



**IO** 500

Creating a Perfect,  
Completely Not Quixotic,  
why would you even think that,  
Storage Benchmark

John Bent  
jbent@ddn.com  
DDN Field CTO  
Data over Distance 2018  
July 19, 2019



# IO500 | My Most Recent Motivation: “How Fast Does a Disk Drive Go?”

- ▶ Firmware engineer: “150 – 160 MB/s”
- ▶ Marketer: “200 MB/s”
- ▶ Performance engineer: “130 MB/s”
- ▶ Salesperson: “100 MB/s”
  
- ▶ User: “Why do I only see 10 MB/s?!?”

# IO500 | My Original Motivation

- ▶ Experience at LANL with hard IO patterns
  - There are easy IO patterns and hard IO patterns
  - Performance divergence is extreme
  - In capacity systems, easy never survives
  - Vendors/community were focused on easy
  - But hard IO patterns are more important

How to get community to embrace the challenge of hard IO patterns?



# IO500 | Motivation Summary

- ▶ **More honesty from vendors**
  - Empathy for users who don't know who to believe
- ▶ **Create realistic expectations**
  - Empathy for users who don't know what to expect
- ▶ **Community repository**
  - Empathy for users who don't know how to tune
- ▶ **More balanced systems**
  - Empathy for users who run on imbalanced systems
- ▶ **Easier RFP writing**
  - Empathy for procurers who struggle to define an ideal system
- ▶ **Better storage**
  - Force vendors to focus on the problems of real users

# IO500 | Demotivation Summary

“Oh, hmm. Thanks John, but you shouldn’t invite us to your BoF. Trust me, you don’t want us coming to your BoF.”

“Dumbest idea ever.”

“Uh, shouldn’t it be called the IO-9? Heh, heh.”

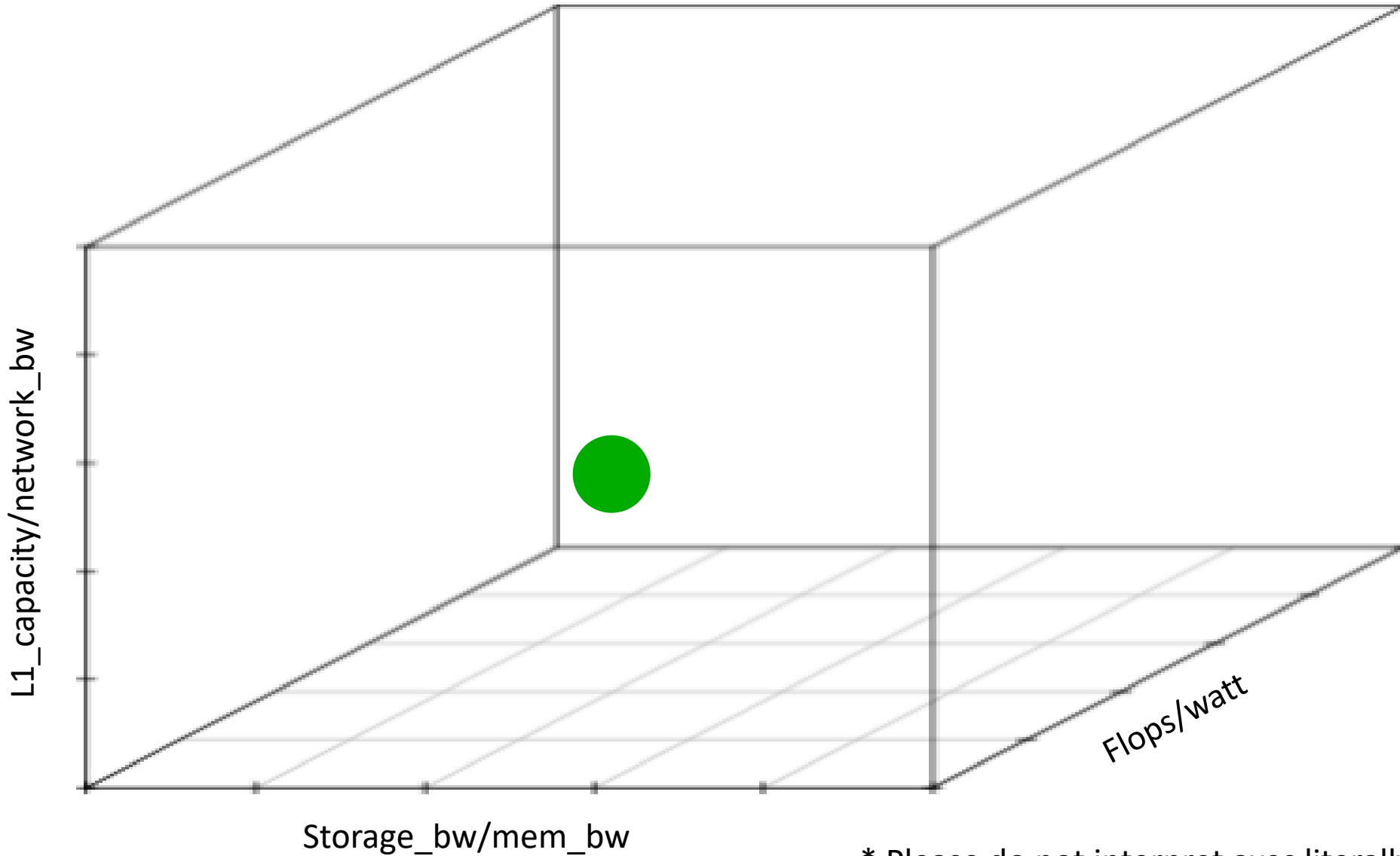
“Last thing the world needs is another overly simple benchmark like Linpack that is driving supercomputing into the toilet.”


“So many better things to do. Please don’t waste your time, and the community’s time, on this.”

“Seriously? You’re joking, right? Ugh, you’re actually serious.”

“We tried 20 years ago to do this. It’s impossible to create a single representative benchmark.”

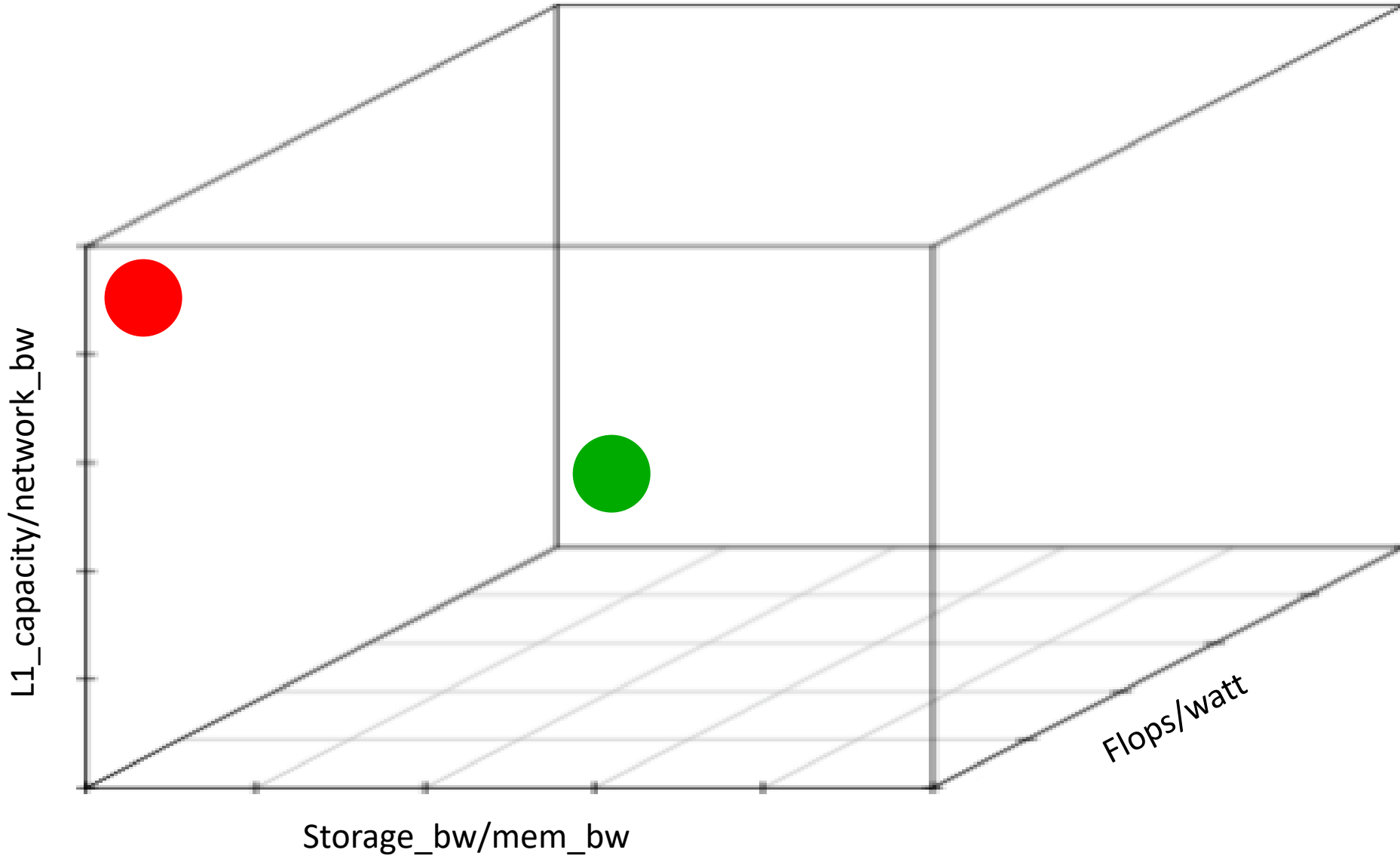
# IO500 | A Legitimate Concern About Linpack



 "Ideal" Supercomputer

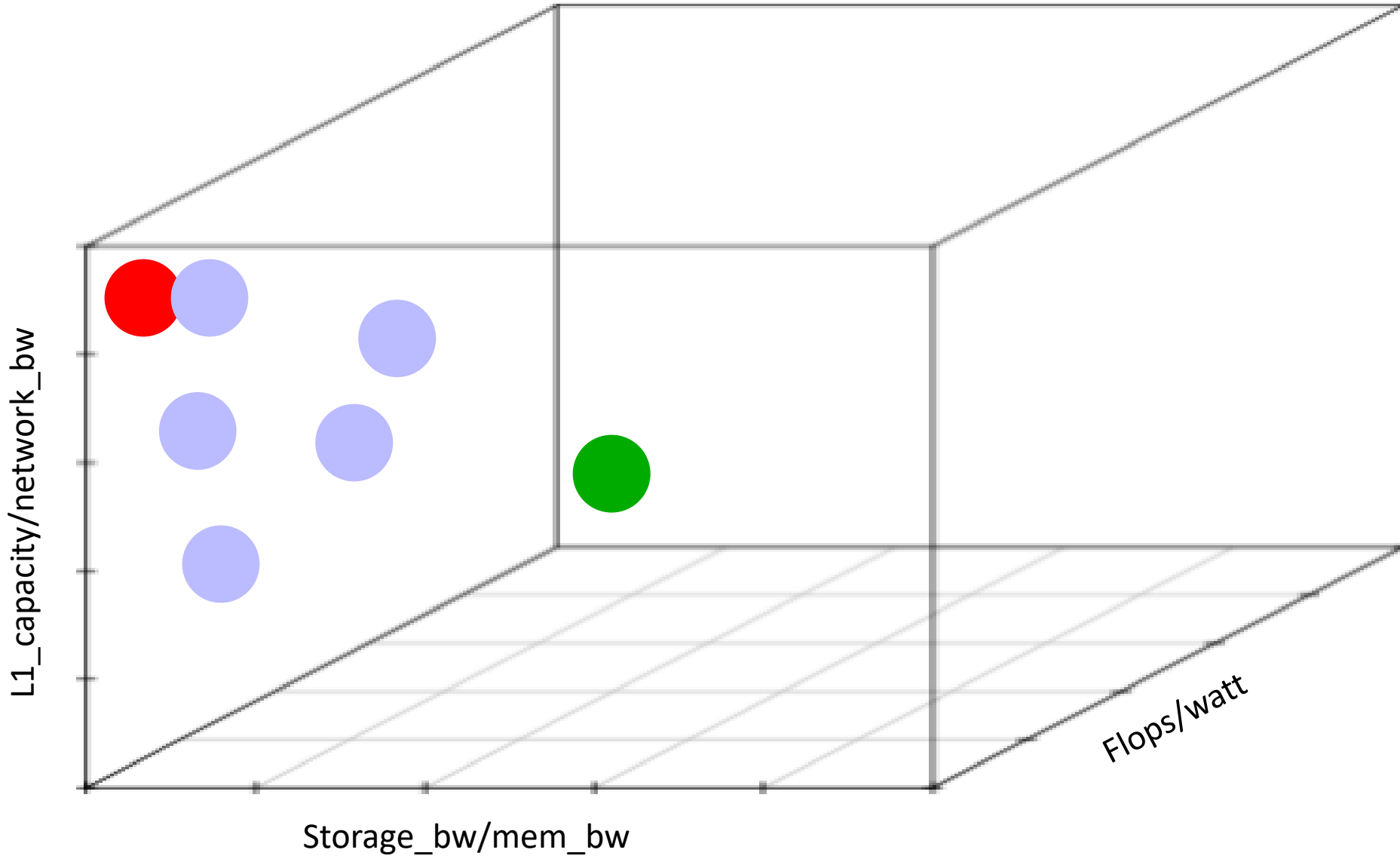
\* Please do not interpret axes literally.  
Just examples illustrating multi-variable complexity.

# IO500 | A Legitimate Concern About Linpack



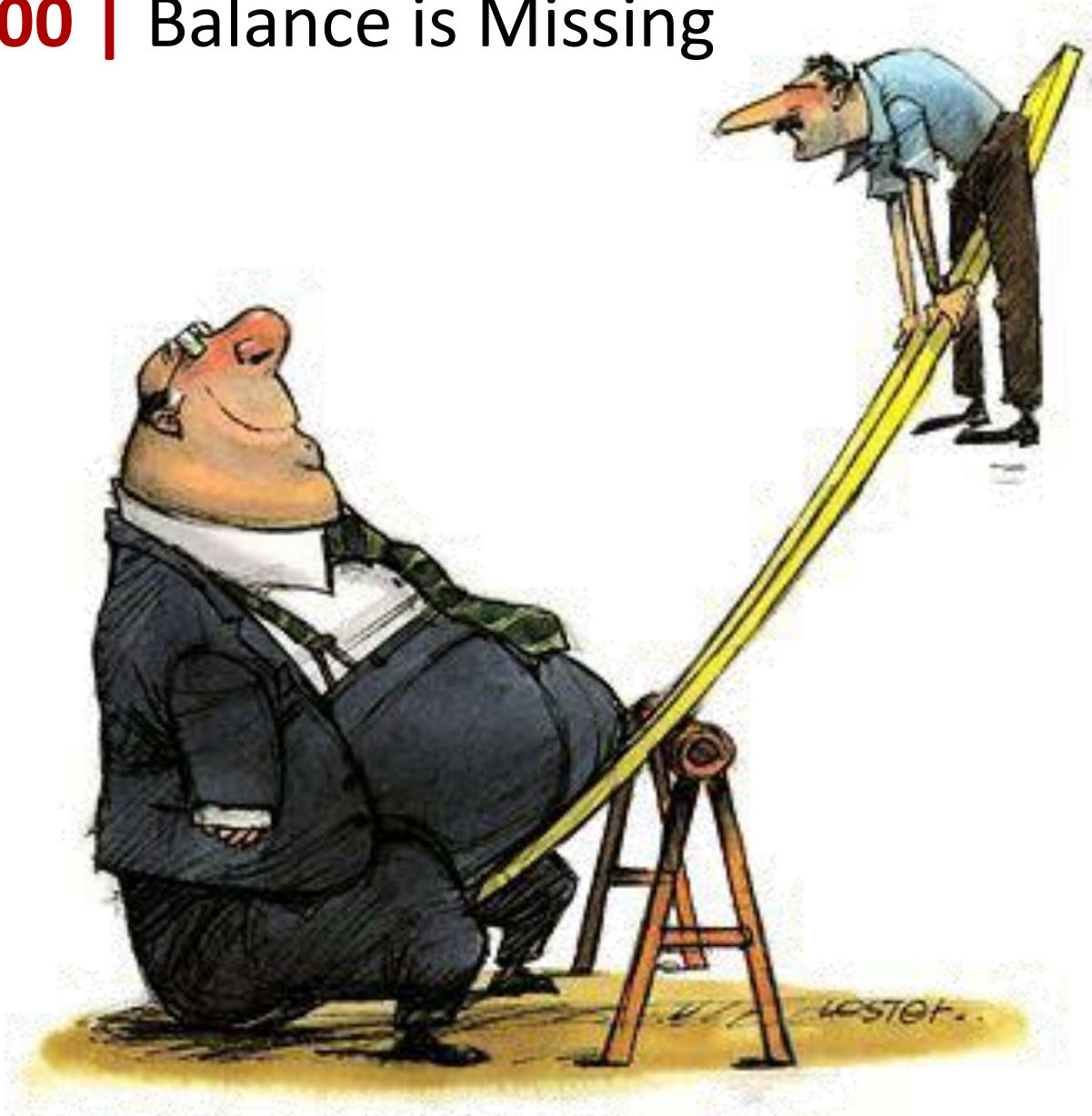
- "Ideal" Supercomputer
- Linpack Supercomputer

# IO500 | A Legitimate Concern About Linpack





# IO500 | Balance is Missing



## Lesson and Goal

IO500 must itself be balanced and, in being so, will help restore balance to supercomputing.

# IO500 | IO500 is Balanced

## ▶ 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

# IO500 | IO500 is Balanced

## ▶ Hero bandwidth

- Write and read

## ▶ Anti-hero bandwidth

- Write and read

## ▶ Hero metadata

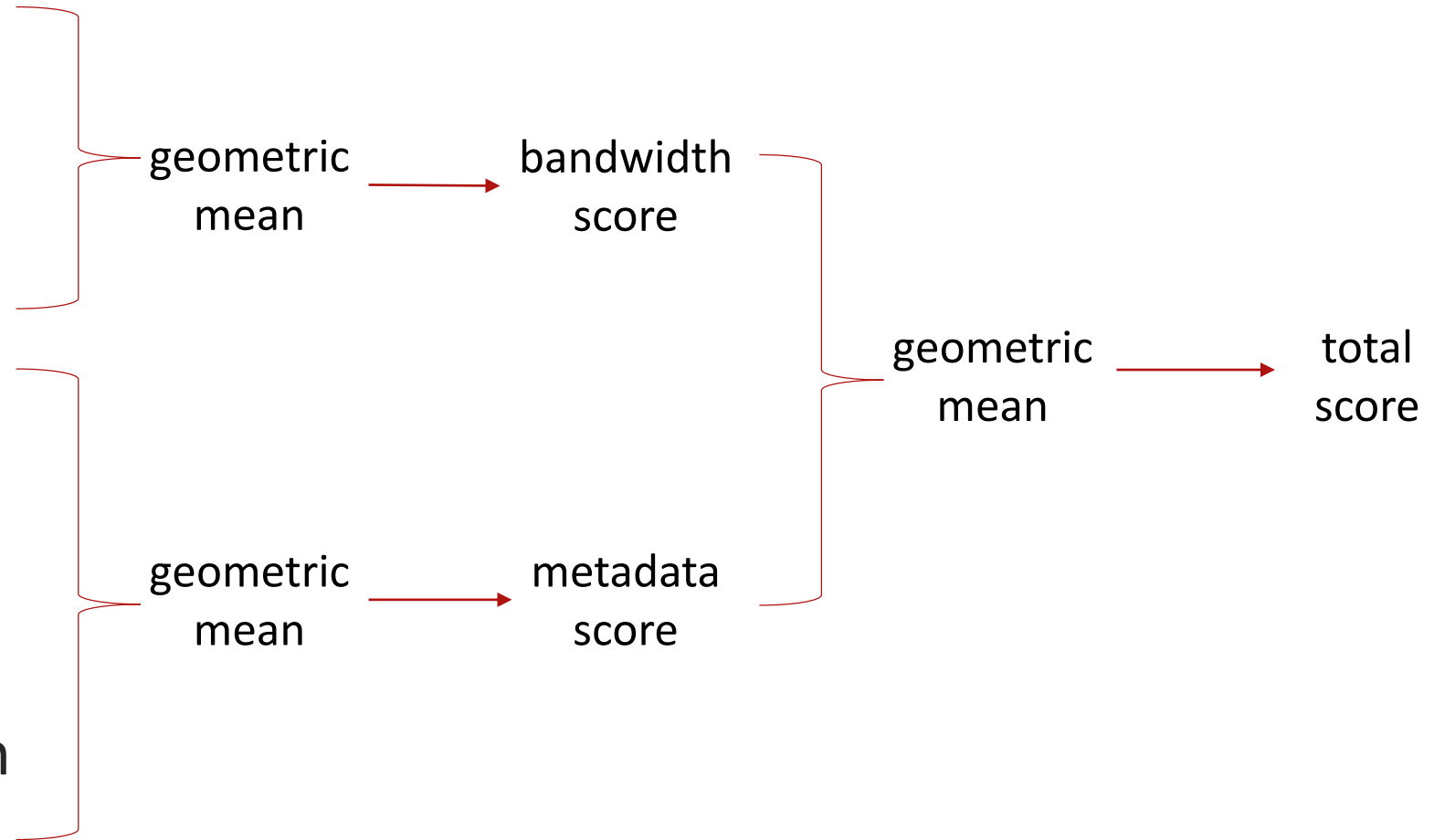
- Create, stat, delete

## ▶ Anti-hero metadata

- Create, stat, read, delete

## ▶ And a namespace search

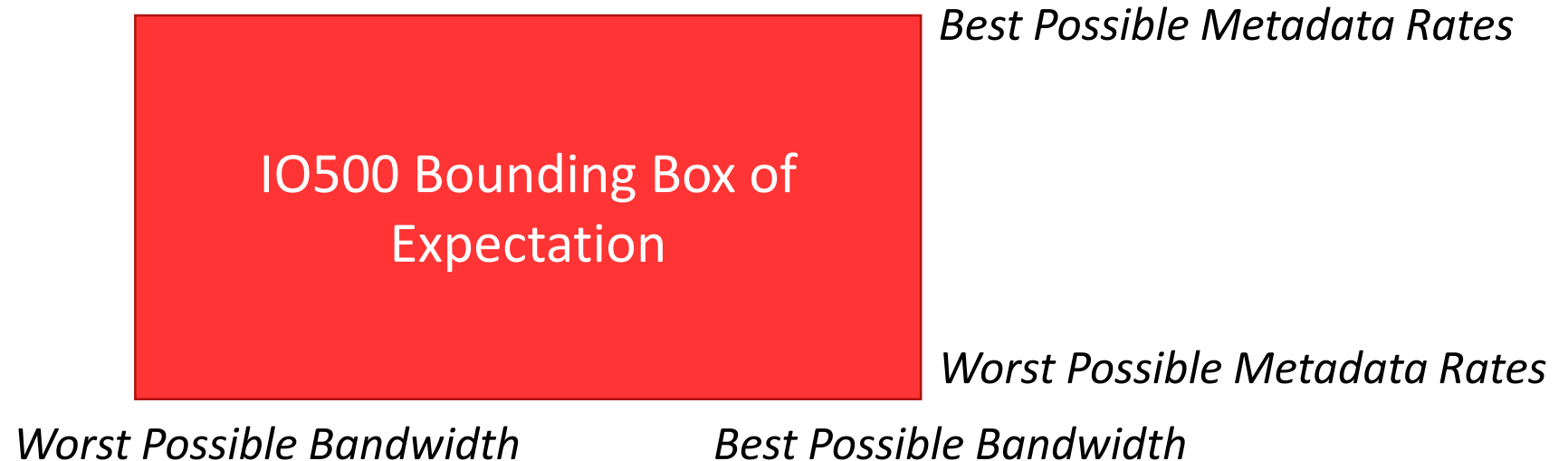
- Search



# IO500 | Bounding Box of Expectation

▶ “We tried 20 years ago. Impossible to create a single representative benchmark.”

- Great point! We won't try. Our bounding box includes them all.



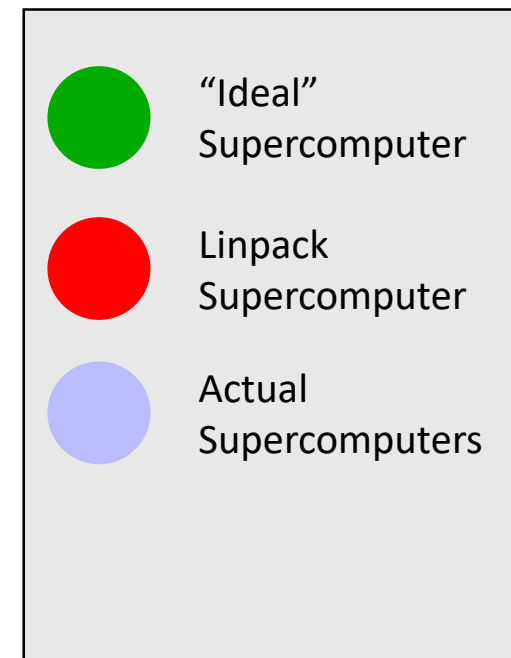
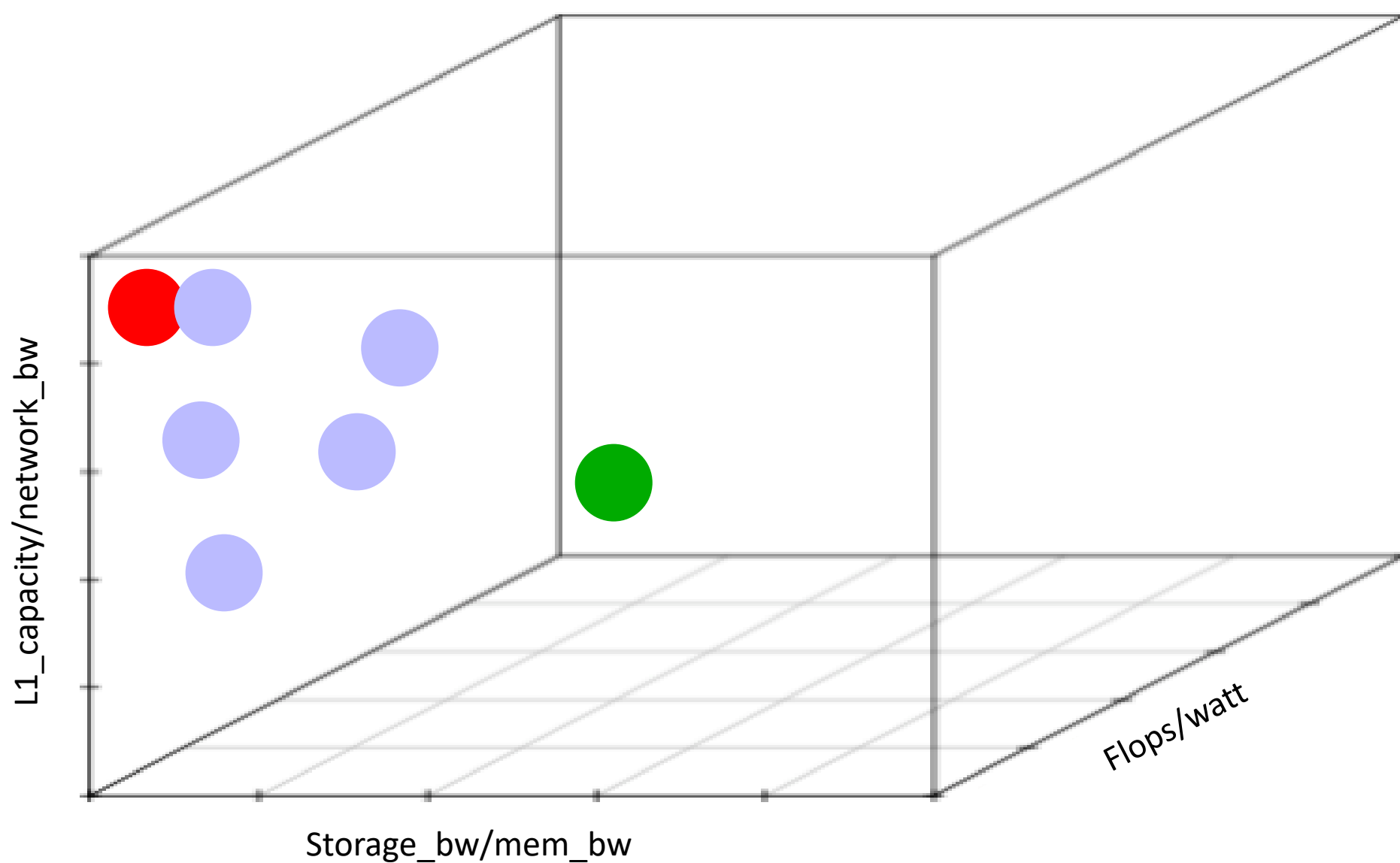
## BOLD CLAIM

IO500 cannot be gamed.

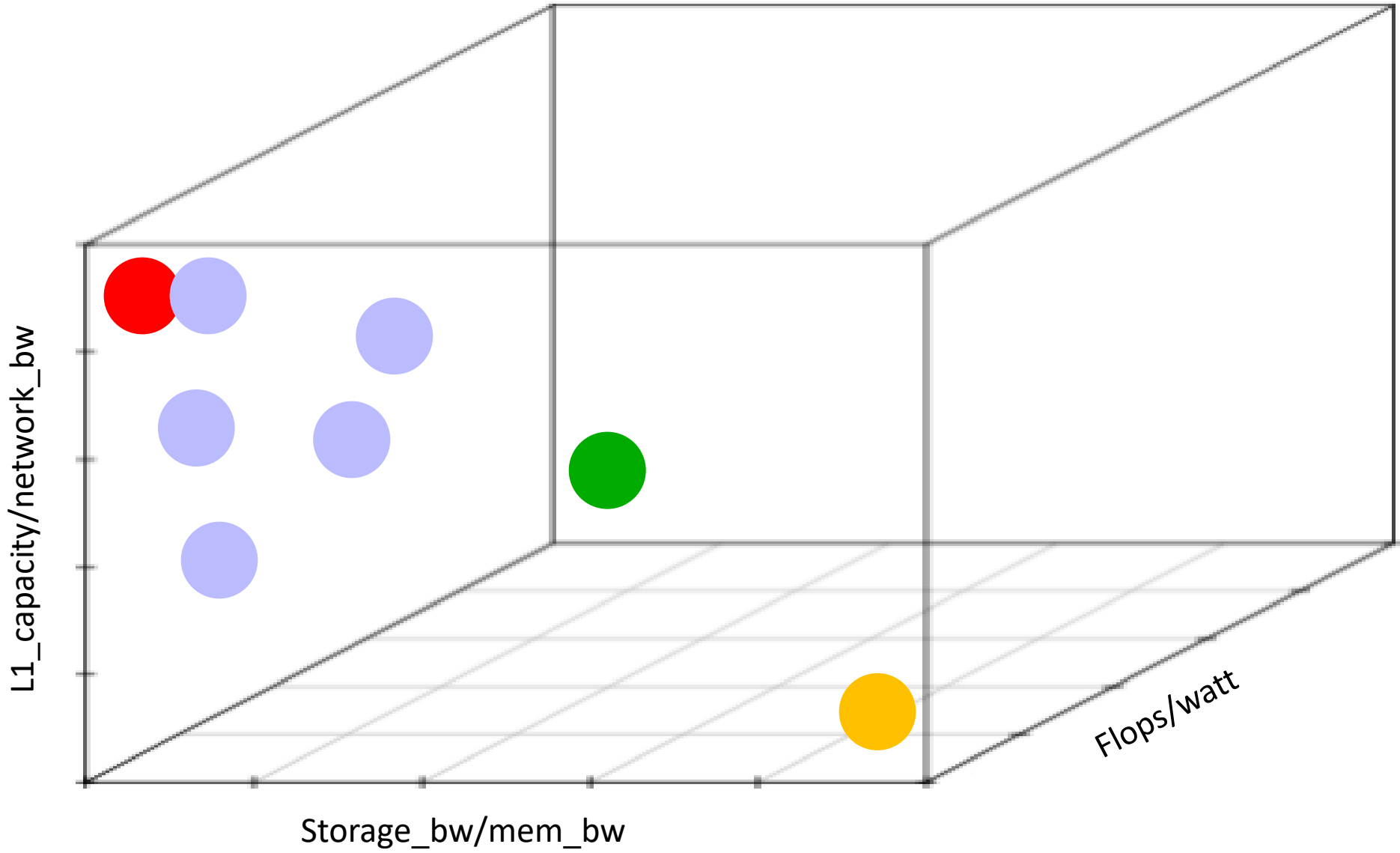
Whatever you do to improve your IO500 score will result in a better storage system for applications.





Prove me wrong. 😊

# IO500 | IO500 Restores Balance



# IO500 | IO500 Restores Balance



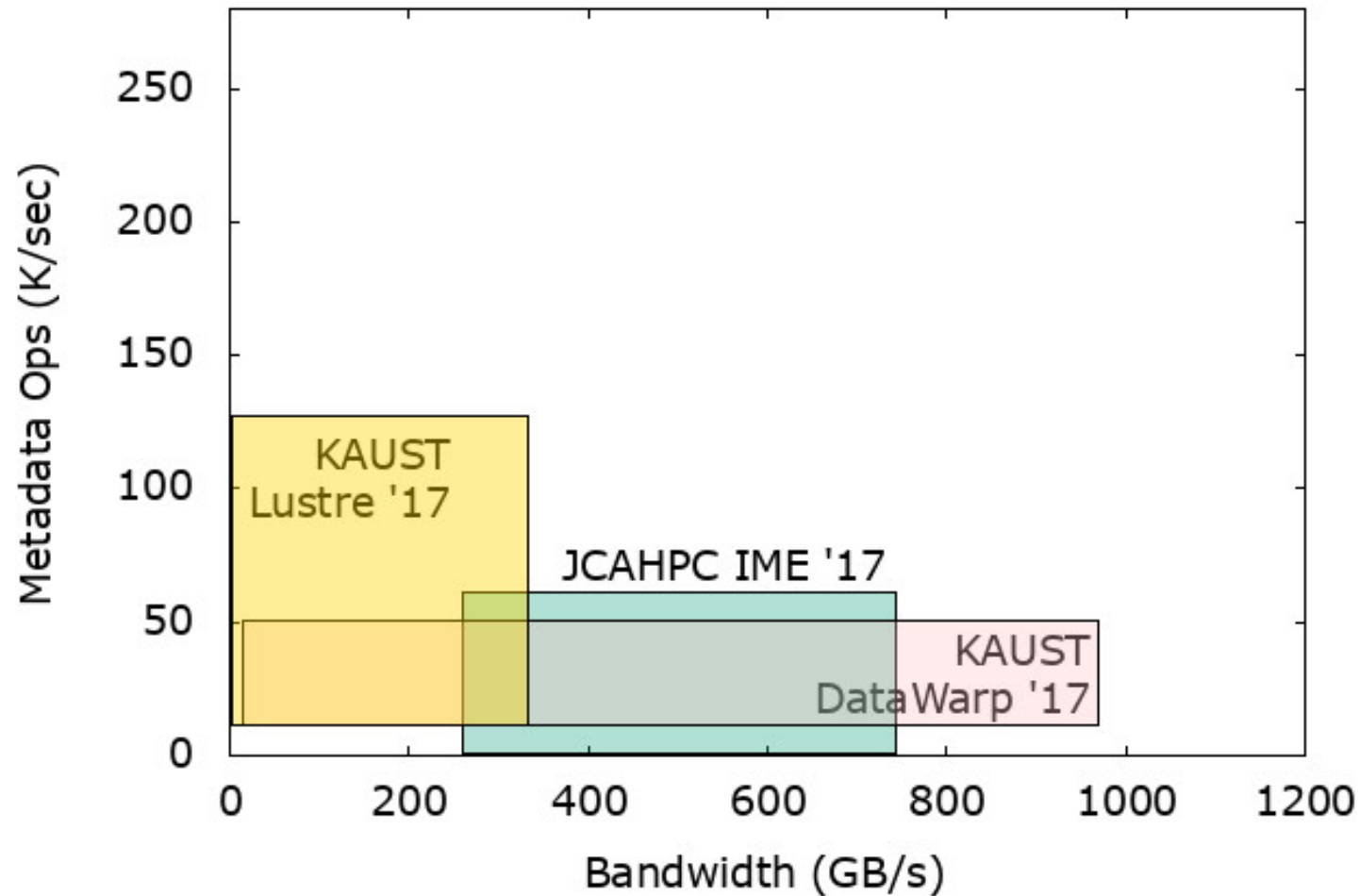
-  "Ideal" Supercomputer
-  Linpack Supercomputer
-  Actual Supercomputers
-  IO500 Supercomputer

# IO500 | First List at SC'17

| # | information    |             |                |              | io500  |        |        | ior        |           |            |           |
|---|----------------|-------------|----------------|--------------|--------|--------|--------|------------|-----------|------------|-----------|
|   | system         | institution | filesystem     | client nodes | score  | bw     | md     | easy write | easy read | hard write | hard read |
|   |                |             |                |              |        | GiB/s  | kIOP/s | GiB/s      | GiB/s     | GiB/s      | GiB/s     |
| 1 | Oakforest-PACS | JCAHPC      | IME            | 2048         | 101.48 | 471.25 | 21.85  | 742.38     | 427.41    | 600.28     | 258.93    |
| 2 | Shaheen        | Kaust       | DataWarp       | 300          | 70.90  | 151.53 | 33.17  | 969.45     | 894.76    | 15.55      | 39.09     |
| 3 | Shaheen        | Kaust       | Lustre         | 1000         | 41.00  | 54.17  | 31.03  | 333.03     | 220.62    | 1.44       | 81.38     |
| 4 | JURON          | JSC         | BeeGFS         | 8            | 35.77  | 14.24  | 89.83  | 30.42      | 48.36     | 1.46       | 19.16     |
| 5 | Mistral        | DKRZ        | Lustre         | 100          | 32.15  | 22.77  | 45.39  | 158.19     | 163.62    | 1.53       | 6.79      |
| 6 | Sonasad        | IBM         | Spectrum Scale | 10           | 21.63  | 4.57   | 102.38 | 34.13      | 32.25     | 0.17       | 2.33      |
| 7 | Seislab        | Fraunhofer  | BeeGFS         | 24           | 18.75  | 5.13   | 68.58  | 18.79      | 22.34     | 0.89       | 1.86      |
| 8 | EMSL Cascade   | PNNL        | Lustre         | 126          | 11.17  | 4.88   | 25.57  | 17.81      | 30.19     | 0.39       | 2.72      |
| 9 | Serrano        | SNL         | Spectrum Scale | 16           | 4.25   | 0.65   | 27.98  | 1.08       | 1.03      | 0.22       | 0.71      |

# IO500 | Some Analysis of First List

IO500 SC2017 Bounding Box of Expectations



|                |     |
|----------------|-----|
| JCAHPC IME     | 101 |
| KAUST DataWarp | 71  |
| KAUST Lustre   | 41  |



# IO500 | We suggest a default ordering, but it's flexible

This is the official list from Supercomputing 2017.

| # | information |                |             |                | io500        |       |       |        |
|---|-------------|----------------|-------------|----------------|--------------|-------|-------|--------|
|   | Equation    | system         | institution | filesystem     | client nodes | score | bw    | md     |
|   |             |                |             |                |              |       | GiB/s | kIOP/s |
| 1 | 4.47        | JURON          | JSC         | BeeGFS         |              |       |       |        |
| 2 | 2.16        | Sonasad        | IBM         | Spectrum Scale |              |       |       |        |
| 3 | 0.78        | Seislab        | Fraunhofer  | BeeGFS         |              |       |       |        |
| 4 | 0.32        | Mistral        | DKRZ        | Lustre         |              |       |       |        |
| 5 | 0.27        | Serrano        | SNL         | Spectrum Scale |              |       |       |        |
| 6 | 0.24        | Shaheen        | Kaust       | DataWarp       |              |       |       |        |
| 7 | 0.09        | EMSL Cascade   | PNNL        | Lustre         |              |       |       |        |
| 8 | 0.05        | Oakforest-PACS | JCAHPC      | IME            |              |       |       |        |
| 9 | 0.04        | Shaheen        | Kaust       | Lustre         |              |       |       |        |



Congrats to @fzj\_jsc team for achieving the best per-client performance with their BeeGFS file system on Juron in the new #IO500 list at #SC17. Whitepaper for Juron is here: [beegfs.io/docs/whitepape...](http://beegfs.io/docs/whitepape...)

## Controls

Equation

Add column

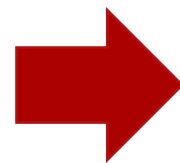
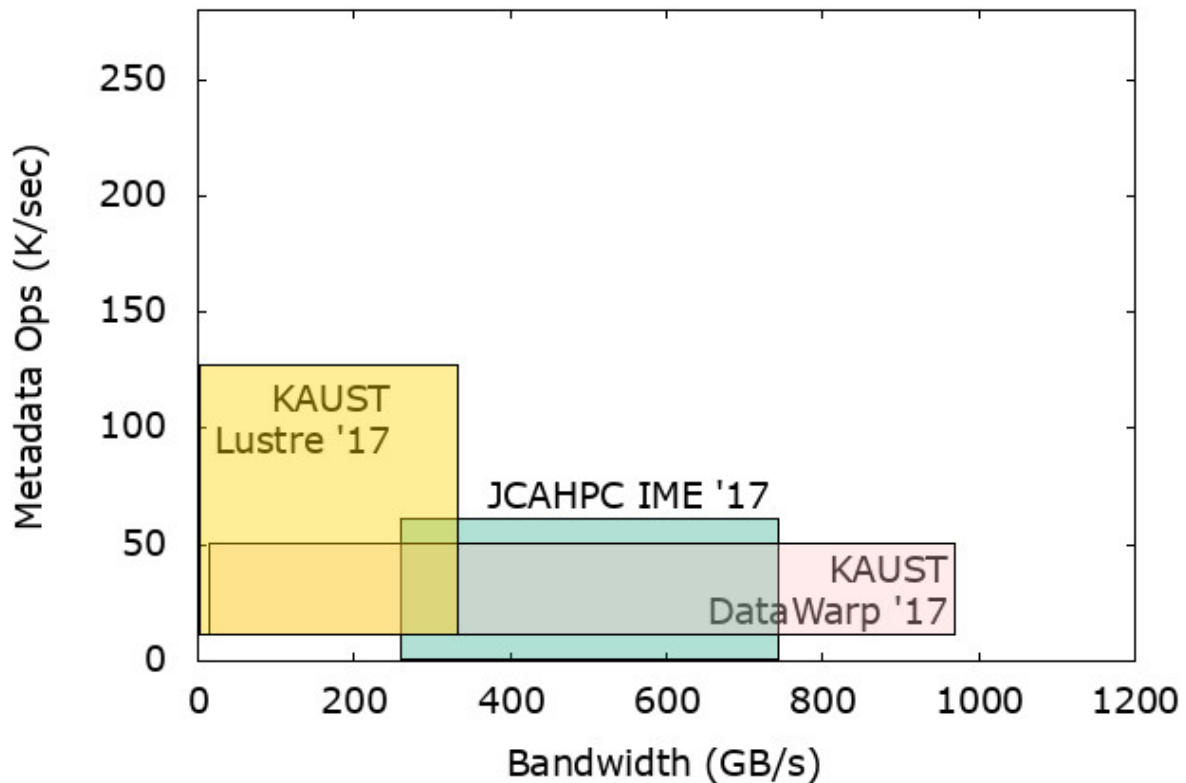
Remove column

# IO500 | Second List at ISC'18

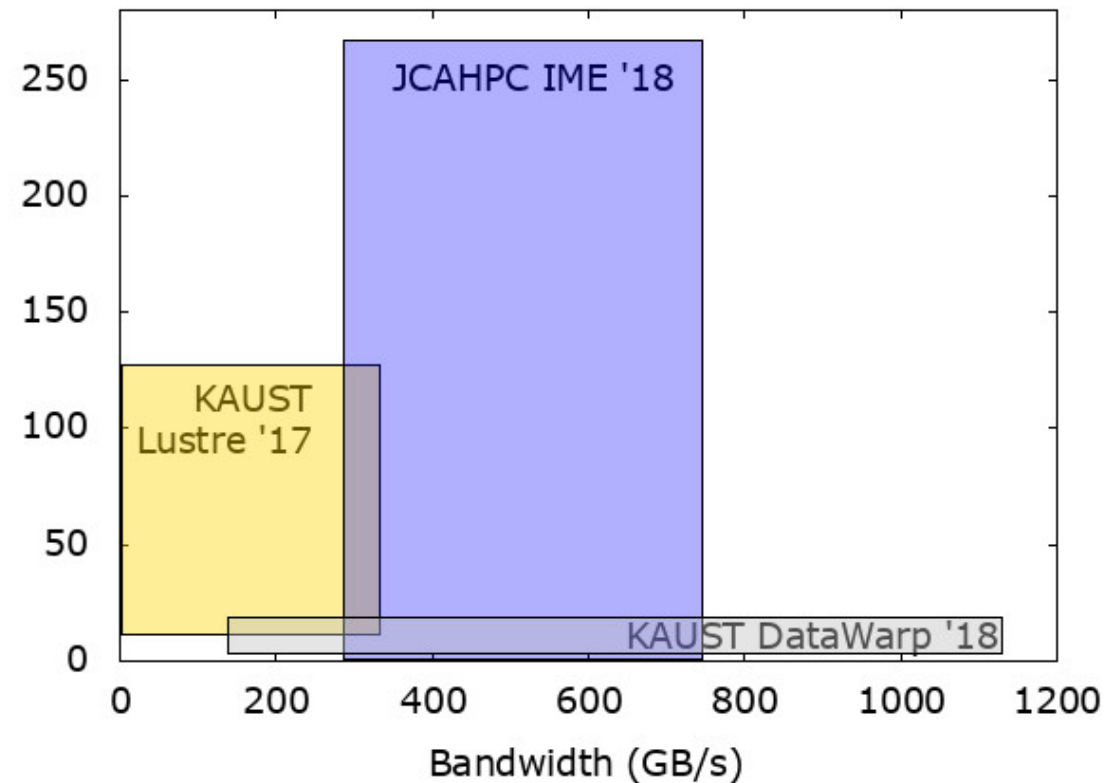
| #  | information    |                                      |                |                |              |      | io500  |        |        |
|----|----------------|--------------------------------------|----------------|----------------|--------------|------|--------|--------|--------|
|    | system         | institution                          | filesystem     | storage vendor | client nodes | data | score  | bw     | md     |
|    |                |                                      |                |                |              |      |        | GiB/s  | kIOP/s |
| 1  | Oakforest-PACS | JCAHPC                               | IME            | DDN            | 2048         | zip  | 137.78 | 560.10 | 33.89  |
| 2  | ShaheenII      | KAUST                                | DataWarp       | Cray           | 1024         | zip  | 77.37  | 496.81 | 12.05  |
| 3  | ShaheenII      | KAUST                                | Lustre         | Cray           | 1000         |      | 41.00* | 54.17  | 31.03* |
| 4  | JURON          | JSC                                  | BeeGFS         | ThinkparQ      | 8            |      | 35.77* | 14.24  | 89.81* |
| 5  | Mistral        | DKRZ                                 | Lustre2        | Seagate        | 100          |      | 32.15  | 22.77  | 45.39  |
| 6  | Sonasad        | IBM                                  | Spectrum Scale | IBM            | 10           | zip  | 24.24  | 4.57   | 128.61 |
| 7  | Seislab        | Fraunhofer                           | BeeGFS         | ThinkparQ      | 24           |      | 16.96  | 5.13   | 56.14  |
| 8  | Mistral        | DKRZ                                 | Lustre1        | Seagate        | 100          | zip  | 15.47  | 12.68  | 18.88  |
| 9  | Govorun        | Joint Institute for Nuclear Research | Lustre         | RSC            | 24           | zip  | 12.08  | 3.34   | 43.65  |
| 10 | EMSL Cascade   | PNNL                                 | Lustre         |                | 126          |      | 11.12  | 4.88   | 25.33  |
| 11 | Serrano        | SNL                                  | Spectrum Scale | IBM            | 16           |      | 4.25*  | 0.65   | 27.98* |
| 12 | Jasmin/Lotus   | STFC                                 | NFS            | Purestorage    | 64           | zip  | 2.33   | 0.26   | 20.93  |

# IO500 | Some Analysis of Second List

2017 Top Three



2018 Top Three



|                |     |                  |
|----------------|-----|------------------|
| JCAHPC IME     | 101 | 138              |
| KAUST DataWarp | 71  | 77               |
| KAUST Lustre   | 41  | Did not resubmit |



Please join us at SC18 BOF and please consider submitting your own results to the list.

**Thank You!**

Keep in touch with us.

- > git clone <https://github.com/VI4IO/io-500-dev>
- > cd io-500-dev
- > ./utilities/prepare.sh
- > ./io500.sh
- > email io-500@vi4io.org –m “Help with tuning please?”
- > tar cz . | email submit@io500.org –m “Submission attached”
- > wget <http://io500.org> | email mom –m “Hey mom, I’m on IO500!”