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] *Cray*

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.

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

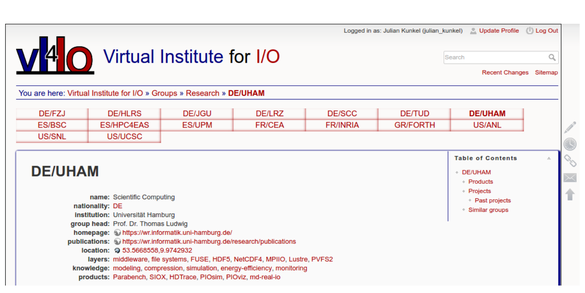
The organization uses a wiki as central hub

- 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

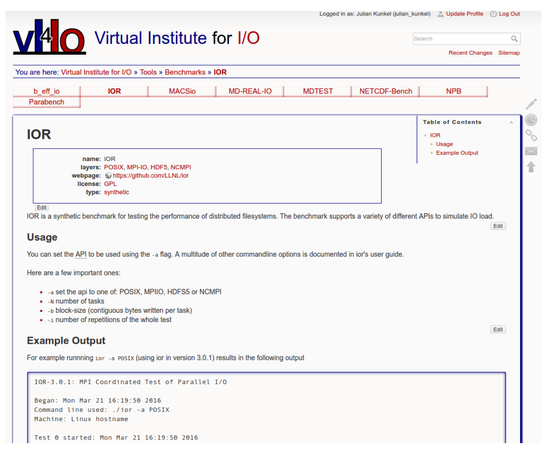
The wiki covers **A) worldwide research groups that address high-performance I/O including:**

- 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

We are discussing the creation of a benchmark to compare facilities and storage systems. This challenge is explored on our task page: http://www.io500.org 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
  - MDEasy: Metadata/small objects
  - MDHard: Small files (3901 bytes) in a shared directory
  - Find: Finding relevant objects based on patterns

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, support their development
- Plugin systems should allow for alternative storage technology
- Reporting one metric per benchmark, use geometric mean to combine them

IO-500 List Nov 2017

The normal list:

#	information				io500			
	system	institution	filesystem	client nodes	score	bw	md	tot iops
						GiB/s	kIOP/s	kIOP/s
1	Oakforest-PACS	JCAHPC	IME	2048	101.48	471.25	21.85	108.00
2	Shaheen	Kaust	DataWarp	300	70.90	151.53	33.17	101.79
3	Shaheen	Kaust	Lustre	1000	41.00	54.17	31.03	67.12
4	JURON	JSC	BeeGFS	8	35.77	14.24	89.83	81.20
5	Mistral	DKRZ	Lustre	100	32.15	22.77	45.39	63.47
6	Sonasad	IBM	Spectrum Scale	10	21.63	4.57	102.38	58.39
7	Seislab	Fraunhofer	BeeGFS	24	18.75	5.13	68.58	47.15
8	EMSL Cascade	PNNL	Lustre	126	11.17	4.88	25.57	24.74
9	Serrano	SNL	Spectrum Scale	16	4.25	0.65	27.98	12.55

All results are available The individual measurements for the benchmarks are stored and can be accessed:

#	information				ior			
	system	institution	filesystem	client nodes	easy write	easy read	hard write	hard read
					GiB/s	GiB/s	GiB/s	GiB/s
1	Oakforest-PACS	JCAHPC	IME	2048	742.38	427.41	600.28	258.93
2	Shaheen	Kaust	DataWarp	300	969.45	894.76	15.55	39.09
3	Shaheen	Kaust	Lustre	1000	333.03	220.62	1.44	81.38
4	JURON	JSC	BeeGFS	8	30.42	48.36	1.46	19.16
5	Mistral	DKRZ	Lustre	100	158.19	163.62	1.53	6.79
6	Sonasad	IBM	Spectrum Scale	10	34.13	32.25	0.17	2.33
7	Seislab	Fraunhofer	BeeGFS	24	18.79	22.34	0.89	1.86
8	EMSL Cascade	PNNL	Lustre	126	17.81	30.19	0.39	2.72
9	Serrano	SNL	Spectrum Scale	16	1.08	1.03	0.22	0.71

#	information		mdtest					
	system	client nodes	easy create	easy stat	easy delete	hard create	hard read	hard stat	hard delete
			kIOP/s	kIOP/s	kIOP/s	kIOP/s	kIOP/s	kIOP/s	kIOP/s
1	EMSL Cascade	126	17.75	61.26	15.63	16.14	23.59	57.04	19.43
2	JURON	8	193.37	718.18	150.61	8.42	0.00	100.85	8.76
3	Mistral	100	18.15	153.05	7.74	17.80	37.58	156.07	8.80
4	Oakforest-PACS	2048	28.29	54.20	35.88	1.51	57.38	61.50	0.95
5	Seislab	24	103.15	433.14	172.95	5.38	13.87	57.40	13.87
6	Serrano	16	32.55	303.02	26.15	2.29	0.00	25.20	26.15
7	Shaheen	300	50.71	49.38	48.89	11.40	0.00	38.73	18.92
8	Shaheen	1000	12.66	120.81	14.96	13.67	0.00	127.32	11.30
9	Sonasad	10	57.22	342.33	47.56	21.57	632.98	529.90	85.34

Flexible equations It supports equations to compute derived metrics, here *easy_create / client_nodes*:

#		information				io500			mdtest
	Equation	system	institution	filesystem	client nodes	score	bw	md	easy create
							GiB/s	kIOP/s	kIOP/s
1	24.17	JURON	JSC	BeeGFS	8	35.77	14.24	89.83	193.37
2	5.72	Sonasad	IBM	Spectrum Scale	10	21.63	4.57	102.38	57.22
3	4.30	Seislab	Fraunhofer	BeeGFS	24	18.75	5.13	68.58	103.15
4	2.03	Serrano	SNL	Spectrum Scale	16	4.25	0.65	27.98	32.55
5	0.18	Mistral	DKRZ	Lustre	100	32.15	22.77	45.39	18.15
6	0.17	Shaheen	Kaust	DataWarp	300	70.90	151.53	33.17	50.71
7	0.14	EMSL Cascade	PNNL	Lustre	126	11.17	4.88	25.57	17.75
8	0.01	Oakforest-PACS	JCAHPC	IME	2048	101.48	471.25	21.85	28.29
9	0.01	Shaheen	Kaust	Lustre	1000	41.00	54.17	31.03	12.66

As we can see, scalability of metadata is limited.

This can be used to create arbitrary new rankings. For example, to rank systems offering most MD performance per IOR: min(easy_create, easy_stat, easy_delete, hard_create, hard_stat, hard_delete, find.hard) / easy_write, e.g., 2 kIOPs per GiB throughput!

#		information					io500
	Equation	system	institution	filesystem	client nodes	nproc	score
1	2.12	Serrano	SNL	Spectrum Scale	16	160	4.25
2	0.88	EMSL Cascade	PNNL	Lustre	126	252	11.17
3	0.63	Sonasad	IBM	Spectrum Scale	10	10	21.63
4	0.29	Seislab	Fraunhofer	BeeGFS	24	24	18.75
5	0.28	JURON	JSC	BeeGFS	8	64	35.77
6	0.05	Mistral	DKRZ	Lustre	100	1000	32.15
7	0.03	Shaheen	Kaust	Lustre	1000	16000	41.00
8	0.01	Shaheen	Kaust	DataWarp	300	2400	70.90
9	0.00	Oakforest-	JCAHPC	IME	2048	16384	101.48

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.

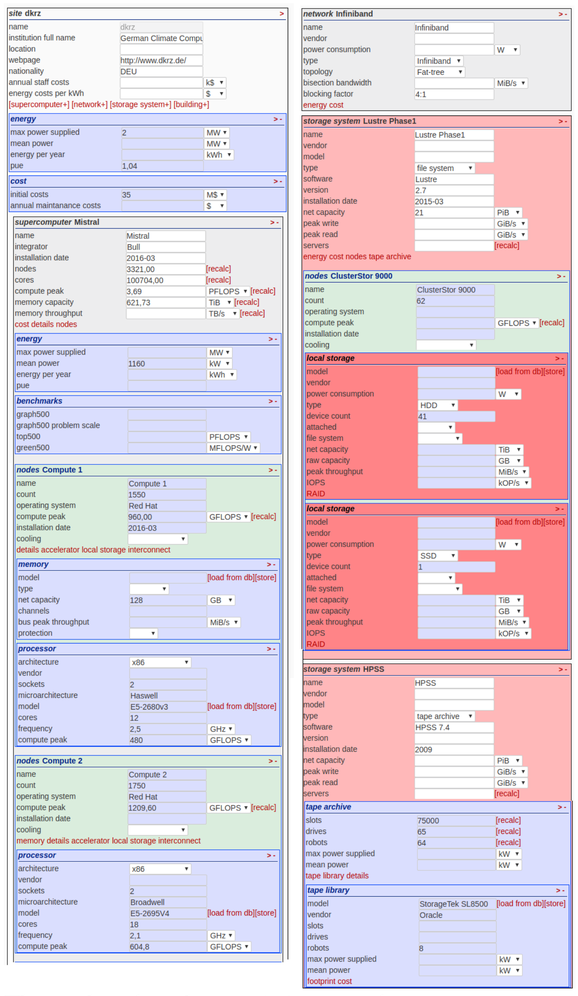
Supported components:

- 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.

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:



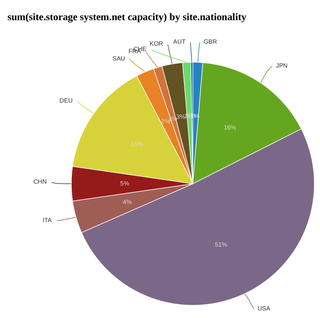
The rules for determining performance are relaxed due to the complexity of I/O measurements, but this is augmented by the IO-500.

HPSL 2018

The current list contains 39 sites:

#	site.name	site.storage system.net capacity	site.supercomputer.compute peak	site.supercomputer.memory capacity
		in PiB	*in PFLOPS*	*in TB*
1	lanl	72.83	11.08	2110.00
2	dkrz	52.00	3.69	683.60
3	llnl	48.85	20.10	1500.00
4	riken	39.77	10.62	1250.00
5	ncar	37.00	5.33	202.75
6	nersc	30.00	4.90	224.30
7	ornl	28.00	27.10	710.00
8	ncsa	27.60	13.40	1649.27
9	gsic	25.84	17.89	275.98
10	jcahpc	24.10	24.91	919.29
11	cineca	23.71	12.93	455.17
12	anl	21.32	10.00	768.00
13	jsc	20.30	6.25	454.15
14	jamstec	19.62	1.31	320.00
15	kma	19.27	2.90	0.00
16	nscc	17.76	125.00	1310.00
17	marcc	17.00	0.87	92.67
18	kaust	16.96	7.20	790.00
19	afrl	15.54	5.61	447.00
20	lrz	15.00	3.58	194.00
21	nscg	14.40	59.60	1286.00
22	nasa	14.21	4.97	664.00
23	tacc	12.43	9.60	270.00
24	erdc_dsrc	10.66	4.57	441.60
25	snl	9.93	0.50	22.10
26	kit	9.57	1.61	222.00
27	hlrs	8.88	7.40	964.00
28	tep	8.17	6.71	54.00
29	cscs	7.73	25.32	521.00
30	eni	6.66	4.60	0.00
31	pgs	5.33	5.37	584.00
32	nu	5.33	3.20	92.00
33	ecmwf	5.33	4.25	0.00
34	arl	4.09	3.70	424.00
35	epcc	3.91	2.55	0.00
36	pnl	2.40	3.40	184.00
37	ndsrc	2.11	2.05	0.00
38	vsc	1.81	0.68	42.18
39	csc	0.75	0.51	77.57

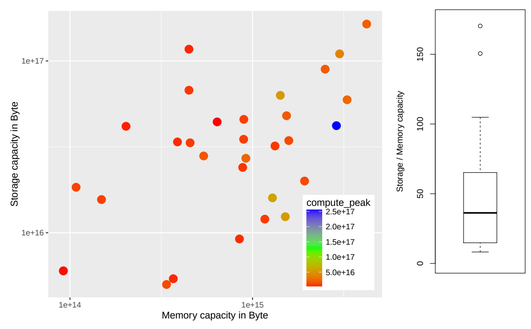
Various views are possible – an example is shown above. Supports flexible data aggregation (below).



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	359.29	132.55	10293.68	322306
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
KOR	19.27	2.90	0.00	0
SAU	16.96	7.20	790.00	6174
GBR	9.24	6.80	0.00	0
FRA	8.17	6.71	54.00	4608
CHE	7.73	25.32	521.00	6751
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

- Supporting standardization efforts
  - IO-500 benchmark
  - Lossy compression interfaces
  - Data center representation
- IO-500 agenda:
  - June'17, proposal for extension rules
- Extending schema
- More 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