BLOCK STORAGE SUBSYSTEM

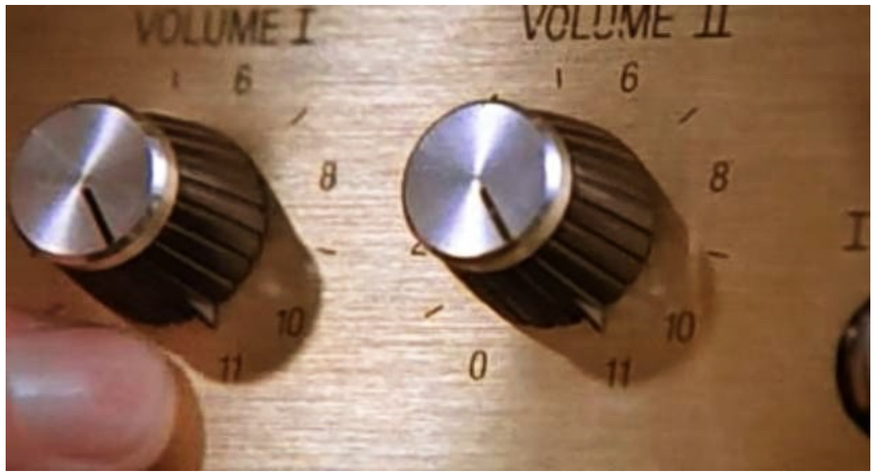
(PERFORMANCE ANALYSIS)

OR TRYING TO OPTIMIZE IOS

Julien Leduc

WHY WOULD YOU LIKE TO DO THIS?

- Understand the IO bottlenecks in your service, if IOWAIT is high it can help...
- If you *need that extra push over the cliff*



WHAT CAN YOU ACHIEVE?

- List your basic IO subsystem architecture limitations
  - RAID level
  - number of disks
  - ...
- Microbenchmark each component of your IO subsystem to identify possible hardware bottlenecks
  - `dd` test to `/dev/null` and from `/dev/zero` for each block device
  - repeat dd at all levels: block device, aggregated device, filesystem... to check that you reach the targetted performance
  - ...

AND NOW?

YOU MAY NEED DEEPER ANALYSIS...

`blktrace` can help you to

- Capture IO requests on block devices during benchmarks
- Graph and analyze those traces later (`seekwatcher`)
- *Replay* those traces on different configuration (`btrecord` and `btreplay`)

SOME EXAMPLES:

REMEMBER THIS:

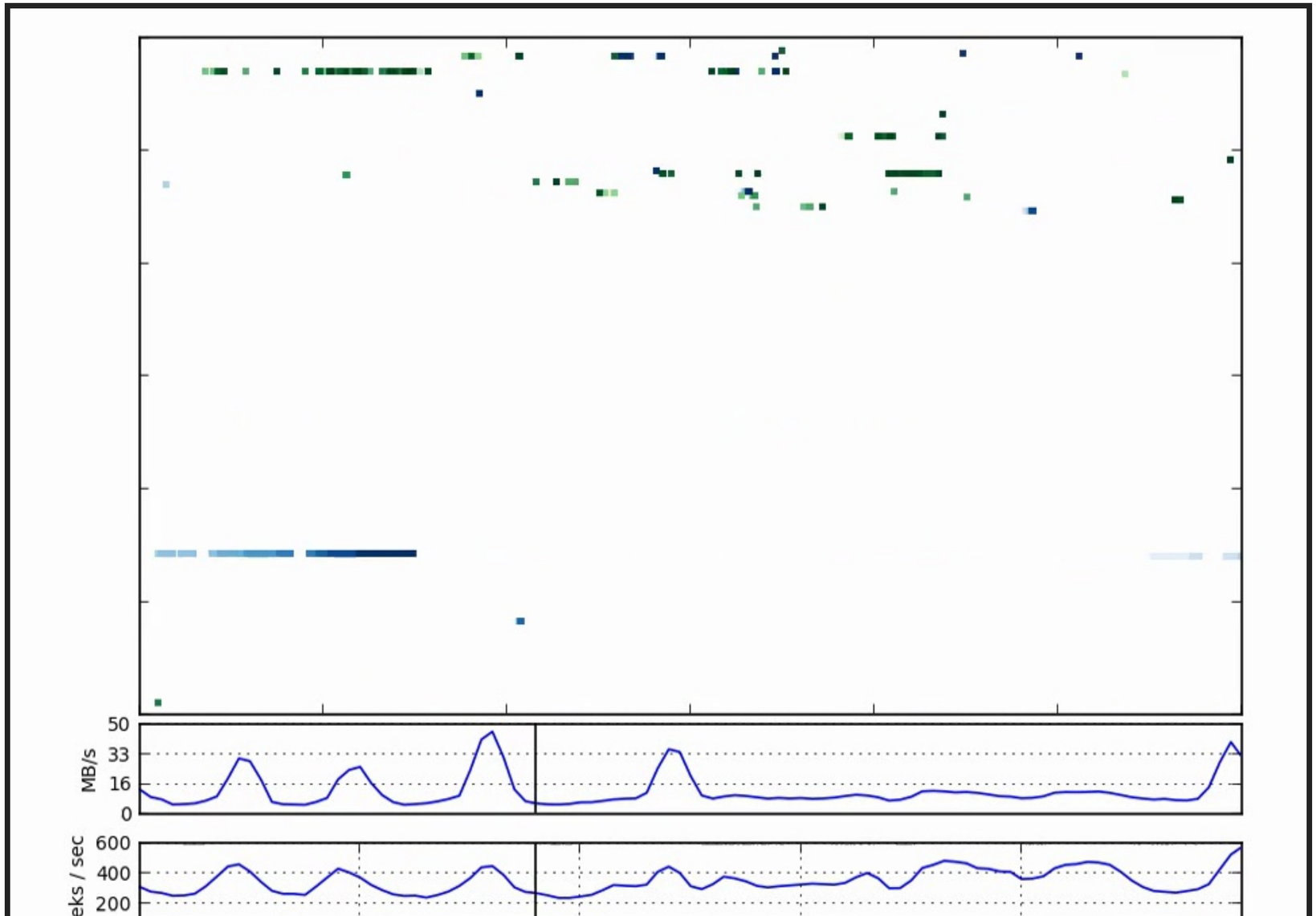
READ/WRITE

TSM REPLICATION EXAMPLE

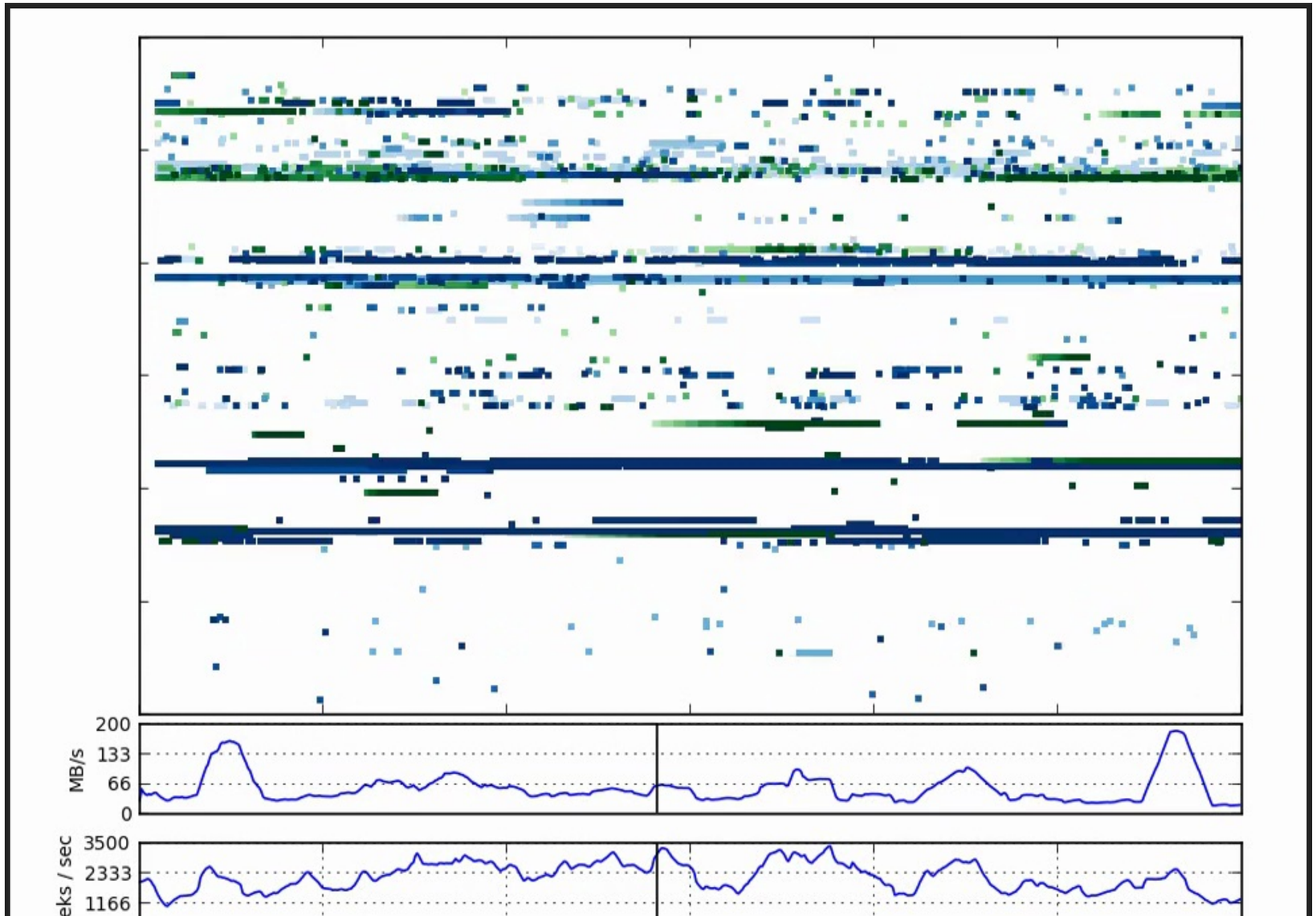
TSM replication process was slow for AFS backups...

- Why?
- Push the volume from 1 to 11 with minimal hardware investment

DB REPLICATION WITH HDDS



DB REPLICATION WITH SSDS

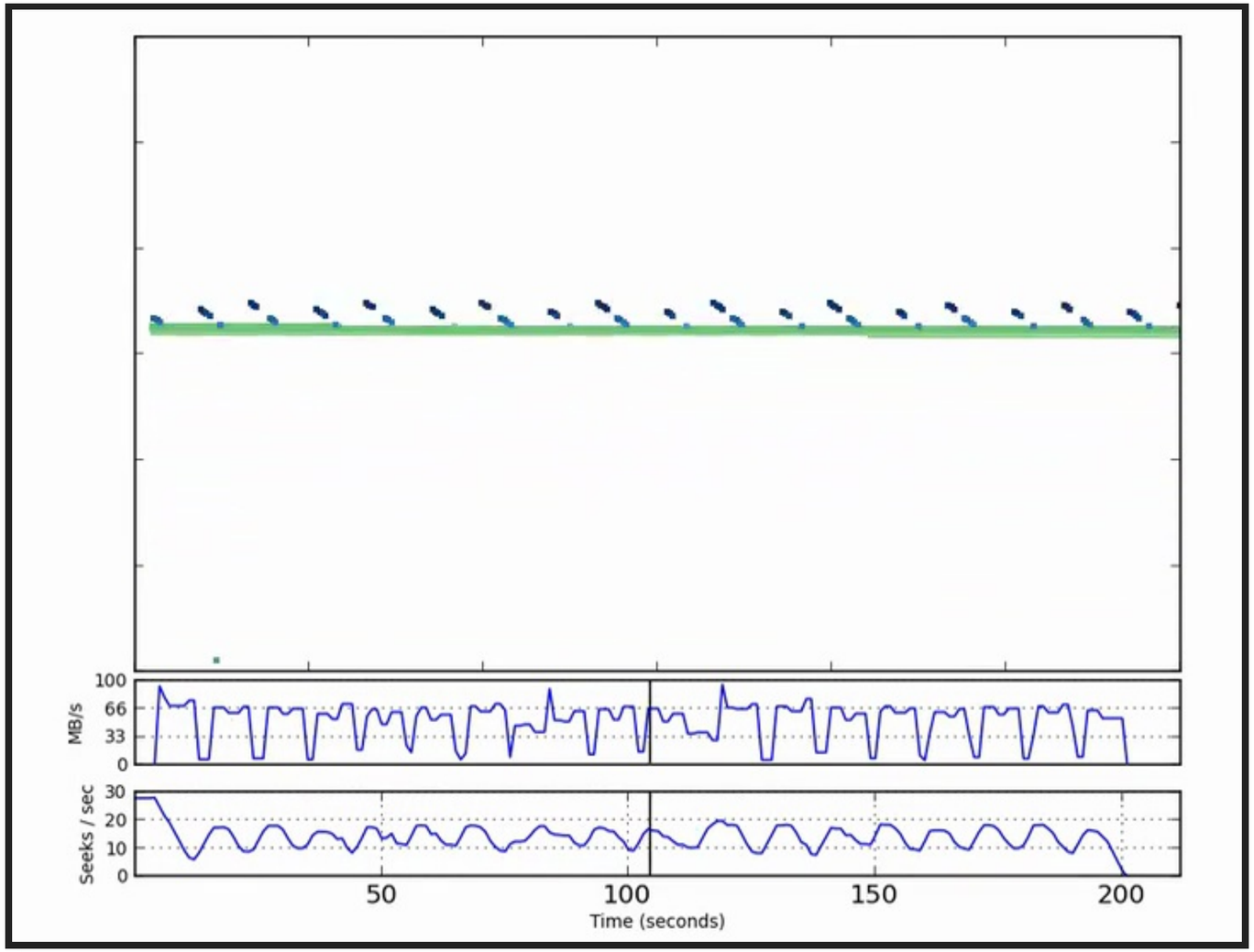


INEFFICIENT STREAM WRITE

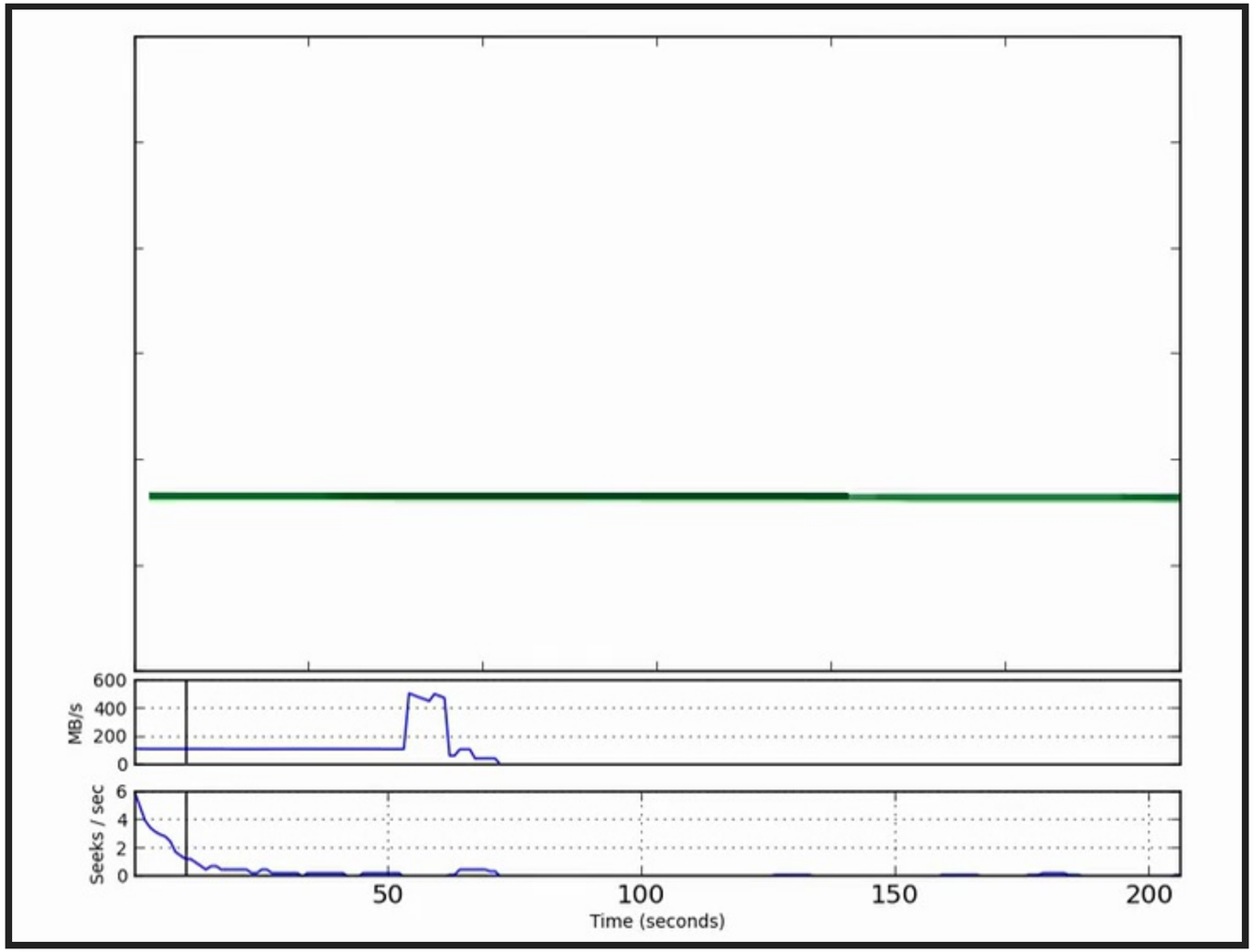
Writing 1 stream on a 2 disk RAID 1E with 38% performance lost from target...

- Why?
- Push the volume from 6 to 10 as expected...

CAPTURE AT 6



BACK ON TARGET



SOLUTION

This was a filesystem alignment issue on the underlying RAID chunks...

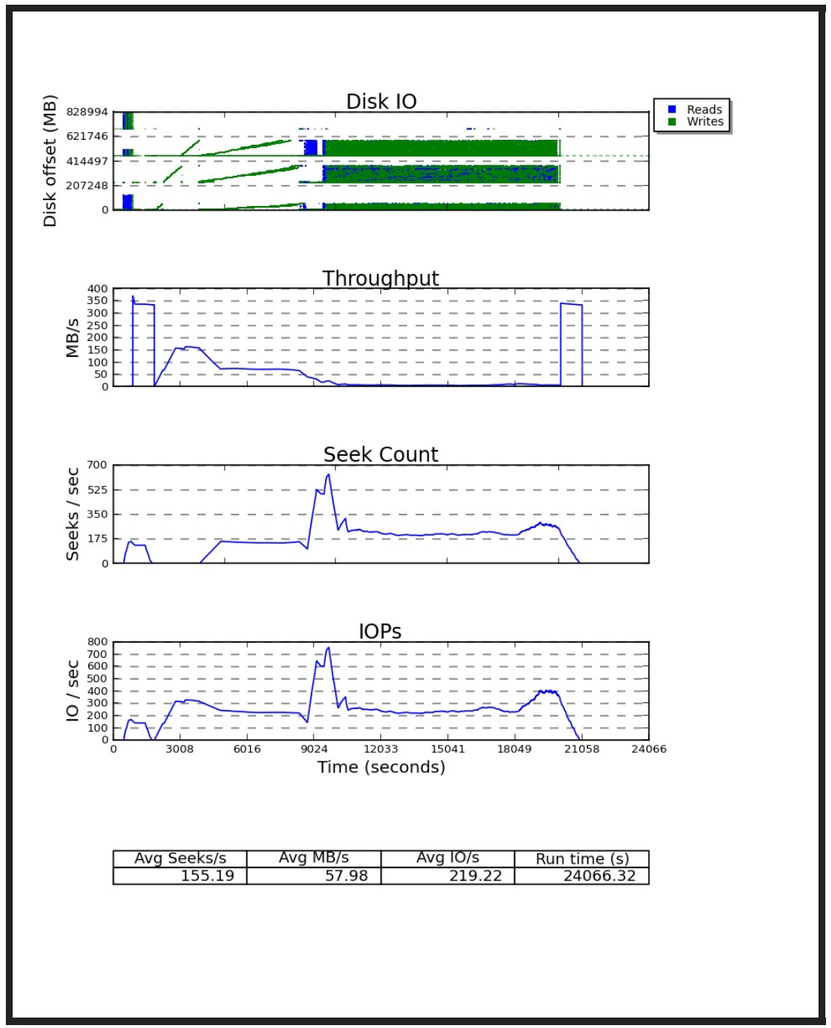
- Hardware cost: 0CHF
- Performance gain: +60%

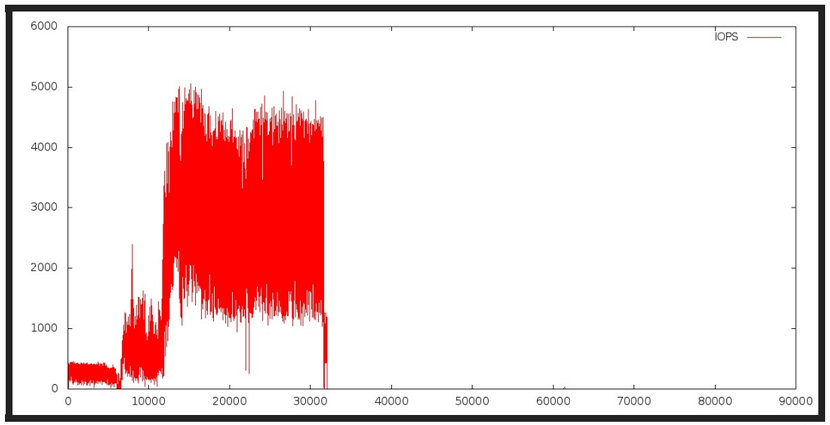
CEPH VS HYBRID DISK/SSD RAID

Traces can be collected from Rados Block Device with the same tools...

The idea was to run a specific benchmark in our proprietary software over RBD and again over local hybrid RAID1 constituted of 1 HDD (set to write mostly) and 1 SSD and compare.

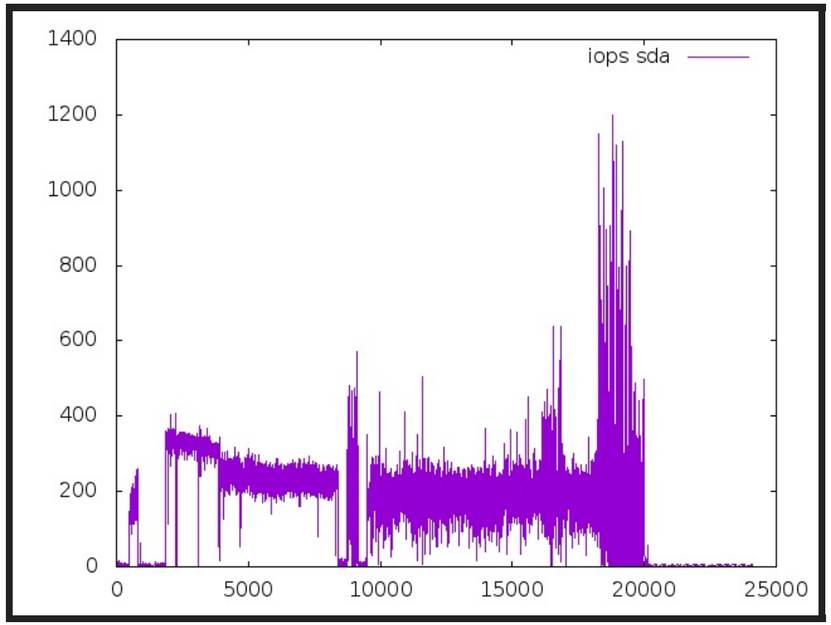
ONE PROBLEM TRACES ARE BIG...



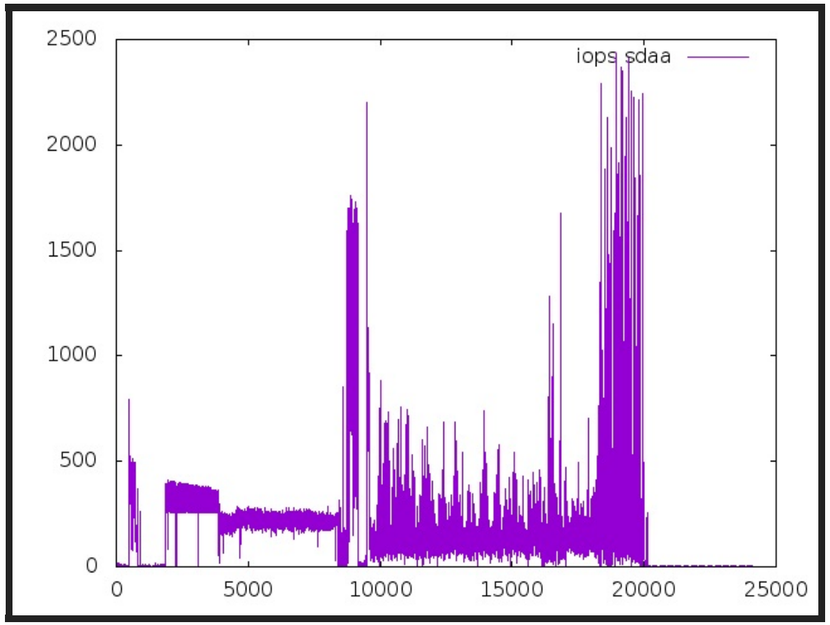


Rados traces were too large for `seekwatcher`...

BUT



We can see that



hybrid write mostly to disk is not satifying...

CONCLUSION

- Do not skip microbenchmarks those are easy means to identify your hardware bottlenecks for later analysis
- A short video is better than a picture
- `systemtap` is your friend too:
  - capture only the traces you need
  - high speed character devices analysis supported
  - realtime and lightweight

THANK YOU FOR YOUR ATTENTION