
CERN

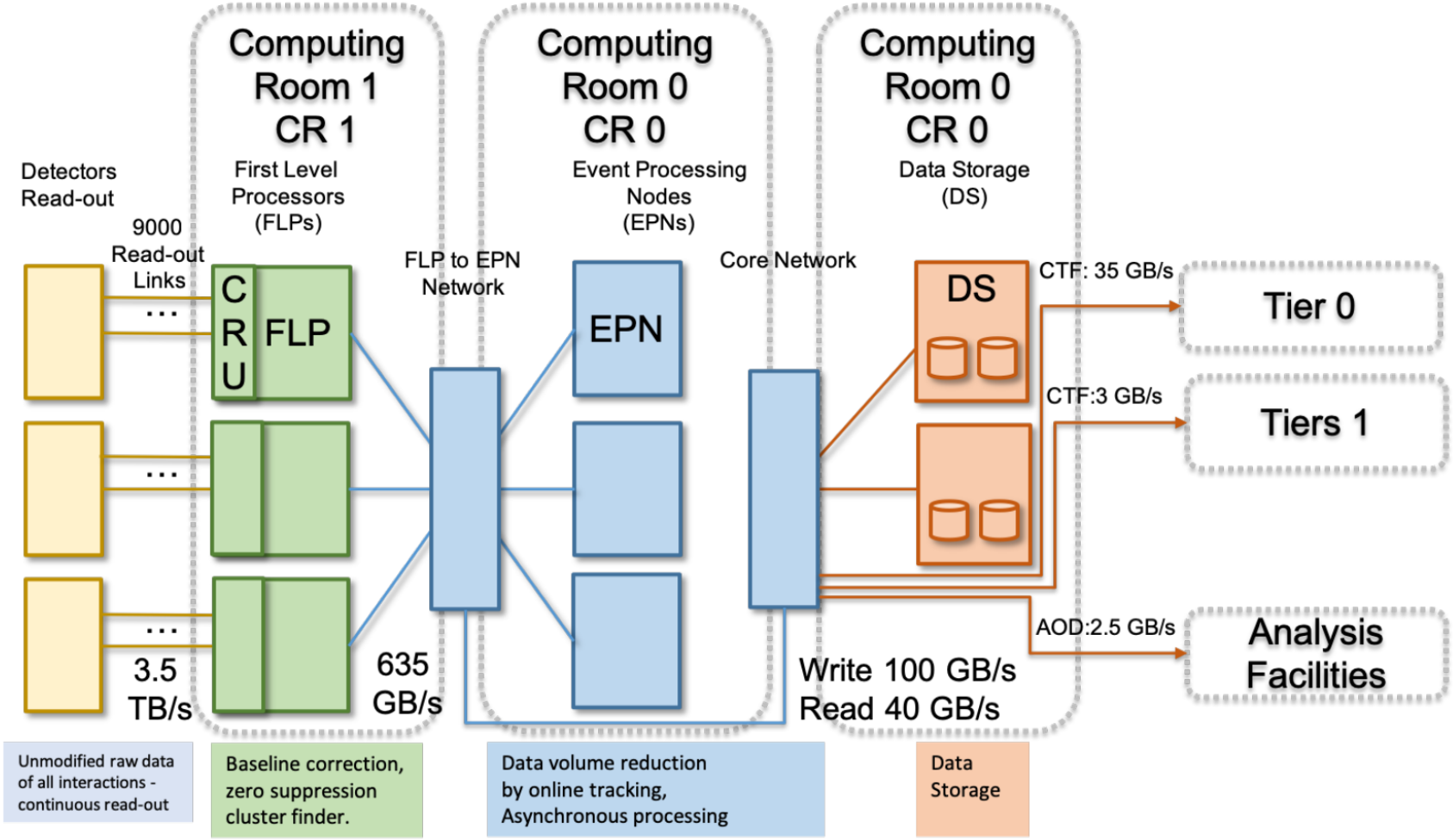
Michal Simon

EOS storage for Alice O2

Outline

- AliceO2 short introduction
- EOS AliceO2 cluster
- Aggregated throughput tests
- Summary

Alice O2


Computing Room 1 CR 1
Computing Room 0 CR 0
Computing Room 0 CR 0
Detectors Read-out
9000 Read-out Links
First Level Processors (FLPs)
Event Processing Nodes (EPNs)
Data Storage (DS)
FLP to EPN Network
Core Network
CRU
FLP
EPN
DS
CTF: 35 GB/s
CTF:3 GB/s
AOD:2.5 GB/s
Tier 0
Tiers 1
Analysis Facilities
3.5 TB/s
635 GB/s
Write 100 GB/s
Read 40 GB/s
Unmodified raw data of all interactions - continuous read-out
Baseline correction, zero suppression cluster finder.
Data volume reduction by online tracking, Asynchronous processing
Data Storage

Alice O2

- 500 EPNs (Event Processing Node), each hosting 4 GPUs, each GPU generating a Time Frame every 40 seconds
  - **2000 data sources** in total
- A Time Frame (TF) corresponds to a single 2GB file in EOS
  - **TF has to be copied to EOS in less than 40 seconds**
- Data sources transfer data to EOS in (kind of) round robin fashion at 20 ms intervals
  - **every 20 ms a new file will be created and 2GB of data transferred**

EOS AliceO2 cluster

- 10 data servers, each 96 HDDs (12TB, 265 MB/s)
- 4 HBA controllers per data server with aggregated throughput of 12GB/s
- each server is equipped with 100 Gigabit ethernet card
- all 10 servers on the same non-blocking switch
- 2 NUMA units, 32 cores (64 threads)

Baseline measurements

IPERF measurements

#streams	node2node one way	node2node bi-directional
1	17.7. Gbit/sec	18.4 Gbit/sec
5	78 Gbit/sec	60 Gbit/sec
10	93.9 Gbit/sec	93 Gbit/sec
20	93.9 GBit/sec	93 Gbit/sec

Single disk performance (Model TOSHIBA MG07ACA1 12TB, bs=1MB)

Stream Write Direct	Stream Read Direct
265 MB/s	265 MB/s

Multistream in direct I/O on one disk

10 streams	100 streams	500 streams
100-120 MB/s	60-80 MB/s	60-70 MB/s

96 disk aggregate throughput

Stream Write Direct	Stream Read Direct
6.48 GB/s	6.35 GB/s

(bs=1MB,count=128)

There's something wrong ... ? It seems we have wrong HBA controllers (x8 PCIe connectivity), after upgrade to SAS HBA with x16 PCIe connectivity:

Stream Write Direct	Stream read Direct
12 GB/s	12 GB/s

Remote I/O

- Write remotely 10GB file per disk
  - 96 disks, so 960 GB in total
  - source takes the data from memory
  - destination writes to disk
- Running 4 vanilla XRootD servers on the destination (one per HBA controller)

aggregate write throughput
5.7 GB/s

- we should be able to write at 12GB/s !!!
  - network is 100Gbit/s ~= 12.5 GB/s
  - aggregate disk throughput is 12GB/s

Debugging remote I/O

- Performance per node obtained with CERN Centos7 far from nominal
  - aggregate disk throughput is: OK
  - network speed: close to nominal
  - when running a full-stack test a ~50% penalty was observed from the expected throughput
- Running the mainline kernel on top of CC7 brought us close to nominal performance (package provided by the ELRepo repository)

aggregate write throughput
10.66 GB/s

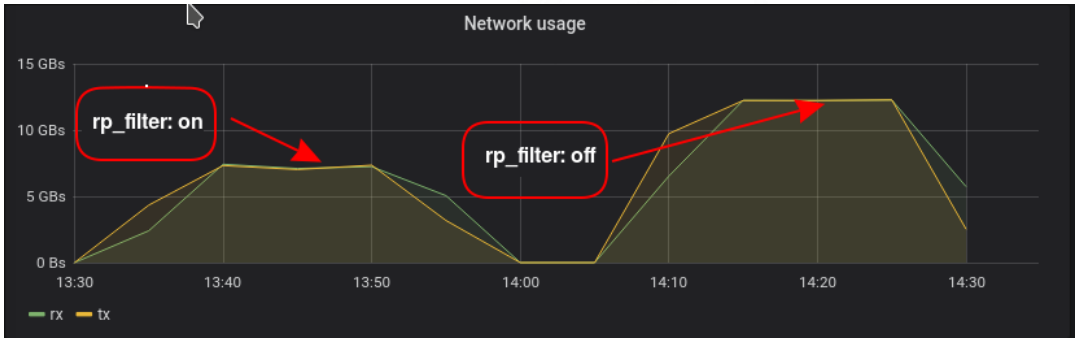
- Reached to the security team and decided that this was not the optional solution
  - timely security patches are not guaranteed by the ELRepo

Debugging remote I/O

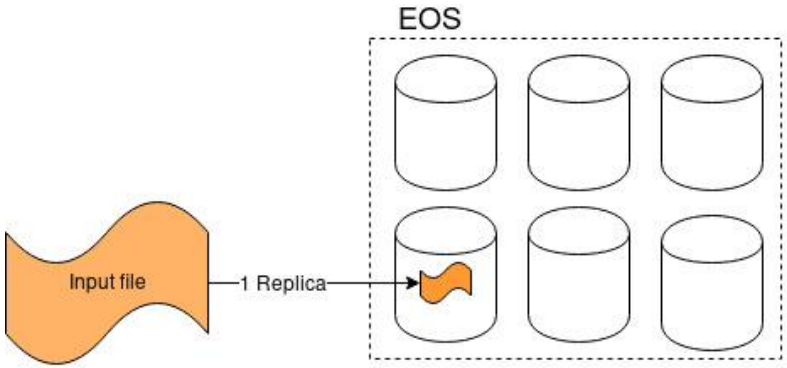
- Storage system upgraded over the summer to Centos 8
- Performance slightly improved (60% of nominal)
  - seemed networked bound this time (iper3 tests confirmed this)
- By process of elimination, we've discovered the culprit for the network 'slowness': IPv6 firewall
  - 60Gbps instead of the nominal 100Gbps

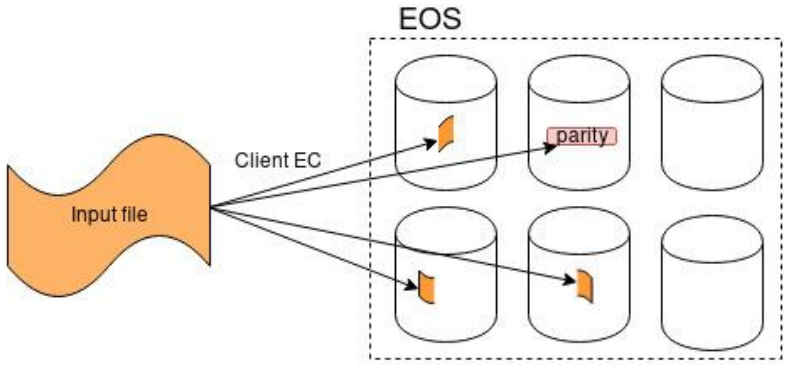
Debugging remote I/O

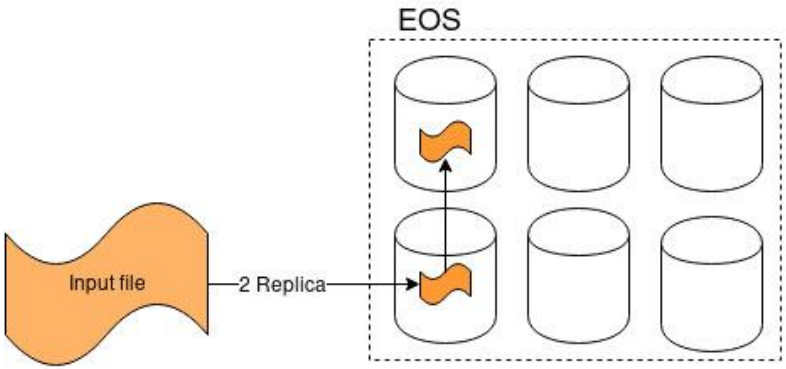
- Debugging further we have discovered that the problem was not linked to the port filtering rules, but rather to the reverse path filtering for IPv6
  - rp_filter ensures a packet would go back on the same interface as the input interface, otherwise it would be dropped
  - Our storage nodes only have one IP address attached to them → the rp_filter is not needed in this case and has been disabled as a workaround → almost full performance obtained (~92Gbps) → OK!

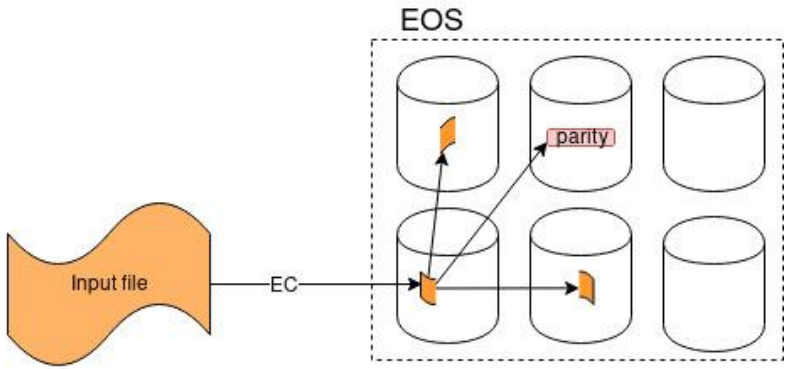


Copy variants



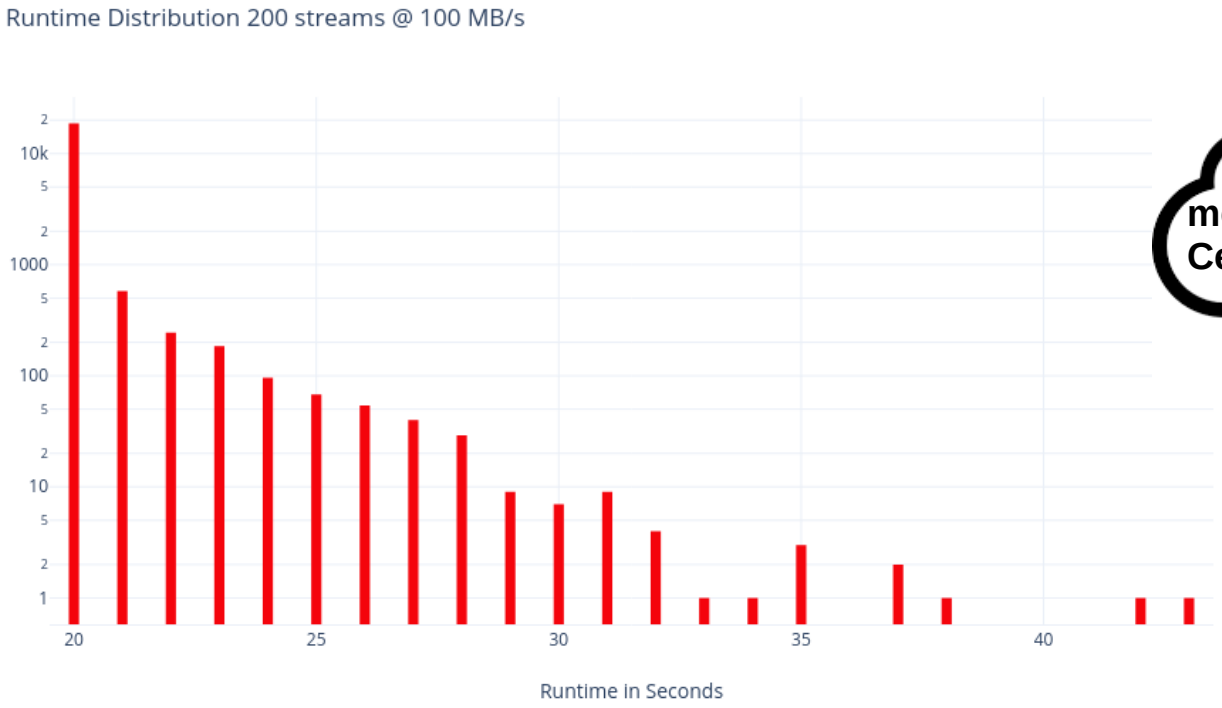






EOS EC

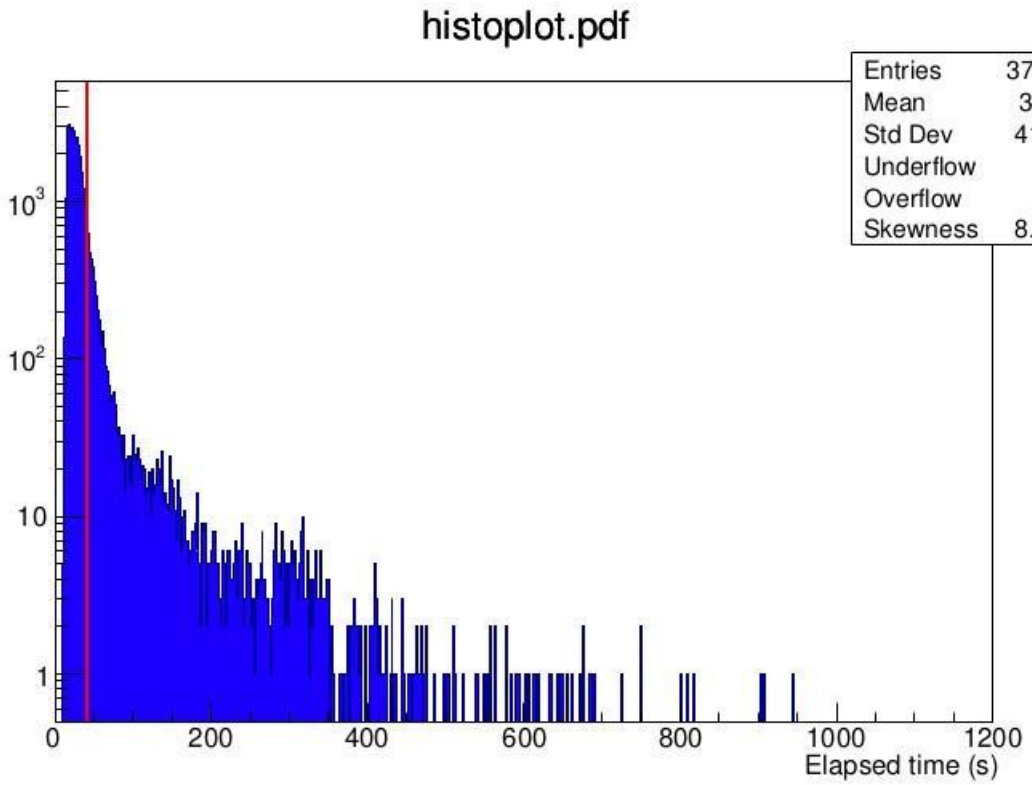
- 200 streams writing continuously 2GB files into EOSALICEO2
  - within the EOSALICEO2 cluster
- 100 MB/s bandwidth limitation (to reduce the jitter)



measured before Centos8 tuning

20% capacity test

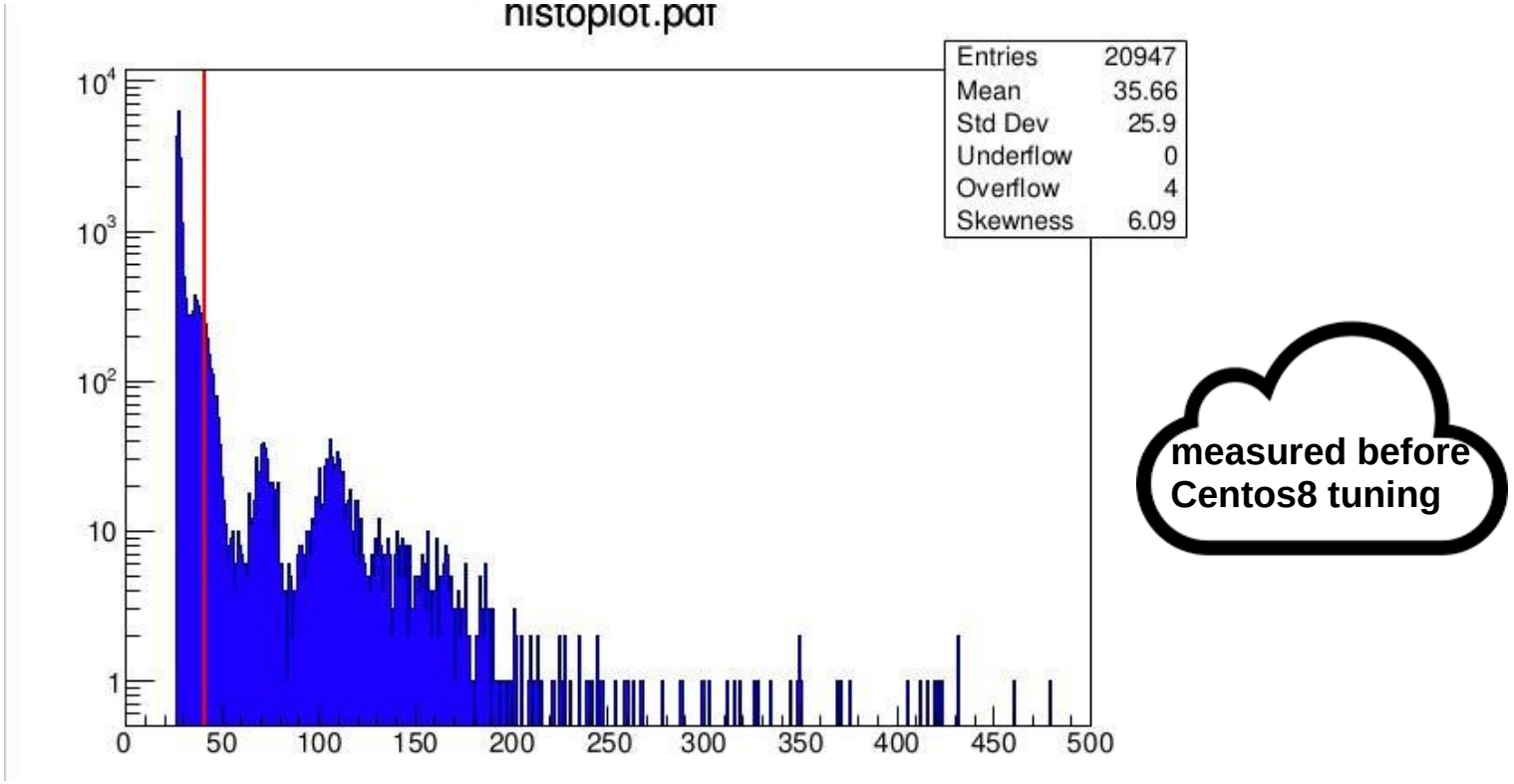
- clients run on the batch farm (simulating **20% of the target production load**), data were recorded on EOSALICEO2 cluster (10 servers, again **~20% of the target production system**)
- the **mean is good**, but we have a long tail of overflow



measured before Centos8 tuning

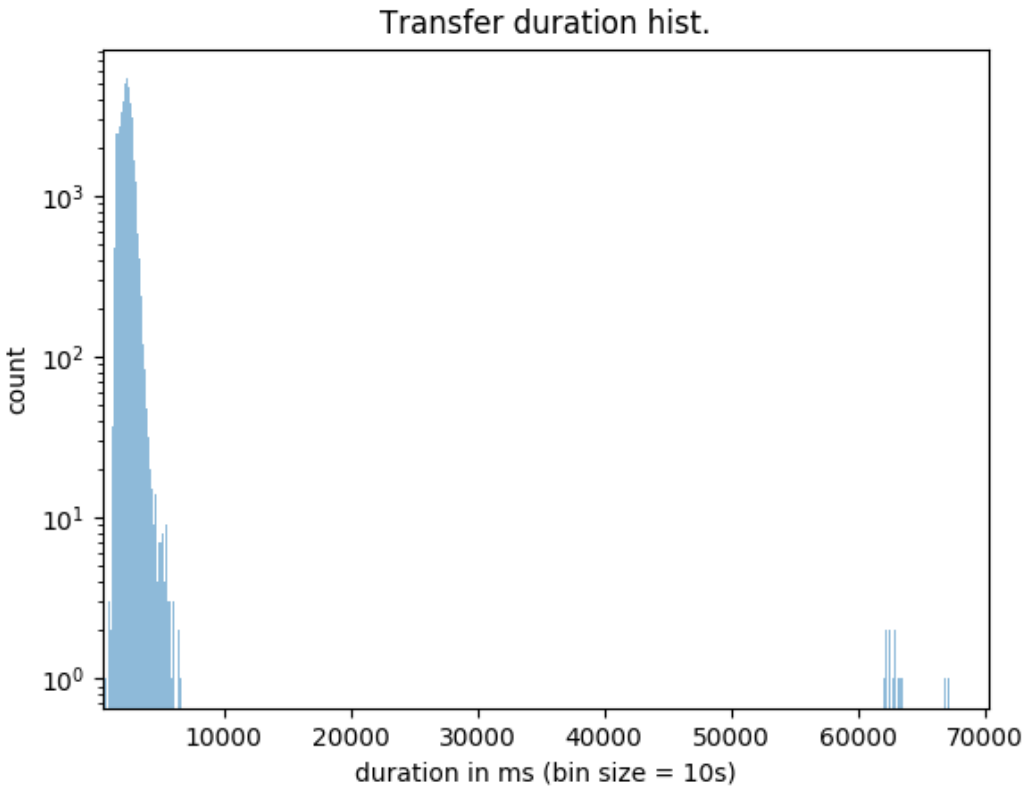
20% capacity test

- in addition to previous test, **80MB/s bandwidth limitation** has been applied (to reduce the overflow)
- source **data were generated with *dd***
- still about **5-6% of the transfers overflow the 40s deadline**



Client side EC

- EPN simulator (client with EC plug-in) run on 4 AliceO2 servers generating **~10% of the target production load**, data were recoreded on 6 EOSALICEO2 servers (running vanilla XRootD), roughly **~10% of the target production system**

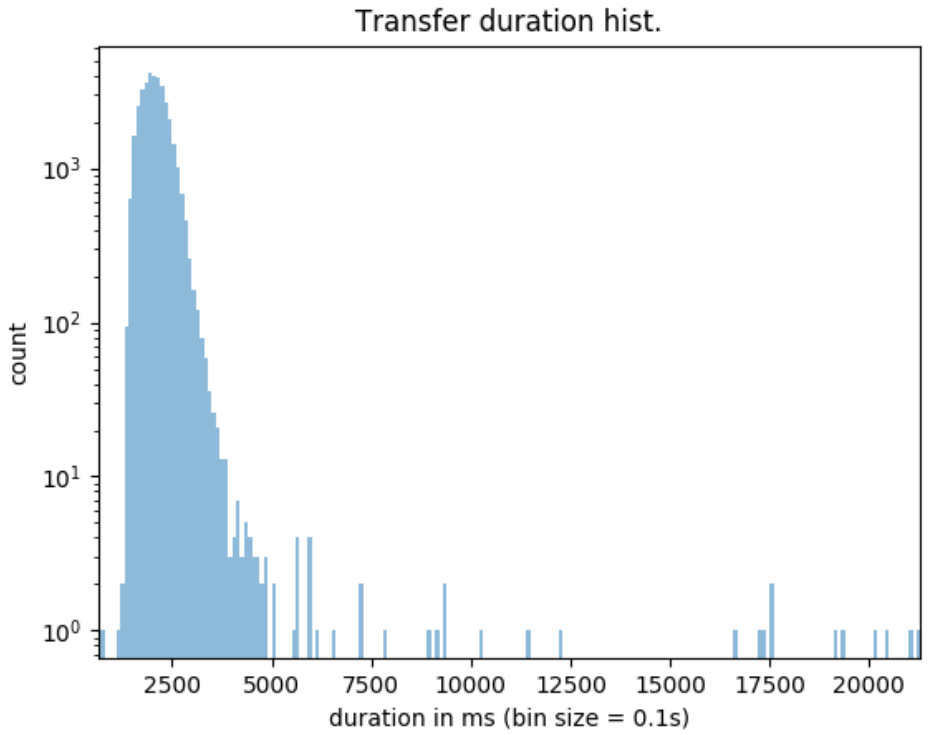


Average data transfer rate	929.47 MB/s
Average transfer duration	2325.34 ms
Transfer duration exceeded 40s	21
Transfer duration exceeded 40s	0.0355%
Data transfer rate standard deviation	203.65
Transfer duration standard deviation	1251.22

measured before Centos8 tuning

Client side EC

- **force stream timeout after 5s**



Average data transfer rate	1007.60 MB/s
Average transfer duration	2098.03 ms
Transfer duration exceeded 40s	0
Transfer duration exceeded 40s	0%
Data transfer rate standard deviation	218.13
Transfer duration standard deviation	1271.70

measured before Centos8 tuning

Summary

- With Centos 8 and and firewall tuning (disabling reverse path filtering) **obtained remote I/O performance is close to nominal**
- The initial tests with the client side erasure coding plug-in yielded very promising results
  - much **faster data transfer rate**, cut the long tail (**no overflows**), better memory usage on EPN side
- Next step
  - redo the throughput tests with Centos8 system
  - **test transfers from Alice P2 to CC**