I/O, Networking, and Storage Working Group

Lee Ward, Luis Martinez; SNL

Rob Ross; ANL

Dick Watson, Chris Morrone, Mark Gary; LLNL

Hal Armstrong, John Bent; LANL

Matt Crawford, FNAL

Working Group Description

- Scope (i.e., major SW stack responsibilities)
  - I/O and file system within machine, archive near the machine, as well as the networking infrastructure used to move data within the NNSA complex.
  - Any portion of the I/O software stack, from the application libraries and supporting middleware to the support file systems
    - Working with hardware storage vendors in support of our mission.
  - The networking portion of our domain is limited to the network outside the compute resource (i.e., connecting to external storage, archive, WAN)
    - We work with the vendors to suggest improvements
- Key dependencies on other working groups
  - VDA: overlap in scope at middleware
  - System software: scheduler
  - Apps: of course
  - Hardware: especially as relates to solid state in the compute system

Exascale Challenges
(includes both HW and SW infrastructure)

- Bulk data movement
  - Eliminating data movement to external storage when possible (e.g., in-system alternatives to checkpointing)
  - Minimizing data movement to external storage when possible
    - Data summarization, differential checkpointing, compression
  - Eliminating unnecessary synchronization in movement to external storage
    - Involves “burst buffers” and other uses of nonvolatile storage in the system
  - Archive as a component of the bulk data movement and storage system
- Input data staging
  - Pre-staging of input datasets onto storage in the system
  - Includes dynamic loading of libraries
- Data management
  - Very large numbers of files (i.e., namespace)
  - Metadata and provenance capture
  - End-to-end data integrity
- Data abstractions
  - Future interfaces and semantics of storage, coordination and synchronization
  - Libraries that provide high productivity interfaces to storage

Assumptions

- Community will continue to depend on and invest in vendor-supplied parallel file systems
- Rotating media capacity and performance will continue on the same curve as from the last five years
- Machine interconnect speeds will reach current expectations
- External to the system, enterprise network core will evolve to a technology more like our machine interconnects

Critical Technologies and Candidate Software

Technology	Candidate Software	Notes
I/O Forwarding, Async I/O, and Aggregation	IOFSL, DVS, ciod, I/O Delegation	Will require porting to hardware platform
File System Transformation	PLFS, ROMIO/LogFS, ADIOS, Giga+	Probably best incorporated into I/O forwarding
Burst Buffer Management	Zest	Probably best incorporated into I/O forwarding
In-System Checkpoint	SCR	
Revolutionary Storage Prototype		
High-level I/O Interface/Library	HDF5, PnetCDF, ADIOS, SciDB, Damasc	Some dependence on MPI-IO (vendor provided?)
Archive	HPSS	
Networking		
End-to-End Integrity		Probably best incorporated into high-level library

Together form solution to bulk data movement issues.

Note: This plan assumes that legacy file systems, archive, networking development is funded elsewhere.

I/O Forwarding, Async I/O, and Aggregation

- I/O middleware to meld legacy file system to exascale MPP
- Initial next steps
  - IOFSL exists, aggregation capabilities implemented, runs on Linux clusters, Cray XT series, and IBM Blue Gene/P
  - Security design complete but needs implementation
  - Further QA needed, ports
- Timeline (Phase I, II or III; Evolutionary or Revolutionary)
  - Phase I/II
- Required Funding: $800K/yr, $4M total
- Required Partnership: ASCR/NNSA IOFSL project
- Risks:

File System Transformation

- I/O middleware supporting huge directories and N:1 performance improvements
- Initial next steps
  - Integrating PLFS with SCR now
  - Planning needed for integrating Giga+ capabilities
- Timeline (Phase I, II or III; Evolutionary or Revolutionary)
  - Prototype in phase I, product in phase II
- Required Funding: $400K/yr, $2M total
- Required Partnership: CMU (Giga+)
- Risks

Burst Buffer Management

- I/O middleware for driving data movement between in-system storage resources and legacy file system
- Initial next steps
  - Evaluate in-progress IOFSL option for suitability
  - Evaluate LANL FTA (data mover) option for suitability
  - Evaluate Zest option for suitability
- Timeline (Phase I, II or III; Evolutionary or Revolutionary)
  - Prototype in phase I, product in phase II
- Required Funding: $400K/yr, $5M total
- Required Partnership: TBD
- Risks

In-System Checkpoint

- I/O middleware to leverage near-compute storage as an alternative storage location for checkpoint data
- Initial next steps
  - PLFS and SCR are continually improving and evolving
  - Integrate PLFS with SCR (in progress)
- Timeline (Phase I, II or III; Evolutionary or Revolutionary)
  - Phase II
- Required Funding: $600K/yr, $3M total
- Required Partnership: (internal to NNSA)
- Risks

Exascale Storage System Prototype

- Integrated storage solution for exascale; this is the revolutionary effort
- Initial next steps
  - Candidate HW/SW architecture design near complete
  - Core infrastructure in progress
- Timeline; Revolutionary
  - Phase II/III
- Required Funding: $5M/yr, $25M total
  - Assuming partnership with ASCR
- Required Partnership: ASCR
- Risks

High-level I/O Interface/Library

- Libraries providing scientific data model support and self-describing data formats atop legacy file systems
- Initial next steps
  - Identify specific scalability and performance issues in current product
  - Negotiate terms of NRE funding agreement
- Timeline (Phase I, II or III; Evolutionary or Revolutionary)
  - Prototype in phase I, product in phase II
- Required Funding: $800K/yr, $4M total
- Required Partnership: library vendor
- Risks

Archive

- HPSS
- Initial next steps
  - Evaluate HPSS design in light of updated exascale storage requirements
  - Consider how migration capabilities and FS connectivity in HPSS address exascale needs
  - Re-evaluate current funding model for required enhancements (e.g., clustered metadata, storage system management)
- Timeline (Phase I, II or III; Evolutionary or Revolutionary)
  - Prototype in phase I, product in phase II
- Required Funding: $1M/yr, $5M total
- Required Partnership: HPSS Collaboration (ASC, ASCR, IBM)
- Risks

Networking

- External network connectivity, including machine room, long haul networking, encryptors for long haul networking
- Initial next steps
  - Anticipate vendor availability of bandwidth at layers 1-3
  - Socialize need for high speed encryptors
- Timeline (Phase I, II or III; Evolutionary or Revolutionary)
  - Encryptors: phase I/II
- Required Funding: encryptors: $2M/yr, $12M total
  - Assumes funding partner
- Required Partnership: DoD MD procurement office
- Risks
  - If we cannot motivate external demand, encryption technology will not be developed in required timeframe

End-to-End Integrity

- Ability to verify correctness of stored data
- Initial next steps
  - Communicate with storage hardware vendors motivating support of standards
  - Interact with OS/runtime team to design OS-level solution
- Timeline (Phase I, II or III; Evolutionary or Revolutionary)
  - Prototype in phase I, product in phase II
- Required Funding: $800K/yr, $4M total
- Required Partnership: Vendor OS I/O team
- Risks

Recommended Co-Design Strategy

- Critical steps/activities
  - Further discussion with hardware, application teams on options for NVRAM integration into system
  - Further discussion with tools, runtime, and analysis teams on infrastructure for data movement within the system
- Working with vendors
  - I'm sorry, I'm under NDA.
  - Discussed earlier.
- Role of skeleton/compact apps
  - Important component of characterizing application requirements/patterns
  - Need these in a timely manner to have greatest impact
- Concerns/suggestions
  - Need to drive towards quantification of key system attributes.
  - See Working with vendors.

Big Picture Issues

- Testbeds
  - This is a cross-cutting issue
  - Some testbeds already exist, not clear that adequate cycles are available
  - Will need another generation of testbeds before we reach exascale, or regular update of existing testbeds
- Simulators
  - Option for scalability testing, algorithmic testing
- Remaining gaps
  - This is a top-down process, we're working our way down, haven't reached understanding on the integrated problem, yet.

Extra Material

(on network)

Luis: what kind of network I/O is really needed for exascale? stalled at 100Gig(abit?) at the moment

Mark Gary: network doesn't raise to the level of challenge as these other things.

- priced yourself out of moving across WAN

- don't move the data

note: there will be communication inside the system related to I/O that is separate from the app. communication traffic

- 200 PBytes of internal traffic (?)

(on SSD)

- integrity checking WRT SSD storage

possibility of data rearrangement during movement to SSD

similarity to disk-tape relationship (with SSD-disk)

soft quotas on archive to limit capacity growth

- - trying to limit unrestrained growth
- - there will still be tape

Q: do you need disk at all?

- - can you go straight from SSD to tape in some cases?

- discussion in breakout of tools for looking at patterns of I/O related to what should go on SSD…

(on checkpointing/bulk output)

- Q: what fraction of the data stored in the system needs to make it to external storage?
- 32-64PB of RAM in the system
  - assume 50% of data in memory is sufficient to restart
  - roughly 3x this in NVRAM
- Q: how often do apps write?
  - Based on 1/day MTTI, checkpoint time, restart time, …
  - Based on analysis needs …
- 4.5TB/sec to external storage (under the following assumptions)
  - checkpoint every hour
  - keep every other one
  - 32PB of data

(on compression, summarization, etc.)

- discussion of summarization vs. compression as mechanisms for dealing with the data requirements of discovery science runs (i.e., where the application teams don't know what they're looking for)

(with tools)

- help developers build, debug, performance tune their applications
- secondarily (?), look into the system software, etc.
- note: scheduler is system software
- tools for I/O stack
  - development tools
  - bottlenecks, fault identification
  - performance tools
- network
  - bottlenecks, fault identification
  - counters, logical comm. to physical links
  - understanding QoS, flow control
- infrastructure for data capture as a service to be used by the IONS team
  - do we go as far as real-time access?
- statistical capture
- tagging of operations to correlate statistical capture (in part), provenance
- idea of interfaces for querying I/O libraries for information, file/storage systems, etc. (e.g., SNMP for these things)
- idea of anomaly detection
- I/O kernels
- and also they generate lots of files at the moment, captured in a thread-centric way, analyzed in a time-centric way, driver for in situ analysis of their data as well

(with analysis)

- their scope: scientific vis, scientific analysis, “data management”
  - mgmt: tracking, organizing, enhancing the use of…
- not a clean line
  - indexing, query
  - in-system data movement
  - metadata, provenance – america competes, data preservation, what should be done first?
  - compression (esp. lossy)
  - infrastructure for applying transformations
- discussion of high-level I/O libraries, somehow that there is an issue here between the two groups (?)
- their challenges:
  - in situ is a mandatory component of analysis
  - quantifiable data reduction techniques
  - new vis and analysis approaches for mult-physics, multi-scale, etc.
  - efficient algorithms for the swimlanes
- Lee: #1 problem today is checkpoint, …
  - Data triage doesn’t help with checkpoint
- from their document:
  - scalable parallel, knowledge infrastructure (?)
- hardware that we can share
- look into experimentation on existing storage products (e.g., LWFS, PVFS, …)

(with HW)

- some apps write more vis data than they have to write to checkpoint
  - jackie's code, as an example, writing modest amounts frequently (1.9MBs/core) every 12 minutes
- what kind of network connectivity do we need to NVRAM devices? do we need gateway nodes?
  - NVRAM mixed in with compute nodes, maybe one per rack, lost of SSD
  - also attached to external network, or maybe not?
    - different use cases change where you'd like to put things
  - security, simple consistency, space partitioning
  - is it a cache or is it a tier?
- back-end, external storage is shared to other systems
- MTTI is a day for the machine
  - some concern that we won't hit that
  - file system specifically called out
- network topology discussion
  - PERCS network…

(with HW)

- SCR discussion, noise
  - simple QoS on high-speed interconnect (nice level/priority, virtual lanes)
  - if you aren't going to drain, you need some other way to get at it. distributed access.
  - discussion of percentages of memory that is really needed, trends over time, …
- SSD API
  - managing as an I/O device?
  - FS as a model that people are accustomed to
  - managing state variables as a possible additional use of SSD, convenience for apps (?)
    - Vetter office of science project here
    - I think we're talking about on-node NVRAM here specifically?
    - IONS is not generating a requirement for on-node NVRAM
- power
  - more discussion that storage power isn't really a big deal
  - future enterprise networks can power down, but they aren't a big component of power
  - currently .2-.3 MW for 10s of PBs today
  - "The motors in our air handlers use more power than our external network switches." – Park Fields, 03/23/2011
  - TODO: POKE EVERYONE TO GET THE DATA ON REAL POWER UTILIZATION.
- Who takes responsibility for external network – storage group?

The Tape Story

- 2009 workshop, HPSS and exascale
  - expect to receive report, saying it can scale
  - summary of report, support needed in three areas: coupling to SDM (e.g., metadata services), multiple metadata servers (happening already out of core support, could use additional funding), storage system management (major work needed)
  - huge challenges not seen here, because evolutionary effort is underway
- pre-staging has requirements on/connections to archive
- no obvious need for revolutionary tape story at this time, perhaps after (or as part of) revolutionary storage plan

misc. discussion from Wed. morning

- clustered metadata in Lustre is considered essential
- networking: working with smaller networking companies for low latency, high bandwidth networks, but there is risk, may need space for special purpose functionality
- machine room networking
  - some concern about availability/cost of 100Gbit networking
  - InfiniBand promising, but want to be agnostic
  - anticipate need for legacy applications, data movement, vis. to require rethinks towards single-sided and RDMA
    - CROSS CUTTING
  - industry will push hardware, some concern that traffic patterns will differ enough that industry solutions will not be applicable to exascale, may need additional software work to work around
- data aware scheduler, uncertainty in the priority there
  - depends on input data set size, frequency of job start
  - driven by application needs; scheduling people might know the answer here.

Path Forward

- Proposed technology
- Initial next steps
- Timeline (Phase I, II or III; Evolutionary or Revolutionary)
- Required Funding ($-$$$$$: $=10M)
- Required Partnership
- Risks

Repeat this slide for each technology proposal - highest priority only & maximum of 5-7 technologies

Critical Technologies and Candidate Software

Technology	Candidate Software	Notes
I/O Forwarding, Async I/O, and Aggregation	IOFSL, DVS, ciod, I/O Delegation	Will require porting to hardware platform
File System Transformation	PLFS, ROMIO/LogFS, ADIOS, Giga+	Probably best incorporated into I/O forwarding
Burst Buffer Management	Zest	Probably best incorporated into I/O forwarding
In-System Checkpoint	SCR	
High-level I/O Interface/Library	HDF5, PnetCDF, ADIOS, SciDB, Damasc	Some dependence on MPI-IO (vendor provided?)
End-to-End Integrity		Probably best incorporated into high-level library
Archive	HPSS	
Networking		
Quality of Service		Dependence on network support
Access Methods		Well-known algorithms
Scientific Data Indexing/Search	FastBit, Damasc	
Provenance Capture	PASS	Probably best incorporated into high-level library

Note: This plan assumes that legacy file systems, archive, networking development is funded elsewhere.