BoF: The Virtual Institute for I/O and the IO-500

Julian M. Kunkel, Jay Lofstead, John Bent

German Climate Computing Center, Sandia National Lab, Seagate

2016-11-17

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

- Treat every member and participant equally
- Allow free participation without any membership fee inclusive to all
- Be 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

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

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

The HPSL contains system characteristics for sites, supercomputer and storage

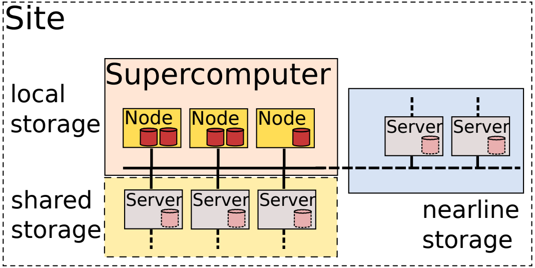
Strategy to overcome certain obstacles

- *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

Components with characteristics

- Site
- Supercomputer
- Storage: shared, local, tape archive



Navigation

- Components are assigned to sites
- The site topology is visualized

Examples: `http://www.vi4io.org/hpsl/2016/de/dkrz/start`
`http://www.vi4io.org/hpsl/2016/jp/jamstec/start`

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 protocoll
- 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

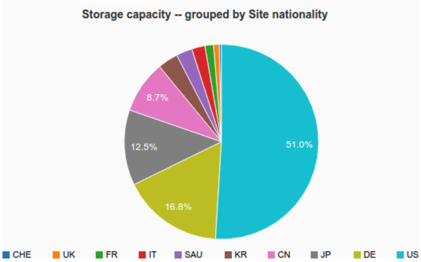
Dealing with hardware upgrades

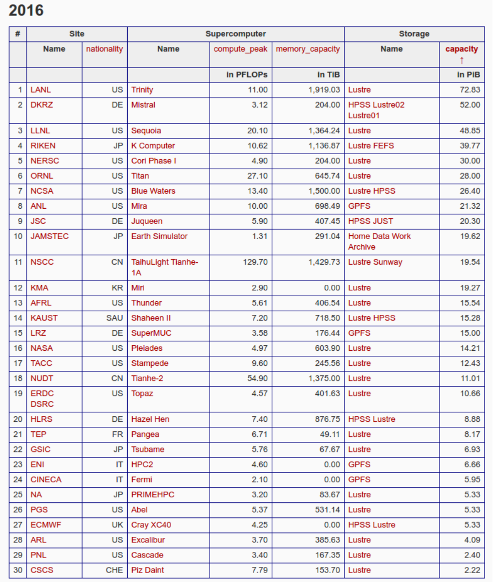
- Procurement in phases: a small system is delivered first, later a big one
  - If both systems work as one big system, you can first add “NAME phase 1”, then later add the system “NAME”
    - Combine the characteristics
  - If not, then you can keep “NAME phase 1” and “NAME phase 2” systems
- Minor 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


Storage capacity -- grouped by Site nationality

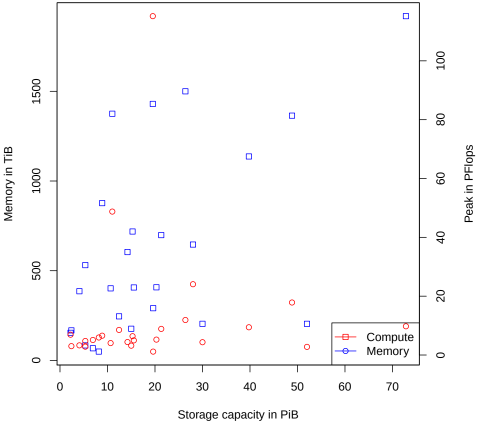


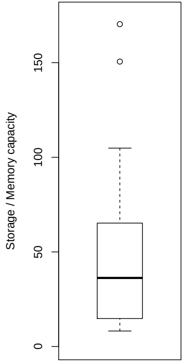
2016

#	Site		Supercomputer			Storage	
	Name	nationality	Name	compute_peak	memory_capacity	Name	capacity ↑
				in PFLOPs	in TiB		in PiB
1	LANL	US	Trinity	11.00	1,919.03	Lustre	72.83
2	DKRZ	DE	Mistral	3.12	204.00	HPSS Lustre02 Lustre01	52.00
3	LLNL	US	Sequoia	20.10	1,364.24	Lustre	48.85
4	RIKEN	JP	K Computer	10.62	1,136.87	Lustre FEFS	39.77
5	NERSC	US	Cori Phase I	4.90	204.00	Lustre	30.00
6	ORNL	US	Titan	27.10	645.74	Lustre	28.00
7	NCSA	US	Blue Waters	13.40	1,500.00	Lustre HPSS	26.40
8	ANL	US	Mira	10.00	698.49	GPFS	21.32
9	JSC	DE	Juqueen	5.90	407.45	HPSS JUST	20.30
10	JAMSTEC	JP	Earth Simulator	1.31	291.04	Home Data Work Archive	19.62
11	NSCC	CN	TaihuLight Tianhe-1A	129.70	1,429.73	Lustre Sunway	19.54
12	KMA	KR	Miri	2.90	0.00	Lustre	19.27
13	AFRL	US	Thunder	5.61	406.54	Lustre	15.54
14	KAUST	SAU	Shaheen II	7.20	718.50	Lustre HPSS	15.28
15	LRZ	DE	SuperMUC	3.58	176.44	GPFS	15.00
16	NASA	US	Pleiades	4.97	603.90	Lustre	14.21
17	TACC	US	Stampede	9.60	245.56	Lustre	12.43
18	NUDT	CN	Tianhe-2	54.90	1,375.00	Lustre	11.01
19	ERDC DSRC	US	Topaz	4.57	401.63	Lustre	10.66
20	HLRS	DE	Hazel Hen	7.40	876.75	HPSS Lustre	8.88
21	TEP	FR	Pangea	6.71	49.11	Lustre	8.17
22	GSIC	JP	Tsubame	5.76	67.67	Lustre	6.93
23	ENI	IT	HPC2	4.60	0.00	GPFS	6.66
24	CINECA	IT	Fermi	2.10	0.00	GPFS	5.95
25	NA	JP	PRIMEHPC	3.20	83.67	Lustre	5.33
26	PGS	US	Abel	5.37	531.14	Lustre	5.33
27	ECMWF	UK	Cray XC40	4.25	0.00	HPSS Lustre	5.33
28	ARL	US	Excalibur	3.70	385.63	Lustre	4.09
29	PNL	US	Cascade	3.40	167.35	Lustre	2.40
30	CSCS	CHE	Piz Daint	7.79	153.70	Lustre	2.22

Some More Analysis: Relationship Storage/Memory Capacity

- On 30 systems that are currently in the list
- Correlation storage cap. vs.
  - memory capacity = 0.58
  - compute peak = 0.04
- Mean(storage/mem capacity) = 54.6





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; everbody 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