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

Julian Kunkel[1], Jay Lofstead[2], John Bent[3]

[1] *Deutsches Klimarechenzentrum (DKRZ)* [2] *Sandia National Laboratory* [3] *Seagate Government Solutions*

Contact: kunkel@dkrz.de

Introduction

The research community in high-performance computing is organized loosely. There are many distinct resources such as homepages of research groups and benchmarks. The Virtual Institute for I/O aims to provide a hub for the community and particularly newcomers to find relevant information in many directions. Additionally, we host the **high-performance storage list**. Similarly to the top500, it contains information about supercomputers and their storage systems. Additionally, in the community, we are working on standardizing an I/O benchmark.

This poster introduces the Virtual Institute for I/O, the high-performance storage list and the effort for the IO-500 which are unfunded community projects.

The Virtual Institute for I/O

Goals of the Virtual Institute for I/O (VI4IO) are

- Provide a platform for I/O researchers and enthusiasts for exchanging information
- Foster training and international collaboration in the field of high-performance I/O
- Track/encourage the deployment of large storage systems by hosting information about high-performance storage systems

The philosophical cornerstones of VI4IO are:

- Treat contributors/participants equally
- Allow free participation without any fee inclusive to all
- Independent of vendors/research facilities

Open Organization

The organization uses a wiki as central hub

- Registered users can edit the content
- Mayor changes should be discussed on the contribute mailing list
- Tag clouds link between similar entities
- Supported by mailing lists, e.g.:
  - Call-for-papers
  - Announcements
  - Contributions / suggestions

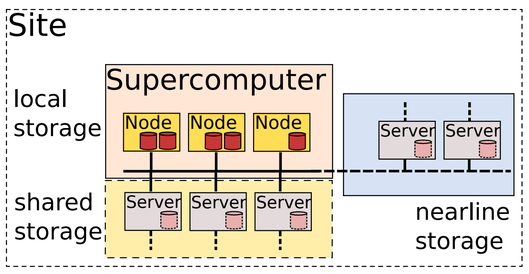
HPSL System Model

The system model describes how characteristics are assigned to components. Storage is difficulty to assign to a single component as it is often shared across supercomputers, therefore, a component based model is used.

Supported components:

- Site: Describes the facility
- Supercomputer: A system
- Storage (shared, local or nearline storage)

Conceptual example:

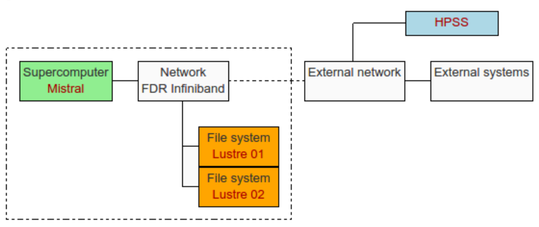


The web page allows the creation of a topology for the facility to indicate the relation between the components. An example for DKRZ:

Site characteristics

Key	Value	Rf
Institution	DKRZ	
Nationality	DE	
Web page	https://www.dkrz.de/	
Energy consumption	2.00 MW	
Power usage effectiveness	1.04	

System architecture



DKRZ hosts the Mistral supercomputer which is tightly coupled with two Lustre file systems. We have some small compute and supporting infrastructure that may access the storage of Mistral and a large HPSS system.

Supercomputers

- Mistral

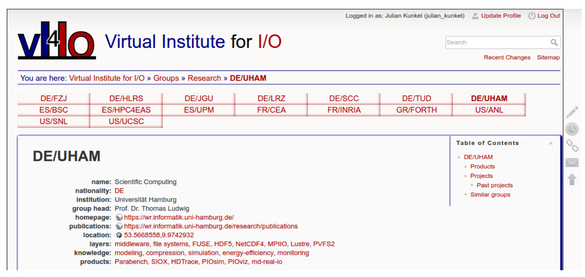
Storage systems

- HPSS
- Lustre02
- Lustre01

Community Content of the VI4IO Wiki

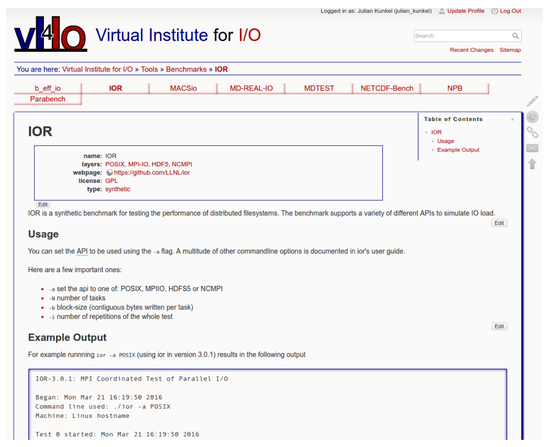
Worldwide research groups that address high-performance I/O including:

- A taglist for available knowledge
- Research products such as file systems
- Ongoing research projects



Everyone is welcome to add (own) group(s)!

Relevant I/O related tools and benchmarks.



High-Performance Storage List

The High-Performance storage list contains the characteristics of site, supercomputer and connected storage (see the box about the system model). The list shown in the box on the right is sortable on the metric of choice. It allows to add/remove metrics (see the list next). Graphs are created based on selectable grouping.

Metrics: Most metrics can be determined without measurement and describe hardware and software characteristics that should be well known to the site and vendor. A few metrics cover actually observed metadata and I/O performance, in this case the measurement procedure must be clear. The list stores data entered in the wiki into a database and converts data to a base unit.

The following list of supported metrics includes a description:

Institution

- institution: The abbreviation of the institution. Note that systems are linked together based on year and institution.
- year: The year for which the data is valid.
- nationality: The international abbreviation for the nationality of the institution.
- web page: The web page of the institution.
- energy consumption: The overall energy consumption of the datacenter.
- power usage effectiveness: The PUE of the datacenter.

Supercomputer

- institution: see above
- year: see above
- vendor: The vendor of the supercomputer.
- software: A list of keywords with relevant software components, e.g., which file system, parallelization software.
- installation: This is the date when the supercomputer has been installed. Multi-phase installations should appear with their last upgrade date.
- compute peak: The theoretical peak performance in FLOPs.
- node count: The number of nodes.
- total cores: The total number of available cores.
- memory capacity: The available memory capacity in Bytes.
- memory bandwidth: The sum of the theoretical memory bandwidth available in B/s.
- memory per node: The memory capacity per node.
- application domain: A list of the main (scientific) domains that use this supercomputer.
- applications: A list of the main applications (if known).
- energy consumption: The energy consumption of the supercomputer (without storage) in Watts – this does not take the PUE into account.
- interconnect: A list of keywords about the interconnect.
- processor: A list of keywords specifying the processor.
- graph500: The achieved performance according to the graph 500 list. This is not the position in the list, as this may change over time.
- graph500 problem scale: according to the graph 500 list.
- top500: The achieved performance according to the top 500.
- green500: The achieved efficiency according to the green 500.
- architecture: A list of keywords covering the system architecture, e.g., i386 64, GPGPU

Storage

- institution: see left
- year: see left
- type: The type of the storage, i.e., tape archive/shared storage
- installation: see left
- energy consumption: The energy consumption of the storage part in Watts – this does not take the PUE into account.
- capacity: The effective capacity that is available to users. It includes overhead of erasure (RAID) coding and potential hot/-cold spares. This value can be easily derived from the number of available storage devices that support the listed file system.
- interconnect: A list of keywords about the interconnect.
- drives: The total number of tape drives for a nearline tape/-MAID archive.
- cache size: The amount of storage cache in a nearline HSM.
- slots: The number of slots in a nearline tape/MAID archive to hold media.
- vendor: The vendor of the storage hardware.
- software: A list of keywords specifying the software further.
- hardware: A list of keywords specifying the hardware further.
- peak: The theoretical peak performance of the storage system. The value is the performance that could theoretically be achieved when transferring data between clients and storage. It is limited by 1) the aggregated network throughput between client and servers, 2) the aggregated (RAID) controller throughput, 3) the network topology.
- *metadata rate*: Metadata throughput. The value can be determined using any I/O benchmark of choice that ensures that client-side and server-side caches are overwhelmed.
- *sustained write*: Best I/O throughput ever measured when accessing files. The read and write values can be determined using any I/O benchmark of choice that ensures that client-side and server-side caches are overwhelmed.
- *sustained read*: (see the description for write)
- servers: The number of storage servers of the storage system.
- hdds: The number of HDDs that belong to the storage system.
- ssds: The number of SSDs that belong to the storage system.

Measurement procedure for sustained performance: Compared to other lists (TOP500, Green500) that have a clear measurement process, the rules for determining sustained performance for the HPSL are relaxed due to the complexity of I/O benchmarks, However it must be clarified how the measurement has been conducted.

IO-500 Effort

We are discussing the creation of a benchmark to compare facilities and storage systems. This challenge is explored on our task page: https://www.vi4io.org/std/io500 and mailing list.

Goals for the benchmark:

- Capture user-experienced performance
- Reported performance is representative for:
  - IOEasy: Applications with well optimized I/O patterns
  - IOHard: Applications that require a random workload
  - IOMD: Usage that depends on metadata/small objects

Challenges:

- Representative: for optimized, naive I/O heavy workloads; and small objects
- Inclusive: cover various storage technology and non-POSIX APIs
- Trustworthy: representative results and prevent cheating
- Cheap: easy to run and short benchmarking time (in the order of minutes)

Strategy:

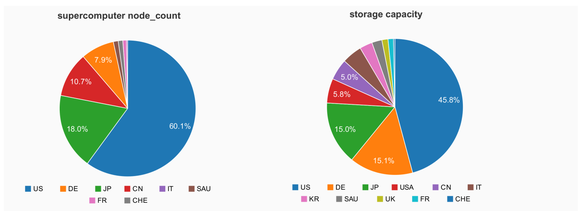
- Build on existing benchmarks
- Plugin systems should allow for alternative storage technology
- Start by reporting one metric per benchmark, decide later about a single number

HPSL 2017

The current list contains 33 sites:

#	Site		Supercomputer			Storage	
	Name	nationality	Name	compute_peak	memory_capacity	Name	capacity ↑
				in PFLOPs	in TiB		in PiB
1	LANL	US	Trinity	11.08	1,919.03	Lustre	72.83
2	DKRZ	DE	Mistral	3.12	204.00	Lustre02 Lustre01 HPSS	52.00
3	LLNL	US	Sequoia	20.10	1,364.24	Grove	48.85
4	RIKEN	JP	K Computer	10.62	1,136.87	Lustre FEFS	39.77
5	NCAR	USA	Cheyenne	5.33	184.40	HPSS GPFS	37.00
6	NERSC	US	Cori Phase I	4.90	204.00	Lustre	30.00
7	ORNL	US	Titan	27.10	645.74	Spider 2	28.00
8	NCSA	US	Blue Waters	13.40	1,500.00	HPSS Lustre	26.40
9	JCAHPC	JP	Oakforest-PACS	24.91	836.09	Lustre Burst Buffer	24.10
10	CINECA	IT	Marconi A2 Fermi	12.93	413.97	GPFS GPFS	23.71
11	ANL	US	Mira	10.00	698.49	GPFS	21.32
12	JSC	DE	Juqueen	5.90	407.45	HPSS JUST	20.30
13	JAMSTEC	JP	Earth Simulator	1.31	291.04	Home Data Work Archive	19.62
14	KMA	KR	Miri	2.90	0.00	Lustre	19.27
15	NSCC	CN	TaihuLight	125.00	1,191.44	Sunway	17.76
16	AFRL	US	Thunder	5.61	406.54	Lustre	15.54
17	KAUST	SAU	Shaheen II	7.20	718.50	Lustre HPSS	15.28
18	LRZ	DE	SuperMUC Phase 2	3.58	176.44	GPFS	15.00
19	NASA	US	Pleiades	4.97	603.90	Lustre	14.21
20	NSCG	CN	Tianhe-2 Tianhe-1A	59.60	1,169.61	Tianhe-2 H2FS Tianhe-2 Lustre Lustre	14.18
21	TACC	US	Stampede	9.60	245.56	Lustre	12.43
22	ERDC DSRC	US	Topaz	4.57	401.63	Lustre	10.66
23	HLRS	DE	Hazel Hen	7.40	876.75	HPSS Lustre	8.88
24	TEP	FR	Pangea	6.71	49.11	Lustre	8.17
25	GSIC	JP	Tsubame 2.5	5.76	67.67	Lustre	6.93
26	ENI	IT	HPC2	4.60	0.00	GPFS	6.66
27	PGS	US	Abel	5.37	531.14	Lustre	5.33
28	Nagoya University	JP	PRIMEHPC	3.20	83.67	Lustre	5.33
29	ECMWF	UK	Cray XC40	4.25	0.00	HPSS Lustre	5.33
30	ARL	US	Excalibur	3.70	385.63	Lustre	4.09
31	EPCC	UK	Archer	2.55	0.00	Lustre	3.91
32	PNL	US	Cascade	3.40	167.35	Lustre	2.40
33	CSCS	CHE	Piz Daint	7.79	153.70	Lustre	2.22

- *storage type*: **all**, local storage, shared storage, tape archive, MAID
- *aggregation*: no, **sum**, avg, max, min – *! removed non-reducable columns*
- columns
  - *site*
    - + energy_consumption, power_usage_effectiveness, initial_facility_costs, annual_staff_costs
    - - nationality
  - *supercomputer*
    - + node_count, total_cores, energy_consumption, graph500, graph500_problem_scale, top500, green500, memory_bandwidth, life_time, annual_procurement_costs, annual_facility_update_costs, annual_tco
    - - compute_peak, memory_capacity
  - *storage*
    - + energy_consumption, drives, cache_size, slots, peak, metadata_rate, sustained_write, sustained_read, servers, hdds, ssds, life_time, annual_procurement_costs, annual_facility_update_costs, annual_tco
    - - capacity



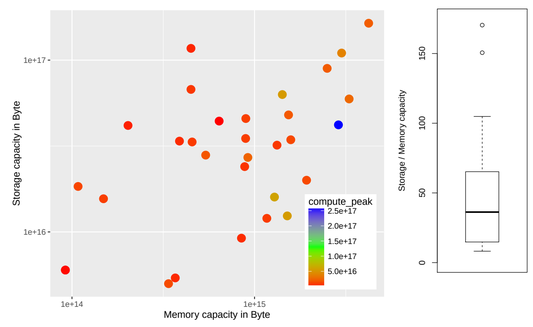
Data table to the graph

site nationality	supercomputer node_count	storage capacity	site	supercomputer	storage
group	sum	sum *PiB*			
US	321184	292.06	AFRL, ANL, ARL, ERDC DSRC, LANL, LLNL, NASA, NCSA, NERSC, ORNL, PGS, PNL, TACC	Thunder, Mira, Excalibur, Topaz, Trinity, Sequoia, Pleiades, Blue Waters, Cori Phase I, Titan, Abel, Cascade, Stampede	Lustre, GPFS, Lustre, Lustre, Lustre, Grove, Lustre, HPSS, Lustre, Lustre, Spider 2, Lustre, Lustre, Lustre
DE	42338	96.18	DKRZ, HLRS, JSC, LRZ	Mistral, Hazel Hen, Juqueen, SuperMUC Phase 2	Lustre02, Lustre01, HPSS, HPSS, Lustre, HPSS, JUST, GPFS
JP	96272	95.75	GSIC, JAMSTEC, JCAHPC, Nagoya University, RIKEN	Tsubame 2.5, Earth Simulator, Oakforest-PACS, PRIMEHPC, K Computer	Lustre, Home, Data, Work, Archive, Lustre, Burst Buffer, Lustre, Lustre, FEFS
USA		37.00	NCAR	Cheyenne	HPSS, GPFS
CN	56960	31.94	NSCC, NSCG	TaihuLight, Tianhe-2, Tianhe-1A	Sunway, Tianhe-2 H2FS, Tianhe-2 Lustre, Lustre
IT	1500	30.38	CINECA, ENI	Marconi A2, Fermi, HPC2	GPFS, GPFS, GPFS
KR		19.27	KMA	Miri	Lustre
SAU	6174	15.28	KAUST	Shaheen II	Lustre, HPSS
UK		9.24	ECMWF, EPCC	Cray XC40, Archer	HPSS, Lustre, Lustre
FR	4608	8.17	TEP	Pangea	Lustre
CHE	5772	2.22	CSCS	Piz Daint	Lustre
all	534808	637.49	33	35	49

- *aggregation*: **sum**, avg, max
- metrics
  - *site*
    - + energy_consumption, power_usage_effectiveness, initial_facility_costs, annual_staff_costs
    - -
  - *supercomputer*
    - + compute_peak, total_cores, memory_capacity, energy_consumption, graph500, graph500_problem_scale, top500, green500, memory_bandwidth, life_time, annual_procurement_costs, annual_facility_update_costs, annual_tco
    - - node_count
  - *storage*
    - + energy_consumption, drives, cache_size, slots, peak, metadata_rate, sustained_write, sustained_read, servers, hdds, ssds, life_time, annual_procurement_costs, annual_facility_update_costs, annual_tco
    - - capacity
- *group by*
  - site: **nationality**, web_page
  - supercomputer: vendor, software, installation, application_domain, applications, interconnect, processor, architecture, memory_per_node
  - storage: type, installation, interconnect, vendor, software, hardware
- *filter by*: US, DE, JP, USA, CN, IT, KR, SAU, UK, FR, CHE, none

Derived Analysis

With the collected data many in-depth analysis becomes possible, for example, the relationship between storage and memory capacity:



- Correlation storage capacity vs.
  - memory capacity = 0.63
  - compute peak = 0.057
- Mean(storage/mem capacity) = 58

Ongoing Work

- Support standardization efforts
  - IO-500 benchmark
  - Lossy compression interfaces
- IO-500 agenda:
  - June'17, proposal for benchmark
  - Benchmark runs on Top-500 sites
  - Nov'17, SC – presentation of results
- Extending benchmarks, HPSL sites
- Support training and teaching for storage

VI4IO and You

Content is under open licenses. You are welcome to join the mailing lists or participate!

https://vi4io.org