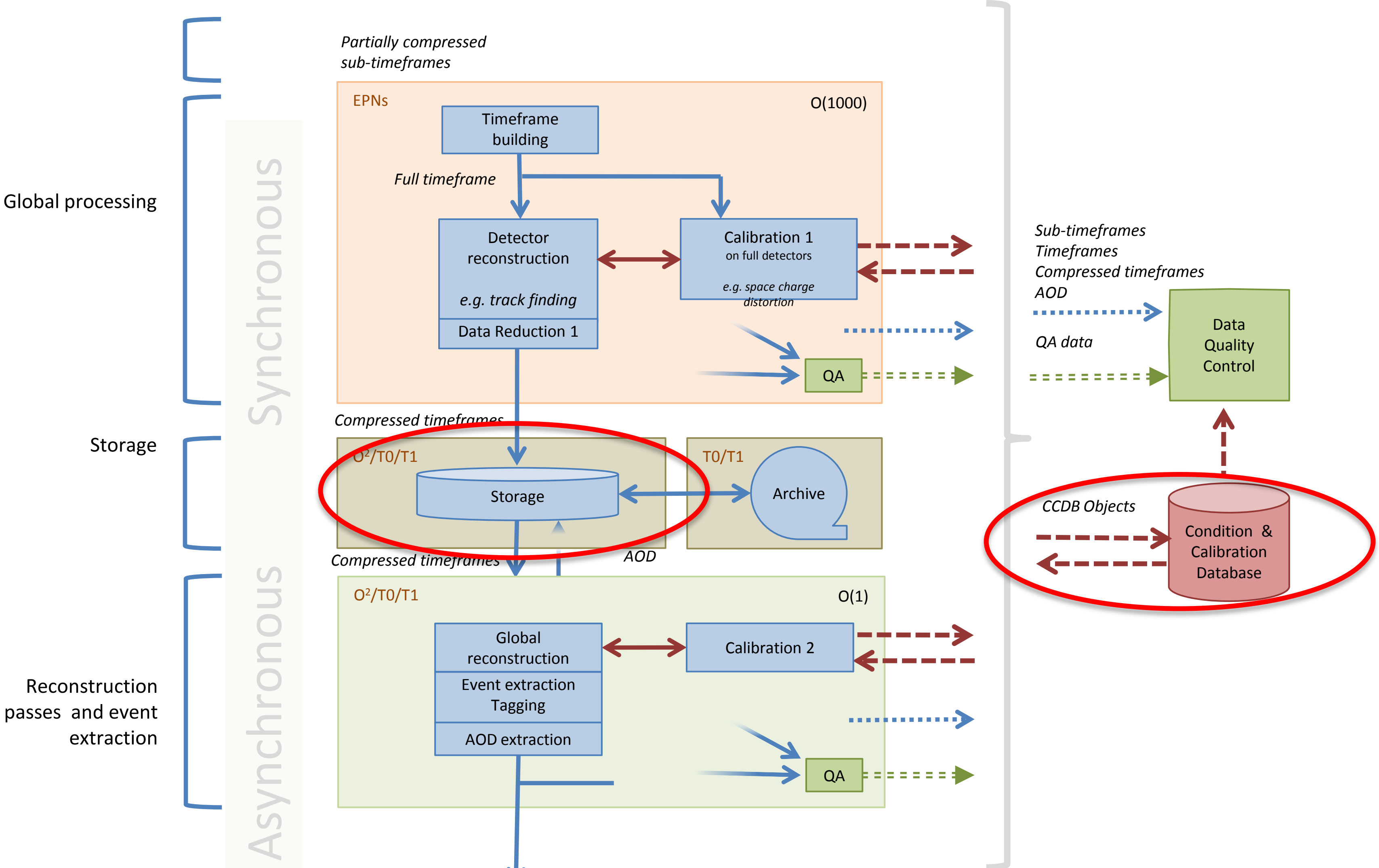
O^2 Project : Data Storage

4th ALICE ITS, MFT and O2 Asian workshop
Pusan, South Korea, 15-16 December 2014

P. Vande Vyvre / CERN-PH

Data Storage

Environment and characteristics



Data Storage

Requirements

	Clients	Total Bandw.	Usage	Storage Unit	Number of Units	Total Capacity
	123456	123456	123456	123456	123456	123456
Compressed Physics Data	EPN OM(1000)	100 GB/s 50 GB/s	Wr once Synch. Sust. Seq. Rd several Asynch Inside O2 & outside	Files of several GB	1.0E+09 files	100 PB
Condition & Calibration Database	FLP+EPN OM(1000)	A few GB/s max	Sust. Direct access Internal	Records of a few kB		A few tens of TB

Data Storage for compressed physics data

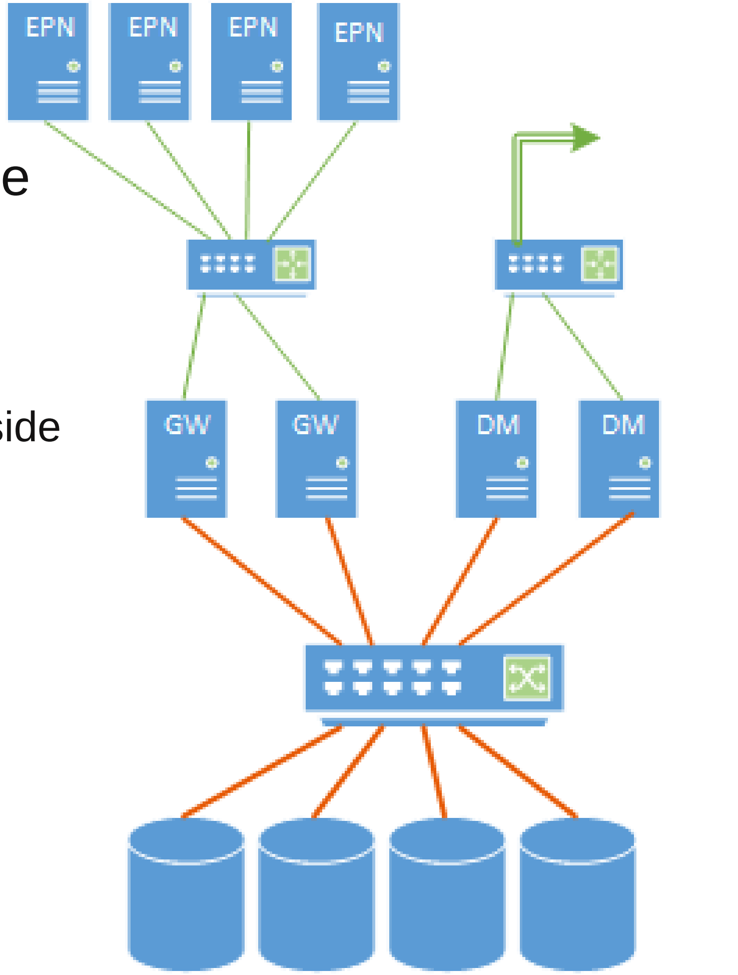
Possible solutions

- Solution used in Run1-2 DAQ system (Clustered file system mounted on all EPNs): possible but expensive
- Local storage on each node: possible but not practical
- Commercial or open clustered file system (e.g. Lustre) with gateways
- Key/value store (e.g. F4 Facebook)
- CERN disk-based storage system (EOS)

Data Storage for compressed physics data

General architecture

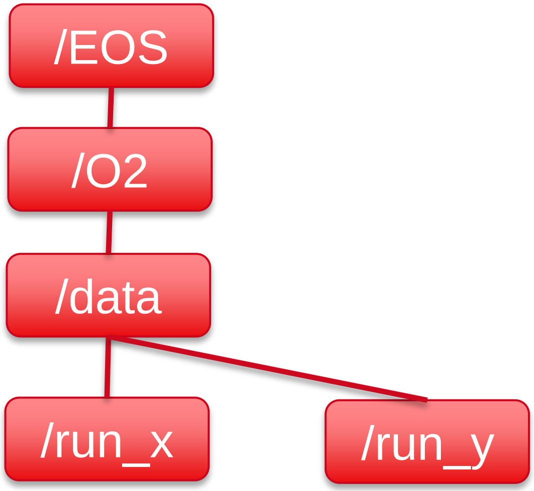
- Bandwidth from each EPN is low and does not justify high speed link to the data storage
- Introduce a few high performance nodes:
  - Gateway (GW) for local data writing/reading
  - Data mover (DM) for data exchange with the outside



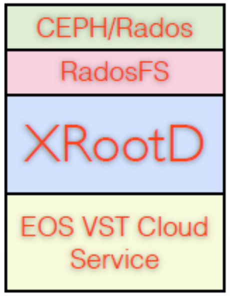
Data Storage for compressed physics data

EOS : Virtual storage cloud

- EOS is the CERN disk-only file storage
- Several instances: CERN, FNAL, SASKE, Subatech, SINICA, RRC Kurchatov, UNAM
- 25.000 disks - 60 PB storage space - 200 Mio files
- Service since 2012 - 1-year availability including scheduled downtimes 99.5% at CERN.
- Virtualized (HTTP enabled) global and cloud storage
- Storage federation: unique namespace and “infinite” storage space



VST
Volume Storage Server



Data Storage for compressed physics data

Key-value store

- Facebook warm BLOB (Binary Large OBject) storage system
- Key-value store for immutable BLOBs (photos, videos, . . .)
- Data organization: 100 GB volumes with in-memory index
- In production since almost 2 years
  - 65 PB, 400 Billion pictures
- WORM access pattern. No delete !

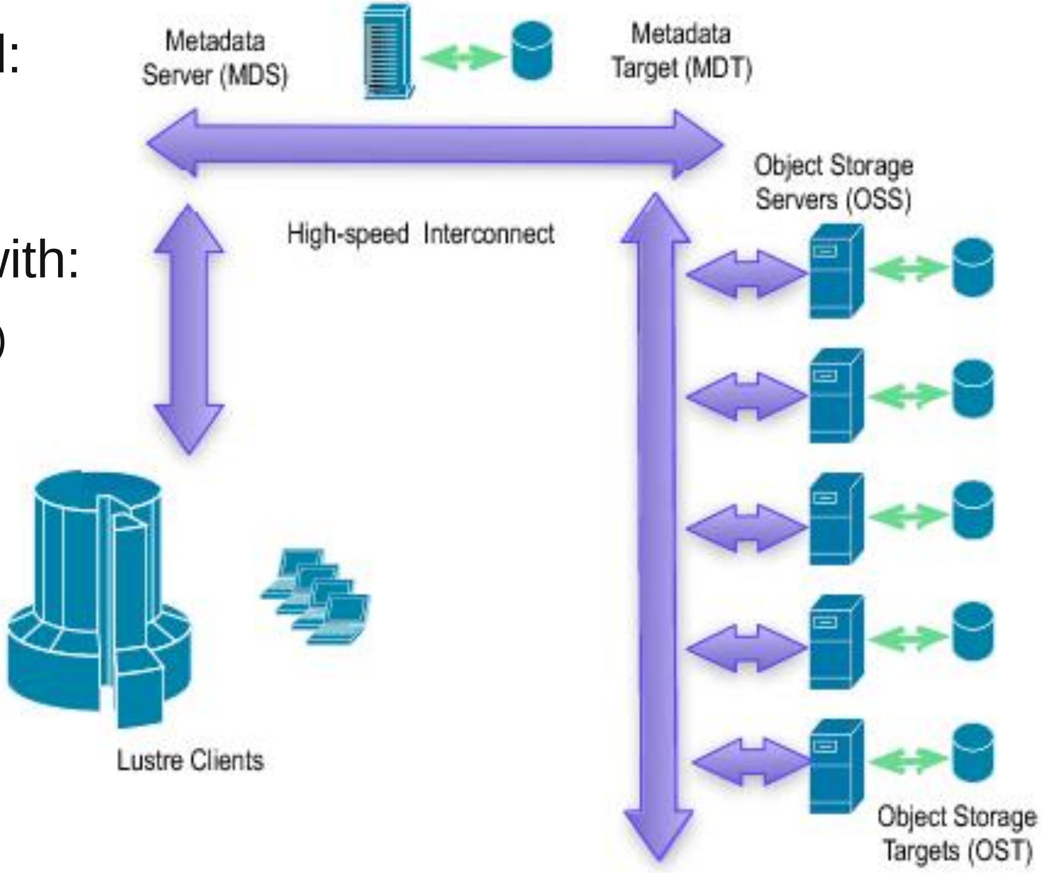
Data Storage for compressed physics data

Lustre: clustered file system

- Many occurrences in the world:

  ~50% of Top500, Titan e.g.

- A proof of concept with done with:
  - Intel (distributing the Lustre CFS)
  - Dell for the hardware
- At
  - DAQ lab
  - Dell technical centre in Germany


Metadata Server (MDS)
Metadata Target (MDT)
Object Storage Servers (OSS)
High-speed Interconnect
Lustre Clients
Object Storage Targets (OST)