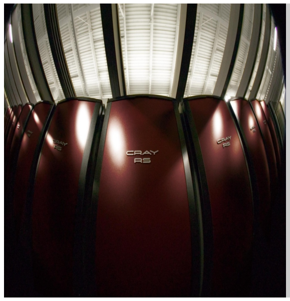

Sandia National Laboratories

Exceptional service in the national interest

NVM Impacts on Storage Design and Management

Jay Lofstead

Scalable System Software
Sandia National Laboratories
Albuquerque, NM, USA
gflofst@sandia.gov

HPC-IODC Discussion Round

June 23, 2016

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.

Problem

- New storage locations
  - On node (DRAM, NVM in various forms and locations)
  - In compute area (burst buffers)
  - Platform parallel file system (local scratch)
  - Data center parallel file system (global scratch)
  - Other data sources/types (EOD or other data stores)

- New storage technologies
  - Flash, PCM, NVM

- Latency hiding storage stacks
  - Burst buffers or IO-forwarding nodes shift latency visibility—but not fully

Memory vs. Storage

- Arbitrary distinction
  - used directly for compute = memory
  - the rest = storage

- Access approach
  - Get/put, read/write, mmap, or something else?

- What about moving data off the platform?
  - Within data center
  - To archive (local or external)

Questions?

- How do we accurately provision (capacity and location) NVM?
- What abstraction is on top of which part of the stack?
- How are these resources managed (allocations per user/job/use intensity)?

Discussion Points Raised

- Use node local storage for job swap ("pre-emption") for more urgent computation
  - What about security?
  - Need NVDIMMs at least as large as RAM
- Desired use of NVM as both slow memory and fast storage
- Should we have explicit or implicit usage?
  - How to guarantee bandwidth?
  - Pmem.io says use one of x that works for scenario
  - Memkind library (HBM and NVM considered)
- No one believes burst buffers will solve NetCDF performance problems
- Need NVM per node to avoid a few users slamming all resources/ interference effects
- Software interface still unclear because hardware specs are still ill-defined