

BoF: VI4IO: The I/O Community Hosting the High-Performance Storage List

Julian M. Kunkel

Scientific Computing
Department of Informatics
University of Hamburg

2016-06-15



Outline

- 1 Overview
- 2 Wiki Content
- 3 High-Performance Storage List
- 4 Discussion
- 5 Summary

Introduction

Goals of the Virtual Institute for I/O

- To provide a platform for I/O researchers and enthusiasts to exchange infos
- To foster international collaboration in the field of high-performance I/O
- To track deployment of large storage systems by hosting a storage list

Web page: <http://www.vi4io.org>



Introduction

Philosophical cornerstones of the institute

- To allow participation of everybody without a membership fee
- To treat every member and participant equally
- To be an independent organization
 - Independent of vendors and research facilities

Open Organization

- The organization uses a wiki as central hub
 - Everybody (registered users) can edit the content
 - Mayor changes should be discussed (see below)
 - The wiki uses tag clouds to link between similar entities
- Supported by mailing lists
 - Call-for-papers
 - Announce list for relevant information
 - Contribute list to discuss and steer organizational issues
- Mayor changes should be discussed on the contribute mailing list
- Members can vote for changes

Everybody is welcome to participate

Supporters

Honoring early contributors

- Jay Lofstead
- Colin McMurtrie
- Hans Ole Hatzel
- Thomas Walther
- Phil Carns

Wiki Content

- Groups involved in high-performance storage
Overview of research groups and industry (companies involved in research)
 - Product development the group is involved in
 - Research projects (with links to their source)
 - Tags for layers, products and knowledge
- Tools: *Overview of relevant tools with small descriptions*
 - Types of tools: analysis, benchmarking, I/O middleware
 - Tags for layers and features
- High-performance storage list (HPSL)
Similar to many other lists, e.g., Top500, Graph500
 - Due to the nature of I/O no simple metric
 - Editable and owned by the community
- Internal section
Provides templates and describes rules for editing the page

Demo and Quick Question Round

`http://www.vi4io.org`

Group Tags

Layers

- Describe the abstraction level in the file system stack
 - block storage, object storage, file system, middleware, tape, grid, cloud
- You may add a specific software as well (MPIIO, ...)

Knowledge

- Orthogonal
 - data management, energy-efficiency, machine learning, compression, deduplication, big data, modeling, virtualization, monitoring, simulation
- You may add a specific software as well (GPFS, HPSS, MPICH)

Products

- Specific software products, e.g., MPICH
- Development of software the group is involved in

High-Performance Storage List

Obstacles

- Storage systems are heterogeneous
 - Storage hardware: SSDs, HDDs, NVRAM
 - Availability of optimizations for random and sequential workloads
 - High-level concepts, e.g., staging (K Computer), burst buffers
- Representativeness of a single metric / benchmark
 - Workloads are very diverse, what do we want to measure?
 - With a fitting benchmark systems extract close to peak performance
 - With another benchmark only 1/100 th of performance
- Runtime for executing a benchmark
 - Executing a specific I/O benchmarks may take quite some time
 - Are you willing to pay for it just to be included on a storage list?

Approach

Strategy

- Community-managed list tracking many characteristics
- List elements: supercomputers and supporting storage infrastructure

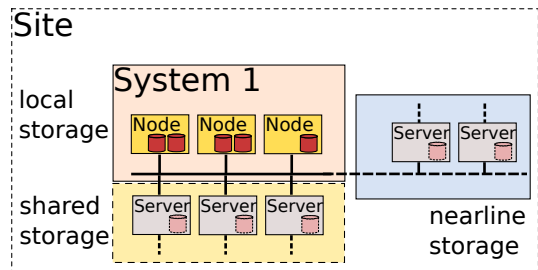
Overcomming obstacles

- *Storage systems are heterogeneous*
 - Communicate a system model that fits most use cases
- *Representativeness of a single metric / benchmark*
 - Rely mostly on theoretic values
 - Allow users to utilize any benchmark/app to determine sustained performance
- *Runtime for executing a benchmark*
 - Optional values: a site can publish computers with a subset of values
 - No overhead, since users can use their own benchmark

System Model

Relevant components

- General system information
 - Energy consumption (system, shared s.)
 - Compute peak and memory capacity
- Local storage: used/offered by compute
- Shared storage: available from all nodes
- Nearline storage: high latency



See: <http://www.vi4io.org/hpsl/metrics>

System Model

- Local storage: individually accessible by only on a node
 - Accumulate characteristics of useable local storage
 - Characteristics for system storage that is not available to users are not counted
 - May cover node-local NVRAM
 - Especially useful for integrated staging solutions (e.g. with K Computer)
 - People may run parallel file systems on top of local storage
- Shared storage: must be accessible from all compute nodes
 - Usually parallel file system/object storage
 - If multiple storage systems are available to one system, aggregate them
 - Only do this if they are similar, otherwise use the best characteristics
 - May use a burst-buffer transparently
- Nearline storage: involves high latency
 - Tape archives, MAID systems, blue ray changer, Amazon Glacier
 - Provides a cache, drives (HDDs for MAID, tape drives)

Collected Information

Peak Performance

- Theoretical value based on hardware limits
 - e.g. network (server) throughput, SATA limits
- Best performance of one server x number of servers.
- Describe in the text how the peak is computed

Sustained Performance

- Actually observed performance with an application or benchmark
- You can use any benchmark and measurement protocol
- Just make sure you are not measuring cache effects
- Describe in the text how the value has been measured

Collected Information

Tags

- Describe hardware and software features individually
- Include coarse grained and fine grained information
 - Lustre, Lustre 2.7, DNE Phase 1
 - Infiniband, FDR-14, fat-tree, blocking 2:2:1
- A taxonomy is needed – but overkill so far
 - Approach: check existing tags and manually fix tag incompatibility

Tracking Data Across Multiple Years

Strategy

- Every begin of a year, systems from the last list are copied over
- Decomission: 5 years after installation, systems are removed from the list

Dealing with hardware upgrades

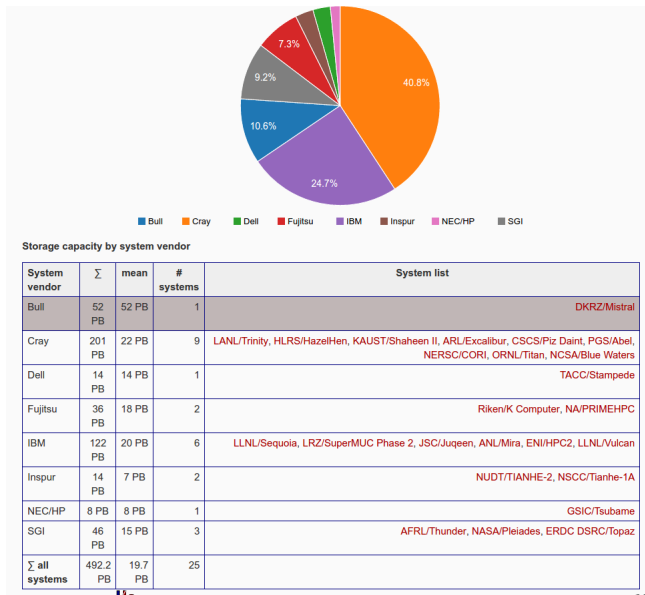
- Procurement in phases: a small system is delivered first, later a big one
 - If both systems work as one big system, you can first add “NAME phase 1”, then later add the system “NAME”
 - Combine the characteristics
 - If not, then you can keep “NAME phase 1” and “NAME phase 2” systems
- Minor upgrades: e.g., more storage, more compute nodes
 - Just update the system characteristics of this year’s supercomputer
 - Keep the older lists as they are

Overview

Wiki features

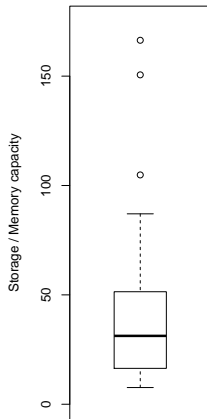
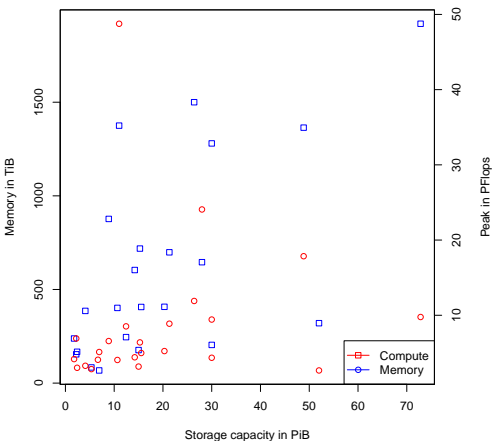
- Table view with selectable columns
- Visualization with flexible metrics selection/aggregation
- More visualizations to come for multi-year analysis

Demo



Some More Analysis: Relationship storage capacity and compute

- On 25 systems that are currently in the list
- Correlation storage cap. vs.
 - memory capacity = 0.63
 - compute peak = 0.13
- Mean(storage/mem capacity) = 45.6



Discussion

- Content provided by the wiki
 - Listing of events (CFP Wiki for storage?)
 - Collecting performance measurements for the individual benchmarks
 - Embed recent publications, link to each group or ResearchGate?
 - Something missing?
 - Taxonomy for tags?
- Steering of the organization
 - Use the contribute mailinglist; everybody can submit suggestions
 - Allow participants to vote on major changes?
 - Should a steering committee be established?

Summary

- The Virtual Institute for I/O is a new community hub
 - Open to everybody and free to join
- It contains information about
 - Tools
 - Research groups
- It hosts the High-Performance Storage List (HPSL)
 - Covers many metrics and allows flexible visualization
 - Will track metrics across years
 - Can be updated by members

You are welcome to participate