EOS as a DAQ back-end buffer for the ProtoDUNE-DP experiment : from tests to production

EOS workshop, CERN, 3-5/02/2020

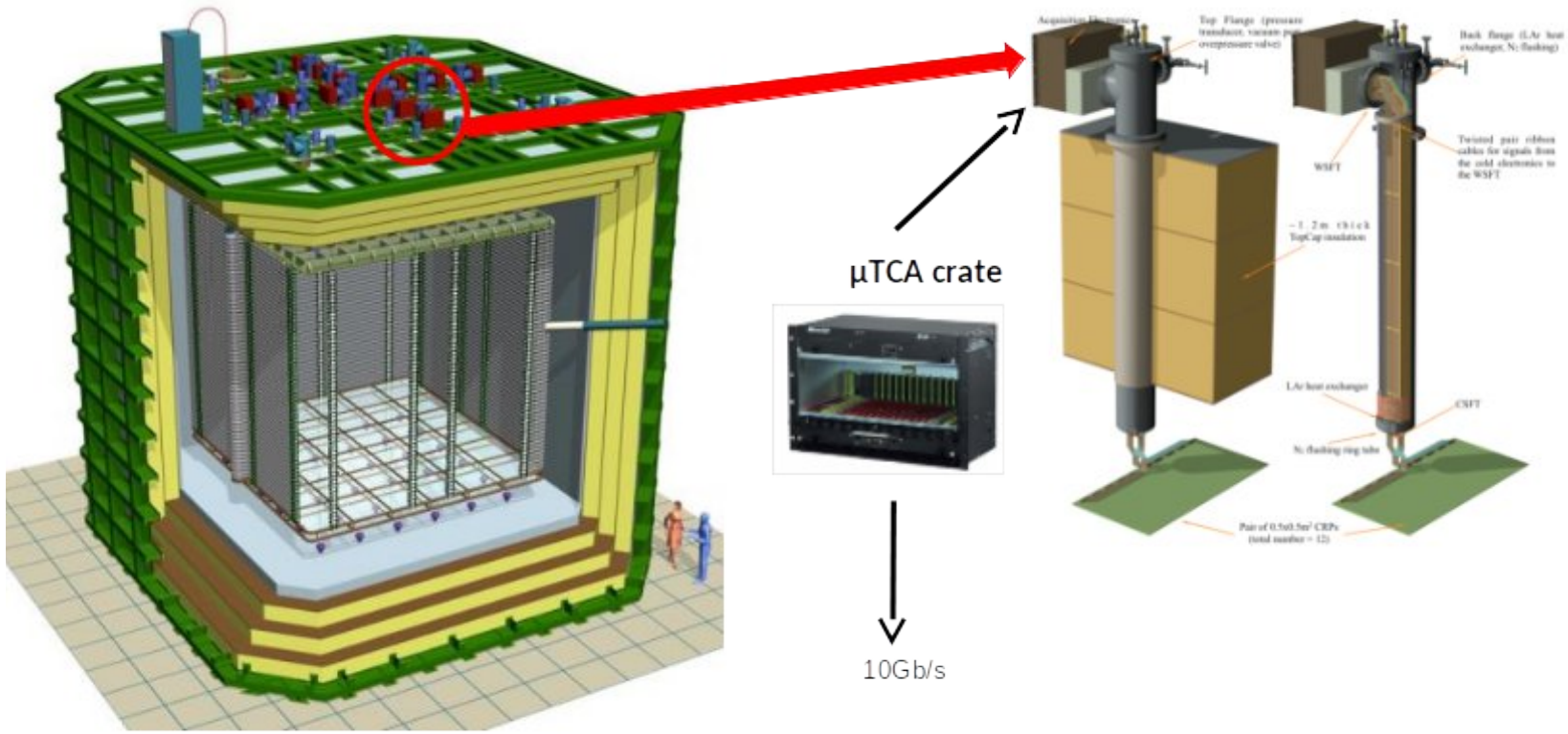
PUGNÈRE Denis

CNRS / IN2P3 / IP2I

2019

Dual-phase protoDUNE DAQ :

6x6x6m^3 active volume (300T liquid argon = 1/20 10kTon LBNO)



Charge: 12 µTCA crates, 10 AMC cards / crate, 64 channels / card => 7680 channels

(12 charge readout + 1 for light readout) * 10 Gb/s links = 13 * 10 Gb/s uplinks to DAQ

ProtoDUNE dual-phase experiment needs

ProtoDUNE **dual-phase** : 146.8MB / event, trigger rate 100Hz

7680 channels, 10 000 samples, 12 bits (2.5Mhz : drift window 4ms) :

=> data rate 130Gb/s

ProtoDUNE dual-phase online DAQ storage buffer specifications :

- ~1 PB (needed to buffer several days of raw data taking)
- It should to store files at a 130Gb/s data rate (raw, no compression)
- It should allow: fast online reconstruction to perform data quality monitoring, and online analysis for assessment of detector performance
- Data moved to the CERN EOSPUBLIC instance via a dedicated 40Gb/s link

Storage system tested (2016)

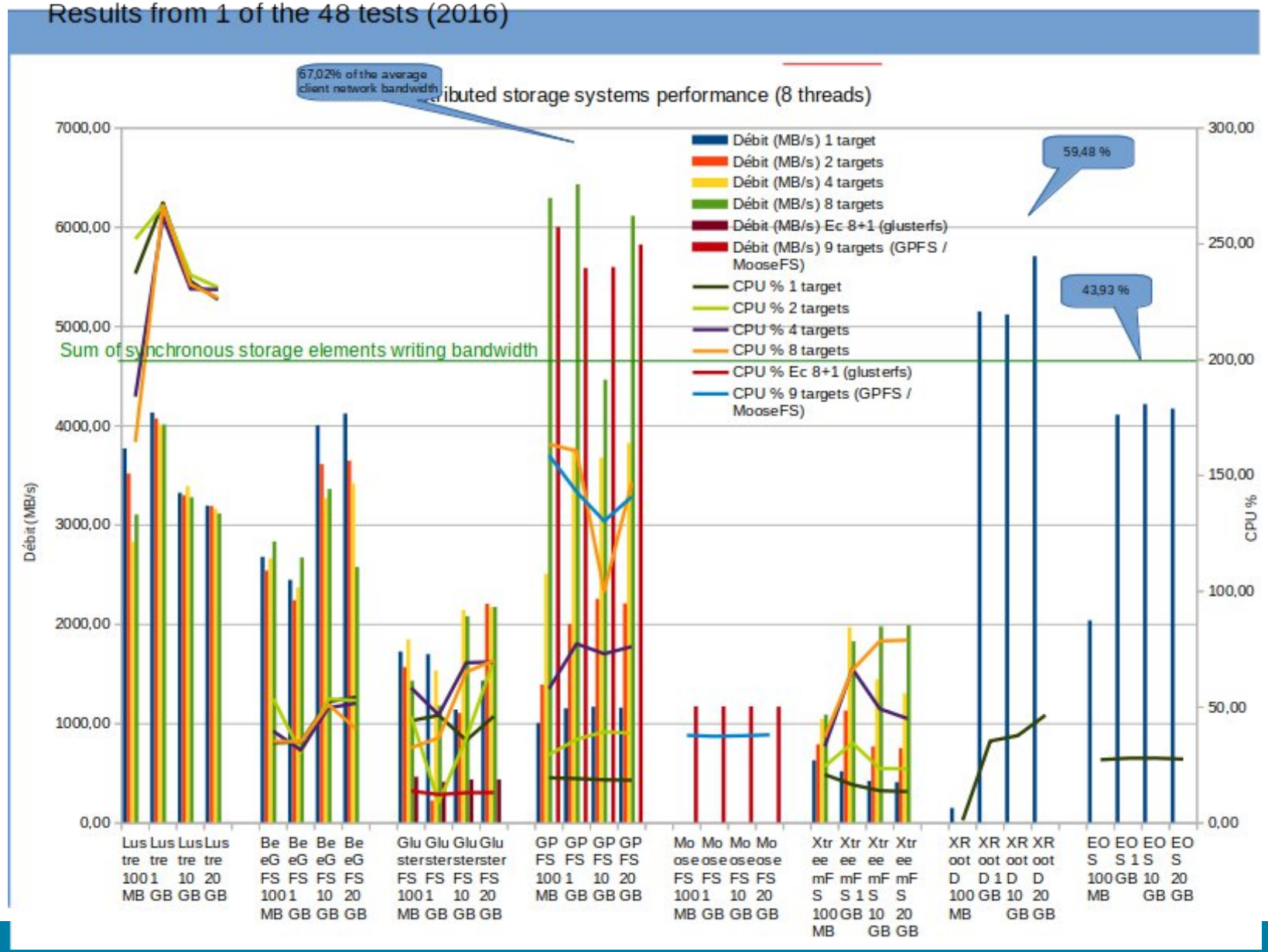
	Lustre	BeeGFS	GlusterFS	GPFS	MooseFS	XtreemFS	XRootD	EOS
Versions	v2.7.0-3	v2015.03.r10	3.7.8-4	v4.2.0-1	2.0.88-1	1.5.1	4.3.0-1	Citrine 4.0.12
POSIX	Yes	Yes	Yes	Yes	Yes	Yes	via FUSE	via FUSE
Open Source	Yes	Client=Yes, Serveur=EULA	Yes	No	Yes	Yes	Yes	Yes
Need for MetaData Server ?	Yes	Metadata + Manager	No	No	Metadata + Manager		Yes	Yes
Support RDMA / Infiniband	Yes	Yes	Yes	Yes	No	No	No	No
Striping	Yes	Yes	Yes	Yes	No	Yes	No	No
Failover	M + D (1)	DR (1)	M + D (1)	M + D (1)	M + DR (1)	M + DR (1)	No	M + D (1)
Quota	Yes	Yes	Yes	Yes	Yes	No	No	Yes
Snapshots	No	No	Yes	Yes	Yes	Yes	No	No
Integrated tool to move data over data servers ?	Yes	Yes	Yes	Yes	No	Yes	No	Yes

(1) : M=Metadata, D=Data, M+D=Metadata+Data, DR=Data Replication

Each file is divided into « chunks »
ditributed over all the storage servers
This is always at the charge of the client CPU (DAQ back-end)

This is now **Yes** with raid6/raiddp

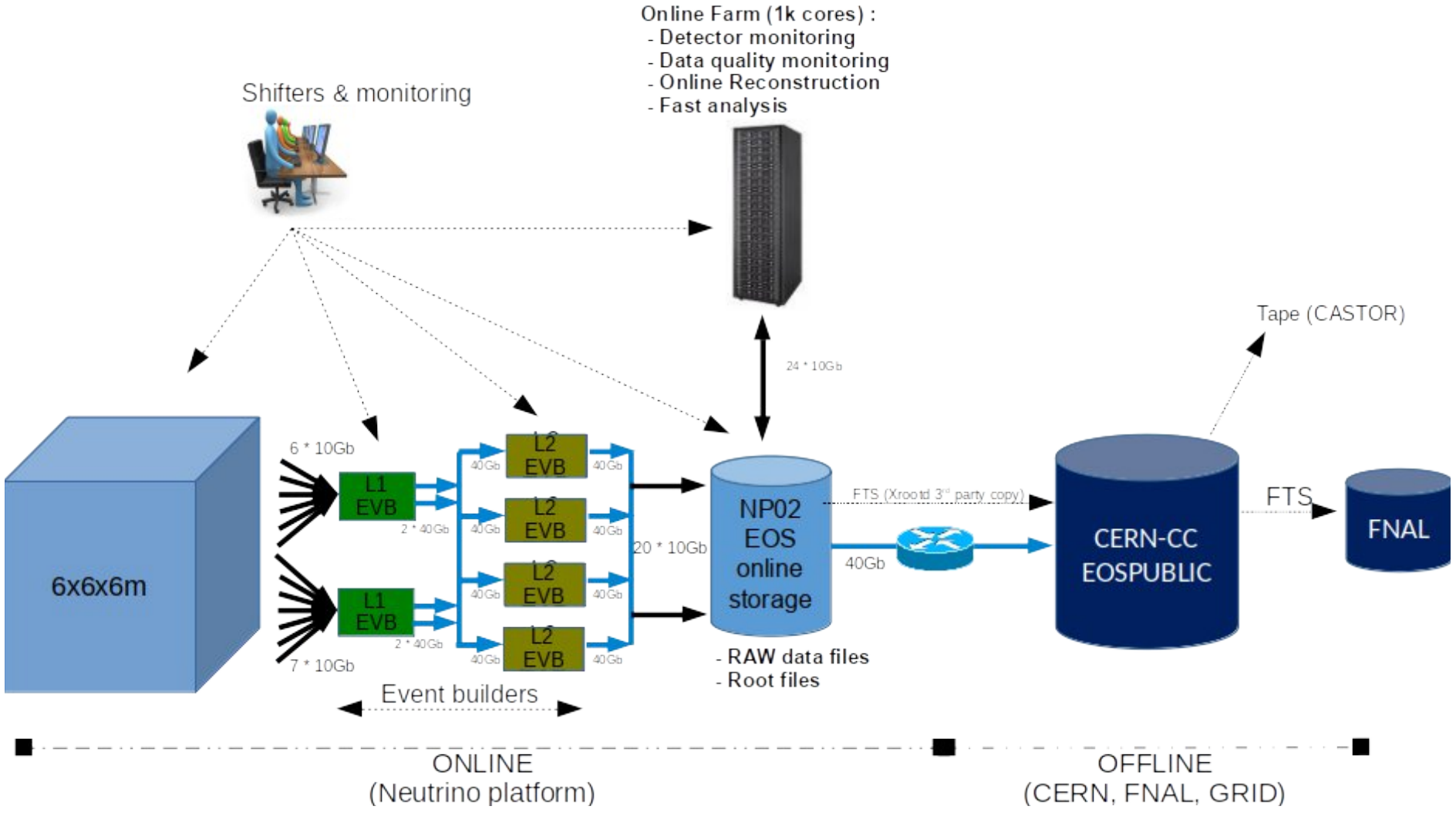
WA105 Technical Board meeting, June 15, 2016 : Results on distributed storage tests
https://indico.fnal.gov/event/12347/contribution/3/material/slides/0.pdf


Results from 1 of the 48 tests (2016)
67,02% of the average client network bandwidth
ributed storage systems performance (8 threads)
Débit (MB/s) 1 target
Débit (MB/s) 2 targets
Débit (MB/s) 4 targets
Débit (MB/s) 8 targets
Débit (MB/s) Ec 8+1 (glusterfs)
Débit (MB/s) 9 targets (GPFS / MooseFS)
CPU % 1 target
CPU % 2 targets
CPU % 4 targets
CPU % 8 targets
CPU % Ec 8+1 (glusterfs)
CPU % 9 targets (GPFS / MooseFS)
59,48 %
43,93 %
Sum of synchronous storage elements writing bandwidth
Débit (MB/s)
CPU %
7000,00
6000,00
5000,00
4000,00
3000,00
2000,00
1000,00
0,00
300,00
250,00
200,00
150,00
100,00
50,00
0,00
Lustre 100 MB
Lustre 1 GB
Lustre 10 GB
Lustre 20 GB
BeeGFS 100 MB
BeeGFS 1 GB
BeeGFS 10 GB
BeeGFS 20 GB
Glusterfs 100 MB
Glusterfs 1 GB
Glusterfs 10 GB
Glusterfs 20 GB
GPFS 100 MB
GPFS 1 GB
GPFS 10 GB
GPFS 20 GB
Moose FS 100 MB
Moose FS 1 GB
Moose FS 10 GB
Moose FS 20 GB
XtreemFS 100 MB
XtreemFS 1 GB
XtreemFS 10 GB
XtreemFS 20 GB
XRootD 100 MB
XRootD 1 GB
XRootD 10 GB
XRootD 20 GB
EOS 100 MB
EOS 1 GB
EOS 10 GB
EOS 20 GB

Storage back-end choice : EOS

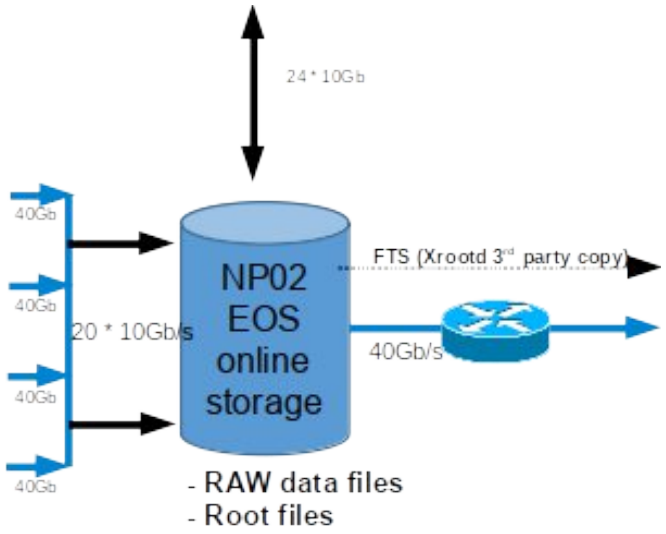
- EOS chosen (after the 2016 tests) :
  - **Low-latency storage**,
  - **Very efficient on the client side** (XrootD based),
  - POSIX, Kerberos, GSI access control,
  - XrootD, POSIX file access protocol,
  - **3rd party-copy support** (used for FTS),
  - Check-sums support,
  - Redundancy (old hardware, remote operating) :
    - Meta-data servers
    - Data server (2 replicas or RAIN raid6/raiddp) <- not yet used
- Data **server life-cycle management** (draining, start/stop operation)

ProtoDUNE Dual-Phase DAQ back-end design



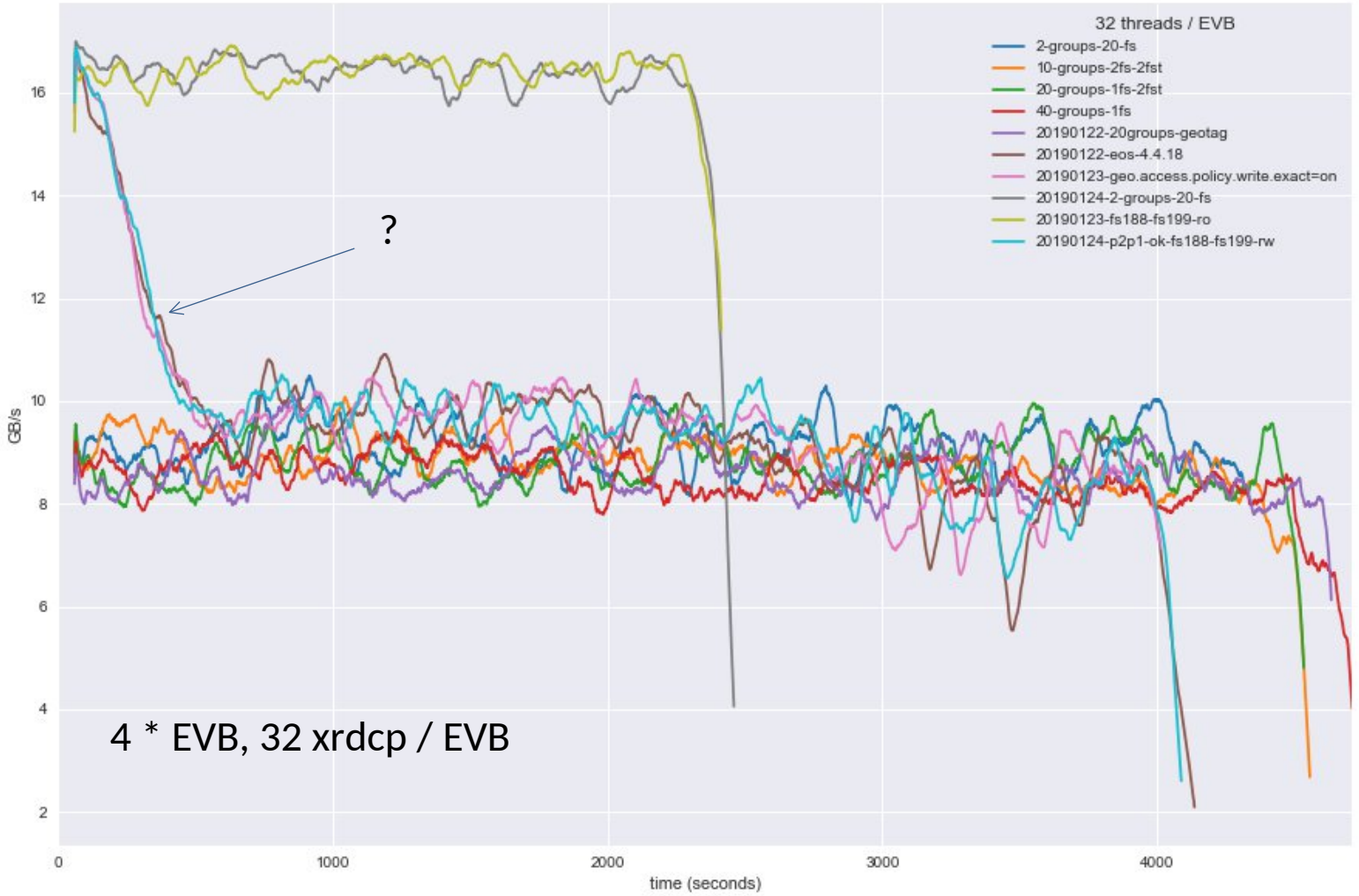
The ProtoDUNE Dual-Phase storage back-end

- NP02 EOS instance :
  - 20 * Data storage servers (= 20 EOS FST)
    - (very) old Dell R510, 2 * CPU E5620, 32 GB RAM) : 12 * 3TB SAS HDD
    - Dell MD1200 : 12 * 3TB SAS HDD
    - 1 * 10Gb/s
  - 2 * EOS Metadata servers (MGM)
    - Dell R610, 2 * CPU E5540, 48 GB RAM
  - 3 * QuarkDB metadata servers (QDB)
    - Dell R610, 2 * CPU E5540, 24 GB RAM, DB on SSDs



The stress-tests before the production

- Until the beginning of 2019 :
  - Various configuration tests to find the optimal layout
  - Various stress-tests to find hot points (MD or FST saturation)
- Current configuration :
  - 20 * FST,
  - 4 * HW RAID 6 (6 HDD / RAID)
  - 4 * FS / FST, 4 groups



The production : ProtoDUNE Dual-Phase first acquisitions

ProtoDUNE-DP operations started on August 28th 2019 : 1.9M events have been collected so far.

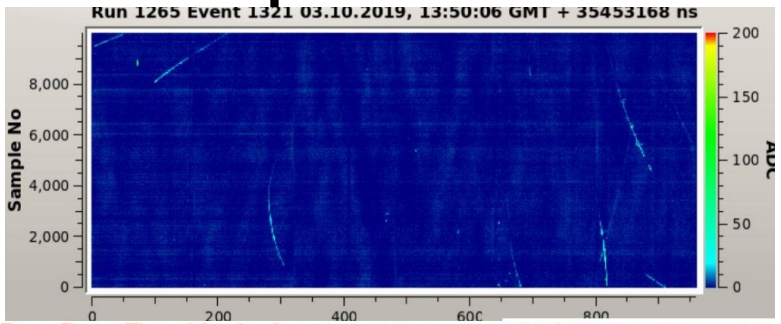
Workflow :

* Raw data file assembly by one (of the 4) L2 Event-Builder), file size = 3 GB (200 compressed events)
* local processing (fast track reconstruction and data quality @ 15 evt/sec)
* FTS3 copies the RAW data & metadata files from local NP02EOS buffer to EOSPUBLIC
* Then FTS3 => FNAL, then RUCIO to the WLCG grid

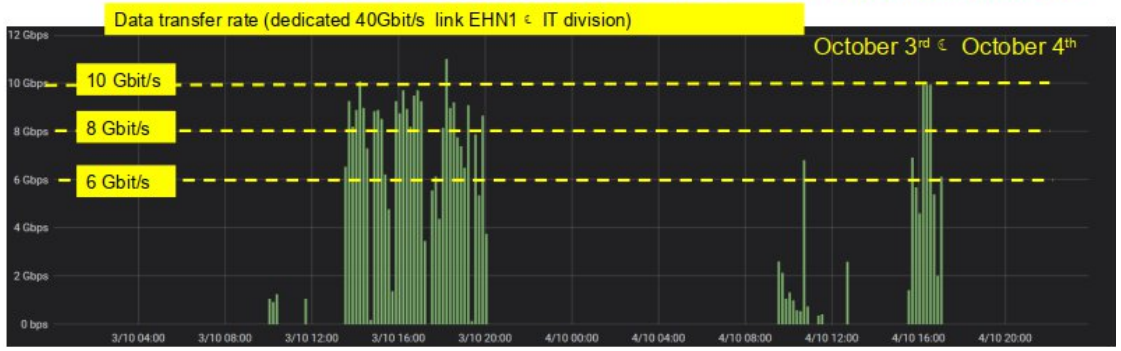
The delay Δt between the creation of a Raw Data file and its availability on EOSPUBLIC is 15 minutes

During the production runs : No bad (lost / empty / check-sum) files in the local EOS buffer !

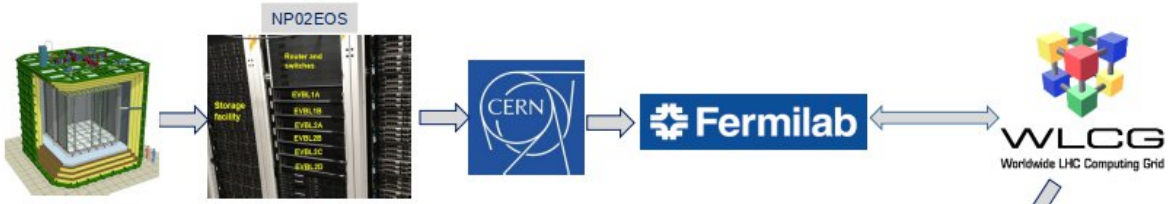
1 RAW event display



Raw Data Flow Monitoring (NP02EOS ↔ EOSPUBLIC) some examples:



The workflow is the following:



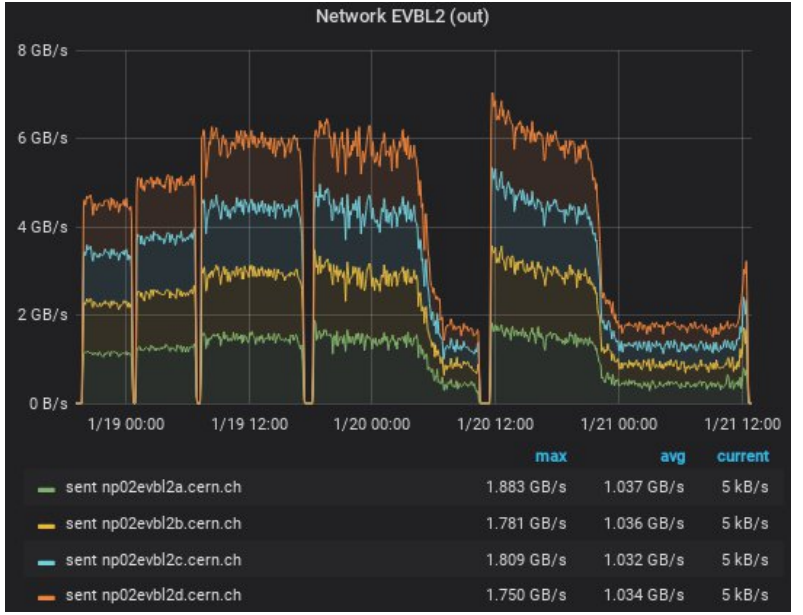
Reconstruction output is stored on tape at CCIN2P3 (resources available)

The stress-tests between 2 production runs

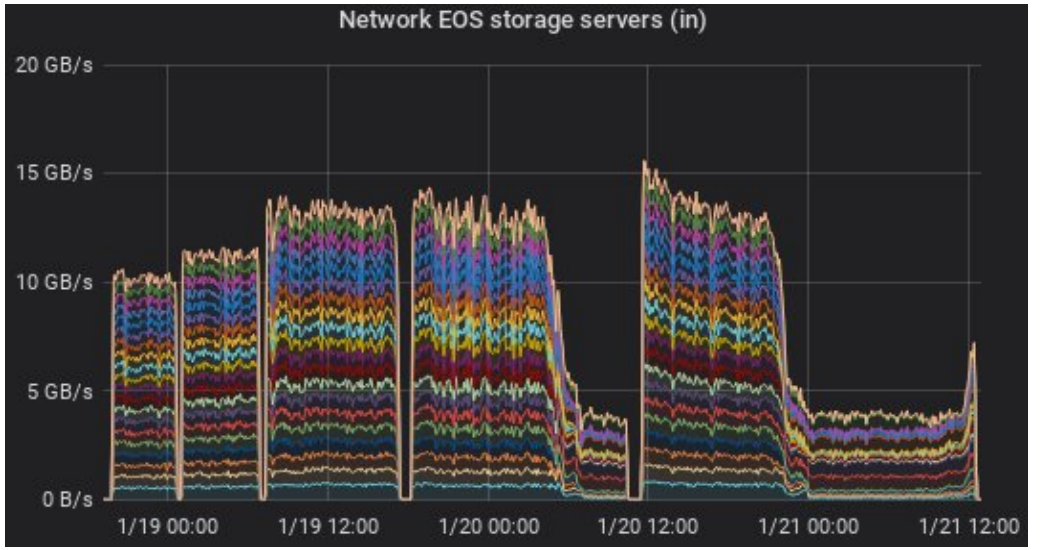
- We are now in a ≠ configuration (Name Space : Memory -> QuarkDB)
- continuing stress-tests

- "plain" layout :
  - On the most high rate tests (128 xrdcp in //) :
    - some problems (< 0,01 % on 128k 3GB files created at a > 17 GB/s continuous rate)
    - some empty files, some files not created
  - no problem at a lower rate
- "RAID6" layout (RAIN) :
  - rate : 80 xrdcp in // (80k * 3GB files) :
    - some problems : < 0,04 % on 80k 3GB files **not** created
  - rate : 128 xrdcp in // (128k * 3GB files) :
    - many problems : **> 23 % on 128k 3GB files not created**
  - no problem at a lower rate

- So we will stay with : plain (no replica, no RAIN) layout



EOS RAID6 tests :
24, 32, 64, 80, 128
// xrdcp, 3GB files



The real life : The ~~daily~~ EOS operation

- No problem during the production. Business as usual :
  - hosts / services monitoring,
  - replacing drives…
  - draining FST for maintenance… see if there is still some stripes remaining on the FST … maintenance .. and then back to 'rw' status
  - this is not a daily task, just a weekly or monthly task, **low human overhead**

- Name-space evolution (memory to QuarkDB transition) :
  - prepared with reading the EOS documentation and Q&A forum https://eos-community.web.cern.ch : **huge help from the EOS team and the community !**
  - some days reading the forum, then building the procedure and finally half a day transition (stressed but DONE! ;-)

- QuarkDB namespace has simplified the active / passive MGM management !

Conclusion

- EOS does the job (thanks EOS team !)
- The ProtoDUNE-DP online storage system is running smoothly [*]
- We are considering still using the "plain" layout, there are too major drawbacks (lower performance, inter FST traffic, lost files) using the RAIN layout for our case.

[*] : It survived from several power-cuts in EHN1 building \o/