facebook

Hybrid XFS: Supercharging HDDs with SSDs

Skanda Shamasunder

Production Engineer (Storage), Facebook

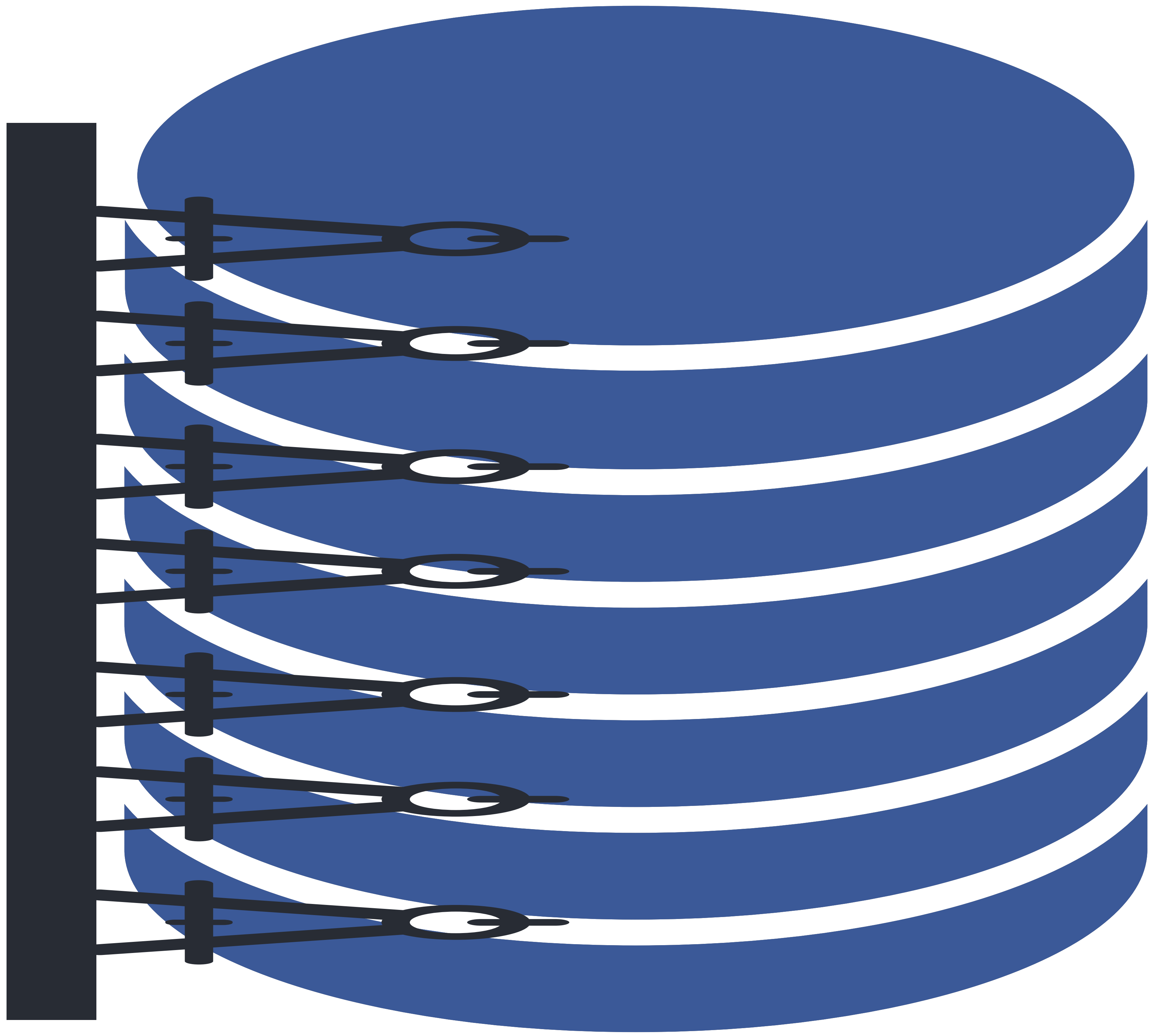
Who am I?

Outline

1. The IO wall
2. The Opportunity
3. The Risks
4. The Analysis
5. Rolling it out
6. Lessons

The IO wall

Image by PublicDomainPictures from Pixabay



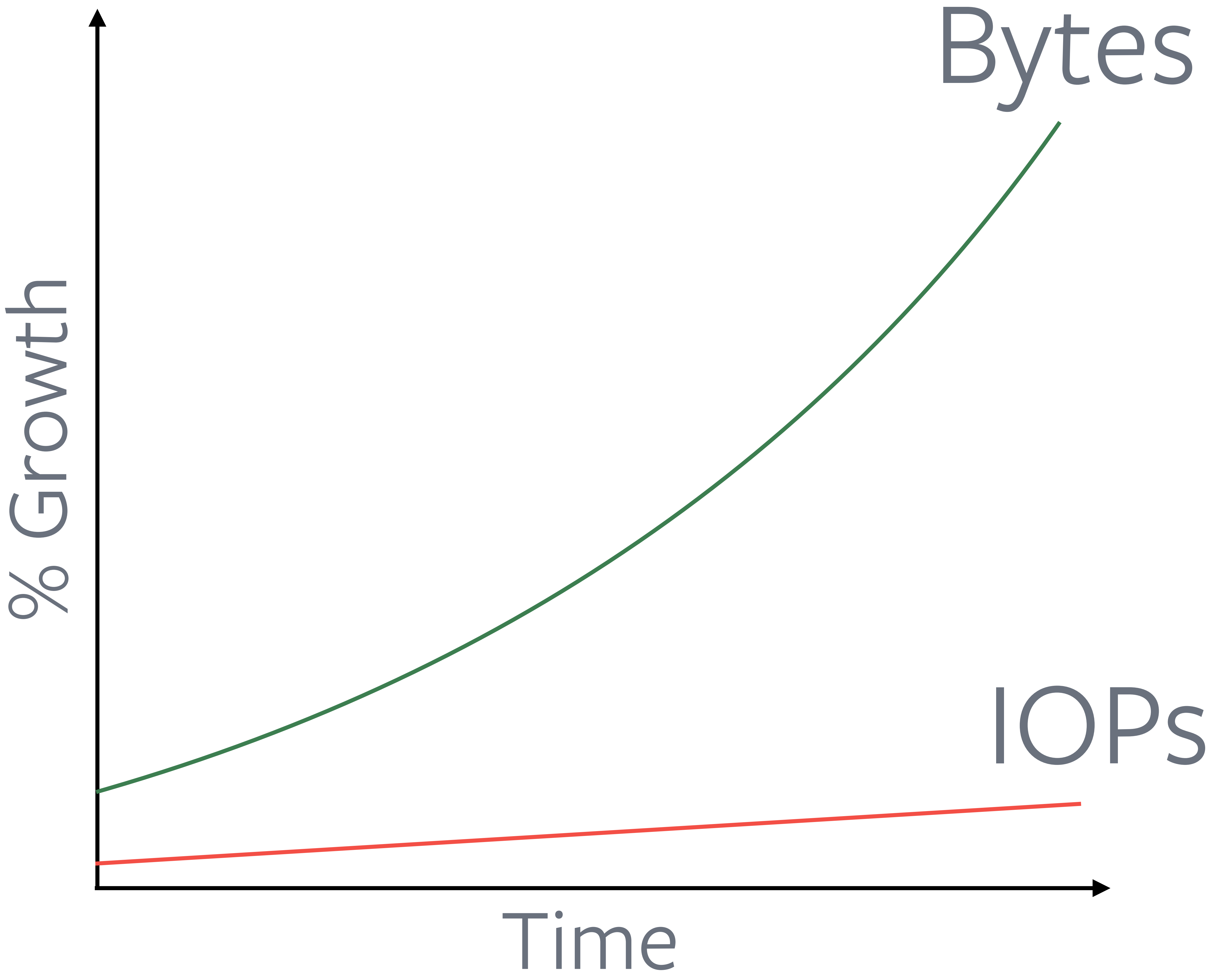
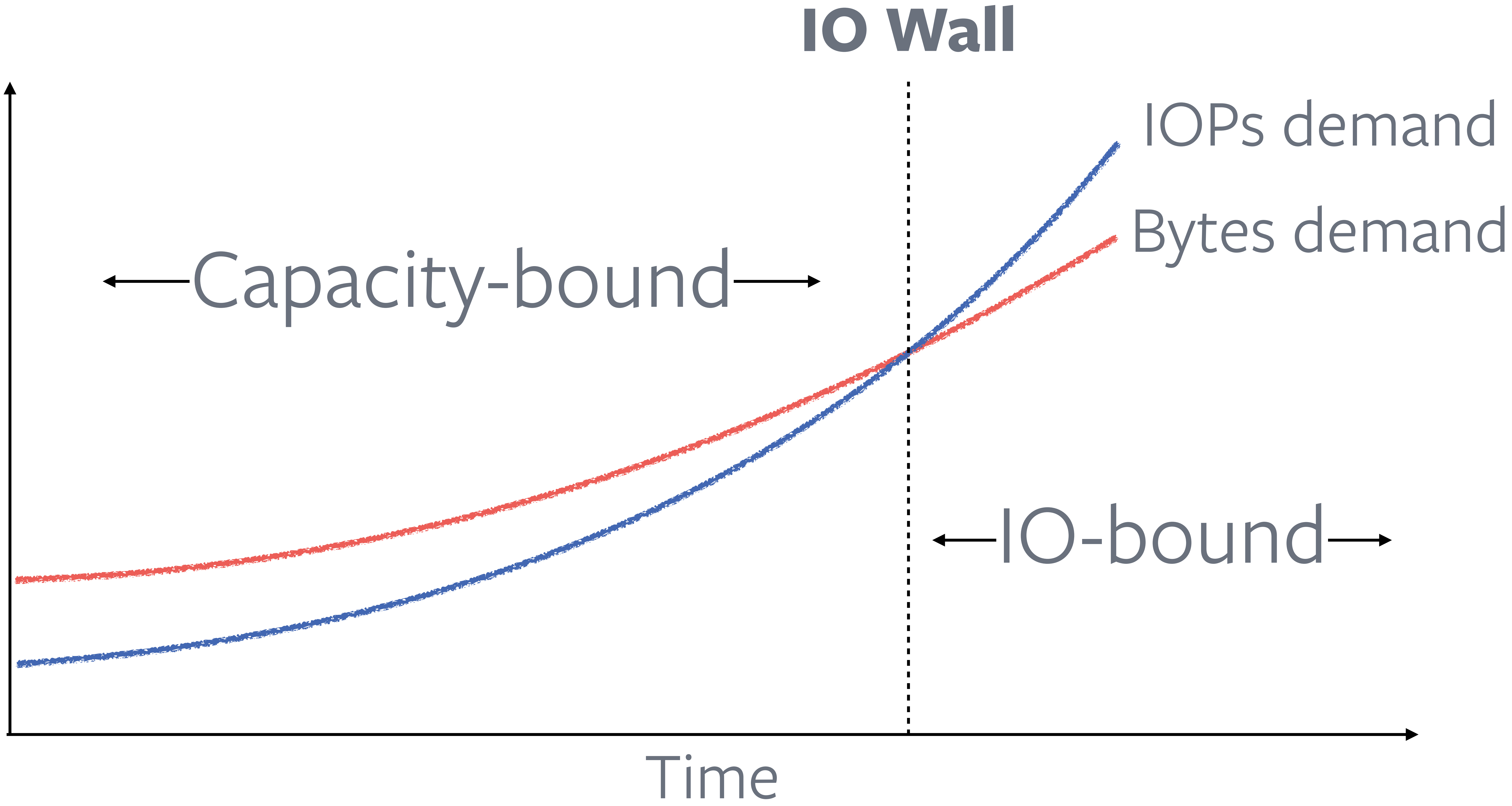


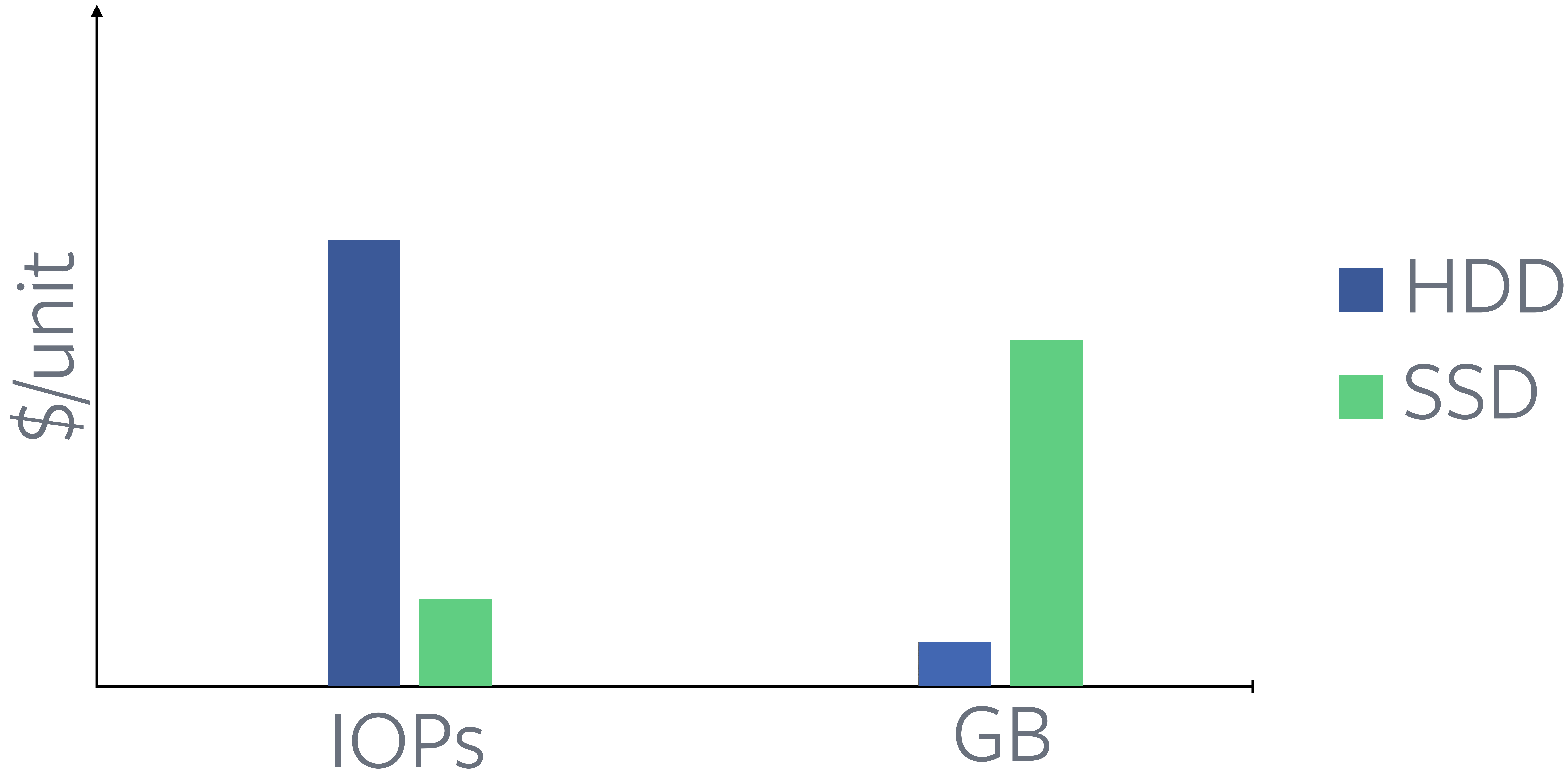
Image by Michael Schwarzenberger from Pixabay

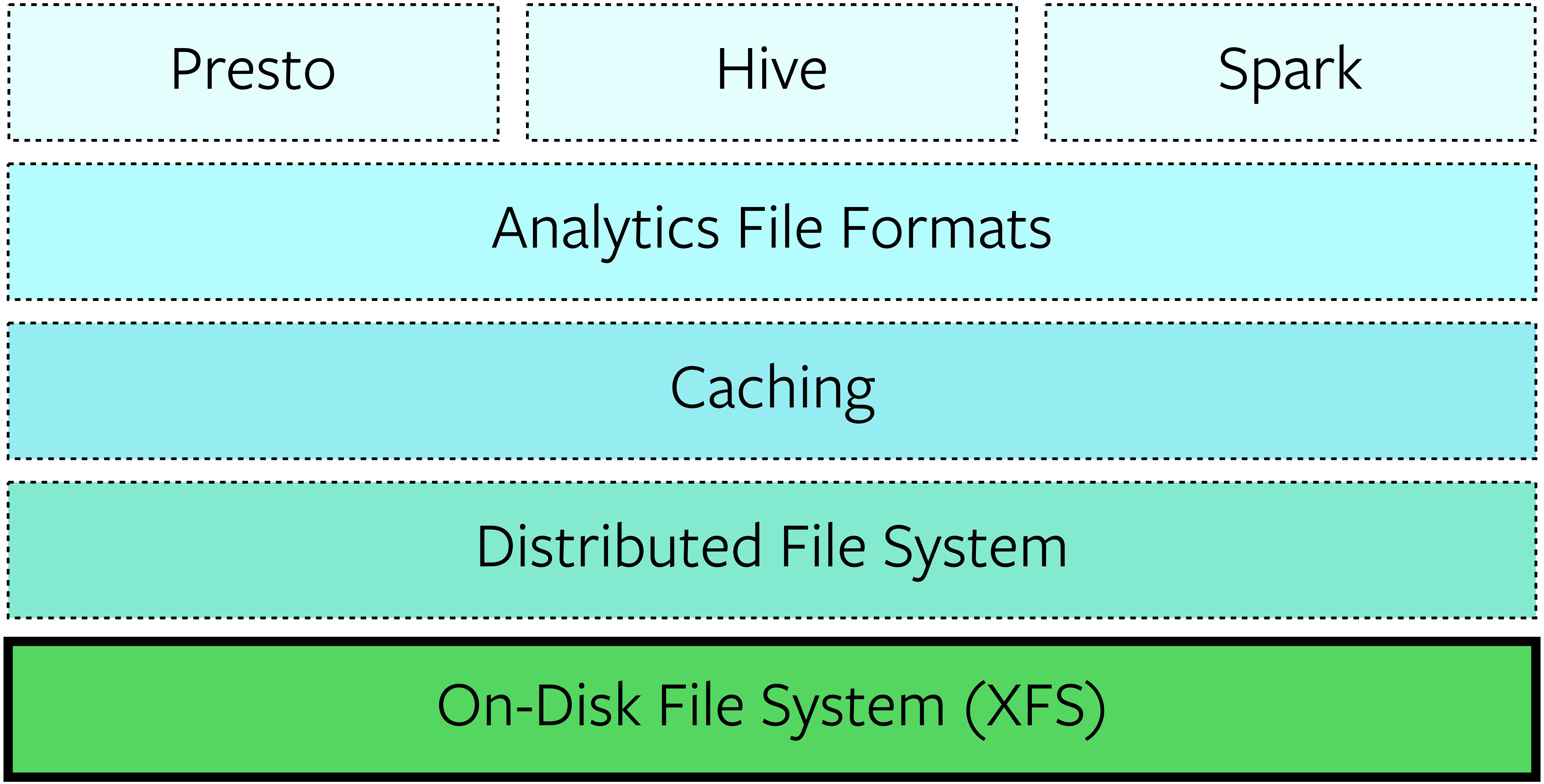
IO workloads are only getting hotter

Thanks, ML

- AI/ML analytics
- Ephemeral photos
- Live Video
- Offloading Cold Data


IO Wall
IOPs demand
Bytes demand
←Capacity-bound→
←IO-bound→
Time


$/unit
IOPs
GB
HDD
SSD



Presto

Hive

Spark

Analytics File Formats

Caching

Distributed File System

On-Disk File System (XFS)

The Opportunity

```
[rwareing@storage001.dc1 /dev/shm/blktrace] blkparse
     41 N     <---- Un-categorized
   8571 R     <---- Data reads
     14 RM    <---- Metadata reads
   4773 WM    <---- Metadata writes
   6901 WS    <---- (Synchronous) Data writes
```

24%

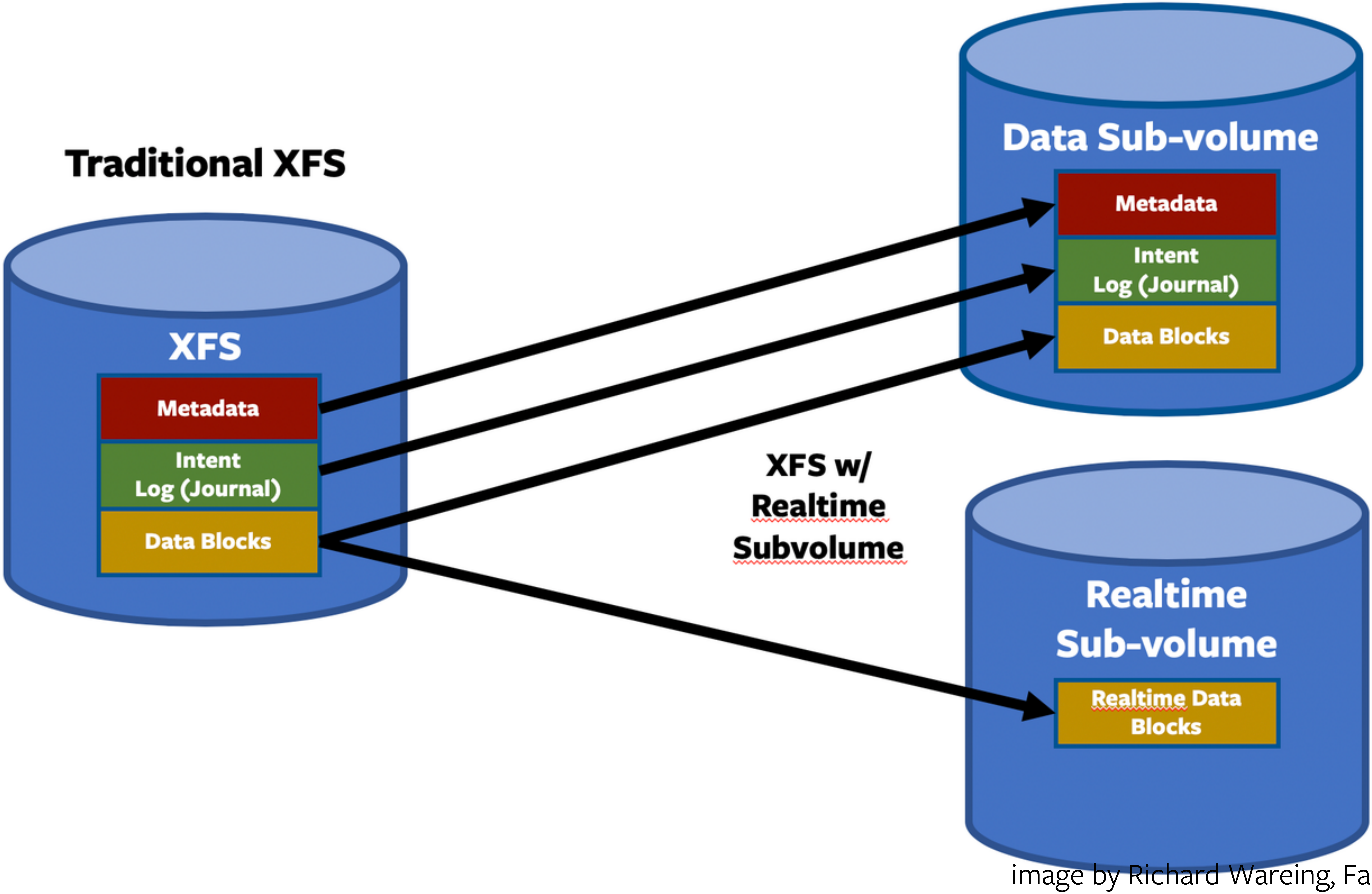
image by Richard Wareing, Facebook

Options to reduce metadata IO

- Write less data
- Smaller metadata
- Fewer flushes for writes
- Create a new filesystem
- Put the metadata elsewhere

XFS Realtime mode - a hidden gem

Image by zombie cygig from Pixabay



Traditional XFS

XFS

Metadata

Intent Log (Journal)

Data Blocks

Data Sub-volume

Metadata

Intent Log (Journal)

Data Blocks

XFS w/ Realtime Subvolume

Realtime Sub-volume

Realtime Data Blocks

image by Richard Wareing, Facebook

Image by PublicDomainPictures from Pixabay

But… we only have one SSD

Partition the heck out of it

And gave it a name - **Hybrid XFS**

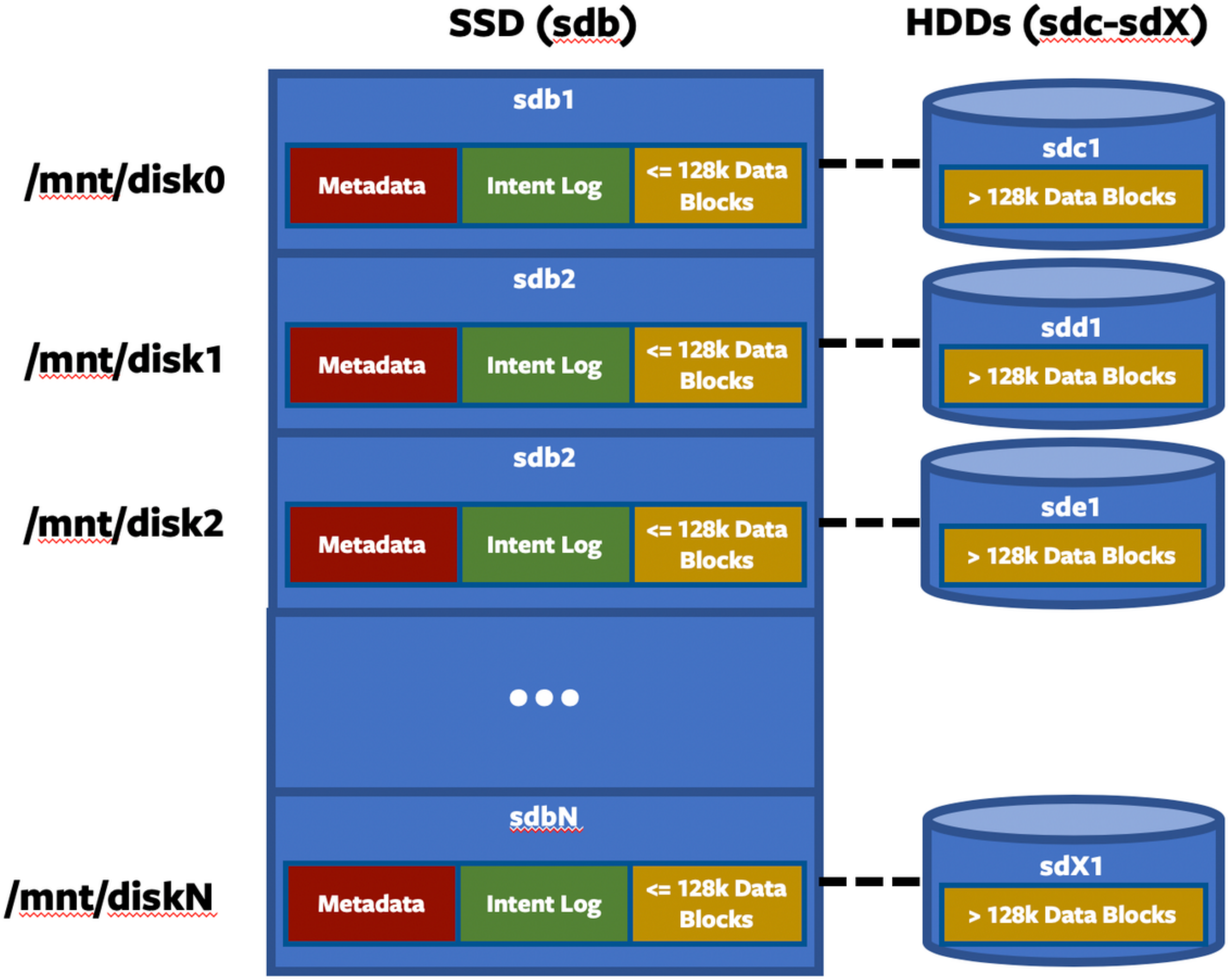
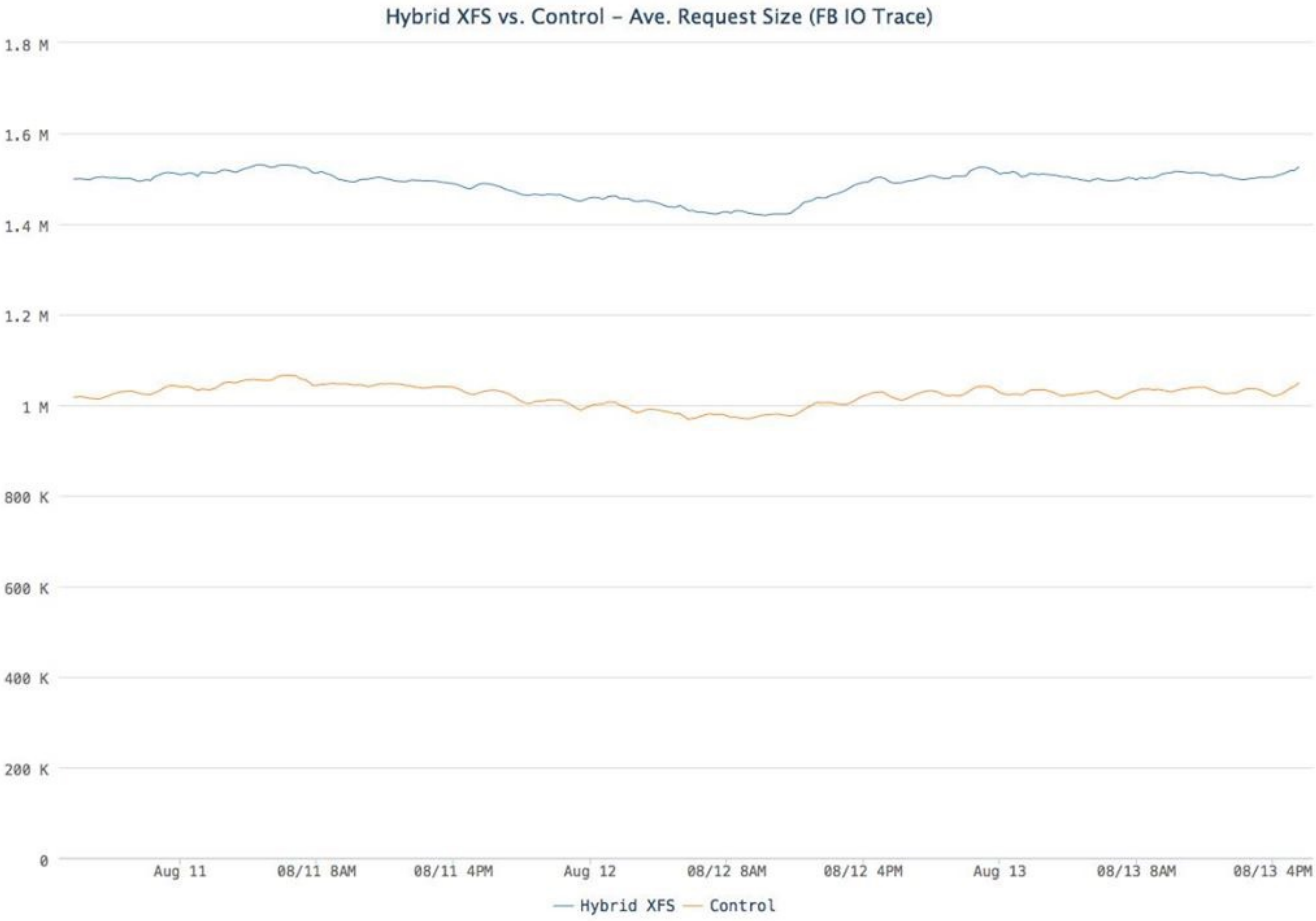


image by Richard Wareing, Facebook

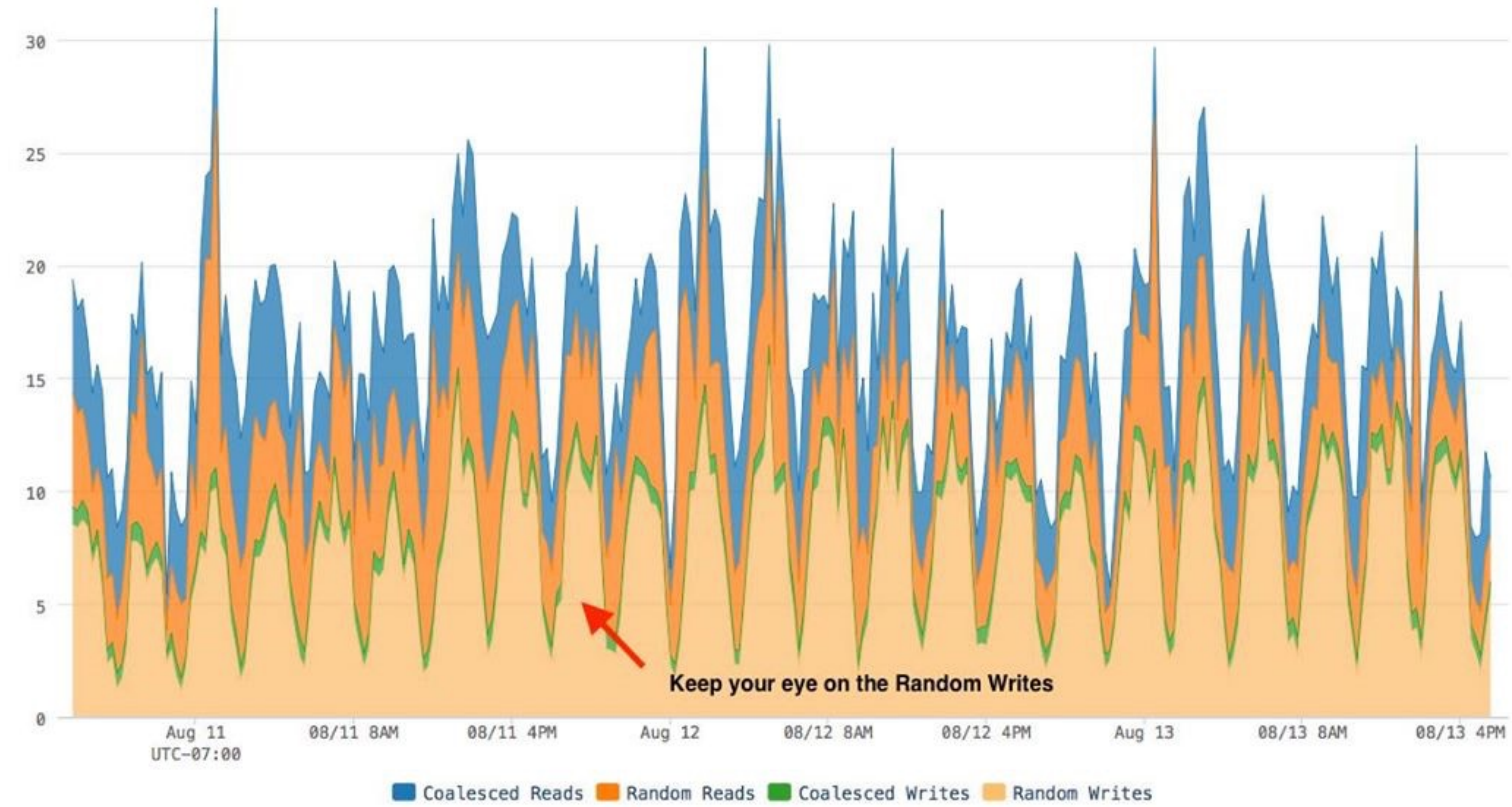
Image by PublicDomainPictures from Pixabay



1.5MB

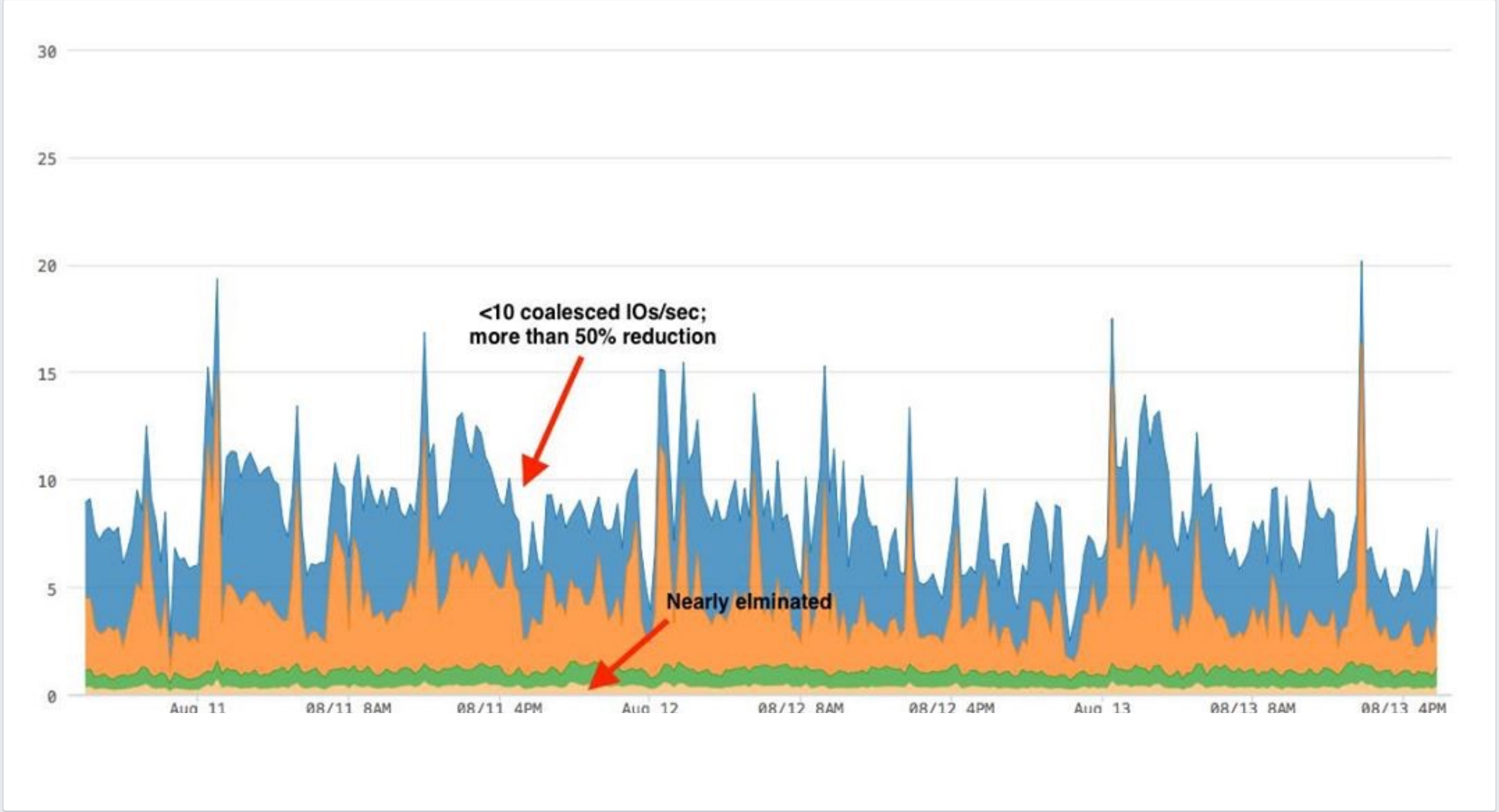
1MB

image by Richard Wareing, Facebook



Keep your eye on the Random Writes

image by Richard Wareing, Facebook



<10 coalesced IOs/sec; more than 50% reduction

Nearly elminated

image by Richard Wareing, Facebook

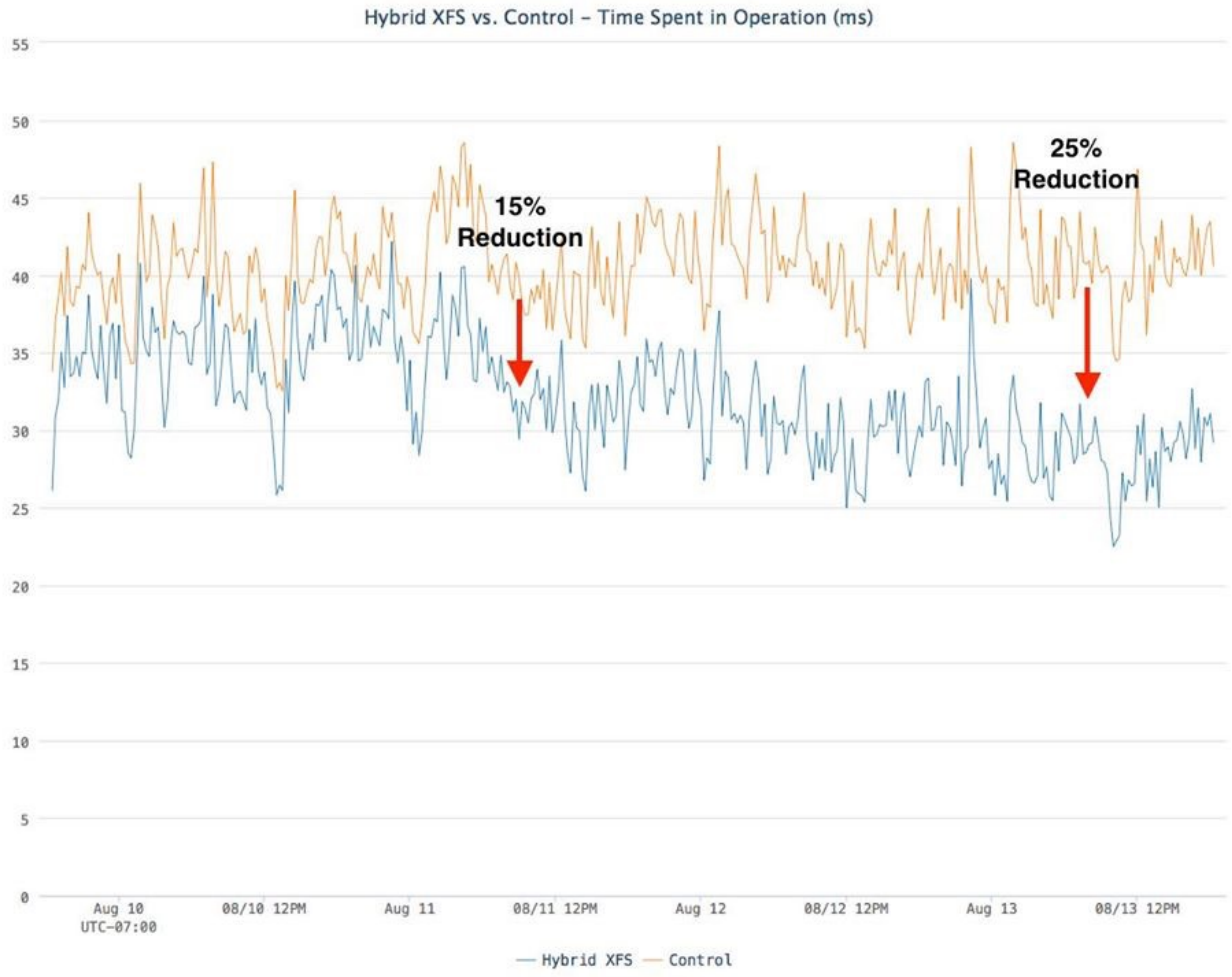


image by Richard Wareing, Facebook

Image by Gerd Altmann from Pixabay

DENIED

The Risks

1. SSD Failures

Image by Magnascan from Pixabay

2. Endurance

Image by Magnascan from Pixabay

3. Hardware changes?

Image by Magnascan from Pixabay

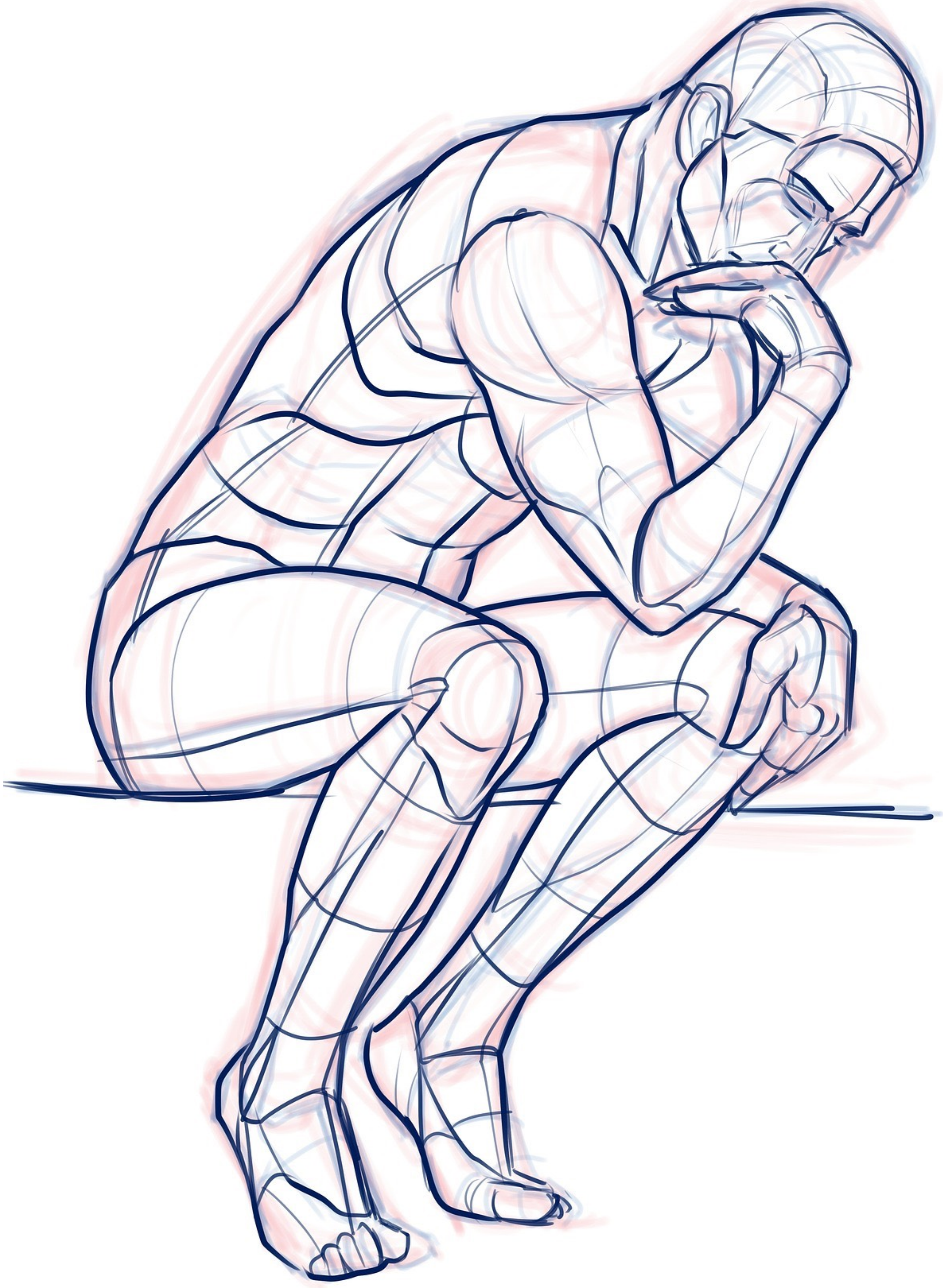
4. Operational headache

Image by Gerd Altmann from Pixabay

Risks.....

Many many risks

- Changes in workload?
- Rollout?
- and...?



Let the Data guide you

Image by skeeze from Pixabay

The Analysis

1. SSD Failures

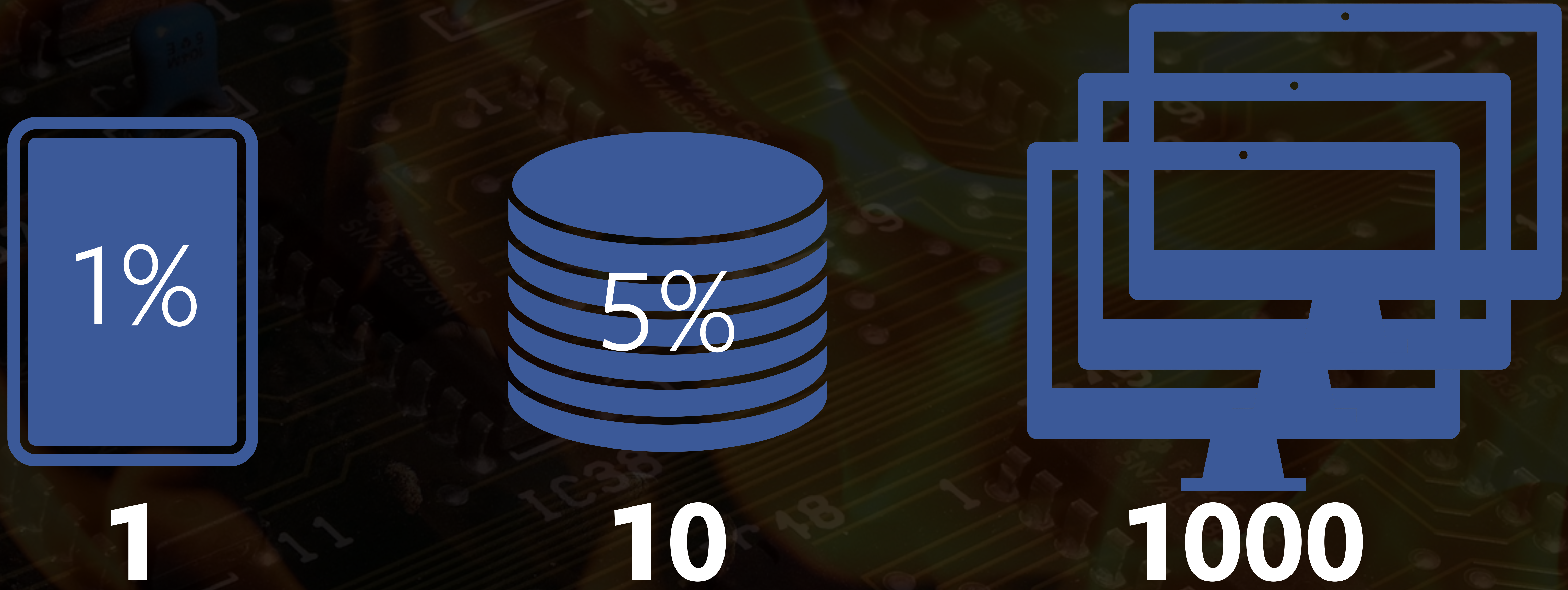


Image by Magnascan from Pixabay

1. SSD Failures

1000 x 1% = 10 SSDs or
100 HDDs

+

10000 x 5% = 500 HDDs

= **600 HDDs**

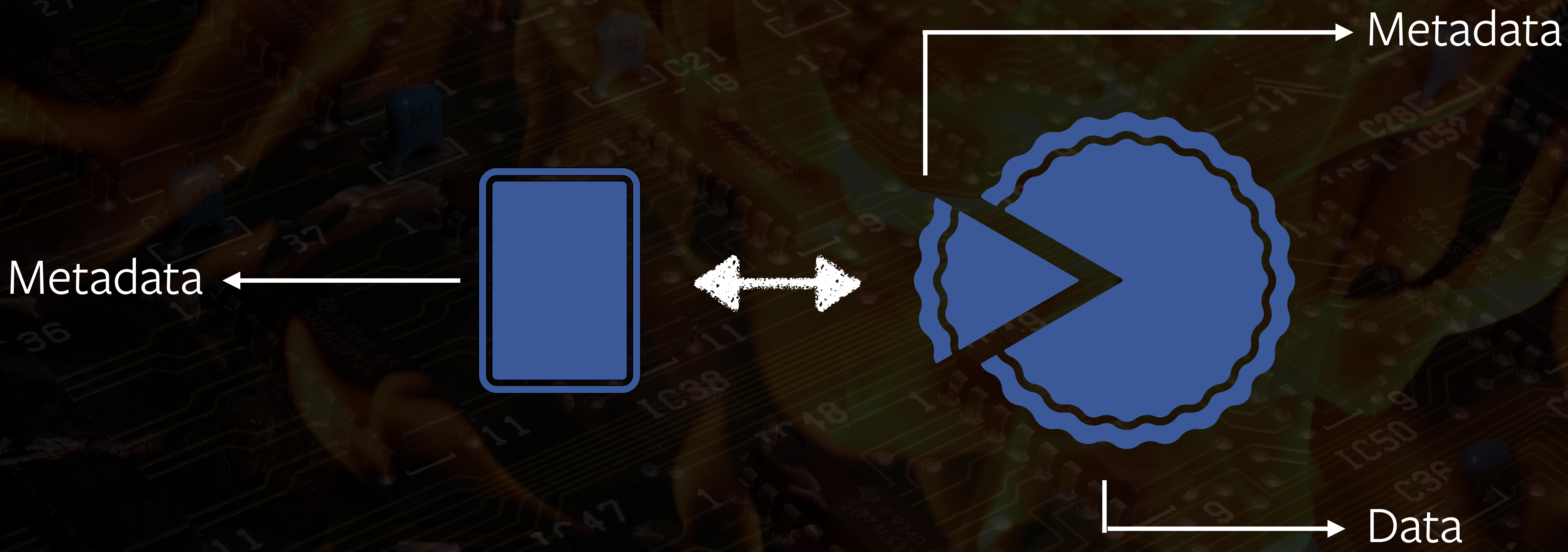
~~**1500 HDDs**~~

Image by Magnascan from Pixabay

1. SSD Failures - en masse!

Image by Magnascan from Pixabay

1. SSD Failures



Metadata

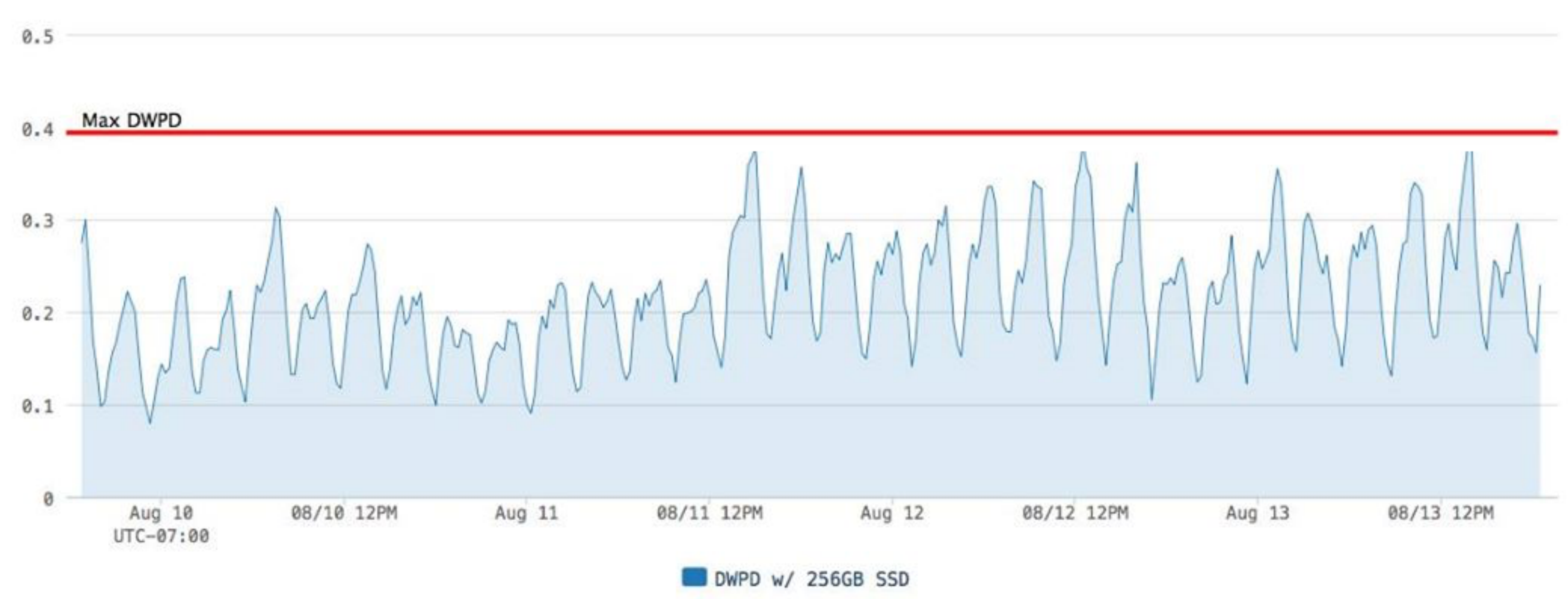
Metadata

Data

Rescue mode

Image by Magnascan from Pixabay
]

2. Endurance



*DWPD = Drive Writes Per Day

Image by Magnascan from Pixabay

3. Will there always be SSDs?

Image by Magnascan from Pixabay

4. Operational headache

- Patched XFS statfs call
- Collect stats for both metadata and data
- Make systemd wait for BOTH devices

Image by Gerd Altmann from Pixabay

Rolling It Out

How do you roll out a destructive change to tens of thousands of hosts?

Image by kathrynlerro from Pixabay

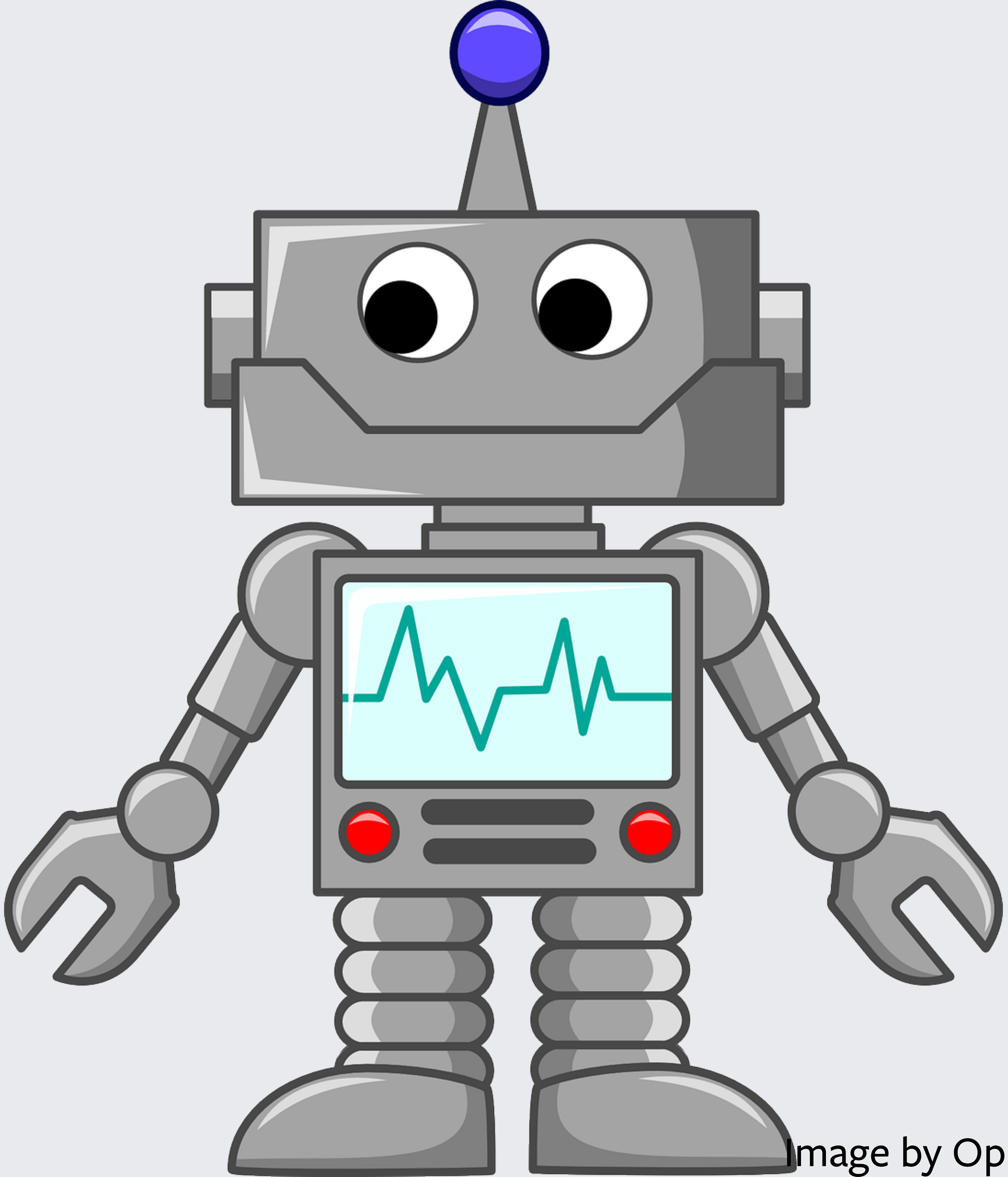


Image by OpenClipart-Vectors from Pixabay

Look out for...

- Impact to:
  - Durability
  - Capacity
  - Performance
- Fallout from failures
- Automation flying blind

Success!

Lessons

“Hard problems can have simple solutions”

“Gut feelings can be wrong”

“Data wins arguments”

“Better safe than sorry”

Thank You!

facebook