = 61 (Rank #1)
2. HEC systems with multimillion way parallelism doing small I/O operations
    1. 3 3 3 3 3 2 2 2 2 2 2 2 1 1 1 1 = 33 (Rank #5)

Potential New R&D Gaps

- Several Simulator requests (what exists ?)
- Power Aware (add to #3 ?)
- POSIX Compliance (add to #1 ?)
- End-to-End
  - Analysis (KDD)
  - Prediction / Optimization
  - Tools
  - Instrumentation
- Data Intensive Computing
- Database technologies (Hide from user)
- What about Next Gen Networks ?

- None of these were officially added

Current R&D Gaps (Notes)

- Understanding file system abstractions – file system architectures
  - There was a lot of discussion concerning living with POSIX. Is there any potential replacement ? How would it be different ? How long would it take to get accepted ? What else might be new ?

- Understanding file system abstractions – naming and organization
  - 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”
  - Some of the discussions concerning number of objects in the filesystem were related to this, but it seems to be another accepted issue. Why are we still stuck on “blocks” smarter protocols (OBSD)++

- Self-assembling, Self-reconfiguration, Self-healing storage components
  - Quite a few people have acknowledged the need for analysis of power aware/efficient systems. This will start impacting even smaller installations. There is some research going on in these areas. The DARPA UHPC (PF in a Rack) is funding some basic research in this area. How can we make what we have more efficient ? Any near term components potentially helpful ?

Current R&D Gaps (Notes)

- Architectures using 10^6 storage components
  - This Gap Area has been integrated into “Understanding file system abstractions – Scalable file system architectures” in Next Generation I/O Arch.
  - This seemed to be an acknowledged problem as we all see that in order to get the level of parallelism required for the HEC/HPC systems, we will have to add components. Even having large quantities of FLASH/PCM, does not remove the quantity issues. Disks will only spin so fast, and Flash/PCM/3D technologies will remain expensive for the next 5 years or so.

- Hybrid architectures leveraging emerging storage technologies
  - Heterogeneous / hybrid computing impact on I/O
  - I/O impact on heterogeneous / hybrid computing
  - As we can see by the number of votes, this was a fairly well discussed and important topic. As we return to hybrid / heterogeneous computing, we will have to answer all of the impact questions.
  - Why are we still stuck on “blocks” smarter protocols (OBSD)++
  - How will stream processing potentially impact I/O (Seismic++)

Current R&D Gaps (Notes)

- HEC systems with multimillion way parallelism doing small I/O operations
  - This was another topic that was accepted as an issue.

- The following items were also discussed (not sure where they fit)
  - Requirements
    - Performance
    - Customer
    - Emerging
    - We do not know what data requirements are needed.
  - New technology (the Google suite)
  - I/O Book ?
  - Get rid of storage. (Karsten)
  - Richer vocabulary for storage separation