Data storage

current status

Outline

Data storage @ LHC experiments (+ CERN/IT)

From local recording to permanent storage

- Requirements
- Architectures
- Implementations (recording and migration)
- Data flow and procedures
- Commonalities and peculiarities
- PROs and CONs
- Experiences
- Future plans

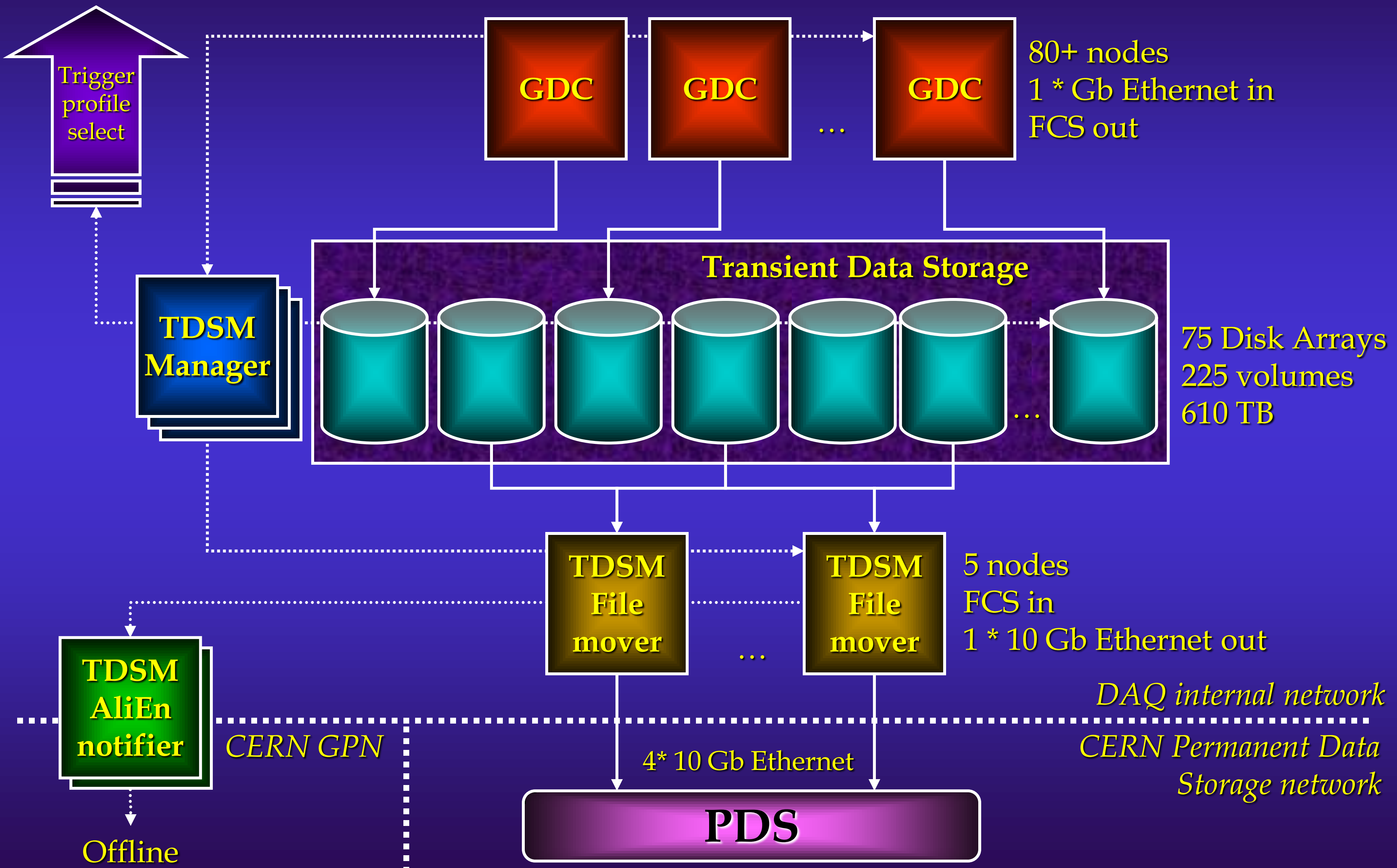
Requirements

- Varying according to beam types, LHC setup, beam luminosity, fill quality…
- Shown here are typical must-be-ready-to-sustain values, maximum achievable figures can be (much) higher

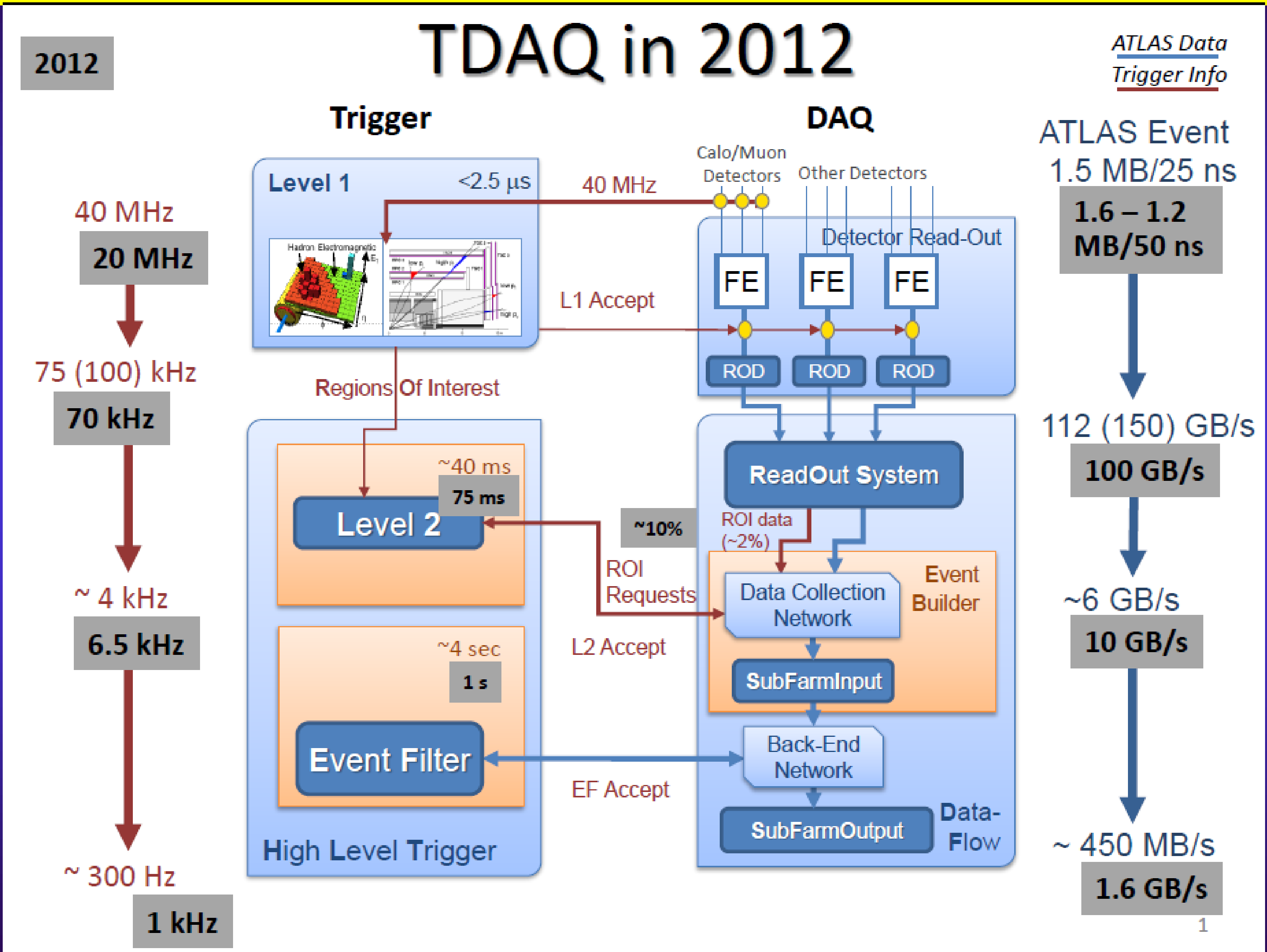
	TP	2010	2011	2012/2013
ALICE	2.5 GB/s 1 kHz		2.2 GB/s 1 kHz (Pb-Pb)	1.4 GB/s 1.2 kHz (pA)
ATLAS	0.3 GB/s 0.2 kHz	0.45 GB/s 0.3 kHz	0.5 GB/s 0.4 kHz (pp)	1.6 GB/s 1 kHz (pp)
CMS	0.4 GB/s 0.3 kHz (pp)	2 GB/s 0.3 kHz (Pb-Pb)	0.3 GB/s 0.4 kHz (pp)	0.8 GB/s 1.5 kHz (pp)
LHCb	0.075 GB/s 1.25 kHz	0.2 GB/s 3 kHz	0.25 GB/s 4 kHz	0.3 GB/s 5 kHz

Architectures

ALICE – Transient Data Storage

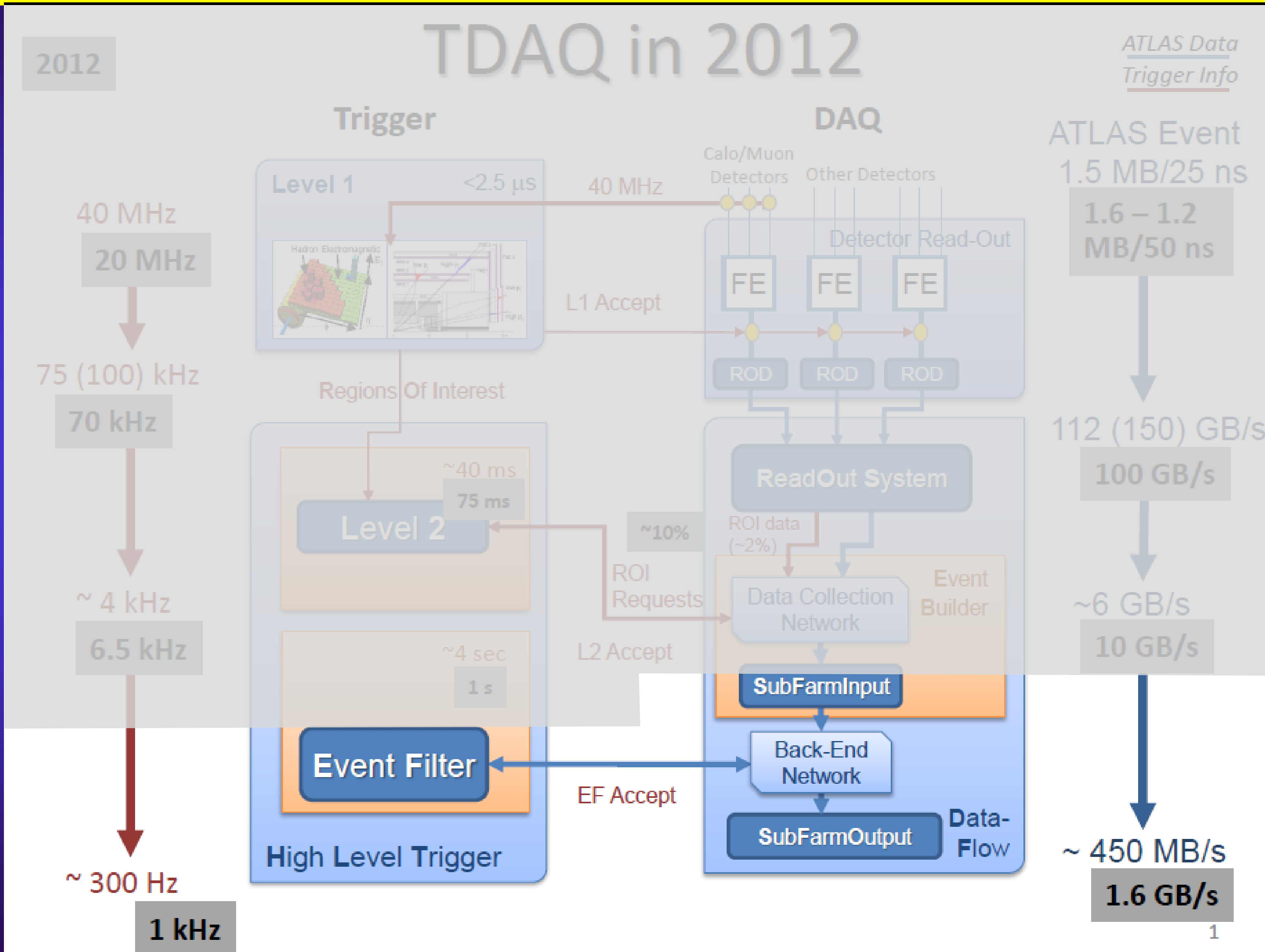


ATLAS - rates

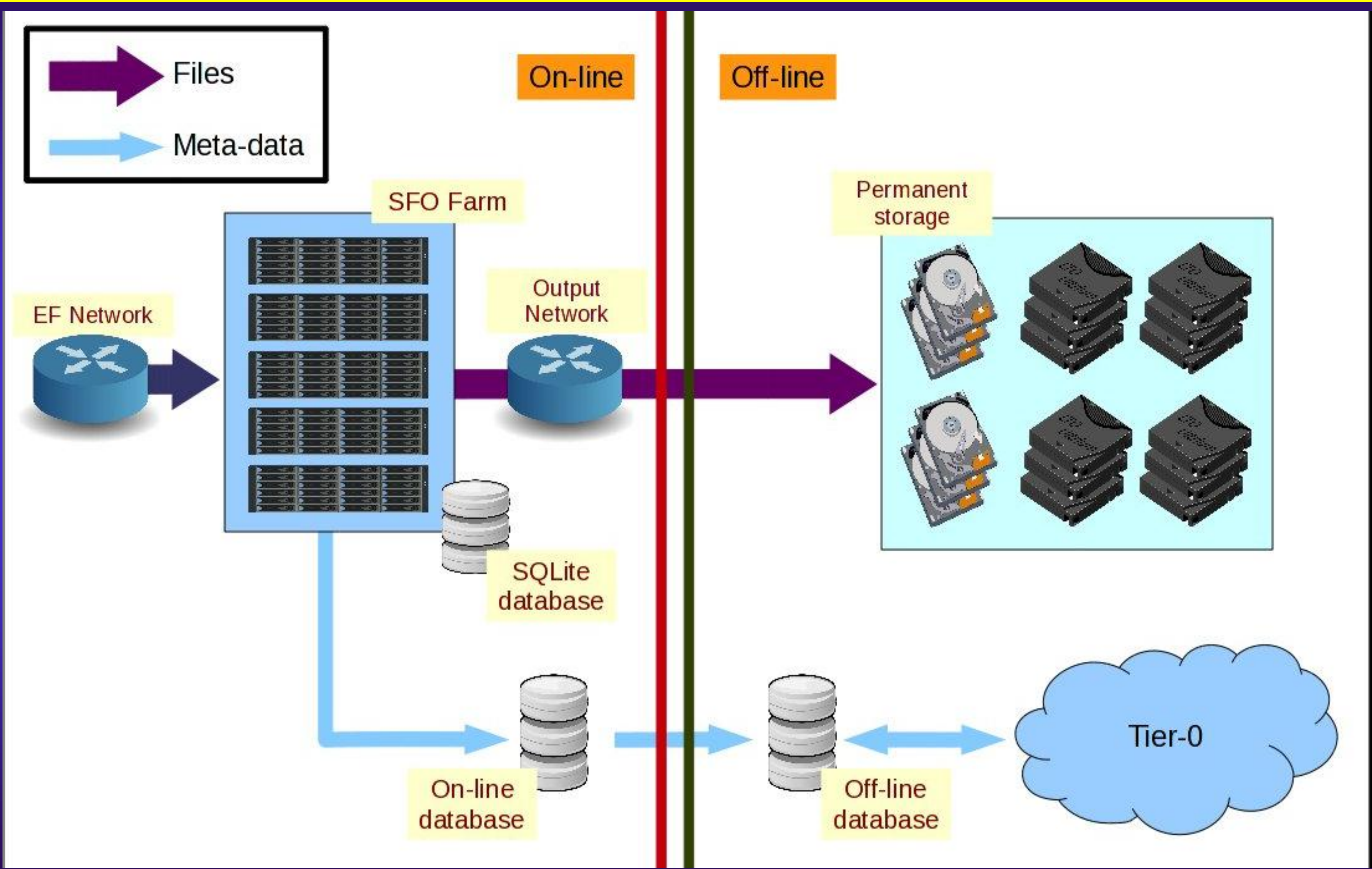


2012
TDAQ in 2012
ATLAS Data
Trigger Info
Trigger
DAQ
ATLAS Event
1.5 MB/25 ns
1.6 – 1.2 MB/50 ns
40 MHz
20 MHz
75 (100) kHz
70 kHz
~ 4 kHz
6.5 kHz
~ 300 Hz
1 kHz
Level 1
<2.5 μs
40 MHz
Calo/Muon Detectors
Other Detectors
Detector Read-Out
FE
FE
FE
L1 Accept
ROD
ROD
ROD
Regions Of Interest
112 (150) GB/s
100 GB/s
~40 ms
75 ms
Level 2
ReadOut System
~10%
ROI data (~2%)
ROI Requests
Data Collection Network
Event Builder
~6 GB/s
10 GB/s
~4 sec
1 s
L2 Accept
SubFarmInput
Event Filter
Back-End Network
EF Accept
Data-Flow
SubFarmOutput
High Level Trigger
~ 450 MB/s
1.6 GB/s
1

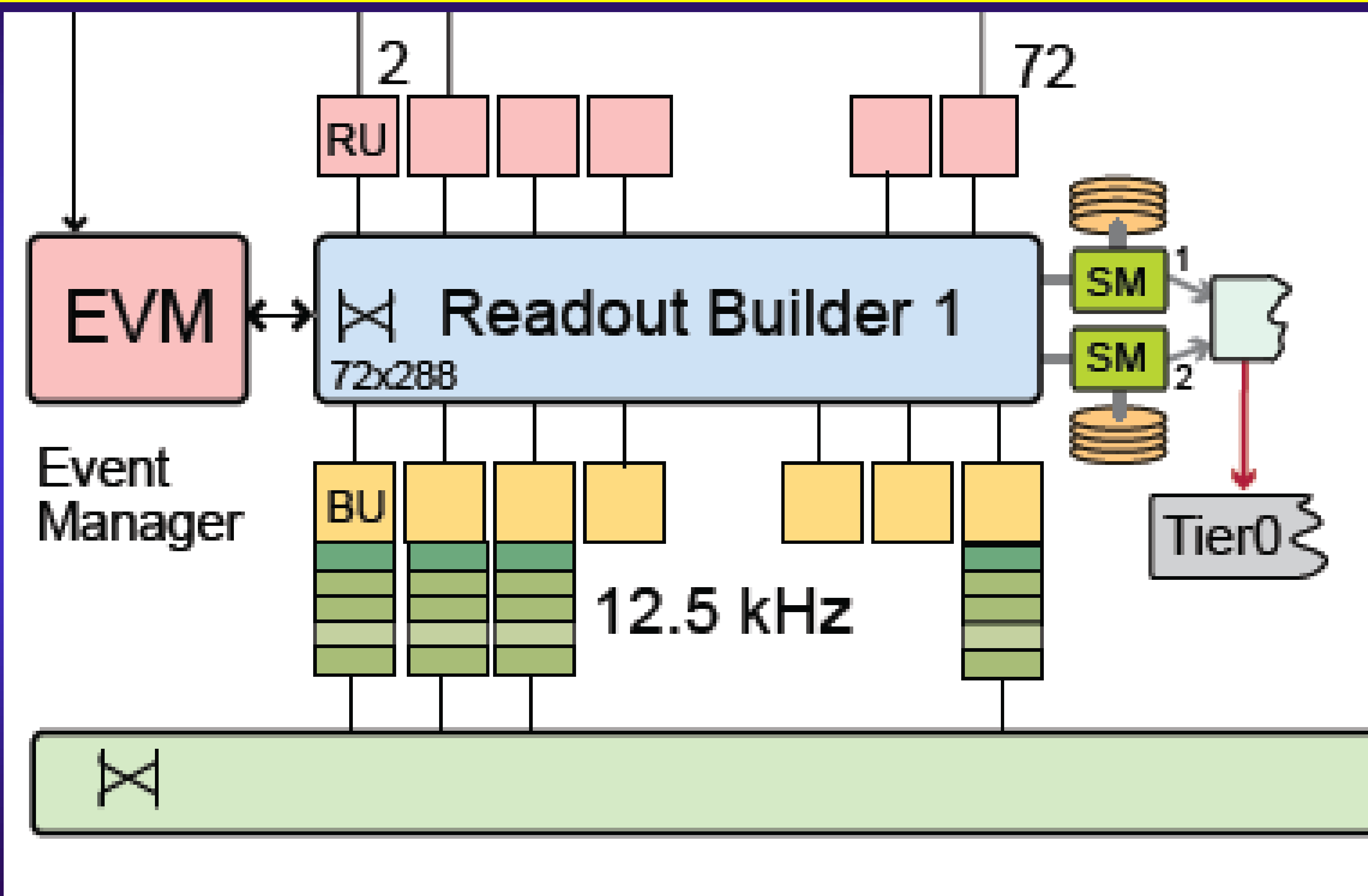
ATLAS - SubFarmOutput



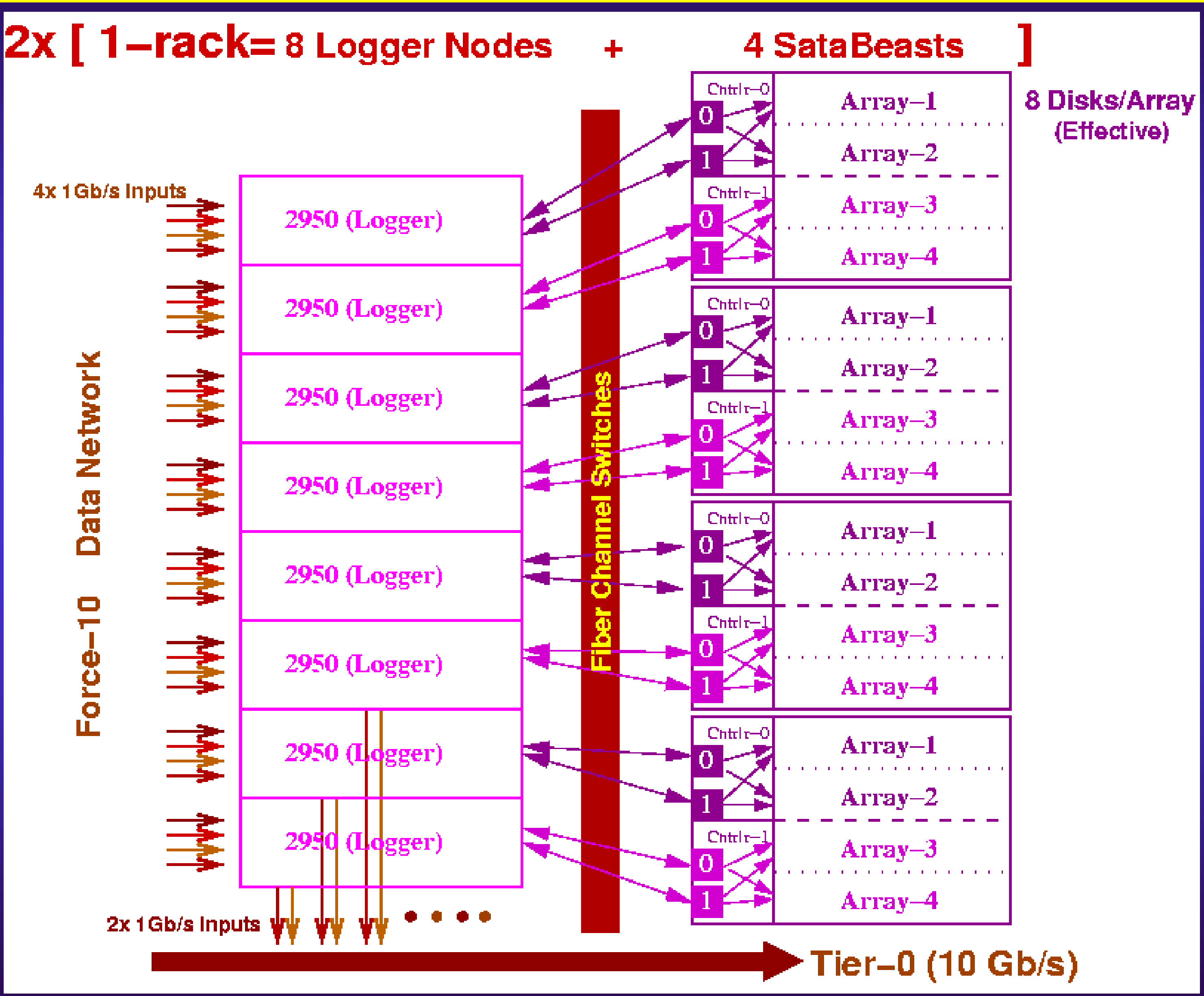
ATLAS – Data flow



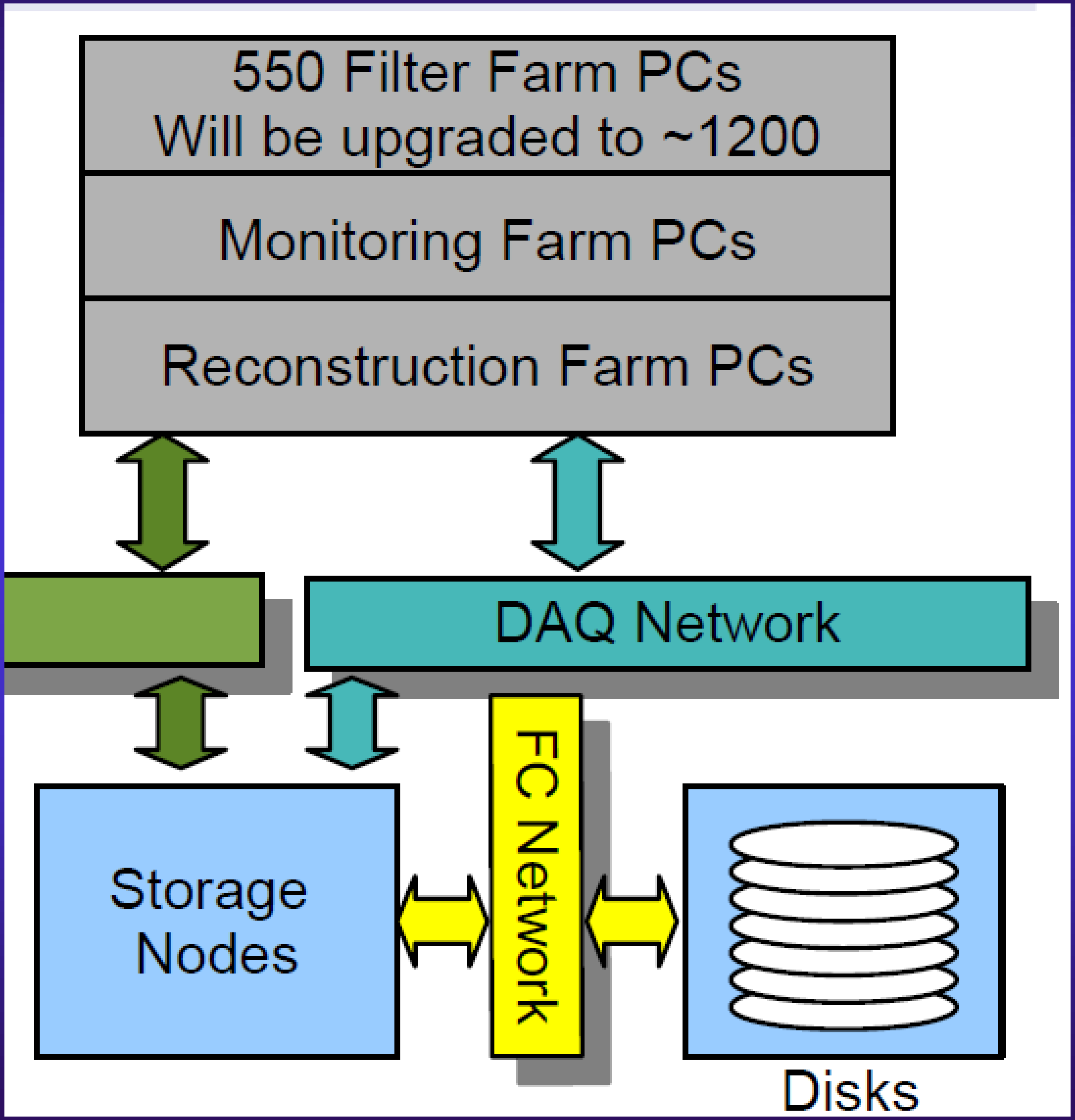
CMS - Architecture



CMS – Storage configuration



LHCb – DAQ/disks networks



Implementations

Local recording

	Streams	Writers	Input @ writers	Links to disks	Disk controllers	# volumes	Usable capacity	RAID	FS	Notes
ALICE	On GDCs	80+ GDCs	1 * 1 GbE	FCS	75 Disk arrays	225	610 TB	RAID 6	StorNext (affinity)	1 exclusive writer/disk GDCs: other tasks
ATLAS	SFOs from HLT	6 new 3 old SFOs	4/2 * 1 GbE bonded	Direct	3 * SATA RAID controllers on SFOs	27 3/SFO	160 TB 7/3.5 TB/SFO	RAID 5	XFS	1 exclusive writer/disk
CMS	HLT to SMs	16 SMs	4 * 1 GbE	FCS	8 * SataBeast 1 * 2 SMs	64 4/SM	225 TB ~14 TB/SM	RAID 6	XFS	Roundrobin non-exclusive
LHCb	Streaming to storage nodes	6 storage nodes	1 * 10 GbE	FCS	1 * Disk array	~20	120 TB data (160 TB total)	RAID 6	StorNext	Non-exclusive (write & read)
CASTOR & EOS	Clients to DS	CASTOR: 637 DSs EOS: 748 DSs	1 * 10 GbE or 1 * 1 GbE	Direct	SATA RAID controllers on DSs	10-17/DS	20-50 TB/DS	RAID 1	XFS	

- GDC: Global Data Collector (event builder)
- SFO: SubFarmOutput
- SM: Storage Manager
- DS: Disk Server
- GbE: Gigabit Ethernet
- FCS: FibreChannel

Migration

	Where	Writers	Protocol	# streams	Links from disks	Output per node	Links to CC	Notes
ALICE	TDSM movers	5 movers	XROOTD	24/mover	FCS	1 * 10 GbE	4 * 10 GbE	Migration rate rarely sustains @ 4 GB/s
ATLAS	SFOs	6 new 3 old SFOs	RFCP	15/SFO	Direct	2 * 1 GbE bonded	2 * 10 GbE	
CMS	SMs	16 SMs	RFCP	5/SM	FCS	2 * 1 GbE	1 + 1 * 10 GbE	Migration can be paused during critical periods
LHCb	Storage nodes	6 nodes	RFCP (XROOTD)	6 total	FCS	1 * 10 GbE	2 * 10 GbE	Shared control with Offline

- TDSM: Transient Data Storage Manager
- SFO: SubFarmOutput
- SM: Storage Manager
- GbE: Gigabit Ethernet
- FCS: FibreChannel

Data flows & procedures

- ALICE: pushed distributed writing (each GDC: ~170 inputs, 1 output volume, 8 AliROOT streams), synchronized reading (multiple – 5 max - movers/volume), migration+CRC via 24*XROOTD streams/mover, synchronization via dedicated manager host (MySQL), auxiliary info via AliEn gateway (SCP) and ALICE logbook (MySQL), alarms via SMTP + dedicated testing procedures.

- ATLAS: data pull by SFOs from Event Filter Dataflows via dedicated protocol, 750 connections/SFO reader, CRC while writing to disk, migration via 15*RFCP/SFO, synchronization via files (names, touch-stamps), auxiliary info via SQLite (local) and Oracle (mirrored to Offline).

- CMS: SMs gets data from HLT via I20, ~80 inputs/SM reader, SM single-write/multiple-reads to/from 2 dedicated arrays/4 dedicated volumes in round-robin (no allocation/dedication), first CRC while writing to disk, migration via 5*RFCP/SM (second CRC), control via CMS DB (Oracle), auxiliary info via Transfer DB (Oracle, replicated to Offline).

- LHCb: streaming nodes push 1 full streams + several auxiliary streams to storage nodes (first CRC), 6 streams/storage node to disk, 6 readers (second CRC), auxiliary info via Oracle DB (located at Computer Center), Disk Array used also for NFS/Samba, files kept on disk until migrated to TAPE in CASTOR.

- CASTOR/EOS
  - Disk servers run CASTOR/EOS daemons, 3 head nodes * (4 + 1), 5 name servers, ORACLE DBs.

Commonalities

- ALICE & ATLAS & CMS & LHCb (& CASTOR/EOS):
  - Early (and multiple) CRC (Cyclic Redundancy Check)
    - ATLAS & CMS & LHCb: on the way to the local disk
    - ALICE & CMS & LHCb: on the way to the Computer Center
    - CASTOR/EOS: as soon as the file is completely transferred
  - Single writer/volume
- ALICE & ATLAS & CMS
  - Multiple readers per volume
- ATLAS & CMS & LHCb:
  - Few dedicated nodes for writing & migration (ATLAS: 9, CMS: 16, LHCb: 6)
- ALICE & ATLAS:
  - Rule of 3 (write, read, spare)
    - ALICE: soft/anonymous, ATLAS: hard/bound to SFO
  - Exclusive writer/volume
- ATLAS & CMS:
  - Static assignment writer/reader nodes ⇔ Volumes

Peculiarities

- ALICE
  - Direct write from main DAQ nodes to disks
  - Writing and reading share almost no resources
- ATLAS
  - No Disk Arrays, only SFOs (as CASTOR/EOS)
- CMS
  - Migration can be paused to boost writing
- LHCb
  - One monolithic Disk Array
  - Migration control shared by Online & Offline

PROs

- ALICE
  - Scalable: writers++ (dynamic), Disk Array++, TDSM movers++
  - Robust, no hotspots, little or no tuning
- ATLAS
  - Scalable: disks++, SFOs++
  - Cheap (~ 9000+ CHF/SFO)
  - SFOs can run other recording-related tasks (e.g. SMART tests, system monitoring etc…)
- CMS
  - Reliable
- LHCb
  - 2nd generation Disk Array: reliable (compared to 1st generation: better striping, p2p disk attachments vs. daisy-chain, better load balancing)
- CASTOR/EOS
  - Modular, easy to handle and operate

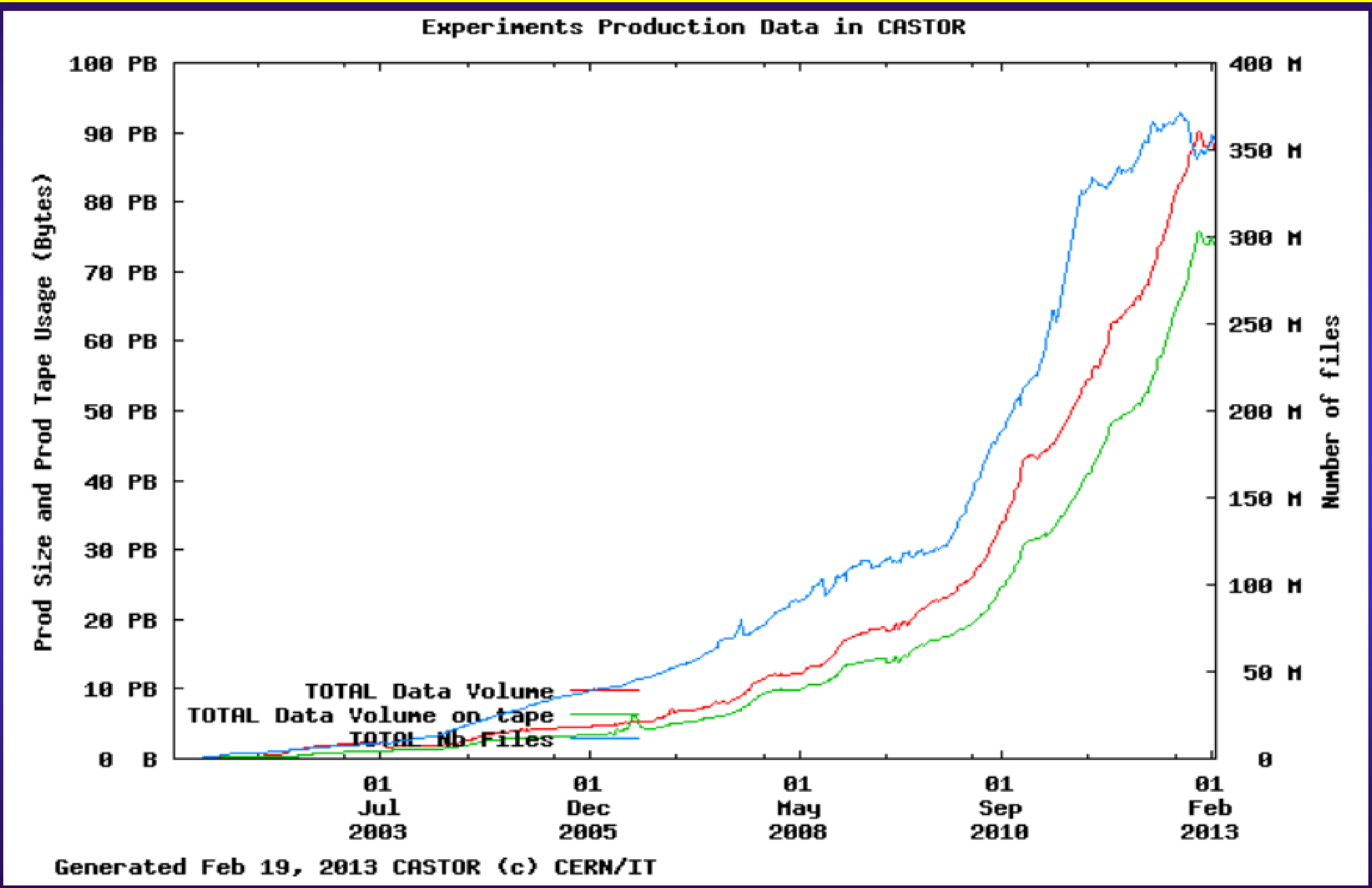
CONs

- ALICE
  - Reconfigurations are often global
  - Disk arrays must be broken up to provide enough volumes (rule of 3)
  - Needs special diagnostics (e.g. one bad volume or slow writer: which one?)
- ATLAS
  - Limited # of volumes & fixed attachment can make turnaround tricky (rule of 3)
  - Big un-breakable entities (7 TB/SFO, 3 volumes/SFO)
  - Parametrizing & tuning critical (network, disk I/O, system calls)
- CMS
  - Loss of 1 SM => loss of ½ HLT (slice) processing power (until reconfig/recovery)
  - Bookkeeping across different SMs challenging
- LHCb
  - Stuck recording stream can block downstream nodes
  - Fault-tolerance not very reliable, compensated by HA technologies
- CASTOR/EOS
  - Oracle is too much of a “black box”
  - In RAID1 multiple (quasi) simultaneous disk failures may lead to data losses

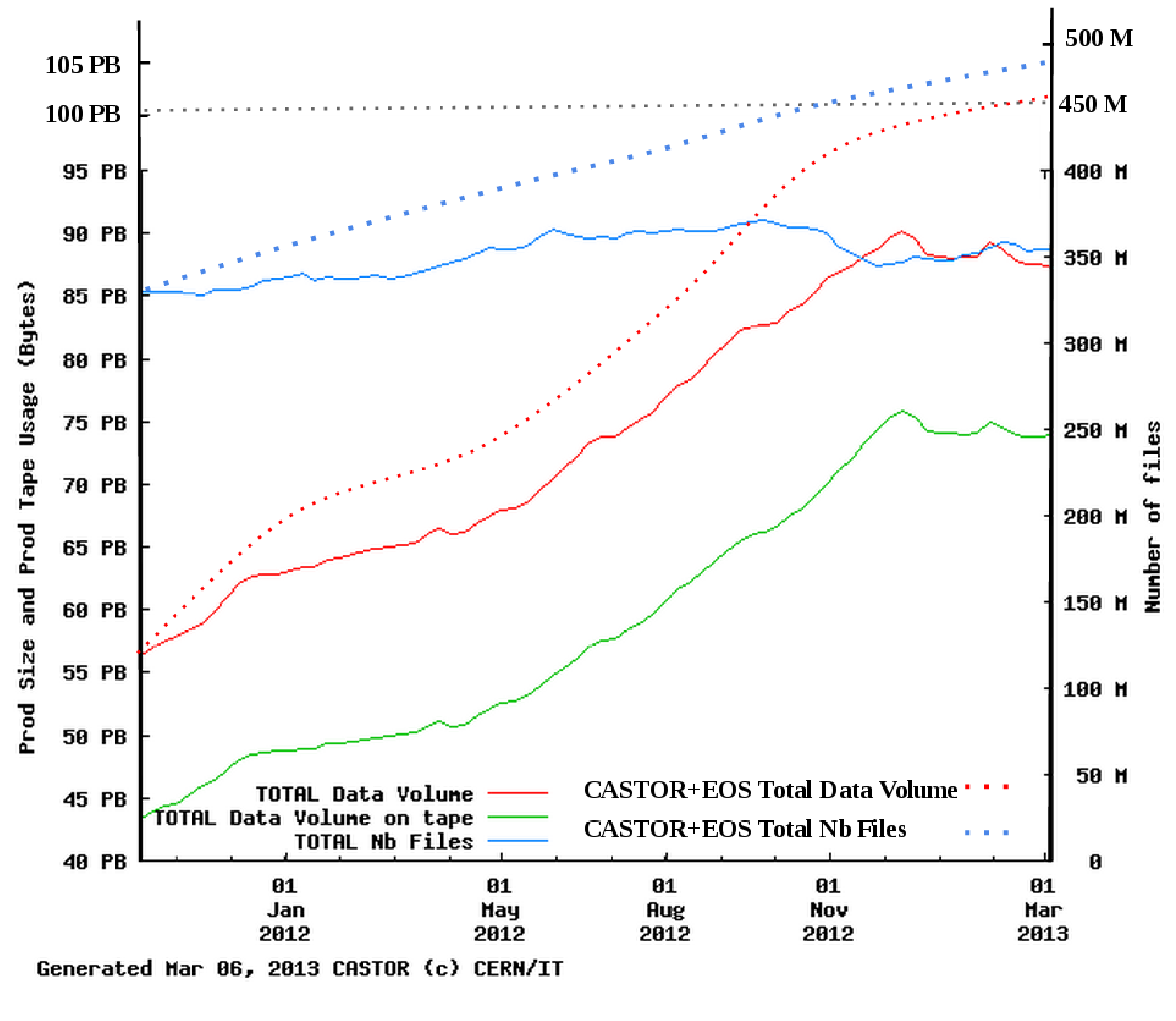
Experiences

- ALICE
  - Few breakdowns (2011-early 2012: increasing # of failures for 6-years-old Disk Arrays, replaced)
  - Writing OK, migration often (but not always) slower than the expected
- ATLAS
  - Reliable & solid (SFOs upgrade in 2010, old SFOs “recycled”: 3+1 years in operation)
- CMS
  - Few breakdowns (more failures @ end of 2012 due to aging)
- LHCb
  - Few breakdowns (Disk Array upgraded in 2010)
- CASTOR/EOS
  - No evident aging effects (HW turn-around time: 3-4 years)

The result (CASTOR only)



CASTOR and EOS



What's coming next (LS1)

- ALICE
  - Expected event & data rates ~equivalent to today's (increased by higher multiplicity)
    - The TRD detector will contribute more, hopefully reduced by the HLT
    - Possible upgrade of the TPC detector may substantially increase the data rates
  - Same architecture
- ATLAS
  - Getting enough headroom to double the data rates
  - Same architecture
- CMS
  - Nothing finalized, so far same requirements
  - Evaluating a radical change in data flow into the MSS: direct disk access from recording nodes which will directly handle the data + metadata files on NAS (no more SMs)
- LHCb
  - Further upgrade of Disk Array
  - More flexible streaming procedure
- CASTOR/EOS
  - Bigger disks, more TB/box
  - Code simplification & cleanup, agile infrastructure for configuration & installation
  - Same architecture

Many thanks to…

- ALICE: myself / Ulrich Fuchs (et. al.)
- ATLAS: Weiner Vandelli
- CMS: Gerry Bauer / Olivier Raginel / Lavinia Darlea
- LHCb: Rainer Schwemmer
- CASTOR/EOS: Xavier Espinal

Questions?