Research Introduction

August 21 2006

Lee Ward

Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company,
for the United States Department of Energy's National Nuclear Security Administration
under contract DE-AC04-94AL85000.

HECIWG Categories of Needed Focus

- **Metadata**
  - Evolution **- Scalability, Extensibility, Archival considerations, ACL's**
  - Revolution **- Scalability, Extensibility, Name Spaces, Archival considerations, Hybrid devices**
- **Measurement and Understanding**
  - Evolution **- Understanding layering contribution, End to end benchmarking and tracing, Visualization**
  - Revolution **- End to end modeling and simulation, VM as tool for large scale simulation**
- **QOS**
  - Evolution **- Determinism with multiple applications and priorities**
  - Revolution **- Adaptive, End to end QOS**
- **Security**
  - Evolution **- Usability, Long term key management, Distributed authentication for file systems, Dealing with security overhead**
  - Revolution **- Novel security as related to file systems and I/O, Novel encryption at rest over time, Key Management, ACL's End to End encryption API**

HECIWG Categories of Needed Focus (cont)

- **Next generation I/O architectures**
  - Evolutionary **- POSIX, Archive considerations, Access aware interfaces, HEC considerations, Small/unaligned I/O, Mixed large and small I/O, Collaborative caching, Impedance matching**
  - Revolutionary **- Redistribution of intelligence and what abstractions we need, Adaptive/reconfigurable stack (application specific perhaps), User space component considerations, File systems that are semantically aware of the data. Novel devices/hybrid devices exploitation**
- **File System related communications and protocols**
  - Evolution **- Exploitation of RDMA/one sided etc., OBSD (transports, security, extensions, applications), NFSv4 (extensions and applications), pNFS (proof of concept, extensions, applications)**
  - Revolution **- Server to server communication**
- **Management and RAS**
  - Evolution - **Reliability and availability at scale end to end and its overhead, Management scaling, Continuous versioning, Power management**
  - Revolution **- Autonomics (adaptive/self healing/predictive), VM as a RAS enabler, Novel devices as enabler**
- **Archive (as it relates to I/O and File Systems)**
  - **Content addressable, Deep archive on disk, Object archives/parallel archives, Scheduling movement/ILM**
- **Assisting research**
  - **Testbeds, Clearing houses (providing traces, reliability info, etc.), Support growth of I/O students**

Sandia National Laboratories

Funding Update

- NSF $11M plus other contributions gave us HECURA call for R&D in FSIO
  - 19 of 62 submissions funded
- DOE SciDAC2 included FSIO
  - 2 FSIO proposals funded
- Thanks
  - NSF; Almadena Chtchelkanova
  - DOE; Fred Johnson, Thuc Hoang, Robert Meisner
  - DARPA; John Grosh (now at LLNL)
  - LANL; Gary Grider

Sandia National Laboratories

HECURA

- 0621393/0621538, Collaborative Research: Petascale I/O for High End Computing; Maccabe, Arthur B/ Schwann, Karsten; UNM/ Georgia Tech Research Corporation – GA Inst of Tech; FSIO
  - Metadata, Next generation I/O architectures
    - Higher level I/O abstractions via I/O graphs
    - Flexible metadata management by metabots
    - Rich metadata
    - Lightweight file systems
- 0621439/0621425, Collaborative Research: Techniques for Streaming File Systems and Databases; Bender, Michael A/Farach-Colton, Martin; SUNY at Stony Brook/Rutgers Univ New Brunswick; FSIO
  - Metadata, Next generation I/O architectures, and File System and related Communication Protocols
    - Streaming B-trees and variants for efficient data layout on disk, databases
- 0621484, Applicability of Object-Based Storage Devices in Parallel File Systems; Wyckoff, Pete; Ohio State Univ Research Foundation; FSIO
  - Metadata, File System and related Communication Protocols, and Next generation IO architectures
    - Objects trade-offs and attributes
    - Applicability of OSDs in parallel file systems
    - Metadata

Sandia National Laboratories

HECURA (Continued)

- **0621526/0621493, Collaborative Research: SAM^2 Toolkit: Scalable and Adaptive Metadata Management for High-End Computing; Jiang, Hong/Zhu, Yifeng; Univ of Nebraska-Lincoln/Univ of Maine; SSDR**
  - **Metadata**
    - **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**
- **0621441, Improving Scalability in Parallel File Systems for High End Computing; Ligon, Walter B; Clemson Univ; FSIO**
  - **Metadata, Management and RAS, and Next generation I/O architectures**
    - **Active caching and buffering**
    - **Server to server and client to client communication**
    - **Autonomics**
    - **Scalable metadata**
    - **Small unaligned data access**
    - **Reliability through redundancy**

Sandia National Laboratories

HECURA (Continued)

- **0621435, HECURA: The Server-Push I/O Architecture for High End Computing; Sun, Xian-He; Illinois Inst of Tech; FSIO**
  - **File systems and related Communication Protocols and Next generation I/O architectures**
    - **Server side push**
    - **Collective I/O aware access patter prediction**
- **0621443/0621402, Collaborative research: Scalable I/O Middleware and File System Optimizations for High-Performance Computing; Choudhary, Alok N/Kandemir, Mahmut T; Northwestern Univ/Pennsylvania State Univ University Park; FSIO**
  - **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**

HECURA (Continued)

- **0621470/0621427/0621478, Collaborative Research: Application-adaptive I/O Stack for Data-intensive Scientific Computing; Ma, Xiaosong/Sivasubramaniam, Anand/Zhou, Yuanyuan; North Carolina State Univ/Pennsylvania State Univ University Park/Univ of Illinois at Urbana-Champaign, FSIO**
  - **Next Generation I/O Architectures and QOS**
    - **Parallel Adaptive I/P (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**
- **0621448, Active Storage Networks for High End Computing; Chandy, John A; Univ of Connecticut; SSDR**
  - **Next Generation I/O Architectures**
    - **Active storage networks-computation at the networks such as reductions and transformations**

Sandia National Laboratories

HECURA (Continued)

- **0621410, Active Data Systems; Reddy, A.L. Narasimha; Texas Engineering Experiment Station; SSDR**
  - **Next generation I/O architectures and QoS**
    - **Broadening active disk applicability by examining running multiple applications at disk concurrently**
    - **Scheduling**
    - **Security**
    - **Sharing**
- **0621512, Quality of Service Guarantee for Scalable Parallel Storage Systems; Chiueh, Tzi-Cker; SUNY at Stony Brook; SSDR**
  - **Next generation I/O architectures, Measurement and Understanding, and QoS**
    - **Platypus – storage system**
    - **QoS trace replay**
    - **Bandwidth guarantees**
    - **Prefetching using decoupled architecture by extracting a prefetch thread from the computation thread**

Sandia National Laboratories

HECURA (Continued)

- **0621472, Concurrent I/O Management for Cluster-based Parallel Storages; Shen, Kai; Univ of Rochester; FSIO**
  - **Next generation I/O architectures**
    - **Concurrent I/O workload**
    - **Disk seek/spin reduction by prefetching and anticipatory I/O scheduling**
    - **Server level coscheduling**
    - **Load adaptive parallel data aggregation**
- **0621457, Performance Models and Systems Optimization for Disk-Bound Applications; Thottethodi, Mithuna S; Purdue Univ; SSDR**
  - **Next generation I/O architectures, Measurement and Understanding, and File System and related Communications Protocols**
    - **Disk array modeling/algorithms**
    - **Network aware placement and migration**
    - **Power and thermal optimization via entropy-aware disk caching**

Sandia National Laboratories

HECURA (Continued)

- **0621429, Exploiting Asymmetry in Performance and Security Requirements for I/O in High-end Computing; Sivasubramaniam, Anand; Pennsylvania State Univ University Park; FSIO**
  - **Security**
    - **Data Vault – security**
    - **Tunable tradeoff between security and performance for site-specific policies**
    - **Visualization dashboard**
- **0621462, Integrated Infrastructure for Secure and Efficient Long-Term Data Management; Odlyzko, Andrew; Univ of Minnesota-Twin Cities; SSDR**
  - **Security, Archive**
    - **Security**
    - **Hierarchical cluster-based archive**
    - **Long-term key management**
- **0621487, NSF 06-503: Formal Failure Analysis for Storage Systems; Arpaci-Dusseau, Remzi H; Univ of Wisconsin-Madison; SSDR**
  - **Management and RAS and Measurement and Understanding**
    - **Formal analysis of failures with Wisconson's Program Analysis of Storage Systems (PASS) program**

Sandia National Laboratories

HECURA (Continued)

- **0621508, Toward Automated Problem Analysis of Large Scale Storage Systems; Narasimhan, Priya; Carnegie-Mellon Univ; SSDR**
  - **Measurement and Understanding and Management and RAS**
    - **Continuous performance and anomaly tracing**
    - **Auto blame assignment and performance diagnosis**
    - **Automated analysis of failure and performance degradation**
- **0621463, File System Tracing, Replaying, Profiling, and Analysis on HEC Systems; Zadok, Erez; SUNY at Stony Brook; FSIO**
  - **Measurement and Understanding, Next generation I/O architectures, and File System and related Communication Protocols**
    - **Visualization**
    - **Tracing and replaying file system activity**
- **0621534, End-to-End Performance Management for Large Distributed Storage; Brandt, Scott A; UC-Santa Cruz; SSDR**
  - **QoS**
    - **QoS server side**
    - **Server I/O scheduling**
    - **Server and client cache management**
    - **Client-to-server network flow control**
    - **Client-to-server connection management**

Sandia National Laboratories

SciDAC2

- **PetaScale Data Storage Institute**
- **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, development and performance tools.**
- **Lead PI: Garth Gibson CMU**
  - CMU, NERSC, PNNL, ORNL, SNL, LANL, U. Mich., UCSC

Outreach	**Apps Performance**	**Storage Dependability**
Protocol/API Extensions	**Novel Mechanisms**	**Storage Automation**

Sandia National Laboratories

SciDAC2 (continued)

- **Scientific Data Management Center for Enabling Technologies (CET)**
- **PI: Arie Shoshani (LBNL)**
  - **ANL, LBNL, LLNL, ORNL, UofUtah, NCSU, NWU, UC Davis, SDSC**
- **The goal of the center is to provide software solutions to problems found in end-to-end data management for computational science. - efficient interaction with storage devices for parallel applications**
  - **Queries, feature discovery, and statistical analysis of very large datasets**
  - **Automating the process of application execution, dataset collection and storage, data post-processing, and analysis and visualization**

Sandia National Laboratories

Summary

- **We heard you**
  - **HECURA IWG initiates FSIO thrust**
    - **2005 workshop to focus topics**
    - **This workshop**
  - **NSF and others provide $11M funding**
  - **DOE responds with SciDAC2**

Sandia National Laboratories