WLCG Workshop

19th Nov 1 B.C.*

LHCb point of view on Storage Evolution

* B.C.: Being Confined

SRM: should I stay or should I go

- **Tape:**
  - nothing can really replace SRM for tape operations
  - consensus in DOMA TPC that it should stay [1] and we agree
  - Only CTA does not provide it, but provides compatible interface
- **Disk:**
  - We can do without, providing an xroot endpoint, a gsiftp endpoint (see TPC slide), the famous json accounting file [2]

- [1]: https://indico.cern.ch/event/962019/
- [2] Storage Space Accounting definition

Tokens

- **Storage issued tokens**
  - Requested that the token request be done in gfal2 [1]
  - Once this is done, transparent to us as soon as sites enable it
- **VO issued tokens**
  - DIRAC tightly coupled to VOMS
  - Requires big rework of the framework
  - Timeline O(year) (not 2021)

[1] https://its.cern.ch/jira/projects/DMC/issues/DMC-1228

Workflows and storage access

- **In general:**
  - Full file read, no sparse read
  - ALWAYS favor LAN over WAN
  - Run where the data is
- **Production jobs:**
  - Download the file on the worker node
- **User jobs, Working Group productions**
  - Remote xroot read (LAN first, failover if file cannot be opened)

Workflows and storage access

- **Download is more reliable than remote read**
  - Histogram merging done with remote read shows non negligible failure rate
  - Flaky connections result in job crashing
- **Latency does not show to be problematic**
  - No IO bound applications
  - May change with the evolution of our new event model

Workflows and storage access

- **In conclusion:**
  - Locality is paramount and key to job efficiency
  - Always favor LAN over WAN access
  - Download files on the worker node when possible
  - Caches are of no use for us

Interest in QoS

- **[1] LHCb presentation QoS WS**
  - LHCb ideas were inline with examples of white paper
  - Mostly interested in reliability (safer disk/tape)
  - QoS transition performance (aka staging) should be taken into account
  - Important that QoS is exposed via “simple” attributes (namespace, hostname)

Storage-less sites

- **Sites used for MC production**
  - Occasionally for user jobs without input data
- **No strategy change foreseen for HL-LHC**
- **Sites with storage are expected to have reliable network connectivity**

Caches

- **No, thanks**

TPC

- **Adding an extra TPC in DIRAC is trivial**
- **LHCb strongly objects the multihop approach**
  - Leads to the need of one protocol supported across WLCG
    - Acknowledged by DOMA TPC, https is put forward (remains the CTA question though)
- **All our TPC are going through FTS**

Data lake model

- **Ideally, a data lake looks just like a single site with a single external interface**
- **But in practice ...**
  - data locality → Lake network has to be as efficient as LAN
  - We lose diagnostic capabilities
    - CERN tests with Clouds/Wigner shows that we can not afford that

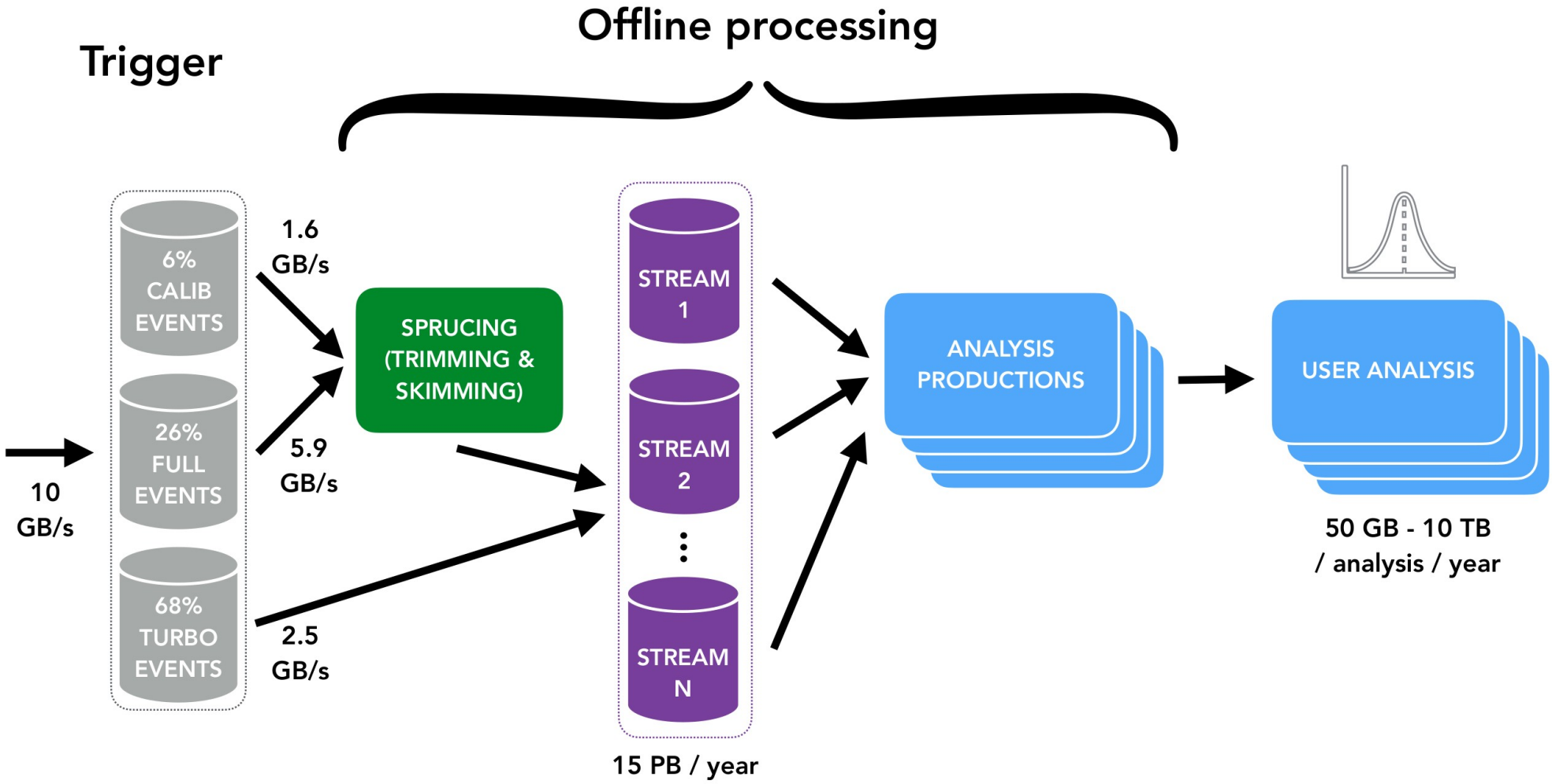
Storage for HPC/Clouds

- **No experience**
- **Plan to stick to MC Simulation only**
  - No input data

User analysis evolution

- **Lot of work ongoing in “Data Processing & Analysis” (DPA) project**
- **General trend is to go towards organized analysis productions**
  - Halfway between plain user jobs and centralized productions

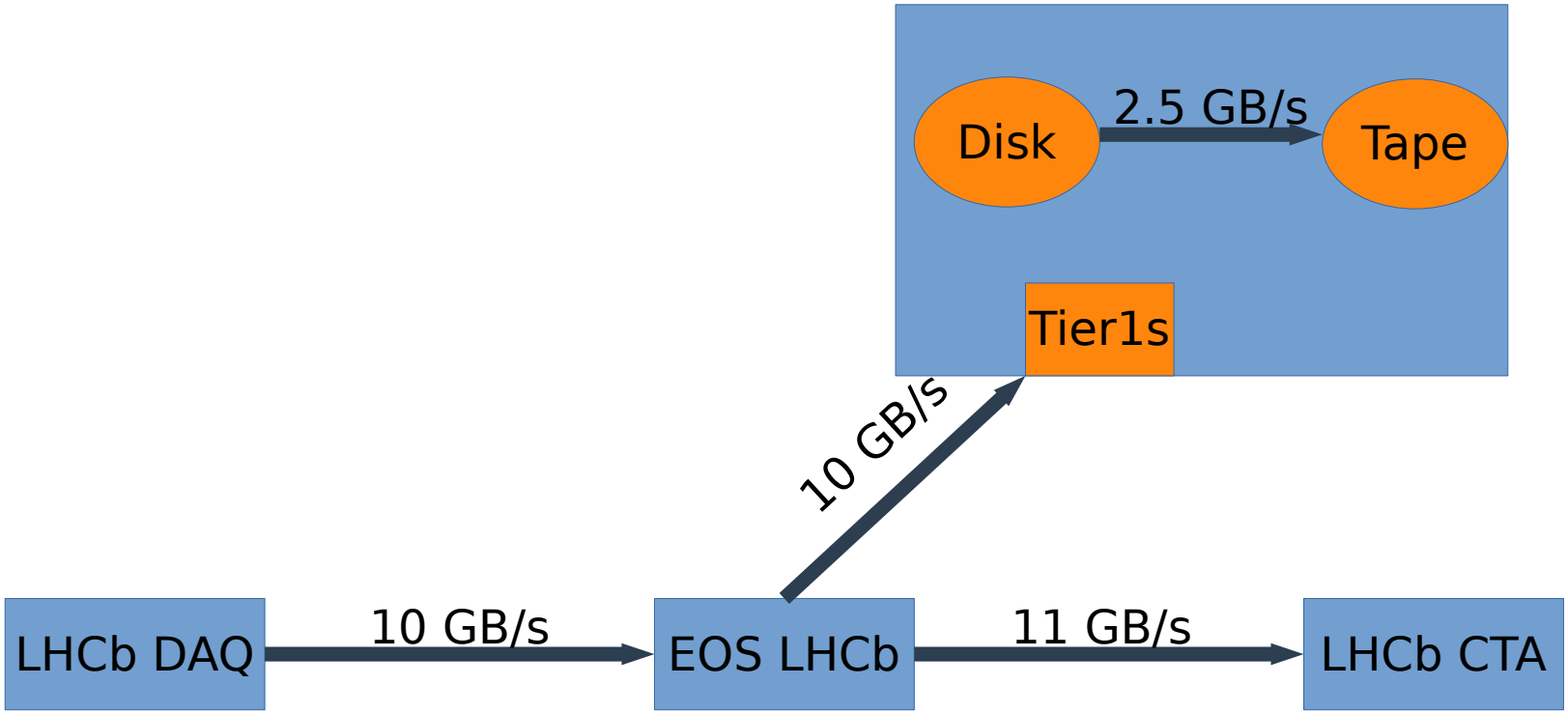
User analysis evolution



Staging speed vs buffer space

Data workflow and throughput to tape during data taking

LHCb computing TDR section 6.1.3



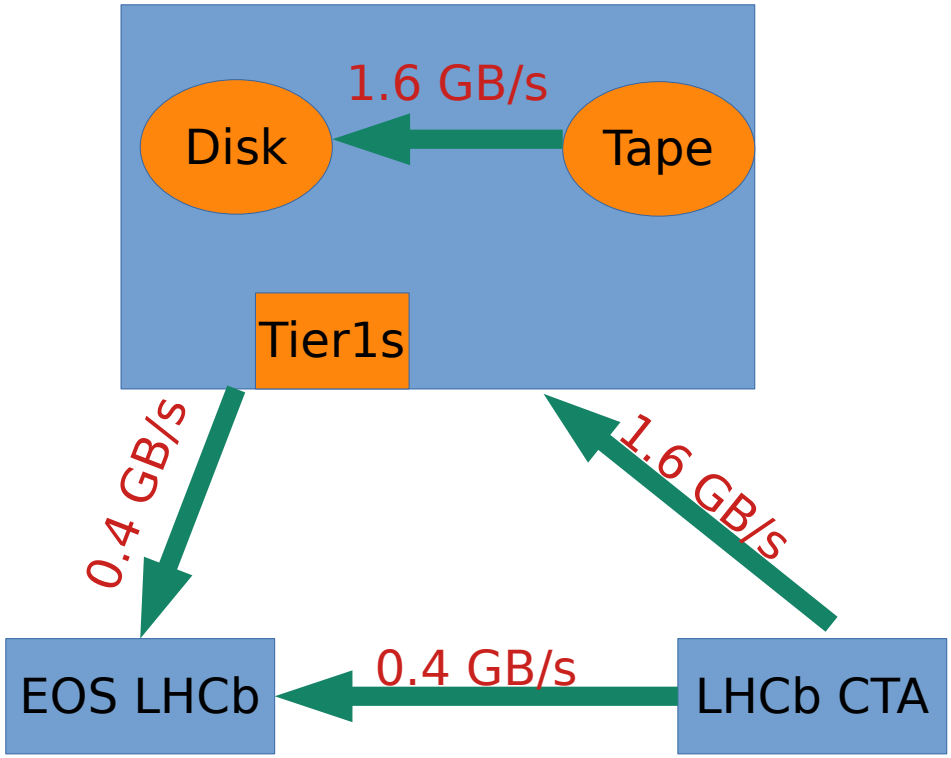
Caution: unit is GB per LHC second

Staging speed vs buffer space

Data workflow during winter shutdown

LHCb computing TDR section 6.1.3

Reprocessing in 4 months
means 4GB/s staging speed



Caution: unit is GB per real second

Staging speed vs buffer space

- **4 months is the maximum time allowed for reprocessing**
  - Can sites do twice as fast (~8GB/s aggregated T1+CERN) ?
  - During Run2, observed aggregated throughput ~1GB/s
- **Staging faster → smaller buffer needed**
- **Note: tape classes show very efficient for massive recall**
- **Conclusion: staging throughput is not to be forgotten**
  - Especially if more experiments start having similar reprocessing strategies (e.g. Data Carousel)

Summary

- We need one TPC protocol available everywhere
- Local file > LAN > WAN
- Run the job where the data is
- No interest in caches
- Storage less sites → MC simulations
- Staging performance is key for Run3