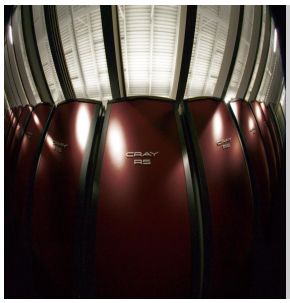

Sandia National Laboratories

Exceptional service in the national interest

Storage Systems and Input/Output to Support Extreme Scale Science Workshop Outcomes

Jay Lofstead

Center for Computing Research
Sandia National Laboratories
Albuquerque, NM, USA
gflofst@sandia.gov

HPC I/O in the Data Center @ ISC 2015

July 16, 2015

U.S. DEPARTMENT OF ENERGY

NNSA National Nuclear Security Administration

Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

Workshop Overview

Sandia National Laboratories

December 8-11, 2014

Gather input from three groups:

- Science/Mission
- Crosscutting Computer Science
- Computer Scientists

Report basis for funding offering that closed on Monday

9 specific findings and 4 research areas identified

- not comprehensive

1. In Situ Data Analytics

Sandia National Laboratories

Factors forcing alternative approaches (maximize value)

- Data too big to write to disk
  - XGC1 (100 PB wanted, 1 PB maximum effective)
- Generated too fast
  - QMC (2 TB every 10 seconds)
- Need to incorporate UQ
- Validate simulation against EOD

1. In Situ Data Analytics

Sandia National Laboratories

New Opportunities

- Growing maturity of ADIOS, GLEAN, and VTK/ParaView, LibSim
  - No standards yet and no strategy seems best
  - In situ, In transit, on node, on machine, off machine
- NVRAM
  - Fewer additional resources needed to support operations
  - Resource management is a BIG unknown

2. Solid State Storage Complicates

Sandia National Laboratories

- Various placement options each have an optimal use case
  - On memory bus
  - on IO bus
  - in compute area
  - burst buffer (PFS node in compute area cache)
  - in PFS (in PFS cache)
- All other system components may be affected by choices

How to use these devices effectively in ANY location still not understood

2. Solid State Storage Complicates

Sandia National Laboratories

Challenges

- Quality of service, scheduling, pre-emption
- RMA possibilities and potentially injected interference
- Offloading IO tasks to separate machine area for asynchronous processing
- Security, sharing among a job set, data migration

3. Unified Storage View Important

Sandia National Laboratories

- What is memory and what is storage?
  - Memory has a get/put interface
  - Storage is block based? Requires different kind of operation?
- What level of direct device access is necessary?
- What can be abstracted away and to what advantage?
- Do we bring tape or other archives back into scratch metadata?

3. Unified Storage View Important

Sandia National Laboratories

- Portability? (Summit/Sierra vs. Trinity)
- Do the abstractions help or hurt big data applications?
- Transactional access support for workflows?
- Auto or manual migration?

4. Metadata and Public Data Access

Sandia National Laboratories

- POSIX interface already problematic
  - Consistency requirements force serial operations
- Validation drives additional provenance annotation
  - Lots of attempts, nothing works well enough to be a standard
- Additional metadata to pull big data through a thin pipe
- Metadata consistency still required (with link web)

4. Metadata and Public Data Access

Sandia National Laboratories

- Metadata support for In situ or workflow operations
- User-defined extensions? Any limits on size or types? What is the scope?
- Still need to maintain security envelope
- What if data migrates?
- How to make data (and metadata) available for public review?

5. New Programming Models

Sandia National Laboratories

- Messaging-based bulk synchronous works, mostly
- Accelerators complicate frequency and I/O operations
- Task-based models quite hot (Legion, Charm++, Uintah, …)
  - Writing to storage (synchronous or not) for a preserved output
  - Data warehouse

6. Data Abstractions/Workflows

Sandia National Laboratories

- Intermediate data storage cannot go to disk anymore
  - Too thin a pipe to force data through
- NVM/SSD can offer disk-like capabilities to support workflows
  - Limited capacities potentially problematic
- Other models more interesting
  - Split nodes hosting simulation and analysis communicate with shared memory
  - Key/Value stores may work better (adopt big data pushing overhead on applications to aid scaling) [Also object store devices on network]
  - Pub/sub between workflow components

6. Data Abstractions/IO Middleware

Sandia National Laboratories

- Staging and in transit processing insufficient
  - Generally cannot use disk for intermediate data
- NVM device management exposed to middleware, but how to expose to applications?
- How to connect different users/jobs for workflows?
- Data models for selecting data sent through
  - Does it fall back to disk if something fails?
  - How do we integrate this in should it happen?

7. Storage Resources Second Class

Sandia National Laboratories

- Current “charged” resource, if any is CPU
  - Storage limited by quota at most
- New devices in new locations with limited capacity
  - Shared, but how allocated and shared?
- Scheduling/charging approach
  - Bandwidth or capacity?
  - How long allowed to stay and at what cost?
  - Limited write endurance as a factor?

8. Storage Profiling Difficult at Best

Sandia National Laboratories

- Darshan + other tools (Vampir, SIOX, CODES, HECIOS) current best
  - Cannot profile when applications use direct POSIX calls (25% on Mira)
  - How to map to storage hierarchy unknown
- IOR, MDTest difficult due to automatic caching
  - TACC paper at SC 2014 came to the wrong conclusion
  - HDF5 has special code if memory available to accelerate processing
- IO MiniApps not necessarily robustly built or still representative
  - Must emulate memory footprint and communication interference to get an accurate picture

9. “Support Ecosystem” Needed

Sandia National Laboratories

- Consider interconnect contention
- Memory issues (too much/not enough free)
- Need robust hardware test platforms
- Need end-to-end data logs for replay
  - Include sufficient info to configure system properly
- Tutorials/workshops to learn how to use robust toolsets

Priority 1: Hierarchy Management

Sandia National Laboratories

- How to manage distributed NVM integrated with RAM, disk, and tape?
- How do we integrate this into the application workflow?
- How do we integrate this into machine management?
- How do we share among machine users (including security)?

Priority 2: Rich Metadata Critical

Sandia National Laboratories

- Traditional POSIX-style Metadata still needed
- Rich, user custom metadata to support applications
- Data provenance should be integrated

Priority 3: Support for EOD

Sandia National Laboratories

- Support simulation workloads at maximum IO bandwidth possible with little variability
- Incorporate including EOD for simulation validation/ enhancement
- Incorporate UQ
- Develop middleware tools to support new workloads

Priority 4: System Characterization

Sandia National Laboratories

- Expand to understand all
- Learn how to expand to new devices
  - At least on a case-by-case basis
- Share traces and workload generators to bolster community

Not Covered in Grant Proposal Call

Sandia National Laboratories

- Data center shared storage arrays
  - Machines have private scratch so deemed unimportant
- Developing new hardware
  - Purely software infrastructure for available/future hardware
- Active Storage
  - Not currently viable
  - In situ/in transit processing replacing this area

Other Related Projects

Sandia National Laboratories

- Exascale OS/Runtime
  - Hobbes and Argo
- Data Management
  - Largely workflow/in situ/in transit processing related
- New File Systems
  - Sirocco (SNL) and Triton (ANL)
  - https://institute.lanl.gov/hec-fsio/workshops/2011/2011/Talks/lee.pdf
  - Soon: http://www.cs.sandia.gov/Scalable_IO/sirocco
- Object Stores and Key Value Stores (and devices)
  - Kelpie, Hop (and others)