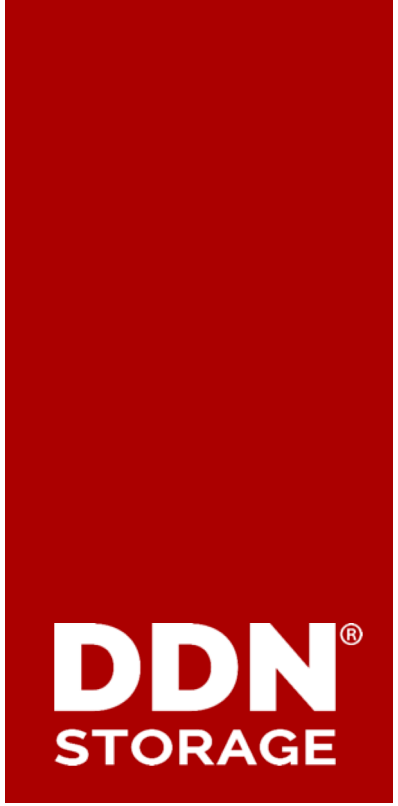


IO-500 | A Storage Benchmark for HPC

... or *Let's Give HPC Storage the Attention It Deserves*

Andreas Dilger
Lustre CTO
Whamcloud

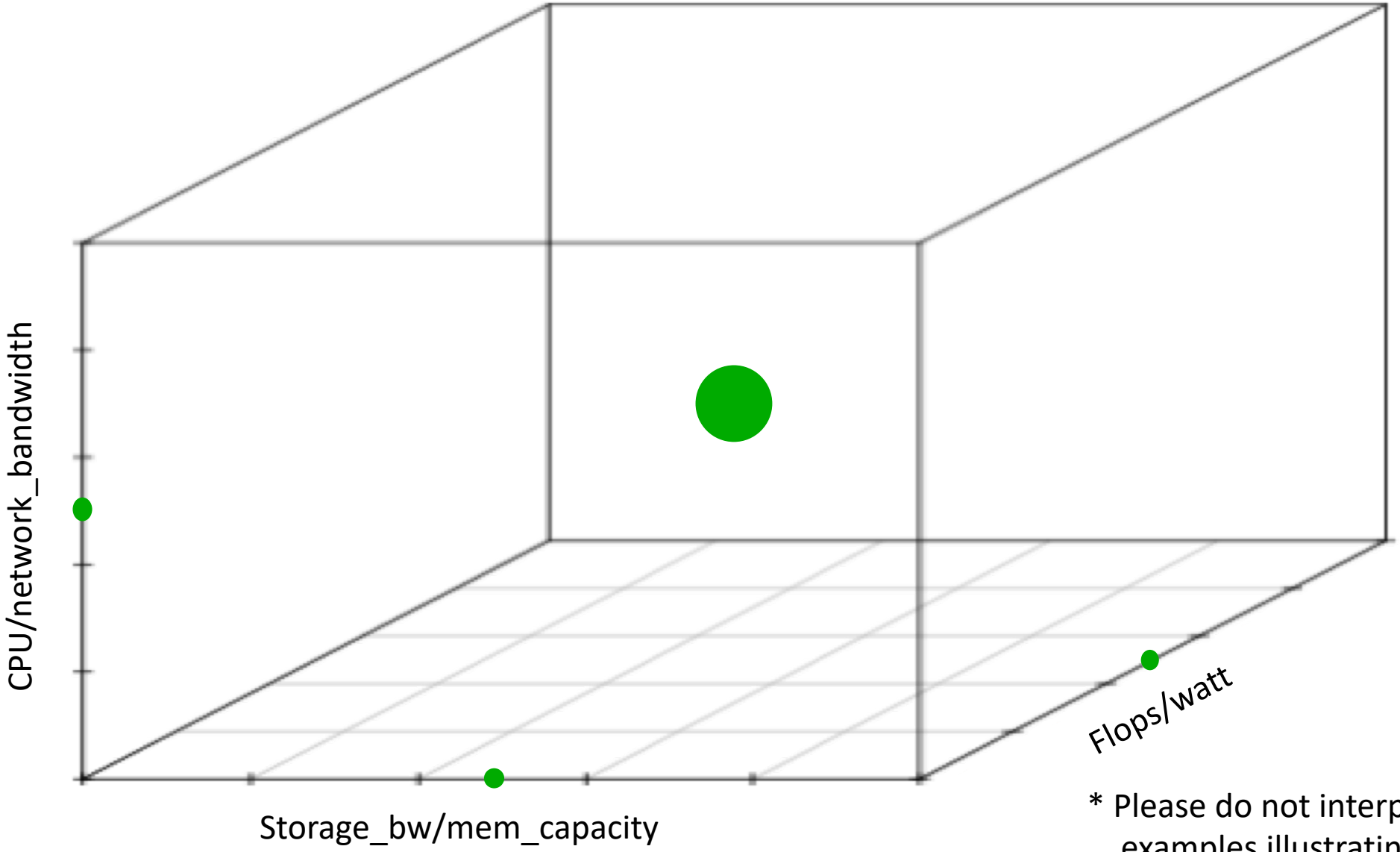
IO-500 | Why Does It Exist?

- Standard way to compare storage performance for HPC systems
  - Not specific to a single filesystem or workload
- Create more realistic user expectations
  - Provide performance bounds for users who don't know what to expect
  - Make both hero **and** anti-hero results available for each test
- Public repository of all results
  - Component results and tunables used to achieve them
  - Record history of storage systems
- Encourage balanced systems
  - Don't focus on just a single metric
- Easier RFP writing
  - Empathy for procurers who struggle to define an "ideal" system
- Better storage products
  - Focus developers and vendors on improving performance seen by end users

IO-500 | …But Why Do We Really Need an IO-500?

- Storage is central to a system's ability to generate science/results
  - Typically under-budgeted compared to CPUs
  - Anything that isn't measured cannot be improved
- Hard to explain storage intricacies and application interactions
- Non-experts have no easy way to compare different storage systems
  - Provide a "single number" for easy/fast comparison …
  - … with additional metrics to allow more in-depth comparisons
- Elevate storage visibility to allow balanced compute platforms
  - Motivator to improve storage with minimal effort
- Leave the world a happier and better place than we found it 😁

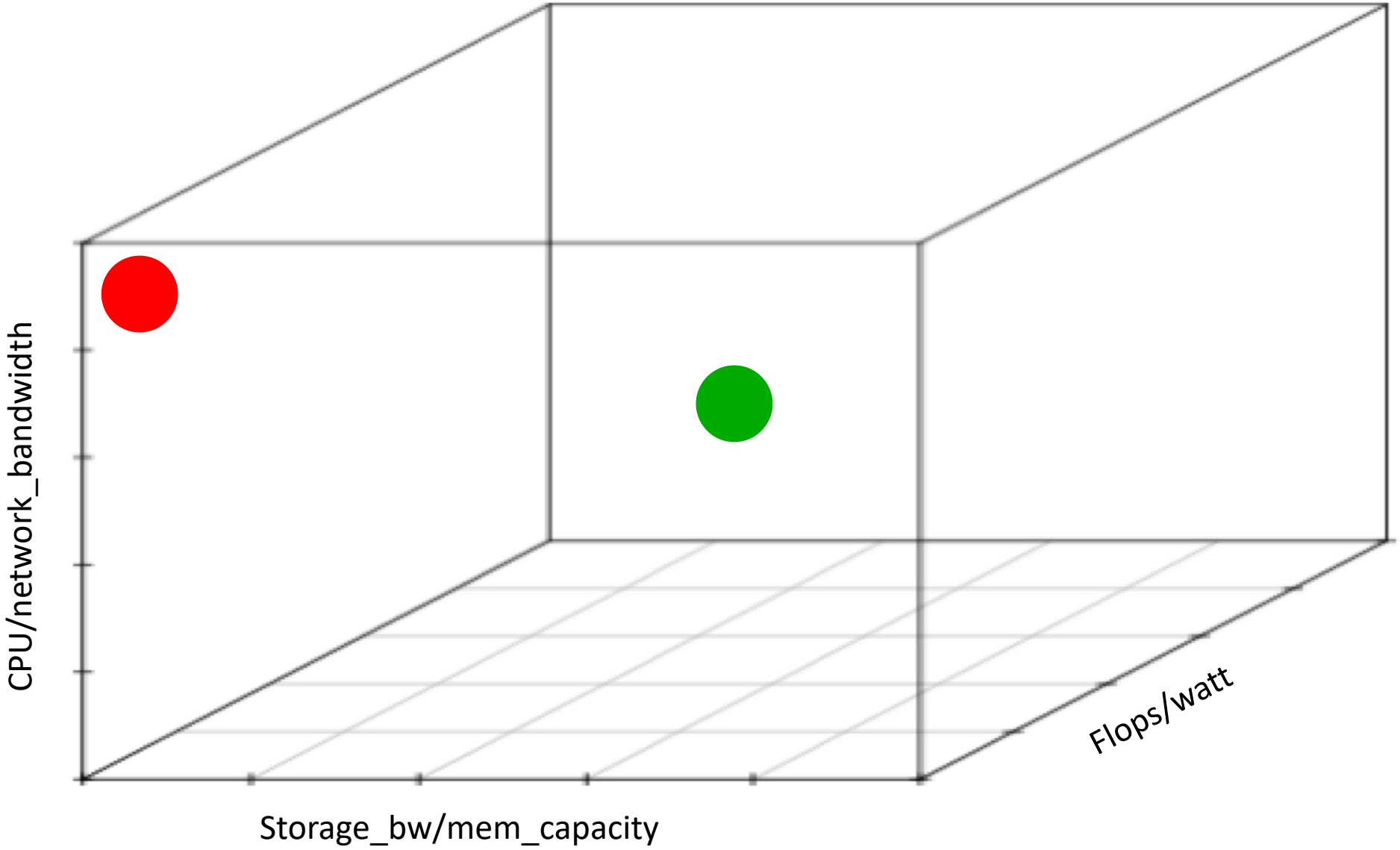
IO-500 | A Legitimate Concern About Linpack

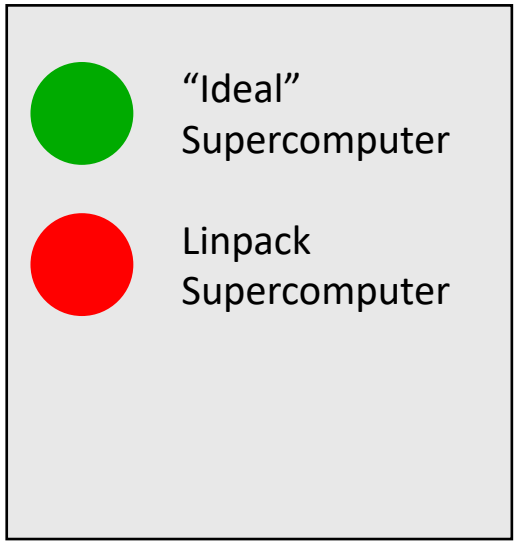


"Ideal" Supercomputer

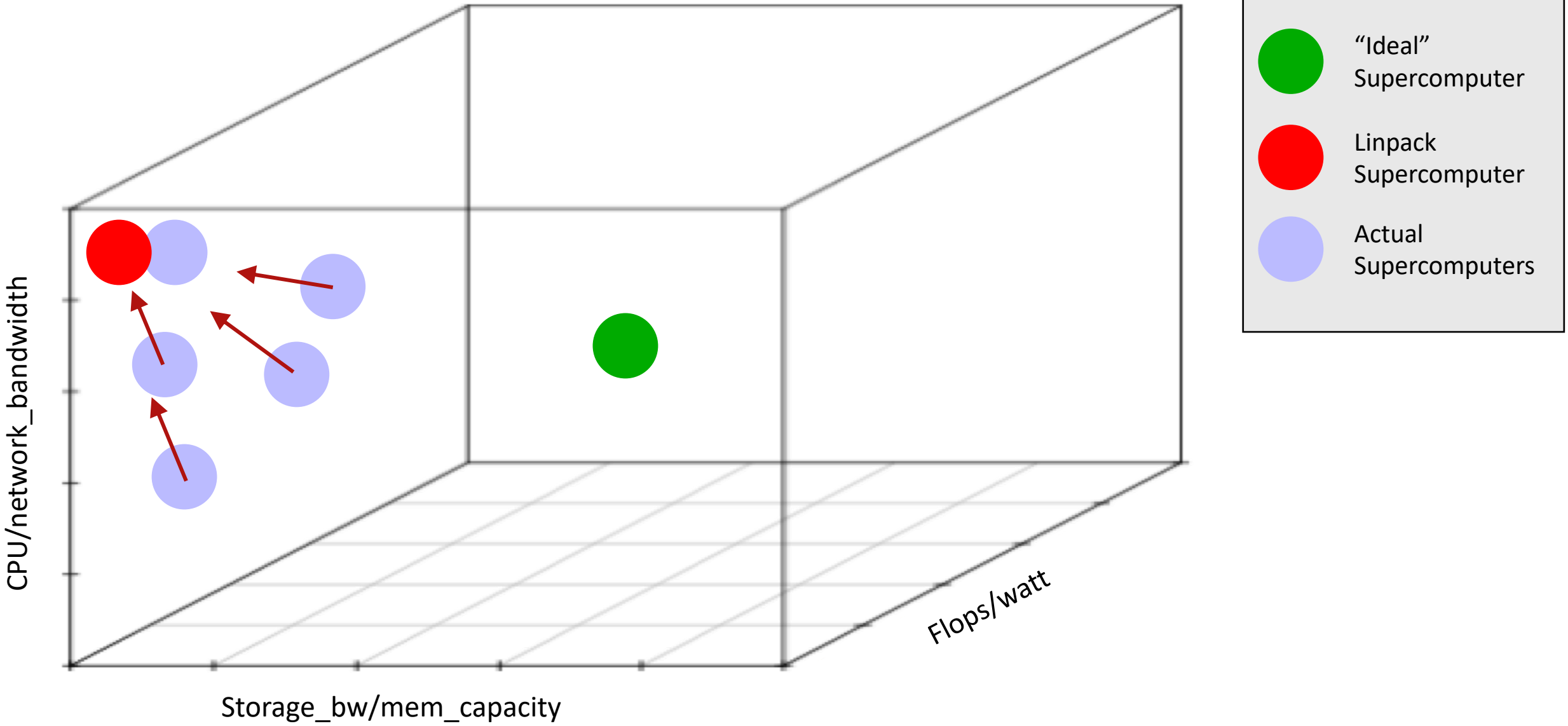
* Please do not interpret axes literally, they are just examples illustrating multi-variable complexity

IO-500 | A Legitimate Concern About Linpack



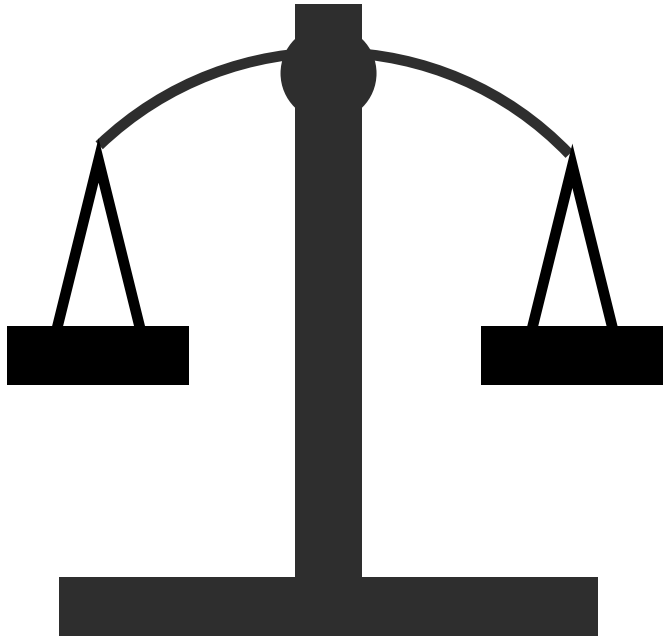


IO-500 | A Legitimate Concern About Linpack



IO-500 | Overall System Balance Is Lacking

IO-500 is designed to be balanced and, in being so, will help restore balance to supercomputing and the storage systems that feed it.

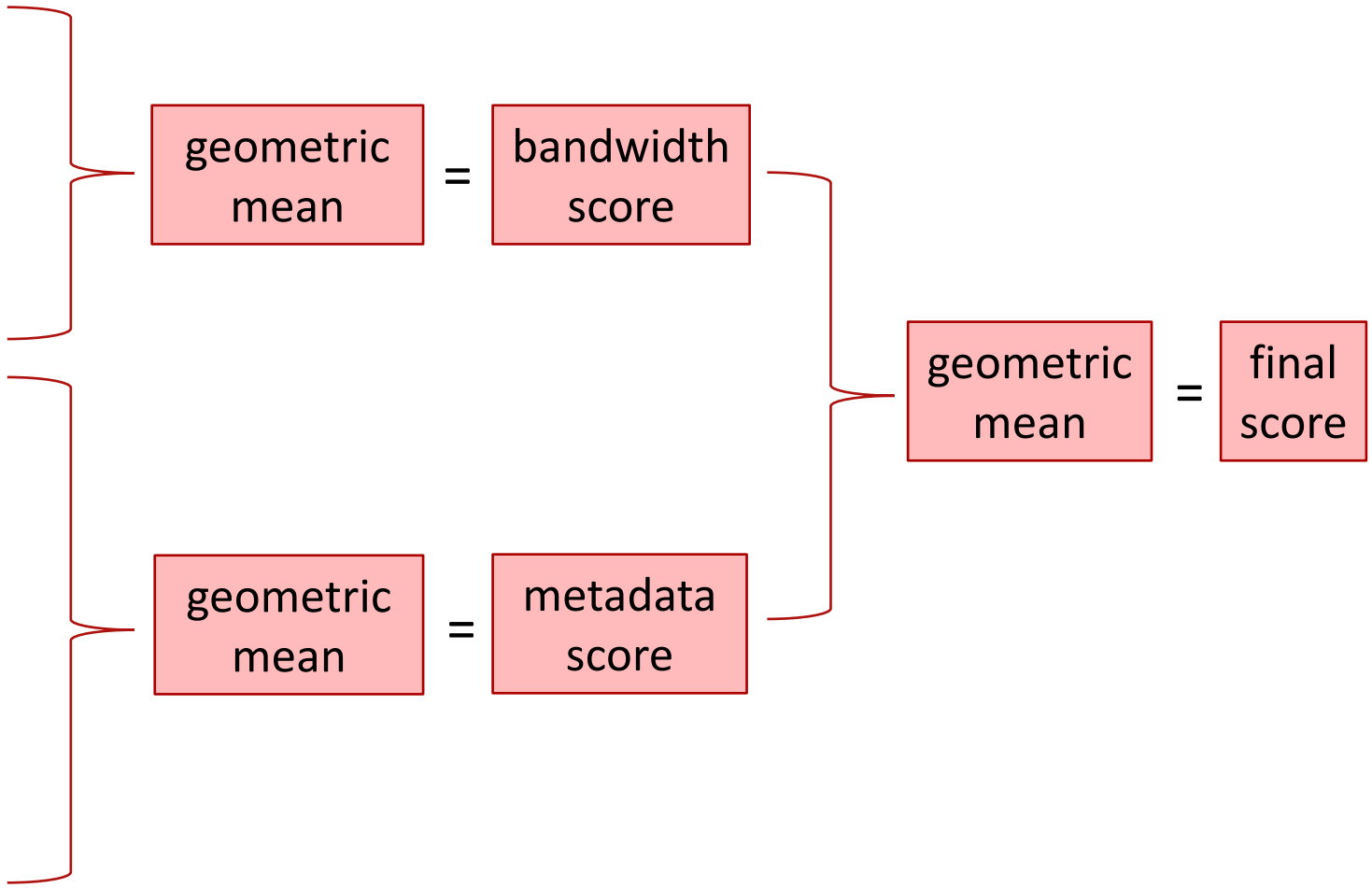


IO-500 | Balance Demonstrated Via Multiple Measurements

- Hero bandwidth
  - Write and read - 5 minutes, typically with large, aligned chunks
- Anti-hero bandwidth
  - Write and read - 5 minutes, unaligned, interleaved, multi-client
- Hero metadata
  - Create, stat, delete - 5 minutes, separate directory per thread
- Anti-hero metadata
  - Create, stat, read, delete - 5 minutes, single directory for all threads
- And a namespace search
  - Search created files - parallel if you have it

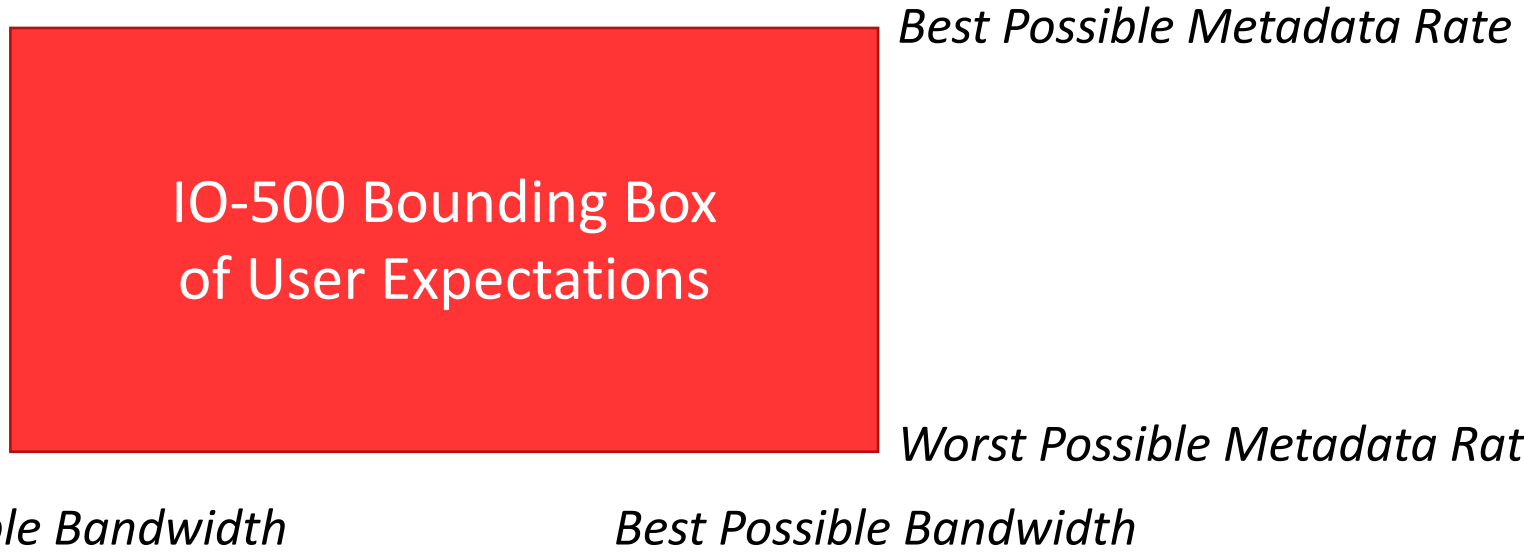
IO-500 | Aggregate Final Score Difficult to Game

- Hero bandwidth
  - Write and read
- Anti-hero bandwidth
  - Write and read
- Hero metadata
  - Create, stat, delete
- Anti-hero metadata
  - Create, stat, read, delete
- And a namespace search
  - Search created files



IO-500 | Bounding Box of Expectation

- "We tried 20 years ago. It's impossible to create a single representative benchmark."
  - Great point! We won't try, our bounding box includes them all.

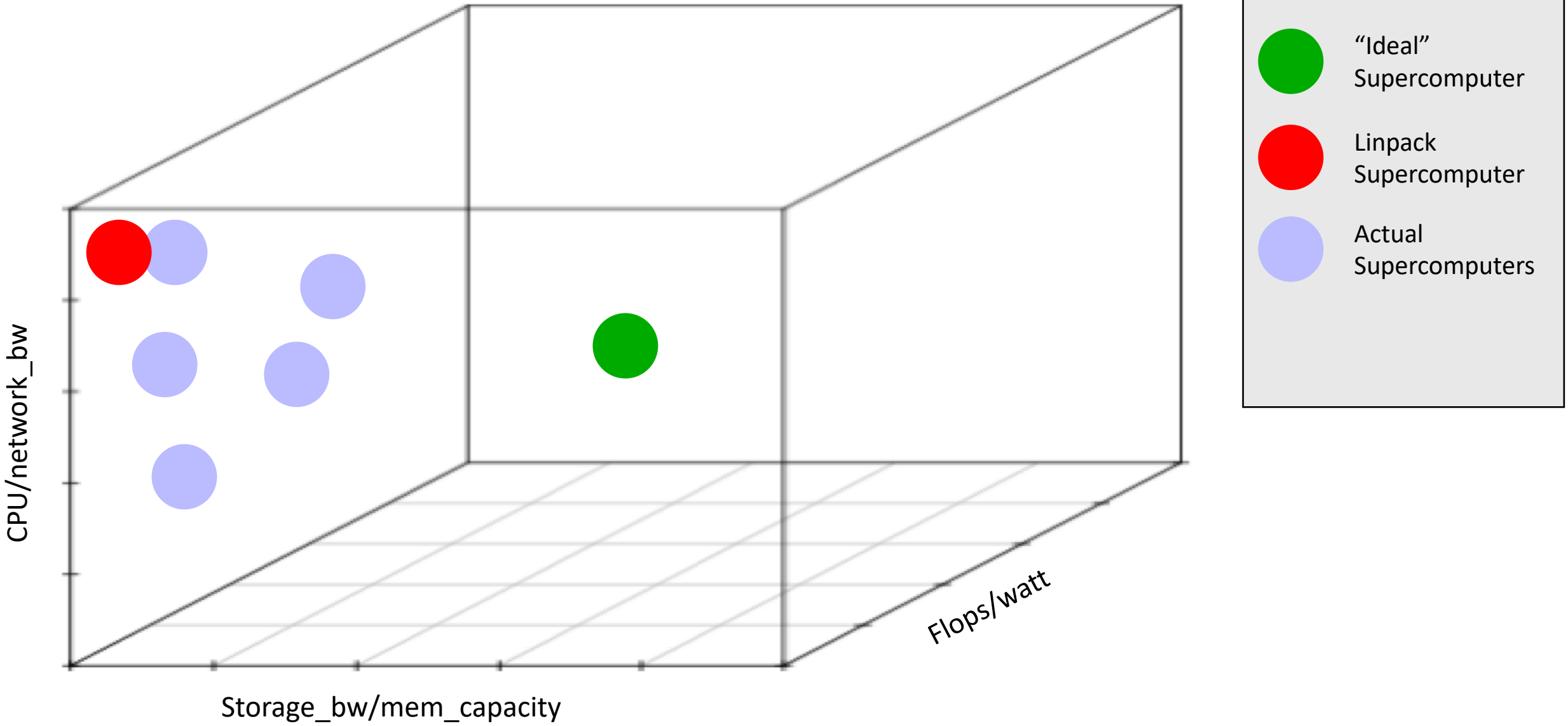


BOLD CLAIM

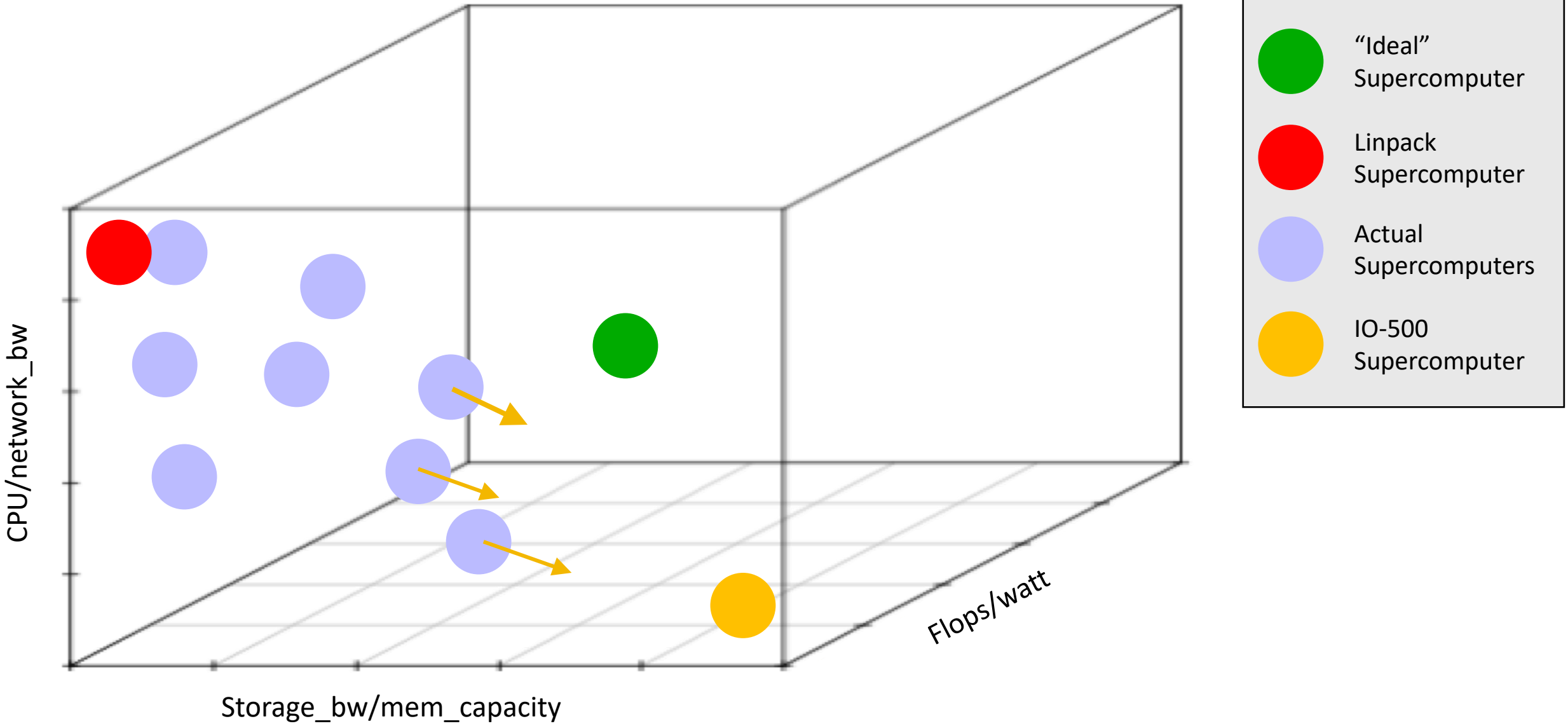
IO-500 cannot be gamed.

Whatever is done to improve IO-500 scores will result in a better storage system for applications.

IO-500 | IO-500 Restores Balance



IO-500 | Brings Balance to New Systems



IO-500 | Third List at SC'18

#	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*

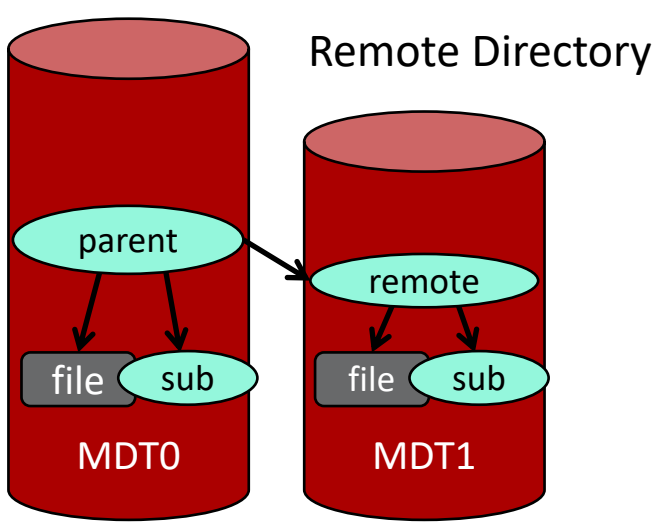
IO-500 | History of Lustre Submissions

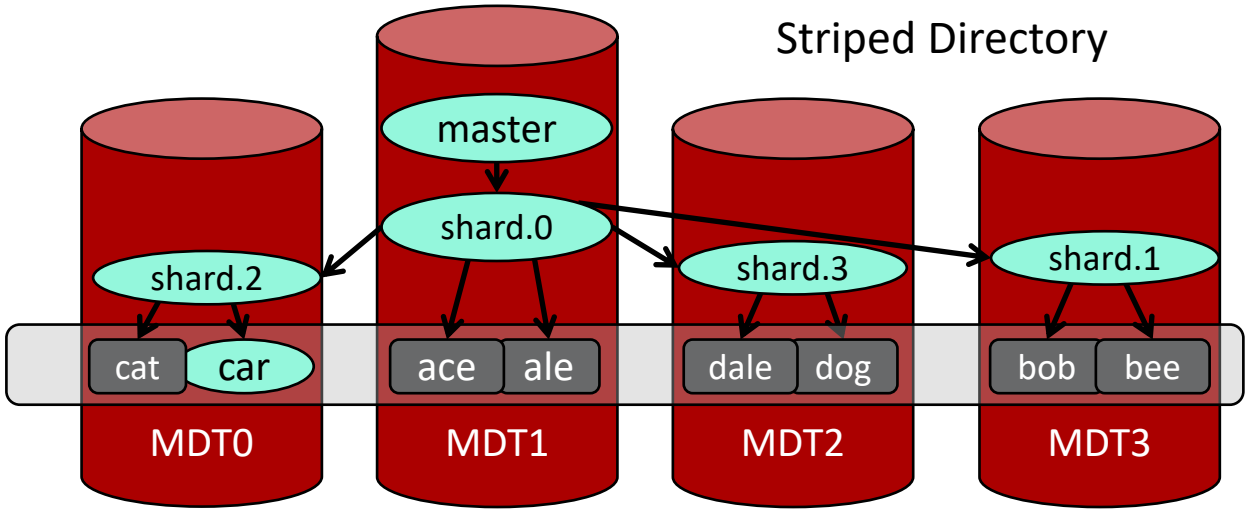
List	Event	Total Site Submissions	Total Lustre Submissions	Highest Lustre Rank
1	SC'17	9	3	3
2	ISC'18	12	5	3
3	SC'18	26	8	2
4	ISC'19	?	More, please	?

- Lustre is about 70% of Top-100 systems, so should make up more than 30-40% of the IO-500
- I encourage sites to submit results for new and old storage systems, 10-Client Challenge
- Sites can submit separate results for different storage tiers, different parameters, etc.
  - Only the top score for each storage system will be a candidate for the bi-annual IO-500 list

IO-500 | Lustre Metadata Tuning Tips

- Lustre supports two types of distributed metadata
  - Remote Directory (DNE1) and Striped Directory (DNE2)





  - Files and subdirectories stay on a local MDT by default, want to use all MDTs
  - Use "`lfs mkdir`" to stripe test directory and inherited default over all MDTs

```
# lfs mkdir -c 4 /scratch0/io500
# lfs mkdir -c 4 -D /scratch0/io500
```


IO-500 | Lustre I/O Tuning Tips

- Client Side setting

```
# lctl set_param \
    osc.*.max_pages_per_rpc=16M \
    osc.*.max_rpcs_in_flight=16 \
    llite.*.max_read_ahead_mb=2048 \
    osc.*.checksums=0
```

Send more aggressive RPCs to server and readahead

- Server Side setting

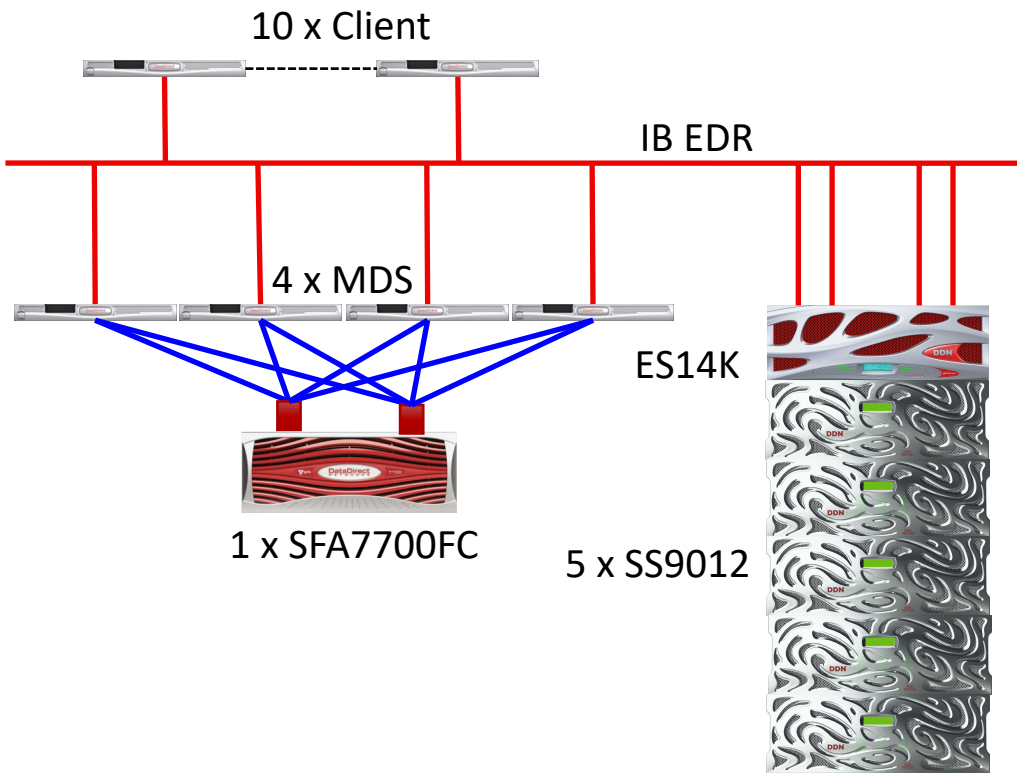
```
# lctl set_param \
    osd-ldiskfs.*.read_cache_enable=0
    obdfilter.*.writethrough_cache_enable=0
```

Avoiding Page Cache on OSS when it flushes data to disk

- That's all changed parameters and configuration from default EXAScaler setting

IO-500 | EXAScaler 5.0-pre Test Configuration (SC'18)

10-Client Challenge "Bancholab"



MDS/MDT	4 x MDS 1x Platinum 8160, 96GB RAM, 1 x IB EDR 1 x SFA7700 FC 2 x MDT (2 x RAID1 **SSD**) per MDS
OSS/OST	1 x ES14K + 5 x SS9012 408 x NL-SAS 10TB **HDD** 8 x DCR POOL (51/MR=1) 4 x vOSS (8 CPU Core, 90GB RAM, 1x IB EDR)
Client	10 x Intel Server 2 x E5-2650v4, 128GB RAM, 1x IB EDR CentOS7.4
Software	EXAScaler 5.0-pre Lustre-2.12-rc1

IO-500 | EXAScaler 5.0-pre Test Results (SC'18)

10-Client Challenge "Bancholab"

```
$ git clone https://github.com/VI4IO/io-500-dev
$ cd io-500-dev
$ ./utilities/prepare.sh
$ vi io500.sh                # provides fairly good directions for what to edit in the file
$ ./io500.sh                 # to run directly, otherwise via batch submission script
```

```
[RESULT] BW   phase 1            ior_easy_write              37.540 GB/s : time 343.38 seconds
[RESULT] IOPS phase 1         mdtest_easy_write             199.685 kiops : time 325.87 seconds
[RESULT] BW   phase 2            ior_hard_write               0.262 GB/s : time 300.21 seconds
[RESULT] IOPS phase 2         mdtest_hard_write              24.348 kiops : time 395.55 seconds
[RESULT] IOPS phase 3                      find            3332.110 kiops : time  21.96 seconds
[RESULT] BW   phase 3             ior_easy_read              35.374 GB/s : time 364.41 seconds
[RESULT] IOPS phase 4          mdtest_easy_stat             527.669 kiops : time 124.08 seconds
[RESULT] BW   phase 4             ior_hard_read               4.627 GB/s : time  17.03 seconds
[RESULT] IOPS phase 5          mdtest_hard_stat              79.476 kiops : time 106.64 seconds
[RESULT] IOPS phase 6        mdtest_easy_delete             226.094 kiops : time 288.22 seconds
[RESULT] IOPS phase 7          mdtest_hard_read              46.141 kiops : time 182.72 seconds
[RESULT] IOPS phase 8        mdtest_hard_delete              58.842 kiops : time 143.64 seconds
[SCORE] Bandwidth 6.33725 GB/s : IOPS 159.413 kiops : TOTAL 31.7843
```


Thank You!

Keep in touch with us.

sales@ddn.com

@ddn_limitless

company/datadirect-networks

9351 Deering Avenue
Chatsworth, CA 91311

1.800.837.2298
1.818.700.4000