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

Julian Kunkel[1], Jay Lofstead[2], John Bent[3], George Markomanolis[4]

[1] *University of Reading* [2] *Sandia National Laboratories* [3] *DDN* [4] *Oak Ridge National Lab.*

Contact: juliankunkel@gmail.com

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. It hosts the **comprehensive data center list (CDCL)**. Similarly to the top500, it contains information about supercomputers and their storage systems.

I/O benchmarking, particularly, the intercomparison of measured performance between sites is tricky as there are more hardware components involved and configurations to take into account. Therefore, together with the community, we standardized an HPC I/O benchmark, the **IO-500** benchmark, for which the first list had been released during supercomputing in Nov. 2017. Such a benchmark is also useful to assess the impact of system issues like the **Meltdown and Spectre* bugs**.

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

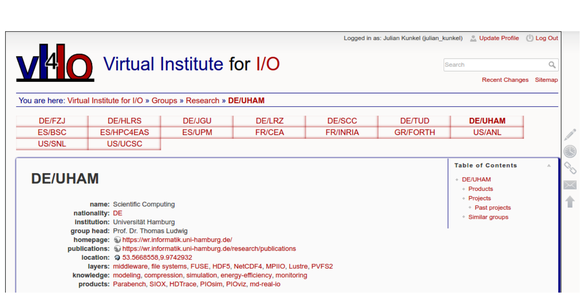
The organization uses a wiki as central hub

- 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

Community Content

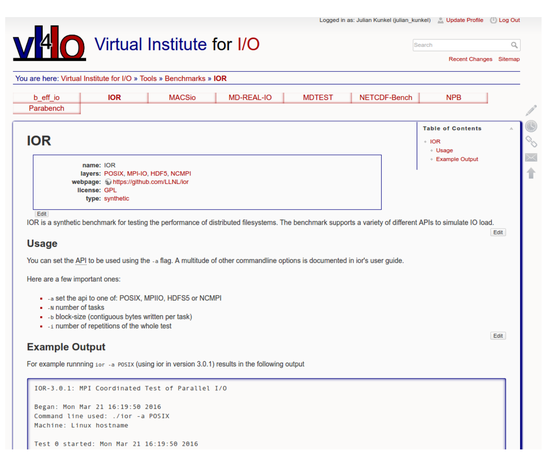
The wiki covers **A) worldwide research groups that address high-performance I/O including:**

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



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

B) Relevant I/O related tools and benchmarks



C) Comprehensive Data Center List
(see the other boxes)

IO-500 Effort

Together with the community, we created the IO-500 benchmark to compare storage systems.

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
  - MDEasy: Metadata/small objects
  - MDHard: Small files (3901 bytes) in a shared directory
  - Find: Finding relevant objects based on patterns

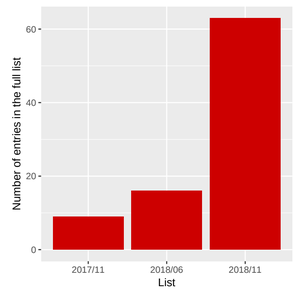


Fig. 1: Growth of the list

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)

Benefit for the community beyond the IO-500:

- Support the development of benchmarks that are used (IO-500 builds on standard benchmarks)
- Feed back best practices of tool usage (e.g., find) and benchmarks
- Aid detailed comparison of individual system characteristics while having a **ranked list**
- Share best-practices to obtain good performance

IO-500 List Nov 2018

There are several lists available, the **full list** contains all the results submitted for comparison while entries can enter a **ranked list** upon user choice and only one solution per system.

There are several ranked lists with awards that stimulate aspects of I/O system development:

- The *IO-500 award* for the fastest system in the ranked IO-500 list
- The *10 node challenge* fosters performance for small-scale runs

Further awards may follow.

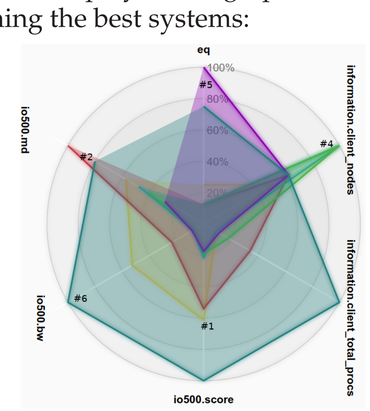
The ranked IO-500 list:

#	information							io500		
	institution	system	storage vendor	filesystem type	client nodes	client total procs	data	score	bw	md
									GiB/s	kIOP/s
1	Oak Ridge National Laboratory	Summit	IBM	Spectrum Scale	504	1008	zip	330.56	88.20	1238.93
2	University of Cambridge	Data Accelerator	Dell EMC	Lustre	528	4224	zip	158.71	71.40	352.75
3	Korea Institute of Science and Technology Information (KISTI)	NURION	DDN	IME	2048	4096	zip	156.91	554.23	44.43
4	JCAHPC	Oakforest-PACS	DDN	IME	2048	16384	zip	137.78	560.10	33.89
5	WekaIO	WekaIO	WekaIO		17	935	zip	78.37	37.39	164.26
6	KAUST	ShaheenII	Cray	DataWarp	1024	8192	zip	77.37	496.81	12.05
7	University of Cambridge	Data Accelerator	Dell EMC	BeeGFS	184	5888	zip	74.58	58.81	94.57
8	Google	Exascaler on GCP	Google	Lustre	120	960	zip	47.23	23.06	96.74
9	JCAHPC	Oakforest-PACS	DDN	Lustre	256	8192	zip	42.18	20.04	88.78
10	KAUST	ShaheenII	Cray	Lustre	1000	16000		41.00*	54.17	31.03*
11	JSC	JURON	ThinkparQ	BeeGFS	8	64		35.77*	14.24	89.81*
12	DKRZ	Mistral	Seagate	Lustre	100	1000		32.15	22.77	45.39
...										
21	CERN	Bytecollider		CephFS	64	64	zip	7.56	2.83	20.16
22	SNL	Serrano	IBM	Spectrum Scale	16	160		4.25*	0.65	27.98*
23	STFC	Jasmin/Lotus	Purestorage	NFS	64	128	zip	2.33	0.26	20.93
24	Clemson University	Palmetto	Dell	OrangeFS	32	32	zip	2.31	1.93	2.77
25	Nemours	nas6	DDN	GPFS	1	2	zip	2.06	1.39	3.04
26	Nemours	nas6	DDN	GPFS	1	2	zip	1.52	0.74	3.11

Flexible equations It supports equations to compute derived metrics, here *mdtest.easy_write / client_nodes* for the full list:

#		information							io500		
	Equation	institution	system	storage vendor	filesystem type	client nodes	client total procs	data	score	bw	md
										GiB/s	kIOP/s
1	24.17	JSC	JURON	ThinkparQ	BeeGFS	8	64		35.77*	14.24	89.81*
2	19.97	DDN	Bancholab	DDN	Lustre	10	240	zip	31.50	6.33	156.69
3	11.34	Clemson University	ofsdev	Dell	BeeGFS	16	64	zip	12.51	2.10	74.72
4	10.36	Clemson University	ofsdev	Dell	BeeGFS	16	128	zip	11.01	1.90	63.81
5	8.48	Clemson University	ofsdev	Dell	BeeGFS	10	80	zip	10.17	2.32	44.67
6	6.87	WekaIO		WekaIO		10	700	zip	58.25	27.05	125.43
7	6.09	Oak Ridge National Laboratory	Summit	IBM	Spectrum Scale	504	1008	zip	330.56	88.20	1238.93
8	5.72	IBM	Sonasad	IBM	Spectrum Scale	10	10	zip	24.24	4.57	128.61
9	5.27	WekaIO	WekaIO	WekaIO		17	935	zip	78.37	37.39	164.26
10	5.19	Clemson University	ofsdev	Dell	BeeGFS	16	32	zip	13.12	3.17	54.21
...											
56	0.07	CERN	Bytecollider		CephFS	64	64	zip	6.01	2.85	12.64
57	0.04	Clemson University	Palmetto	Dell	OrangeFS	48	48	zip	2.15	1.80	2.57
58	0.02	Korea Institute of Science and Technology Information (KISTI)	NURION	DDN	Lustre	2048	4096	zip	35.97	14.06	92.03
59	0.02	JCAHPC	Oakforest-PACS	DDN	IME	2048	16384	zip	137.78	560.10	33.89
60	0.01	Korea Institute of Science and Technology Information (KISTI)	NURION	DDN	IME	2048	4096	zip	156.91	554.23	44.43
61	0.01	KAUST	ShaheenII	Cray	DataWarp	1024	8192	zip	77.37	496.81	12.05
62	0.01	KAUST	ShaheenII	Cray	Lustre	1000	16000		41.00*	54.17	31.03*
63	0.01	JCAHPC	Oakforest-PACS	DDN	IME	2048	16384	zip	101.48	471.25	21.85

The system displays a net graph for further distinguishing the best systems:



As we can see, this can be used to create arbitrary new rankings and investigate the data.

All results are available The individual submission scripts and results for the benchmarks are preserved and can be accessed. The data is also available as CSV file for offline analysis.

Data Center List

The comprehensive data center list with its 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 flexible component based model is used.

Supported components:

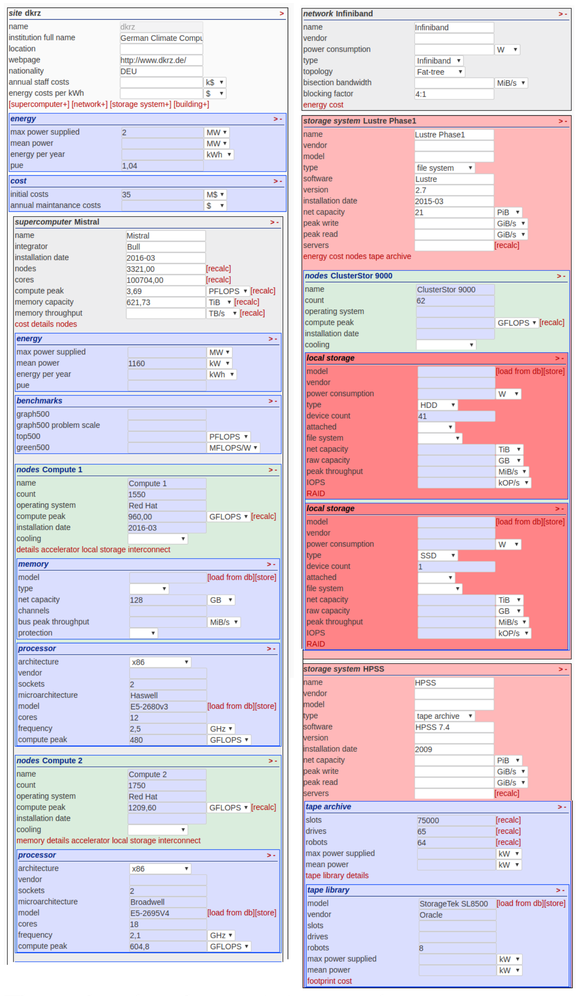
- Site: Describes the facility
- Supercomputer: A system
- Storage system
- Nodes
- Network
- Building

The schema is under active development – we aim to describe data center characteristics. The web page allows the creation of a topology for the facility to indicate the relation between the components – ultimately multiple views will be created to show, e.g.:

- Logical network connectivity
- Physical layout in racks
- Building organization

Metrics: Most metrics can be determined without measurement and describe hardware and software characteristics that should be 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 defined. The list stores data entered in the wiki into a database and converts data to a base unit.

The following is an example of the schema for the DKRZ system:



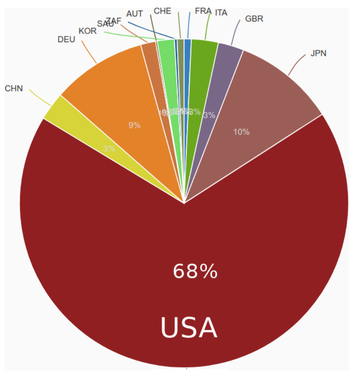
The rules for determining performance are relaxed due to the complexity of I/O measurements, but this is augmented by the IO-500.

HPSL 2019

The current list contains 43 sites:

#	site.institution	site.storage system.net capacity	site.supercomputer.compute peak	site.supercomputer.memory capacity
		in PiB	*in PFLOPS*	*in TB*
1	Oak Ridge National Laboratory	250.04	220.64	3511.66
2	National Energy Research Scientific Computing Center	197.65	37.71	857.03
3	Los Alamos National Laboratory	72.83	11.08	2110.00
4	German Climate Computing Center	52.00	3.69	683.60
5	Lawrence Livermore National Laboratory	48.85	20.10	1500.00
6	RIKEN Advanced Institute for Computational Science	39.77	10.62	1250.00
7	National Center for Atmospheric Research	37.00	5.33	202.75
8	National Center for Supercomputing Applications	27.60	13.40	1649.27
9	Global Scientific Information and Computing Center	25.84	17.89	275.98
10	Joint Center for Advanced HPC	24.10	24.91	919.29
...				
37	Pacific Northwest National Laboratory	2.40	3.40	184.00
38	Navy DoD Supercomputer Resource Center	2.11	2.05	0.00
39	Vienna Scientific Cluster	1.81	0.68	42.18
40	Center for Scientific Computing	0.75	0.51	77.57
41	University of Bristol	0.44	0.38	43.01

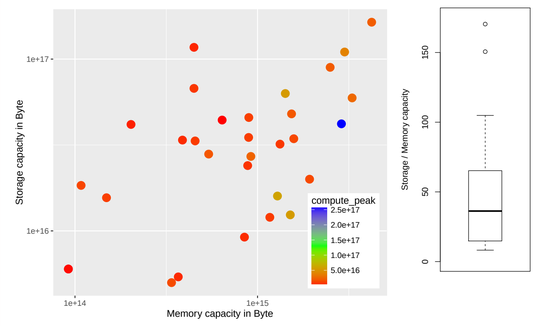
Various views are possible, moreover, it supports flexible data aggregation, e.g., capacity by country:



site.nationality	site.storage system.net capacity	site.supercomputer.compute peak	site.supercomputer.memory capacity	site.supercomputer.nodes
	in PiB	*in PFLOPS*	*in TB*	
USA	776.95	358.90	13728.07	344706
JPN	114.66	57.93	2857.27	98254
DEU	106.50	23.05	2595.32	46232
CHN	32.16	184.60	2596.00	56960
ITA	30.38	17.53	455.17	1500
GBR	29.02	13.05	371.01	485
KOR	19.27	2.90	0.00	0
SAU	16.96	7.20	790.00	6174
FRA	8.17	6.71	54.00	4608
CHE	7.73	25.32	521.00	6751
ZAF	2.98	0.00	8.86	100
AUT	1.81	0.68	42.18	0

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

- IO-500:
  - Clarified execution rules
  - Procedures to adapt IO-500
  - Integration of optional benchmarks
  - Continuous integration deployment Including performance regression
  - Finalize vendor engagement program
- VI4IO standardization efforts
  - Data center representation
  - Next-generation interfaces (NGI)
- CDCL list
  - Extended schema and alternative views
  - More CDCL sites
  - Better link between IO-500 and CDCL
- Support training and teaching for storage

VI4IO, IO-500, and You

You are welcome to join the mailing lists or our slack channel and participate!

Join us on Slack:

The content is under open licenses.

More details on:

- https://vi4io.org
- http://io-500.org