
CERN

@

Hervé Rousseau & Cristian Contescu, on behalf of the EOS Operations Team

Outline

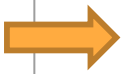
- Glimpse into 2018
- Challenges & improvements
  - New catalogue
  - Splitting instances
- New hardware
- What lies ahead…

2018 vs. 2019 in numbers

	2018	2019
Nodes	~1400	~1400
Disks	~50000	~60000
Raw capacity	~250PB	~270PB
# of files	~2.6 billion (March)	~3.8 billion (January)

2018 vs. 2019 in numbers

	2018	2019
Nodes	~1400	~1400
Disks	~50000	~60000
Raw capacity	~250PB	~270PB
# of files	~2.6 billion (March)	~3.8 billion (January)



+50%

Glimpse into 2018

- Focus on the service stability
- New catalogue deployment started
- Heavy Ion data taking
- The year of the migration:
  - EOSUSER ⇨ EOSHOME(s)
  - Client side: FUSE ⇨ FUSEx

Glimpse into 2018

- Focus on the service stability
- New catalogue deployment started
- Heavy Ion data taking
- The year of the migration:
  - EOSUSER ⇨ EOSHOME(s)
  - Client side: FUSE ⇨ FUSEx

} still ongoing...

Service challenges

- Expectations: availability & reliability
- …or at least a quick recovery after an incident
- No system is perfect

Service challenges

- Expectations: availability & reliability
- …or at least a quick recovery after an incident
- No system is perfect, but it can be improved

Respond to expectations

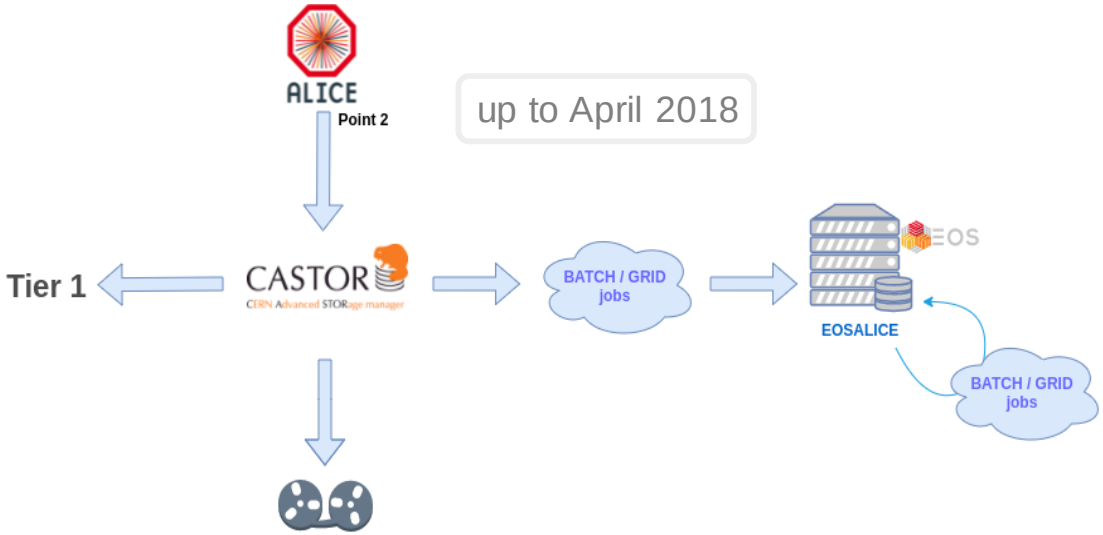
- Improve software stability
- New component developed allowing for quicker service recovery
  - In-memory catalogue => QuarkDB
    - RocksDB + raft consensus algorithm for high availability
    - Reality check: boot time from ~1000s to 1s
    - RAM no longer a limiting factor
  - Deployment started in 2018, to be cont'd

Respond to expectations

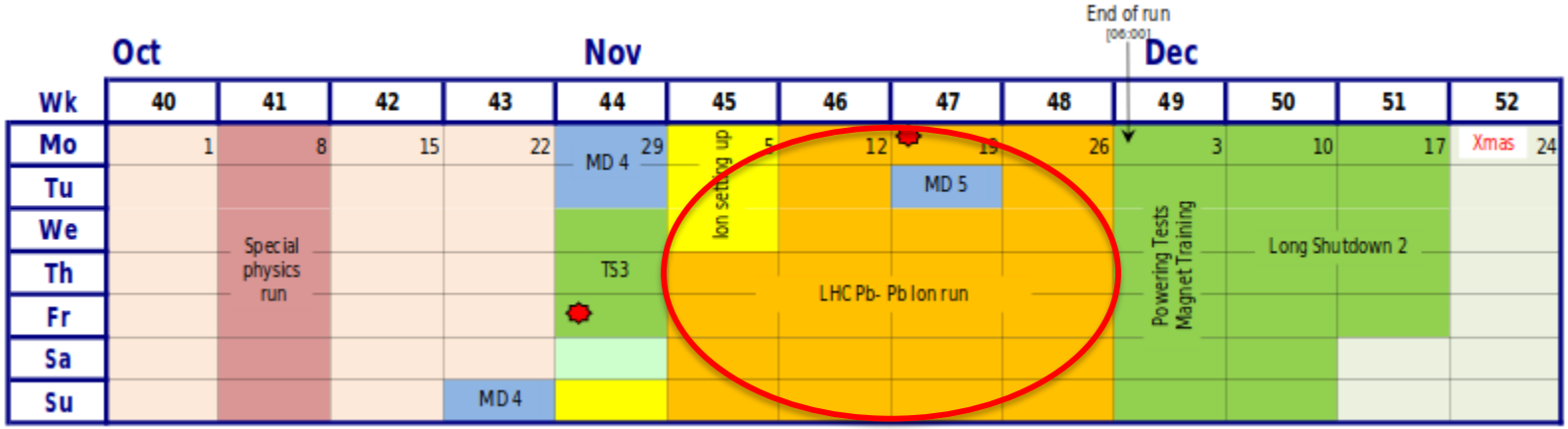
- Scale horizontally:
  - Split instances where possible
    - Faster recovery
    - Less entities affected by downtime
    - Allows staged upgrades

The ALICE DAQ case

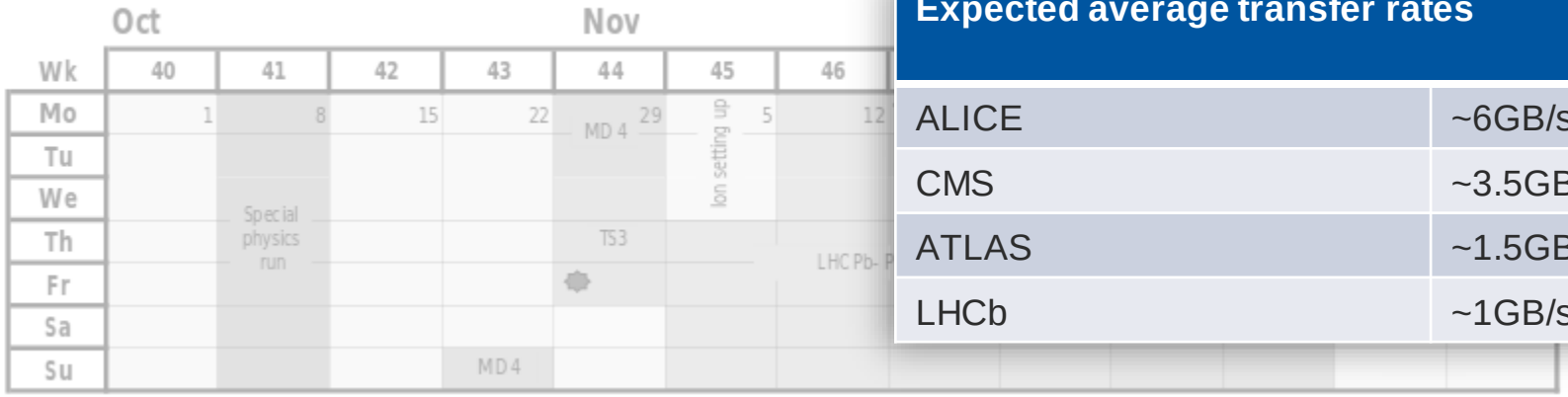
- Initial dataflow setup:
  - ALICE P2 => CASTOR
  - CASTOR =>
    - Tape
    - Tier 1
    - Reco => EOSALICE
  - Further analysis on EOSALICE



The ALICE DAQ case



The ALICE DAQ case

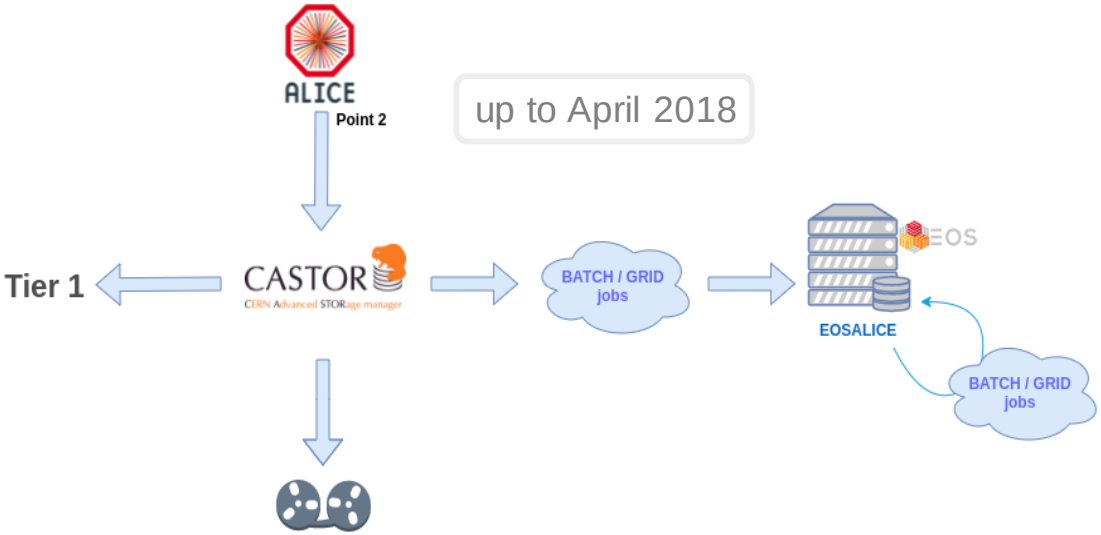


Expected average transfer rates	
ALICE	~6GB/s
CMS	~3.5GB/s
ATLAS	~1.5GB/s
LHCb	~1GB/s

Larger event sizes => increased data rates => (Network & Storage challenges)++

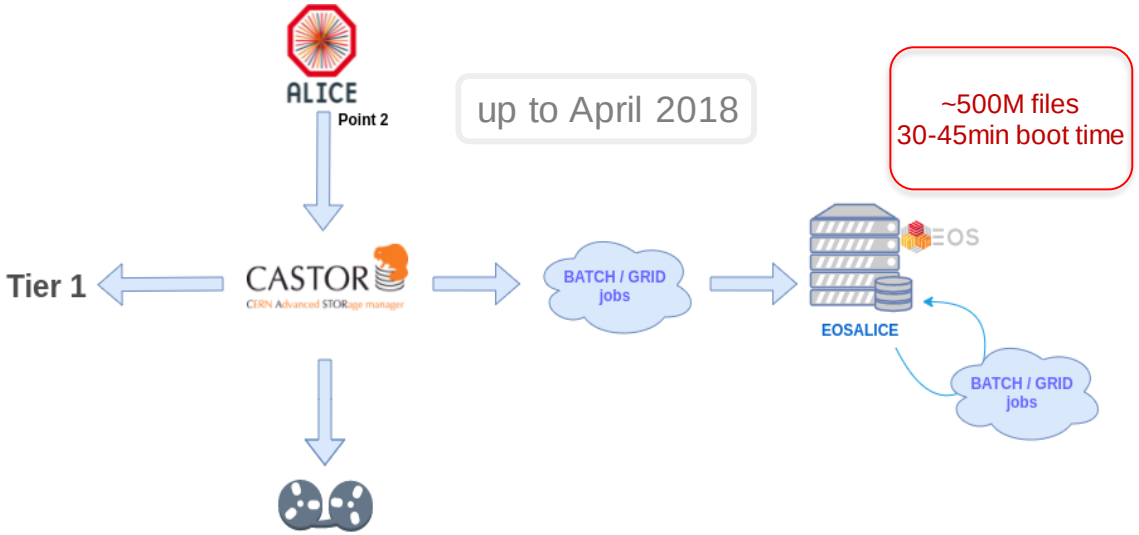
The ALICE DAQ case

- Initial dataflow setup:
  - ALICE P2 => CASTOR
  - CASTOR =>
    - Tape
    - Tier 1
    - Reco => EOSALICE
  - Further analysis on EOSALICE



The ALICE DAQ case

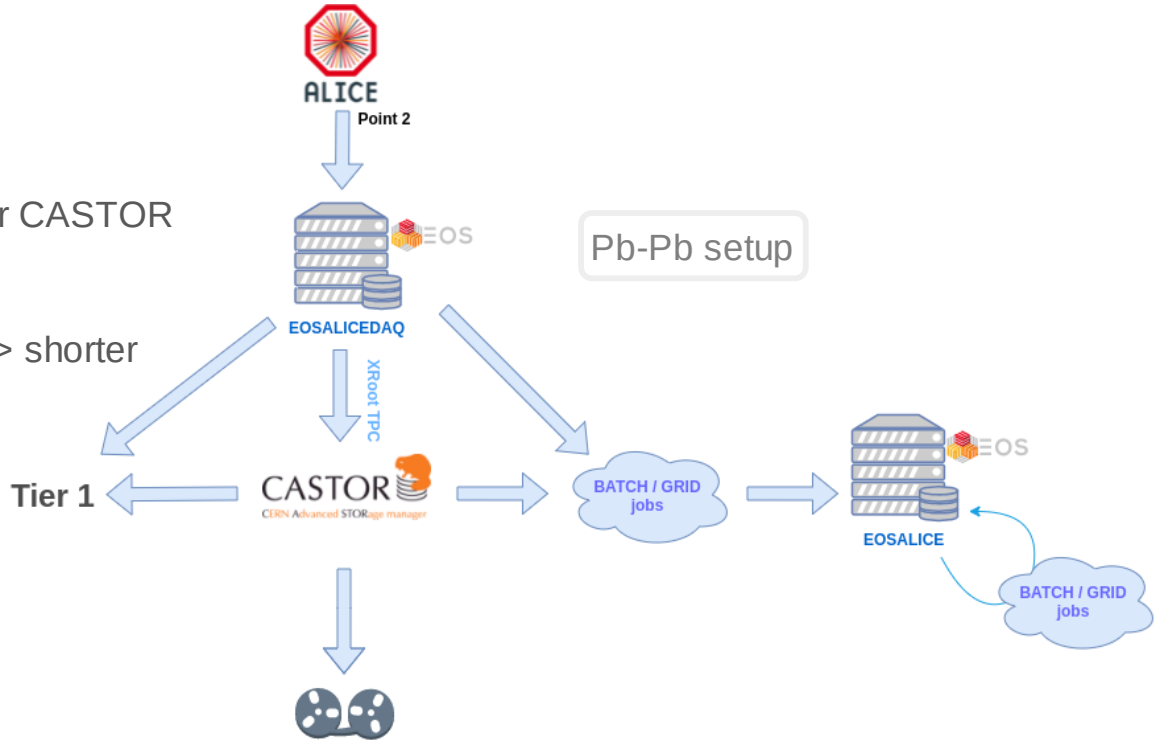
- Initial dataflow setup:
  - ALICE P2 => CASTOR
  - CASTOR =>
    - Tape
    - Tier 1
    - Reco => EOSALICE
  - Further analysis on EOSALICE



The ALICE DAQ case

HI dataflow setup:

- ALICE => EOSALICEDAQ
- EOSALICEDAQ as extra disk buffer for CASTOR
- Quick(er) data recall if needed
- Less files stored on EOSALICEDAQ => shorter recovery time



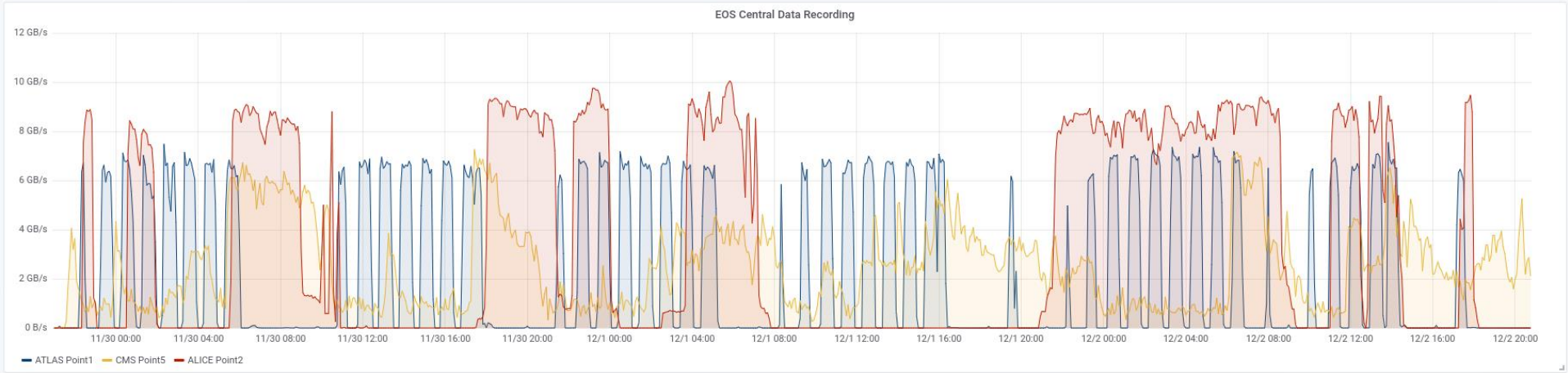
The ALICE DAQ case – data challenge

- Setup validated during mid-September combined data challenge (ALICE+ATLAS+CMS)

- Interesting finding during the DC:
  - If requested to xrdcp, MD5 checksum was computed by reading twice the source file
    - performance impact
  - Disk scheduler issue observed on CC7 storage nodes; deadline vs. cfq; fixed

- At last successful
  - Expected data rates look sustainable (after fixing the quirks above)
  - No Alice <=> CMS interference
  - More details: https://indico.cern.ch/event/758256/

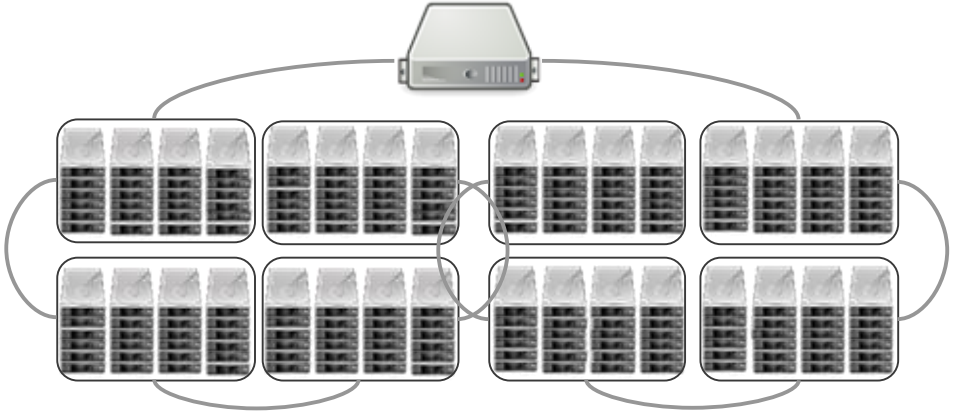
The ALICE DAQ case – real data taking



New hardware (aka ‘the monsters’)

- The monster machines are now in PROD

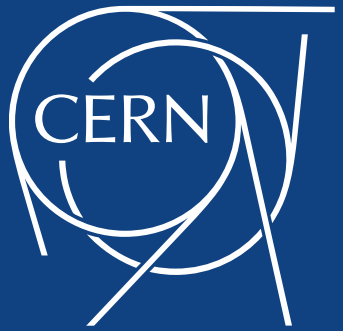
used-bytes	max-bytes
328.43 TB	2.30 PB
326.07 TB	2.30 PB
317.73 TB	2.30 PB



What lies ahead…

- Finish the migration to the new catalogue
- Wigner decommissioning
- New EOSPROJECT instance(s)
  - replacing the project areas on EOSUSER and AFS
- Hardware retirements

Thank you !



home.cern