+

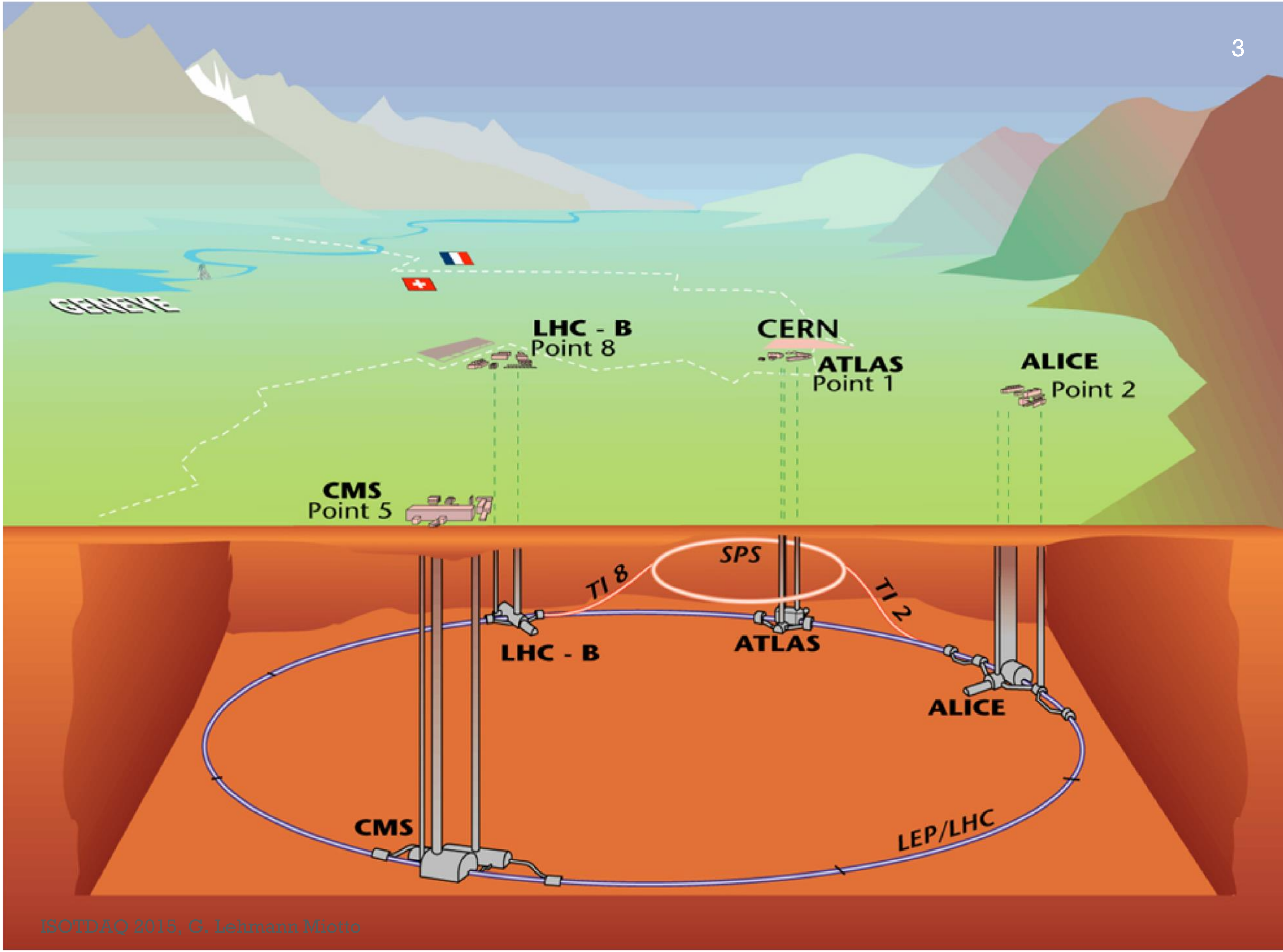
Acknowledgements:
S. Baron
N. Neufeld
C. Schwick
W. Smith
W. Vanelli
P. Vande Vyvre

TDAQ at the LHC Experiments

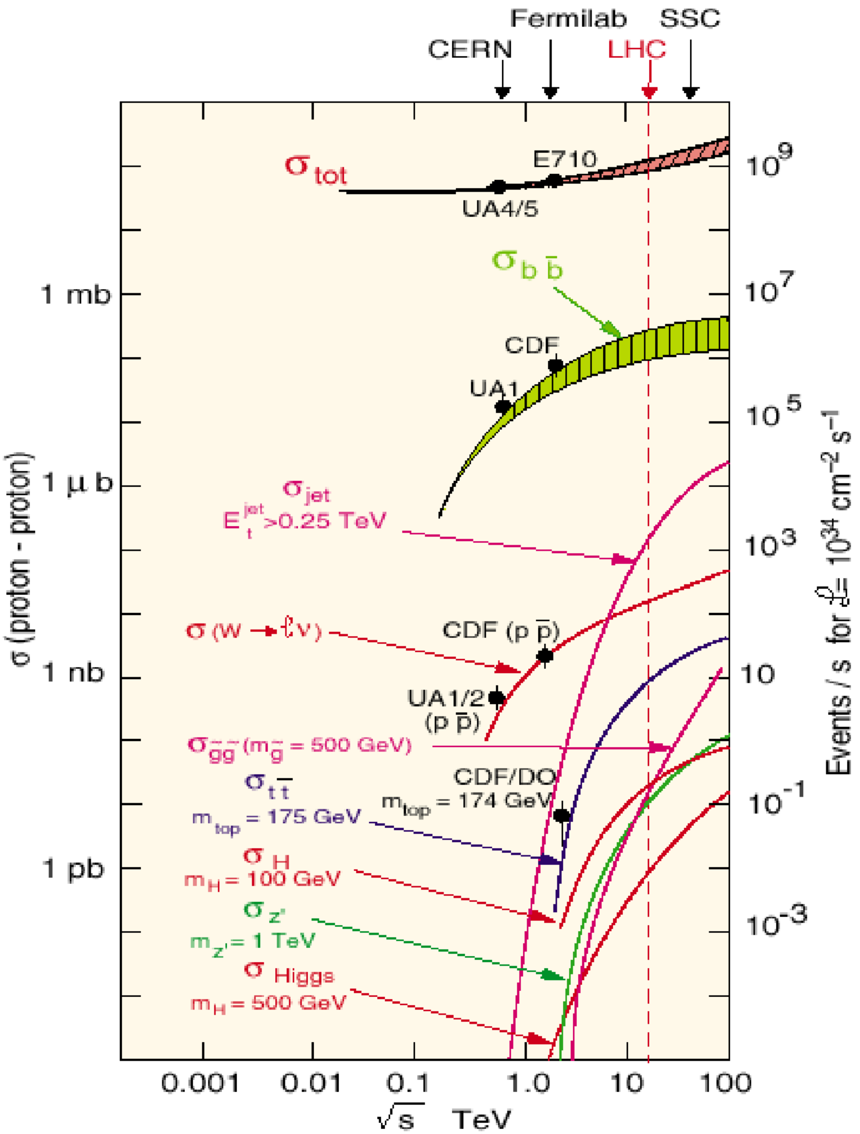
G. Lehmann Miotto

Outline

- The 4 LHC experiments
- The original TDAQ
  - Constraints and architectures
- Evolving TDAQ systems
  - Physics requirements
  - Technology progress
  - Interesting fields of R&D


GENEVE
LHC - B
Point 8
CERN
ATLAS
Point 1
ALICE
Point 2
CMS
Point 5
SPS
TI 8
TI 2
LHC - B
ATLAS
ALICE
CMS
LEP/LHC

Interesting Physics at the LHC

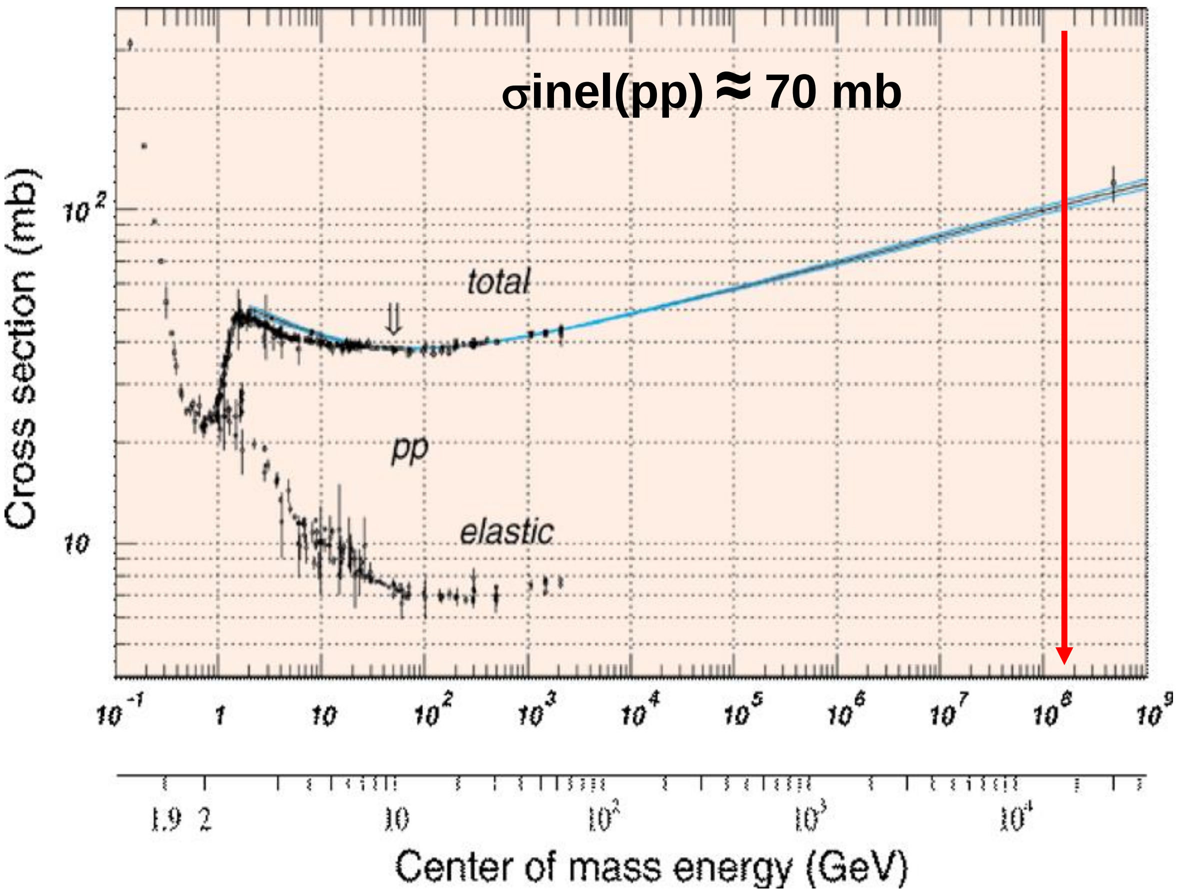


$\sigma_{tot} \approx 100$ mb

1 : 1 000 000 000

σ_{Higgs} is down here!

LHC Experimental Environment



$L = 10^{34}\ cm^{-2}s^{-1}$

$\sigma_{inel} \approx 70\ mb = 70x10^{-27}\ cm^2$

Event Rate => $7x10^8$ Hz

Bunch crossings every 25 ns
=> 17.5 events/bc

Filled bunches (2835/3564)
=> 23 events/bc

In Run 1:

BC= 50 ns
$L_{max}= 7.7x10^{33}\ cm^{-2}s^{-1}$
different fill scheme

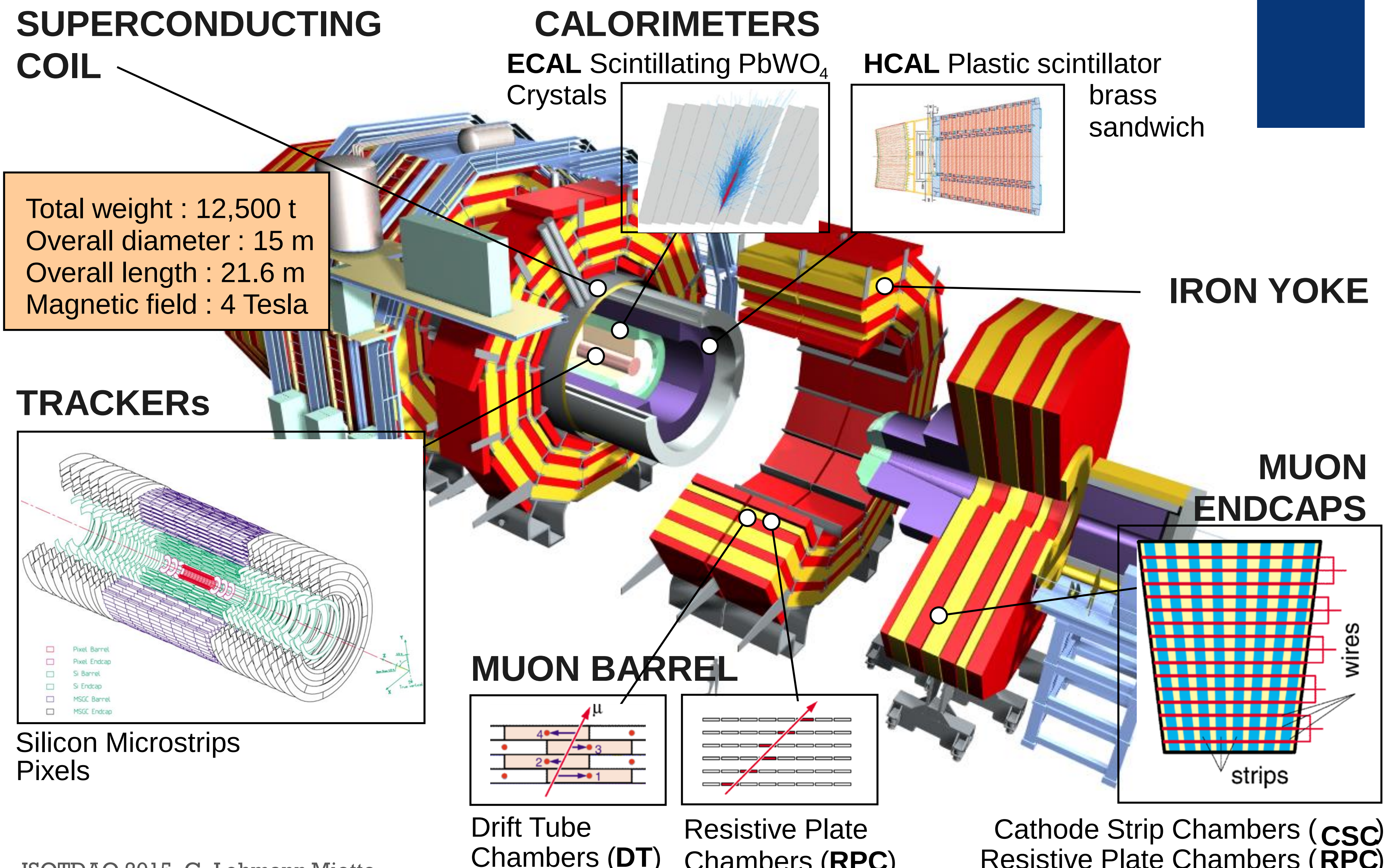
=> 30-35 events/bc

LHC Roadmap

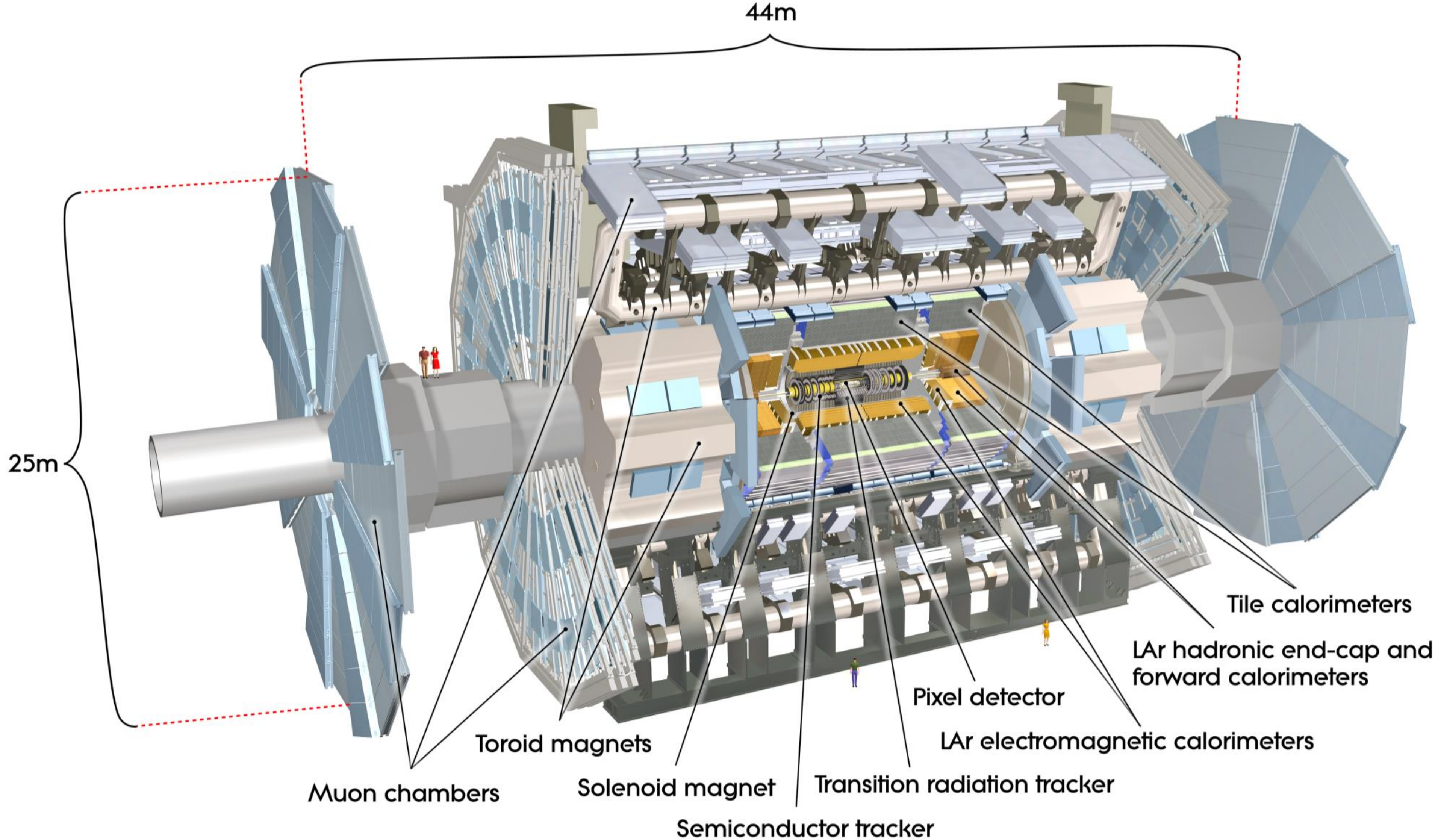
2009		LHC startup, $\sqrt{s}$ 900 GeV	
2010–2012		$\sqrt{s}$=7+8 TeV, L~6x10^{33}cm^{-2}s^{-1}, bunch spacing 50ns	Run 1 ~25 fb^{-1}
2013–2014	LS1	Go to design energy, nominal luminosity - Phase 0	
2015–2017		$\sqrt{s}$=13~14 TeV, L~1x10^{34}cm^{-2}s^{-1}, bunch spacing 25ns	~75-100 fb^{-1}
2018	LS2	Injector + LHC Phase I upgrade to ultimate design luminosity	
2019–2021		$\sqrt{s}$=14 TeV, L~2x10^{34}cm^{-2}s^{-1}, bunch spacing 25ns	~350 fb^{-1}
2022	LS3	HL-LHC Phase II upgrade: Interaction Region, crab cavities?	
2023 … 2030?		$\sqrt{s}$=14 TeV, L~5x10^{34}cm^{-2}s^{-1}, luminosity levelling	~3000 fb^{-1}

+ The 4 Large LHC Experiments

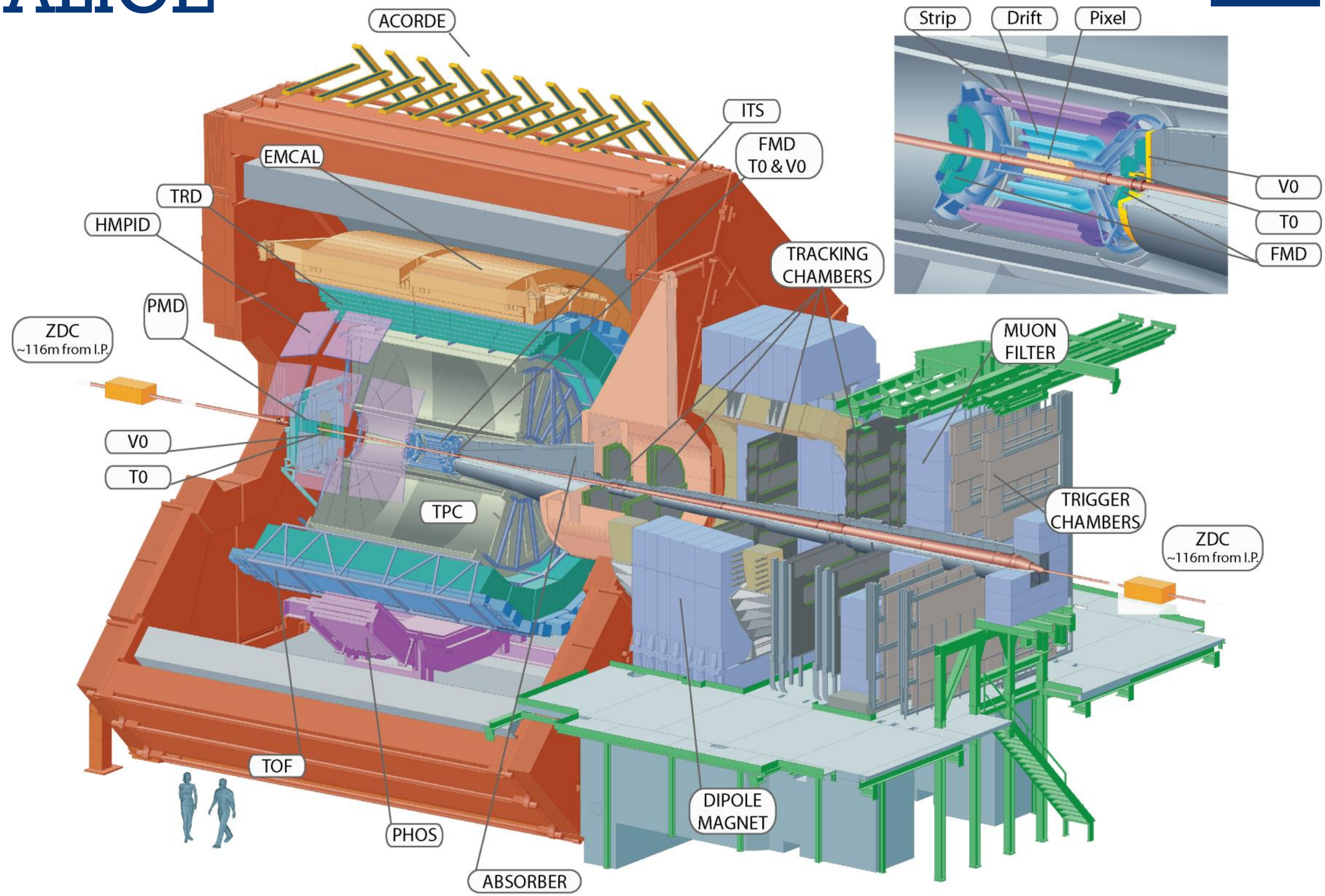
CMS


SUPERCONDUCTING COIL
CALORIMETERS
ECAL Scintillating $PbWO_4$ Crystals
HCAL Plastic scintillator brass sandwich
Total weight : 12,500 t
Overall diameter : 15 m
Overall length : 21.6 m
Magnetic field : 4 Tesla
IRON YOKE
TRACKERs
Silicon Microstrips
Pixels
MUON ENDCAPS
MUON BARREL
μ
wires
strips
Drift Tube Chambers (**DT**)
Resistive Plate Chambers (**RPC**)
Cathode Strip Chambers (**CSC**)
Resistive Plate Chambers (**RPC**)

ATLAS

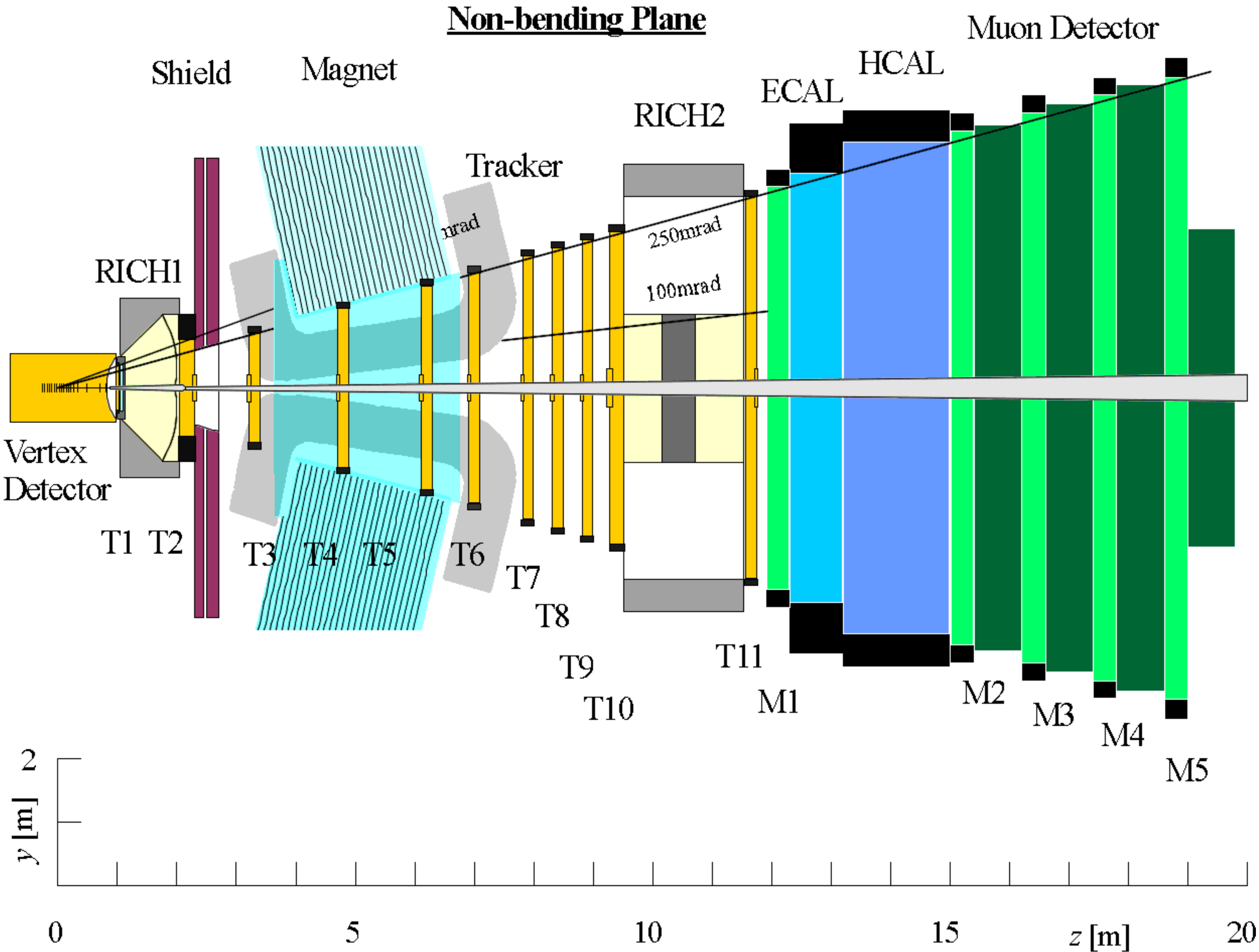

44m
25m
Muon chambers
Toroid magnets
Solenoid magnet
Semiconductor tracker
Transition radiation tracker
Pixel detector
LAr electromagnetic calorimeters
LAr hadronic end-cap and forward calorimeters
Tile calorimeters

ALICE



ACORDE
ITS
FMD
T0 & V0
EMCAL
TRD
HMPID
PMD
ZDC
~116m from I.P.
V0
T0
TPC
TRACKING
CHAMBERS
MUON
FILTER
TRIGGER
CHAMBERS
ZDC
~116m from I.P.
TOF
PHOS
ABSORBER
DIPOLE
MAGNET
Strip
Drift
Pixel
V0
T0
FMD

LHCb

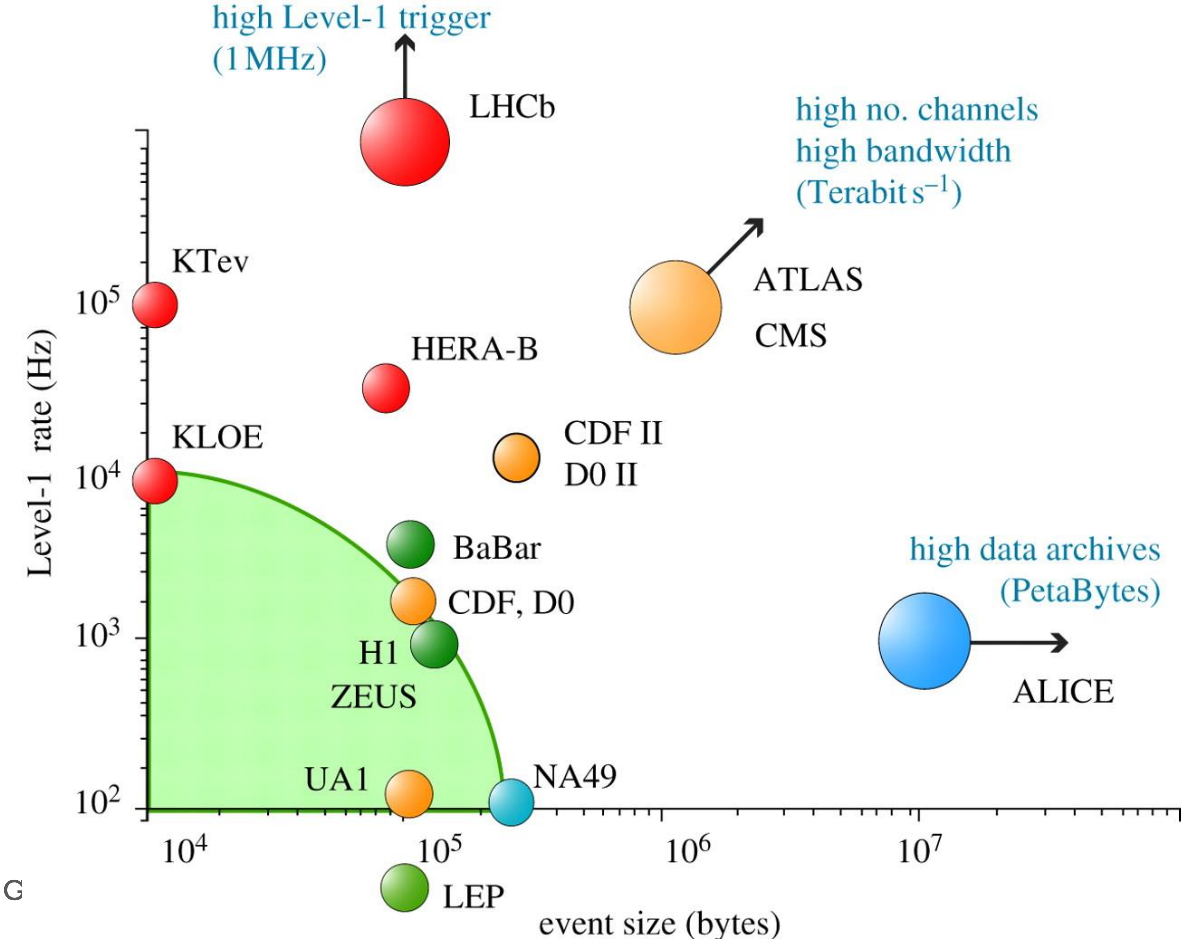

Non-bending Plane
Shield
Magnet
Muon Detector
HCAL
ECAL
RICH2
Tracker
250mrad
100mrad
RICH1
Vertex Detector
T1 T2
T3 T4 T5 T6
T7
T8
T9
T10
T11
M1
M2
M3
M4
M5
2
y [m]
0
5
10
15
z [m]
20

TDAQ Systems at the LHC

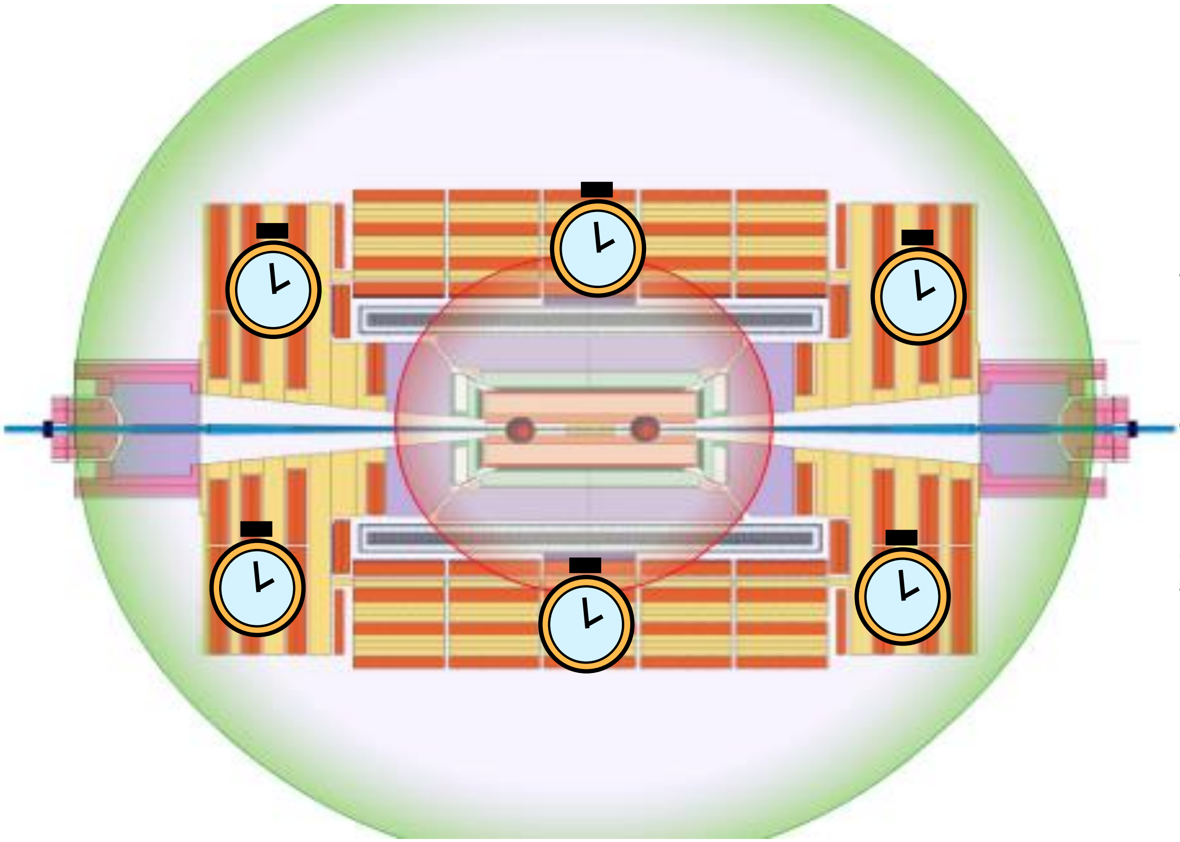
A story about how they were designed originally and how they are evolving…

Initial Design Parameters

- When LHC experiments were designed back in the 90'
  - Raw data storage capped at ~ 1 PB / year per experiment



+ Synchronization



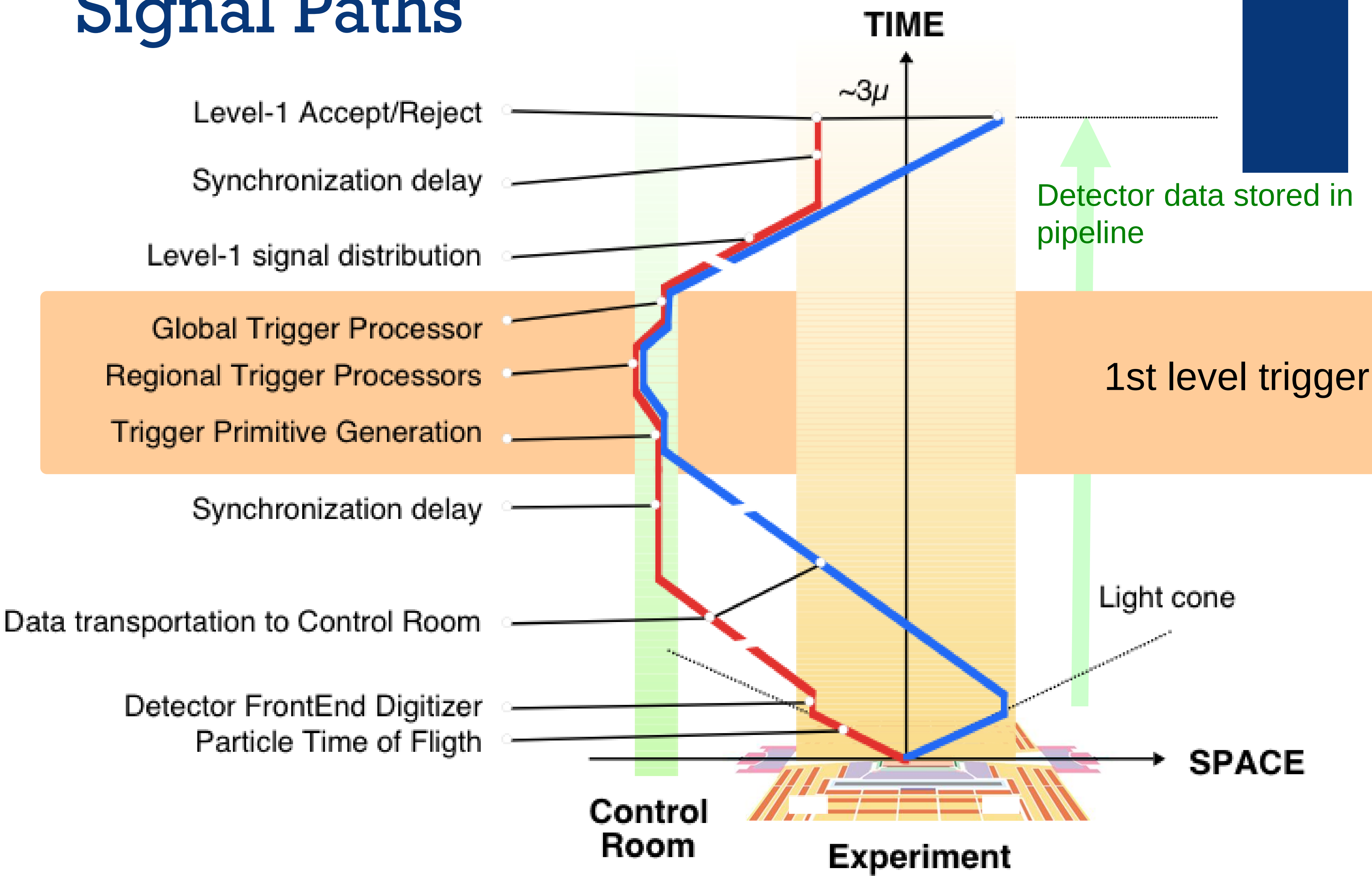
Data corresponding to the same bunch crossing must be processed together.

But:
Particle TOF >> 25ns
(25 ns ≈ 7.5m)
Cable delay >> 25ns (
$v_{signal} \approx 1/3\ c$)
Electronic delays

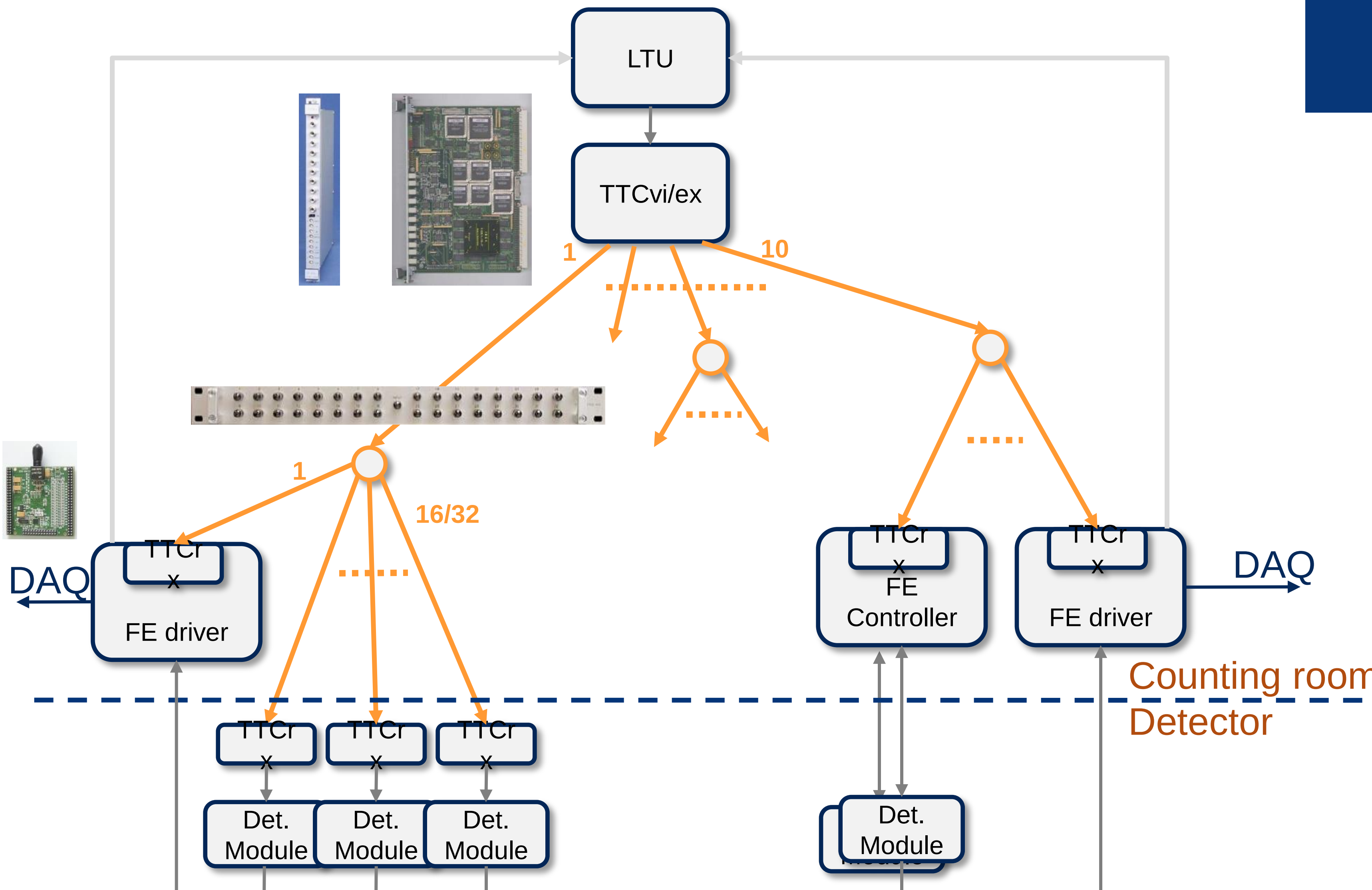
Need to:
Synchronize signals with programmable delays.

Provide tools to perform synchronization (TDCs, pulsers, LHC beam with few buckets filled...)

Signal Paths

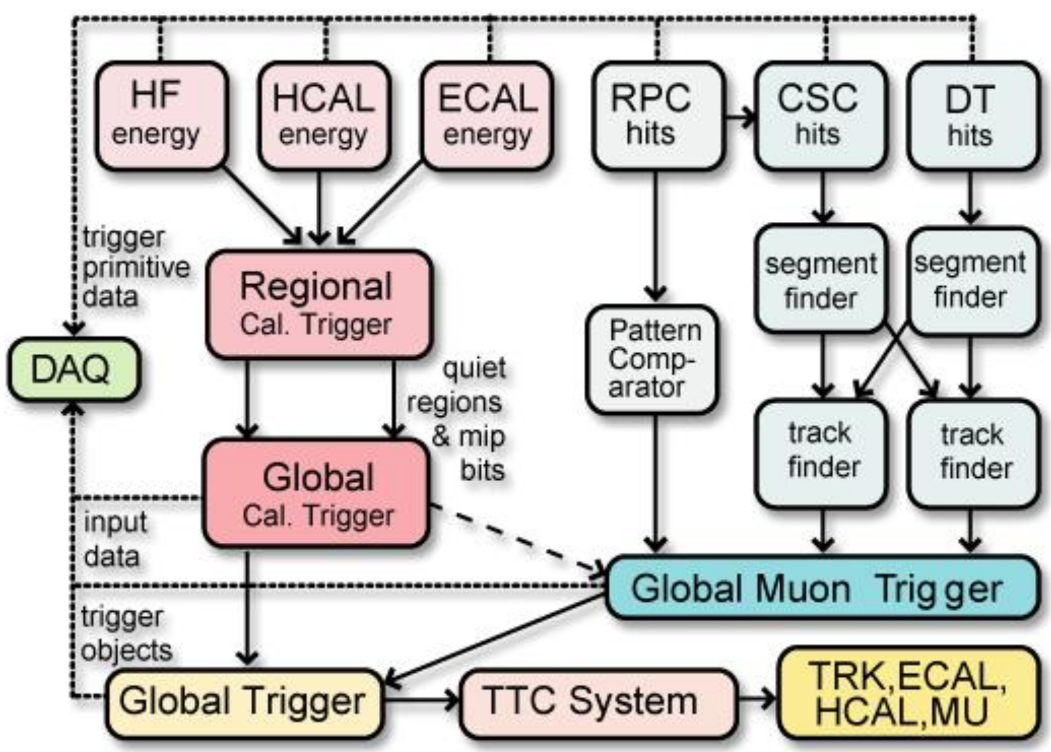


Timing, Trigger & Control System



HW Triggers

- Driven by: physics, trigger detectors readout capabilities, on detector buffering capabilities, overall readout capabilities



In ATLAS/CMS :

- latency budget of ~3 μs
- Max readout 100 kHz

In ALICE:

Detectors with very different latencies in delivering data & in requiring signal

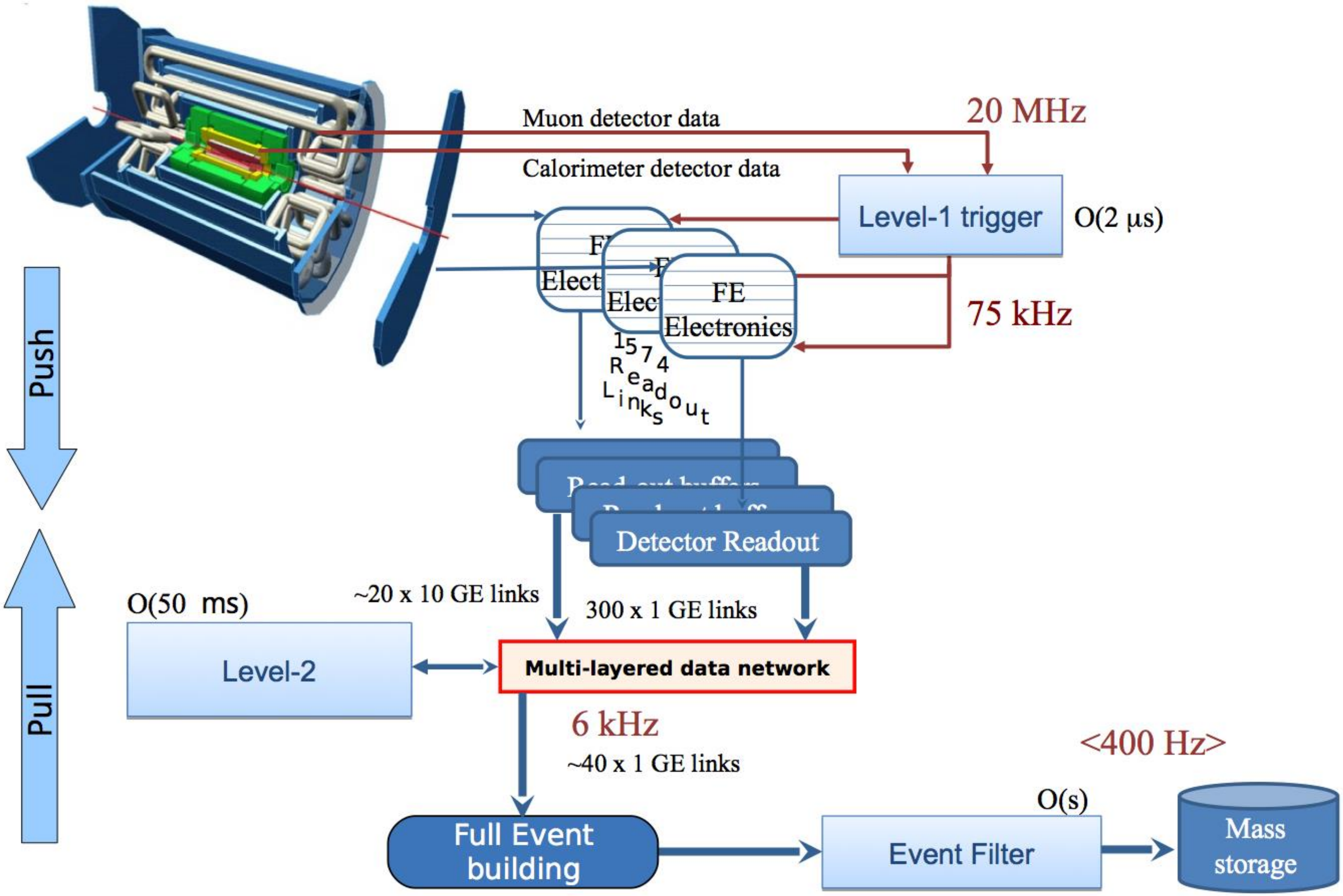
⇒ Multi-level HW trigger

⇒ Pile-up protection for TPC

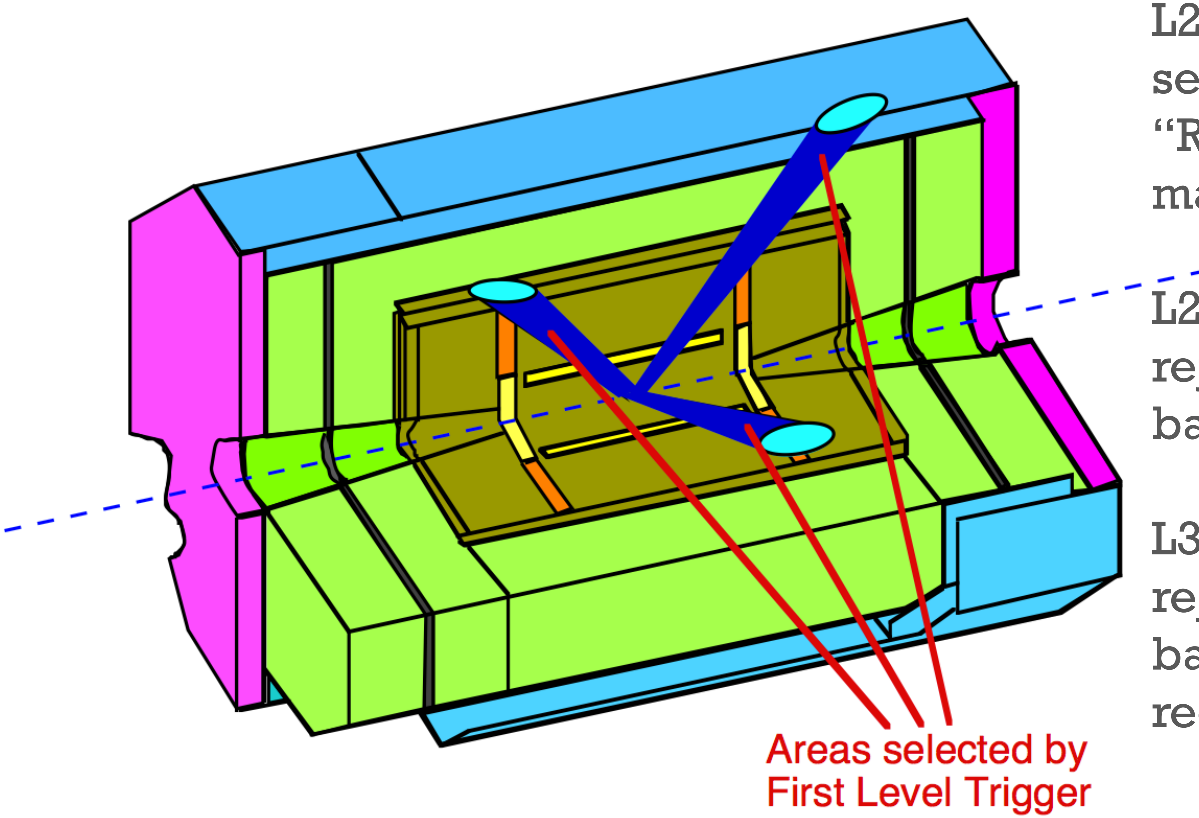
In LHBb:

- Max readout at 1 MHz
- Luminosity kept artificially low

+ Technical Solutions: ATLAS



+ ATLAS: The Clever Idea



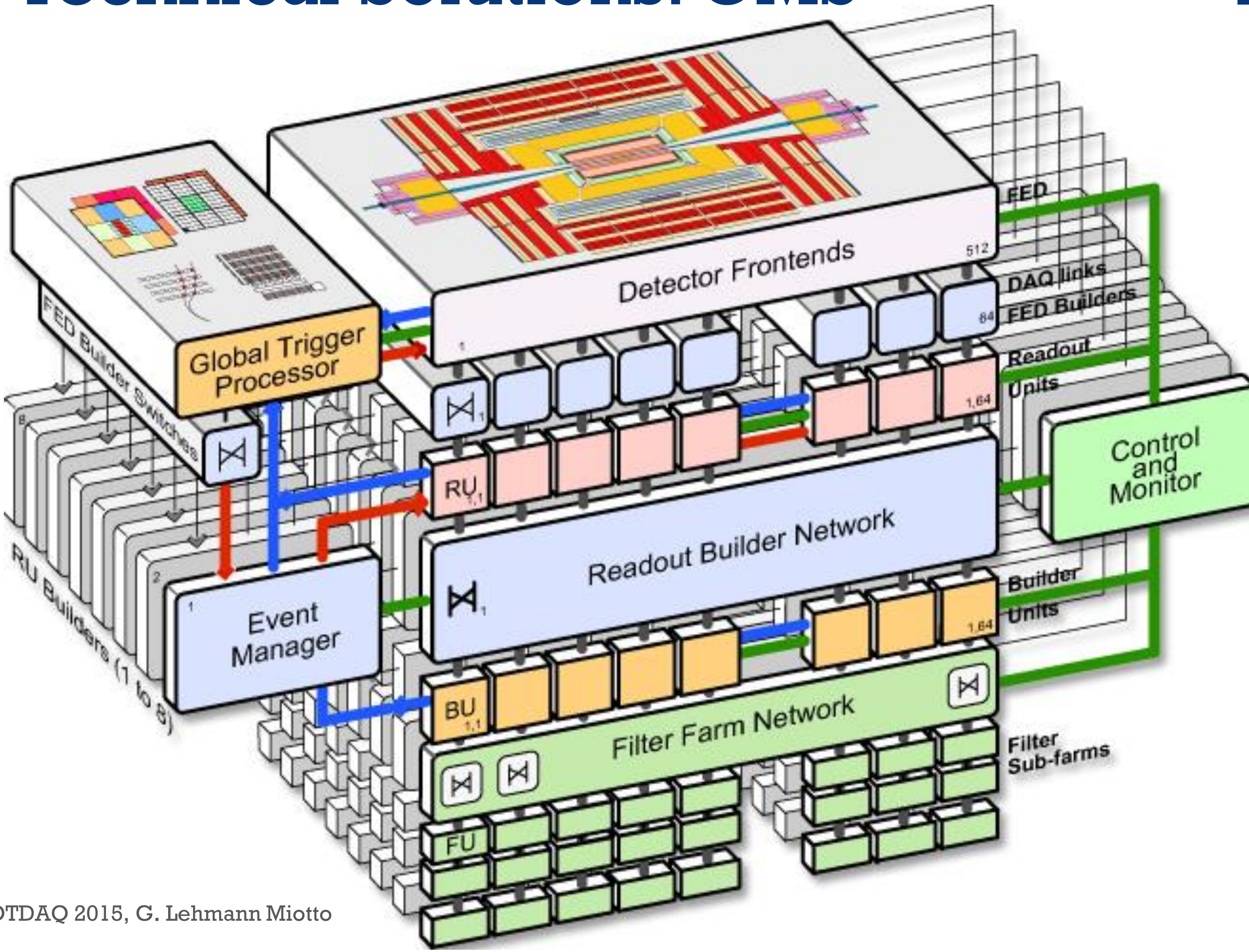
L2 trigger only selects data based on "Regions of Interest" marked by L1

L2 runs at 75 kHz, rejects > 90% of events based on ~10% of data

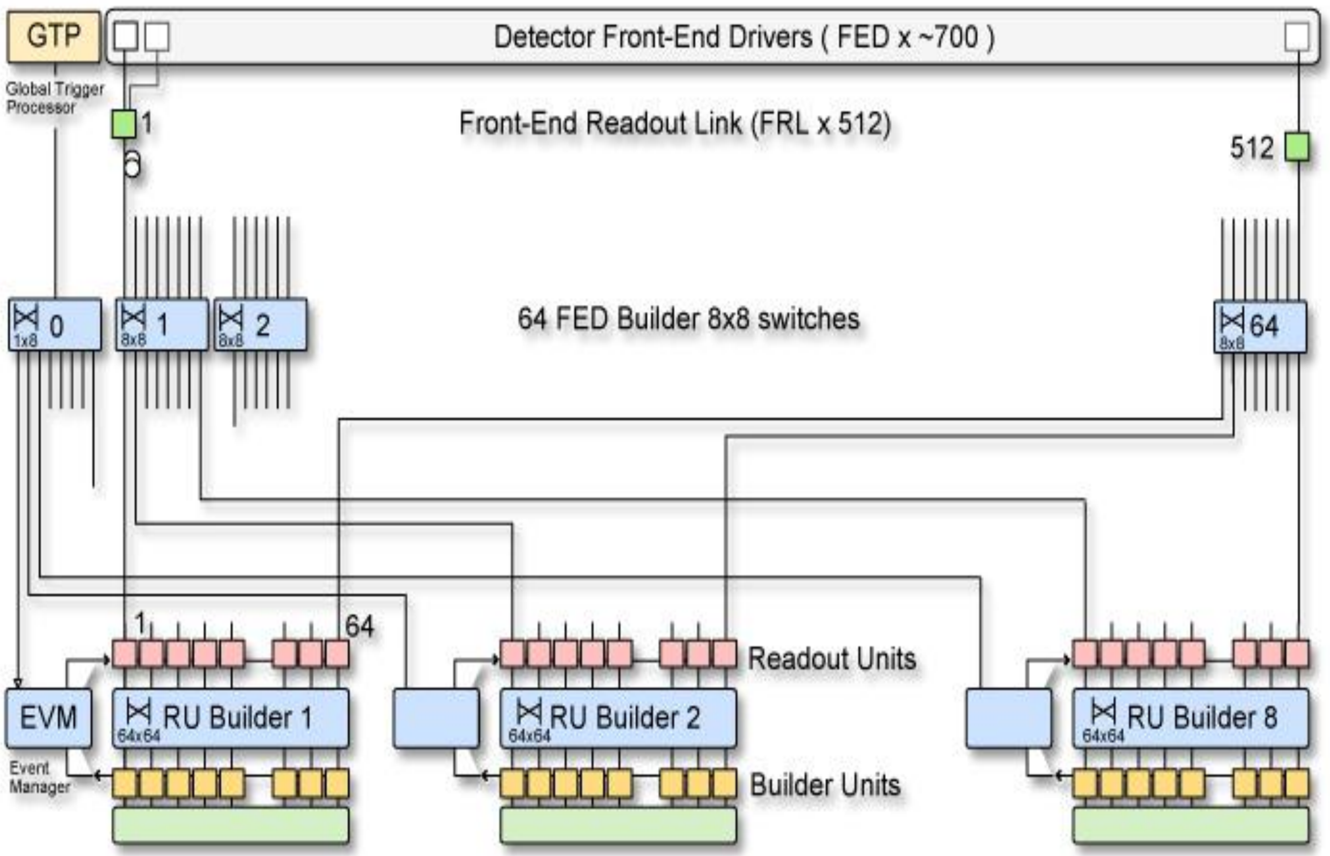
L3 runs at ~6 kHz, rejects >90% of events based on full reconstruction

Overall network bandwidth: ~10 GB/s !

+ Technical Solutions: CMS



CMS: The Clever Idea



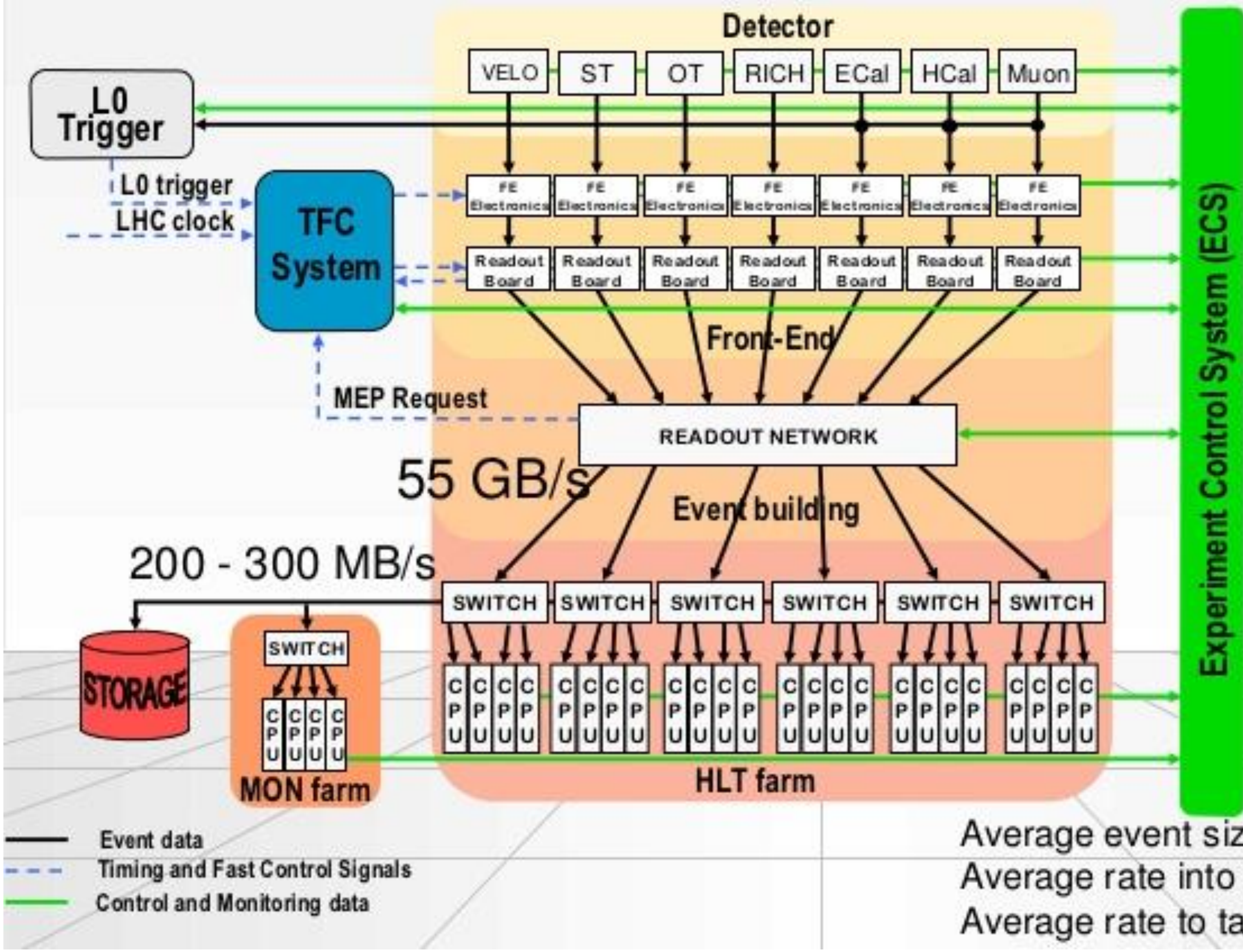
2 stage event building!

1st stage:
builds 1 fragment out of 8 at 75 kHz, sends it to one RU builder

2nd stage:
Works at 10 kHz, serves complete events to trigger farm.

Each RU builder needs ~10 GB/s aggregate bandwidth!

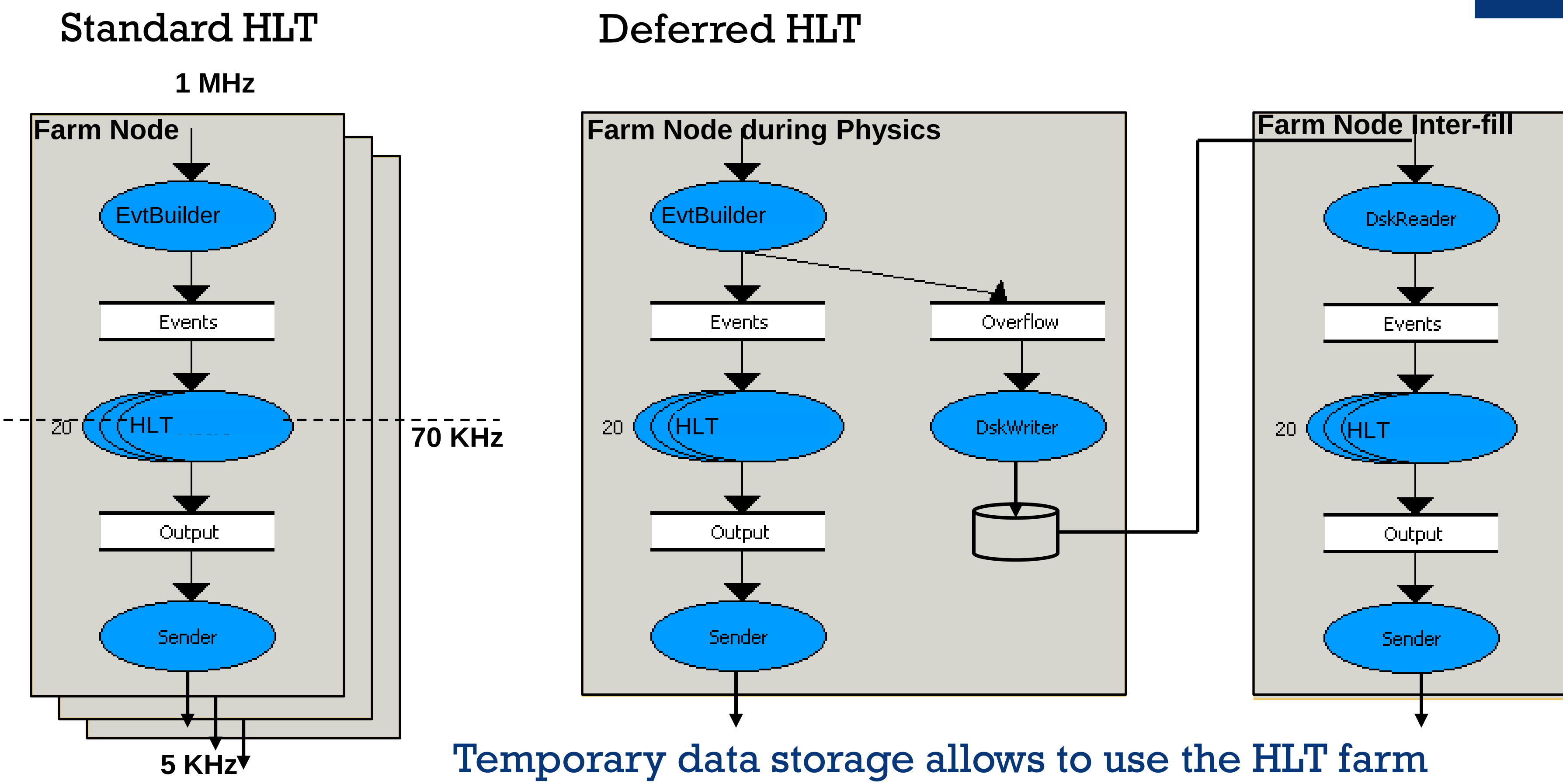
Technical Solutions: LHCb



Events very small: coalesced to reduce overheads and message rate (~80 kHz)

Average event size 55 kB
Average rate into farm 1 MHz
Average rate to tape 4 – 5 kHz

LHCb: The Clever Idea



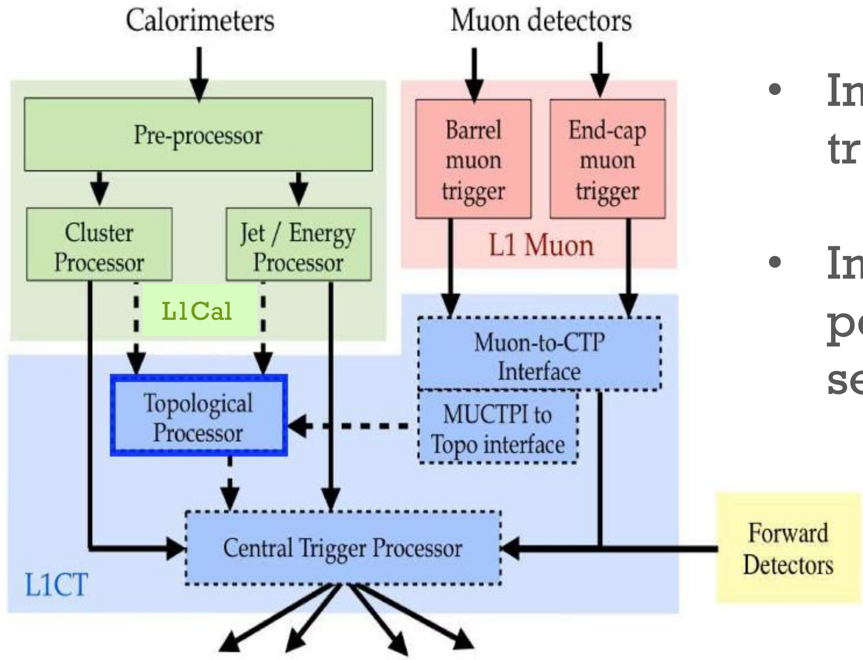
Temporary data storage allows to use the HLT farm continuously and run at higher L1 rate!

Run 1 – Reality Check

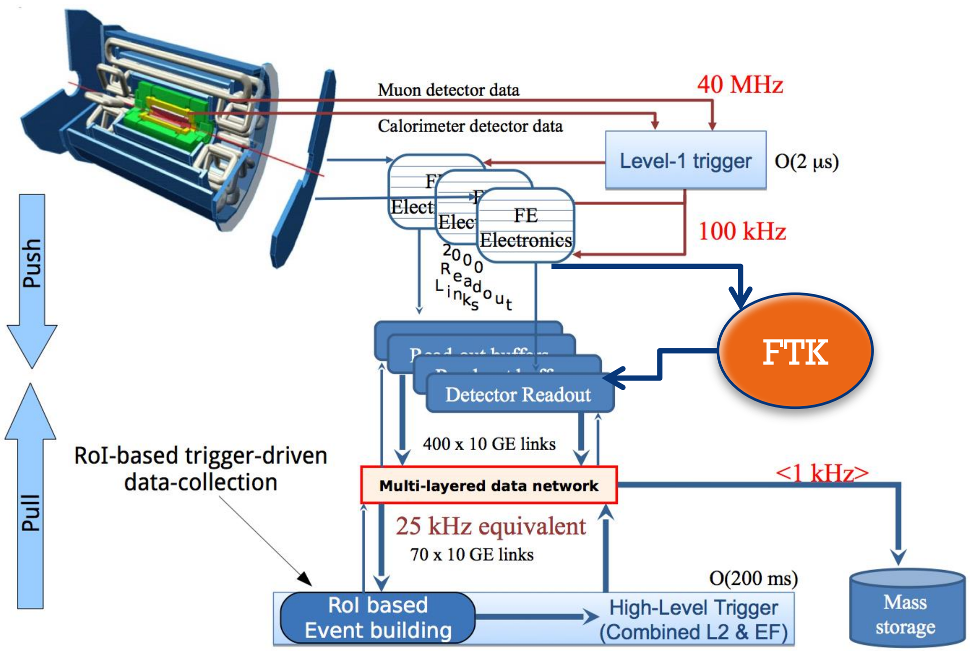
- Some constraints had been over-estimated and some others under-estimated
  - ATLAS/CMS did not manage to run at 100 kHz L1 accept rate
    - Both experiments capped at ~75 kHz
  - Rejection power of HLT was a bit over-estimated
    - Or there is more interesting Physics out there…
  - Storage capacity for raw data was under-estimated
    - In ATLAS 6 PB were stored in 2012 alone
- Technology evolved in our favor
  - Network bandwidth
  - Power of FPGAs

=> Technical stop before Run 2 (2013-2014) was an occasion to redo a lot of things!

Changes in ATLAS

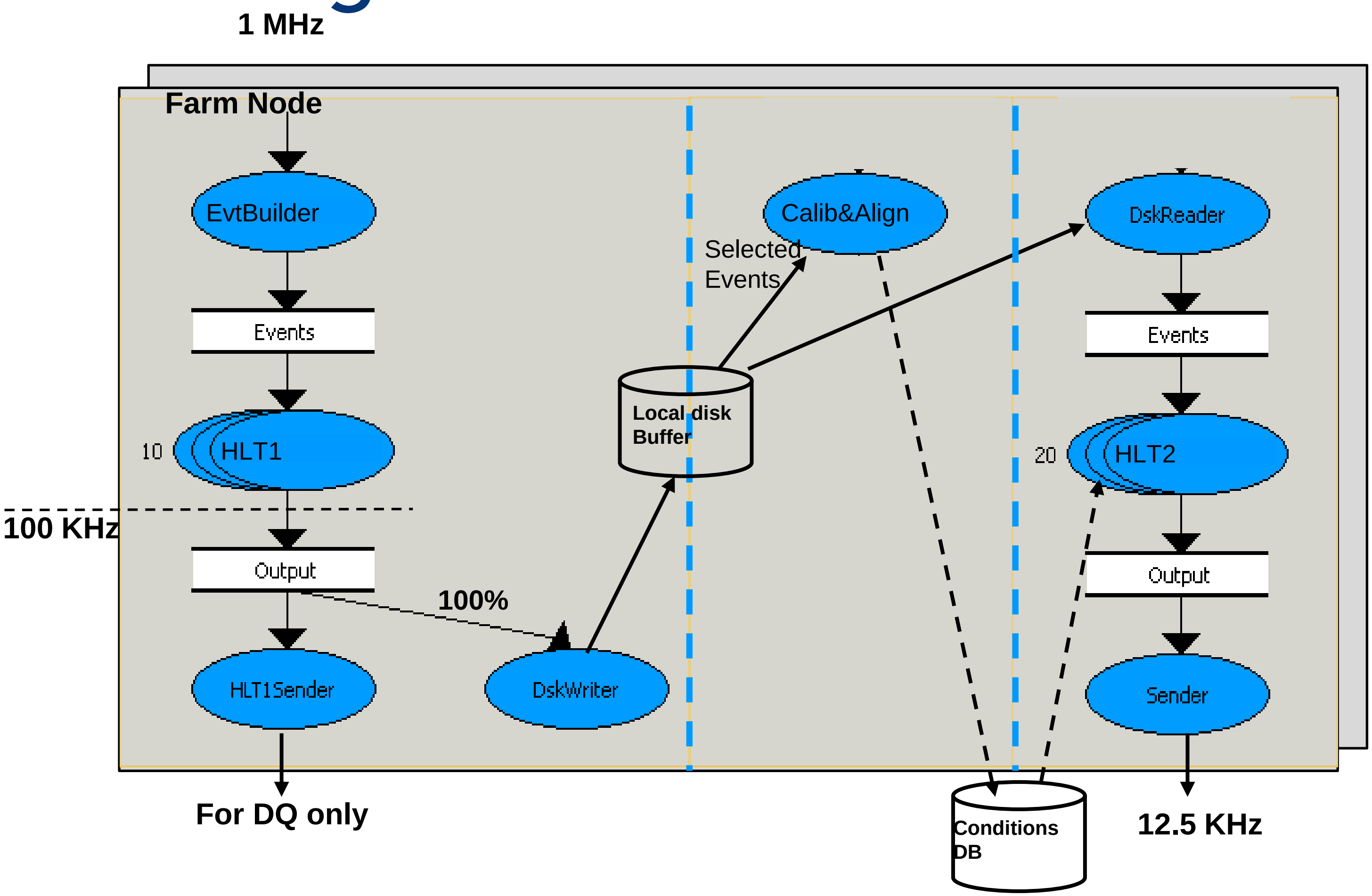


- Introduce Topological trigger at L1 to improve selection
- Introduce Fast Tracker (FTK) to be able to perform fast tracking information to sw selection algorithms (25 μs)



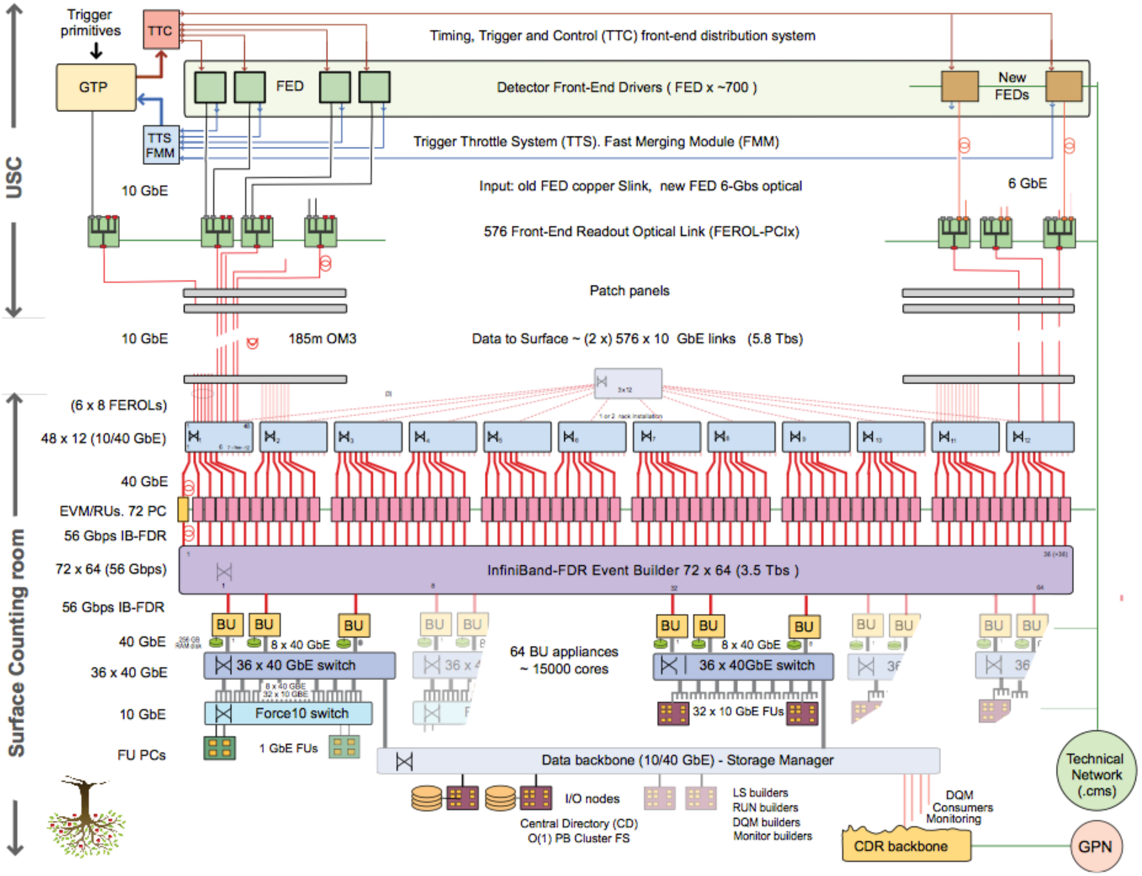
- Use single data network (100 GB/s) and HLT farm
=> simplified data flow
- Increase rate of data to permanent storage

Changes in LHCb



HLT decoupled from data flow via local temporary storage!

Changes in CMS



CMS DAQ completely refurbished:

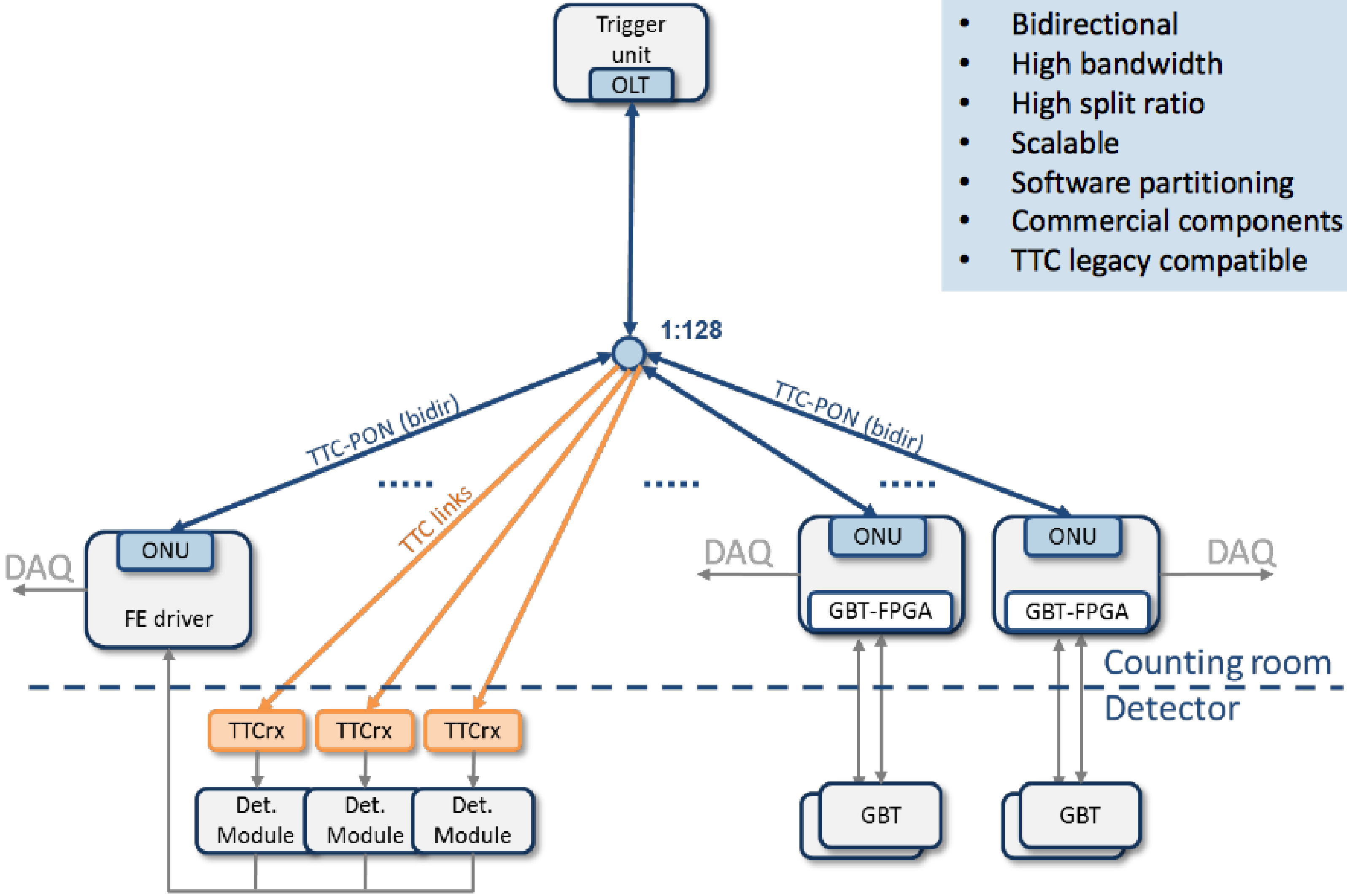
- Elimination of myrinet
- Single event builder InfiniBand Clos network (200 GB/s)

DAQ and HLT decoupled via intermediate shared temporary storage!

Towards the Future

- Experiments upgrade every time the conditions provided by the accelerator change
  - Preparations start well in advance
  - The 4 LHC TDAQ systems are already planning major upgrades
    - ALICE & LCHb will upgrade for Run 3
    - CMS and ATLAS will mainly upgrade for Run 4
- Guiding Principles
  - Physics goals
  - Accelerator conditions
  - Technology reach
  - Cost
- The constraints being fixed, it's always a matter of finding the clever idea(s)…

Synchronization – From TTC to PON

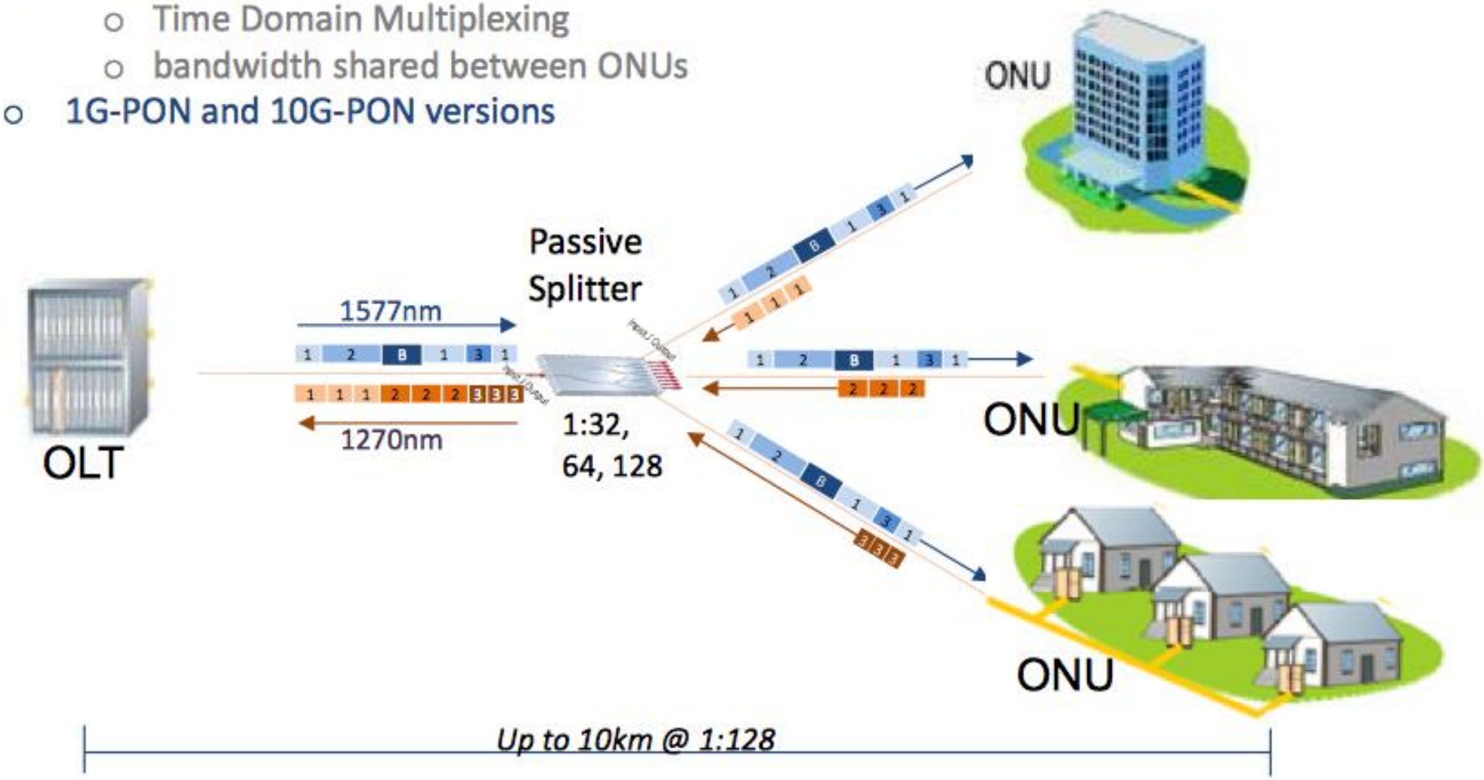


- Bidirectional
- High bandwidth
- High split ratio
- Scalable
- Software partitioning
- Commercial components
- TTC legacy compatible

The PON Principle

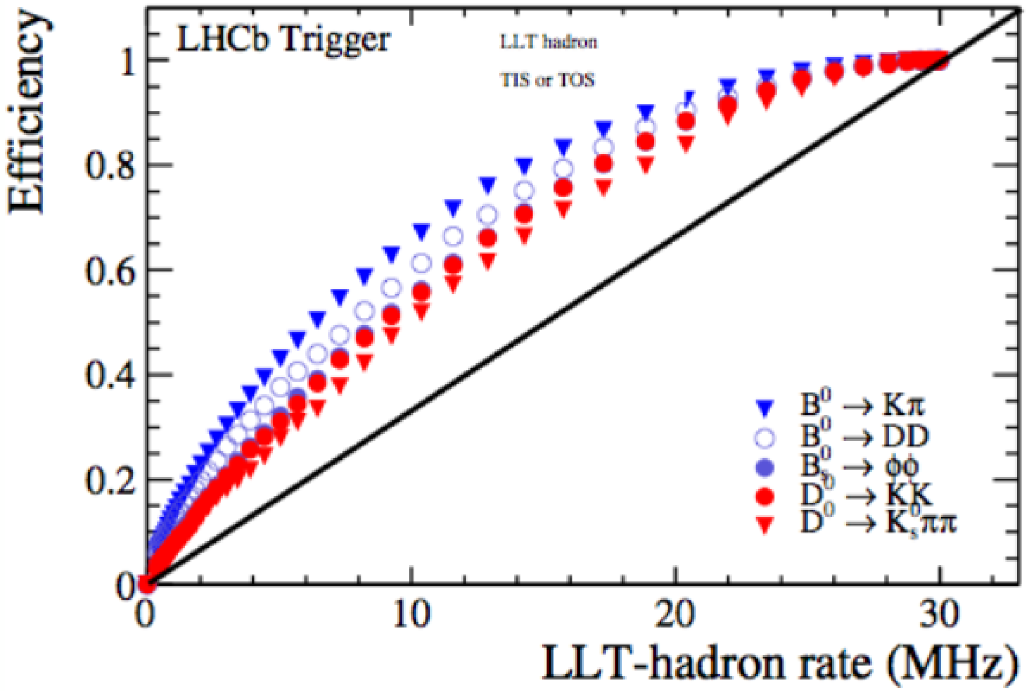
- PON=Passive Optical Network
  - Fiber To The Home (FTTH) technology
  - 1 single fiber, 2 directions
  - 2 wavelengths (one up, one down)
  - Downstream (OLT -> ONUs) :
    - high bandwidth broadcast
  - Upstream (ONUs -> OLT) :
    - Time Domain Multiplexing
    - bandwidth shared between ONUs
  - 1G-PON and 10G-PON versions

ONU = Optical Network Unit
OLT = Optical Line Terminal



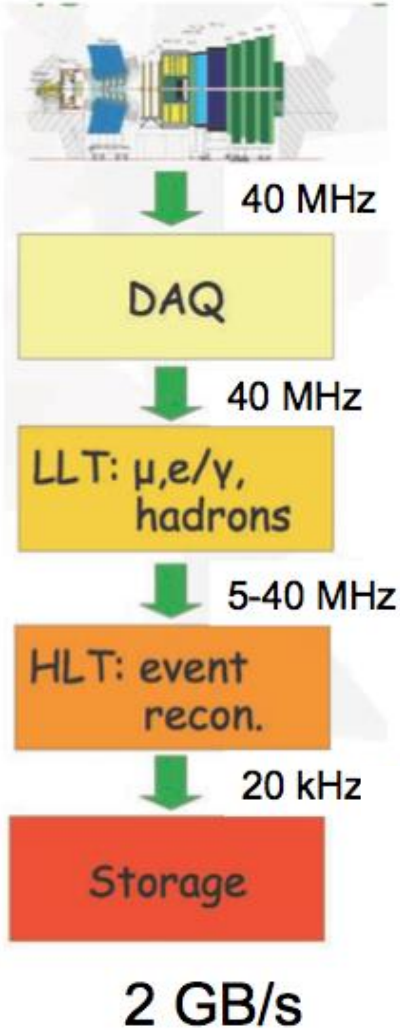
+ LHCb

- Substantial increase in physics reach only possible with massive increase in read-out rate
- Geometry (spectrometer) and comparatively small event-size make it possible – and the easiest solution – to run trigger-free, reading every bunch-crossing
- Note:
  - Any increase beyond 1 MHz requires change of all front-end electronics
  - To keep data-size reasonable, all detectors must zero-suppress at the front-end

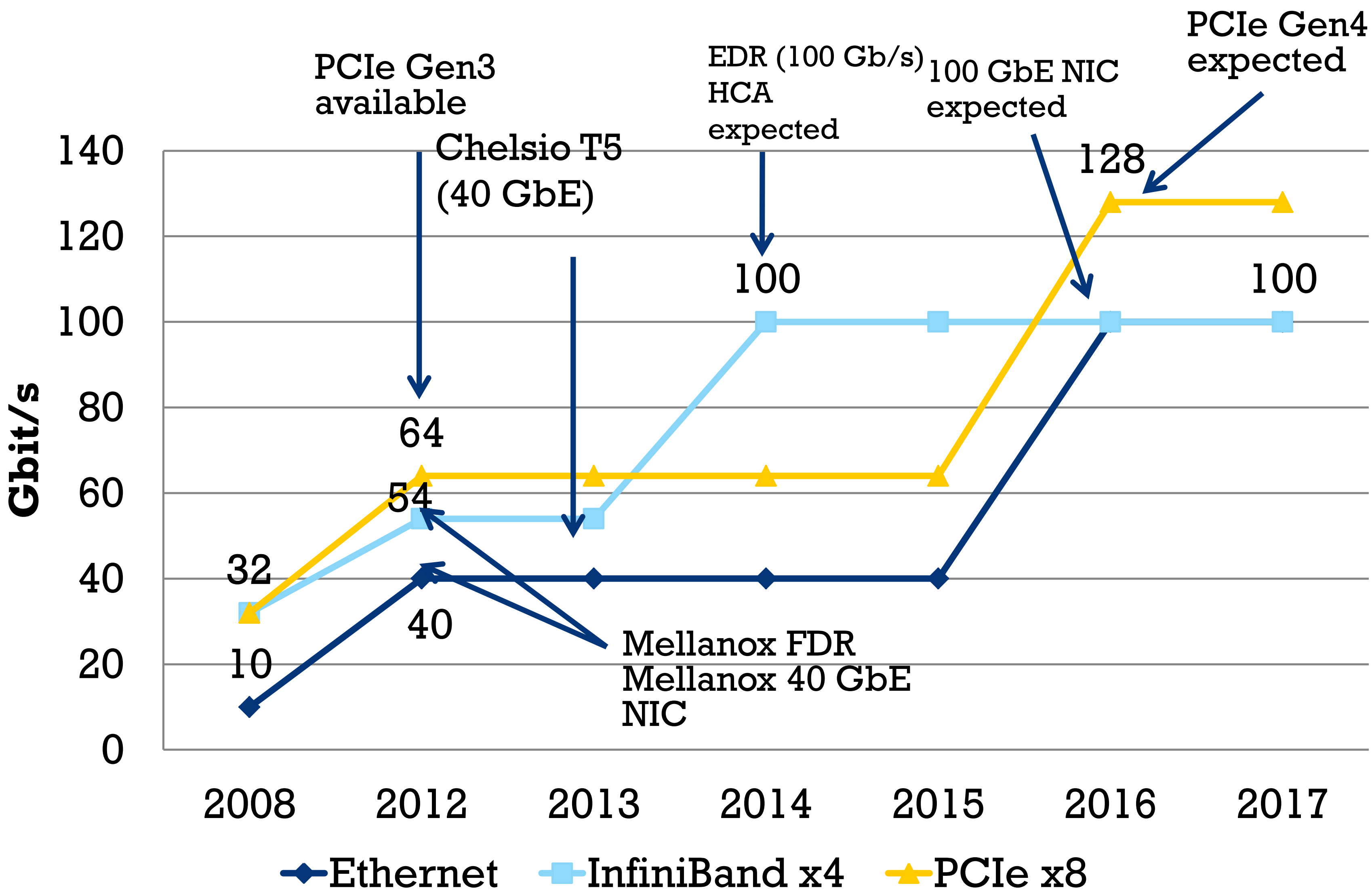


+ LHCb – Requirements for Run 3

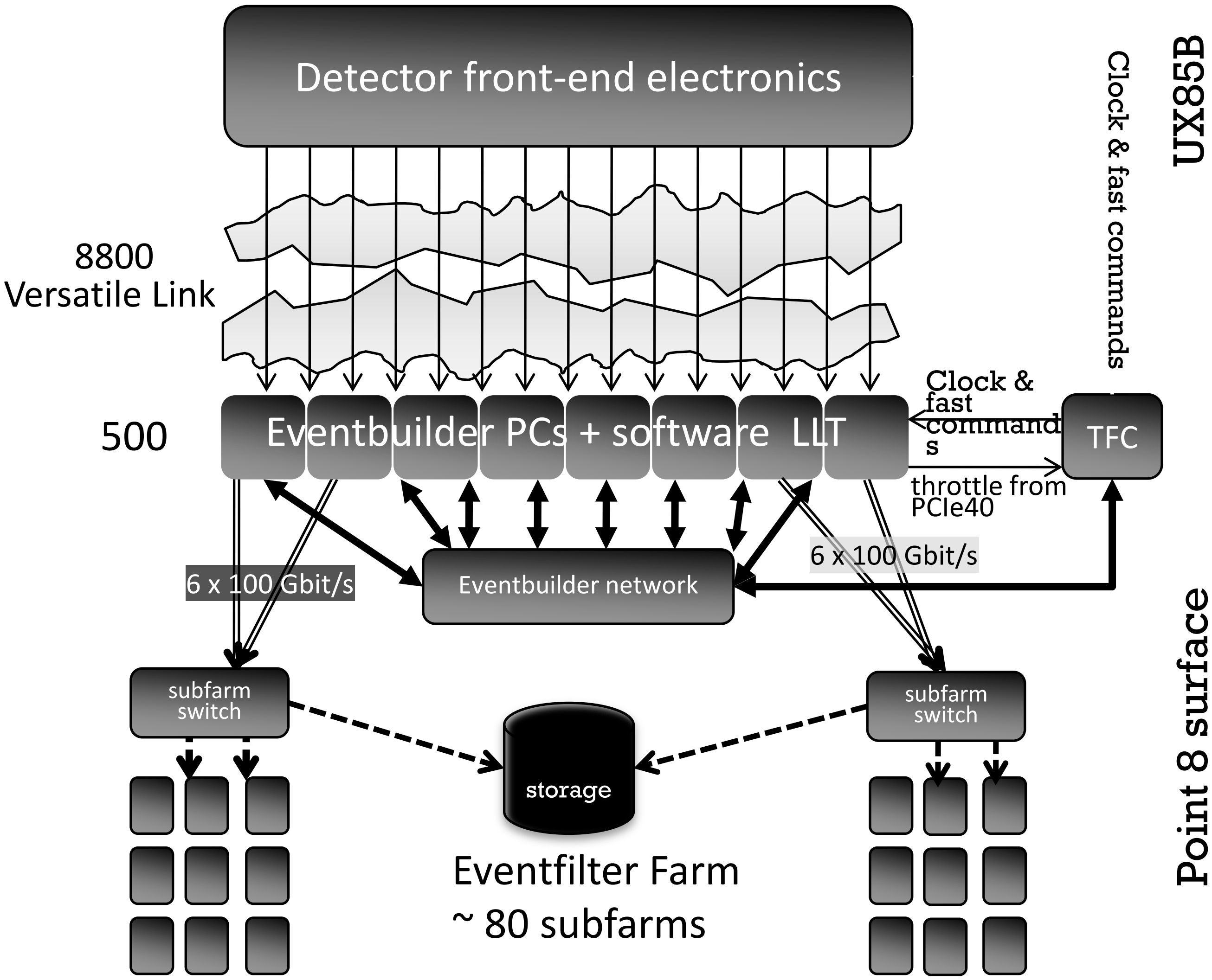
- Event rate 40 MHz
  - of which ~ 30 MHz have protons
- Mean nominal event size 100 kBytes
- Readout board bandwidth up to 100 Gbits/s
  - to match DAQ links of 2018
- CPU nodes up to 4000
  - actual requirements are probably less, but provide for sufficient power, cooling and connectivity to accommodate a wide range of implementations
- Output rate to permanent storage 20 to 100 kHz
- In one number:
  - 8800 (# VL) * 4.48 Gbit/s (wide mode) => 40 Tbps



The evolution of Network Interconnects



Readout Architecture

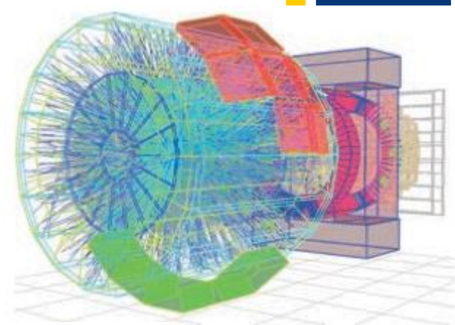


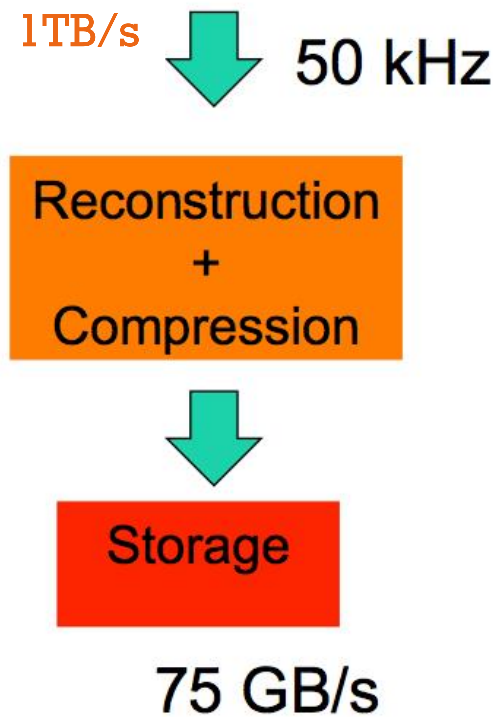
LHCb: Summary

- The trigger-free readout of the LHCb detector requires
  - new, zero-suppressing front-end electronics
  - a 40 Tbit/s DAQ system
- This will be realized by
  - a single, high performance, custom-designed FPGA card (PCIe40)
  - A PC based event-builder using 100 Gbit/s technology and data centre-switches
- LHCb is confident that all inherent challenges can be met at a reasonable cost
  - R&D ongoing on network, versatile links, …

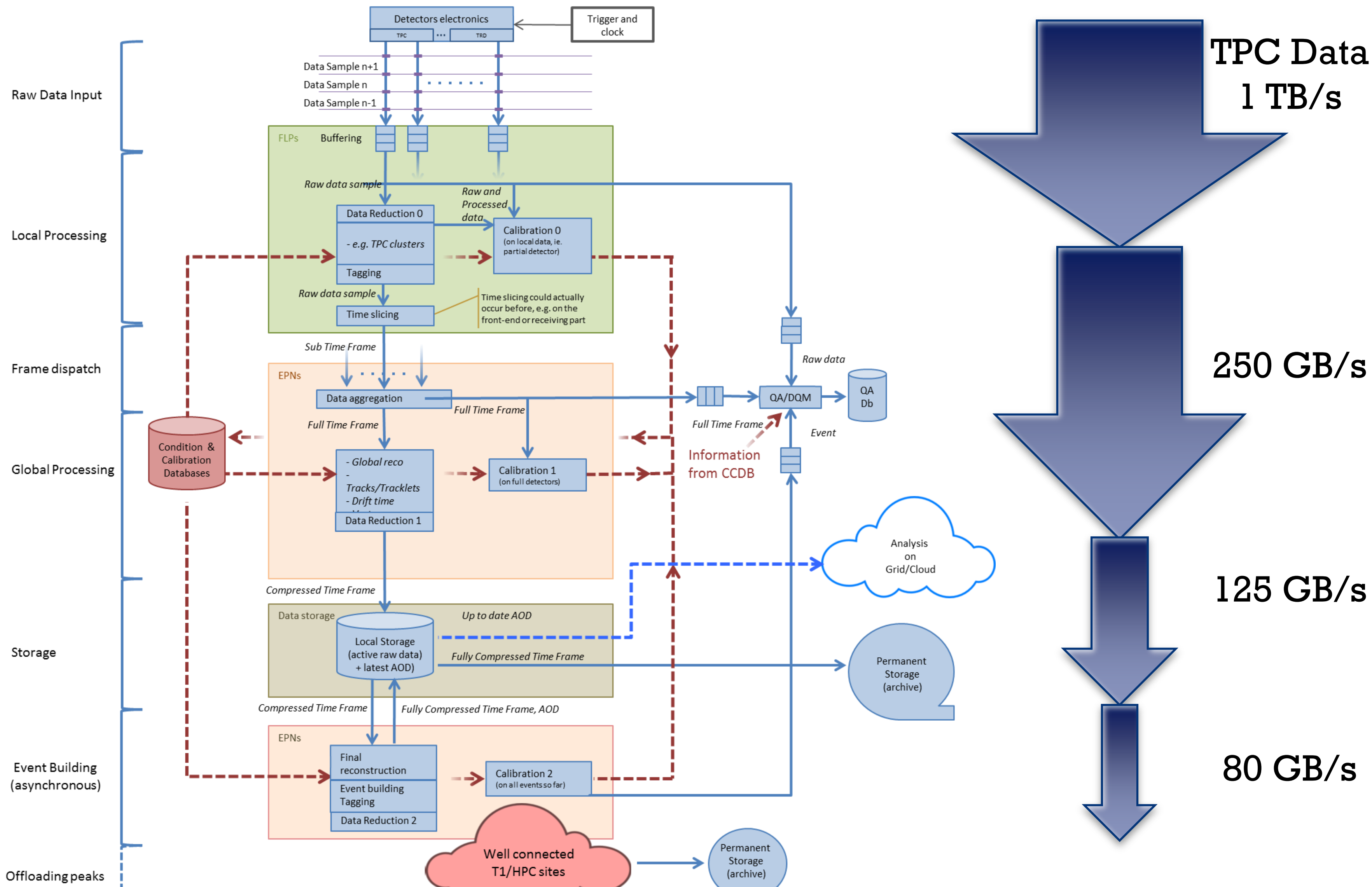
+ ALICE in Run 3

- Focus of ALICE upgrade on physics probes requiring high statistics
- Target Luminosity
  - Pb-Pb recorded luminosity ≥ 10 nb^{-1} (50 kHz)
  - pp (@5.5 TeV) recorded luminosity ≥ 6 pb^{-1} (200 kHz)
  - Minimum bias physics: x100
  - Triggered physics : x10
- Optimize use of detectors:
  - Continuous readout
  - Different busy times
  - Different latency times





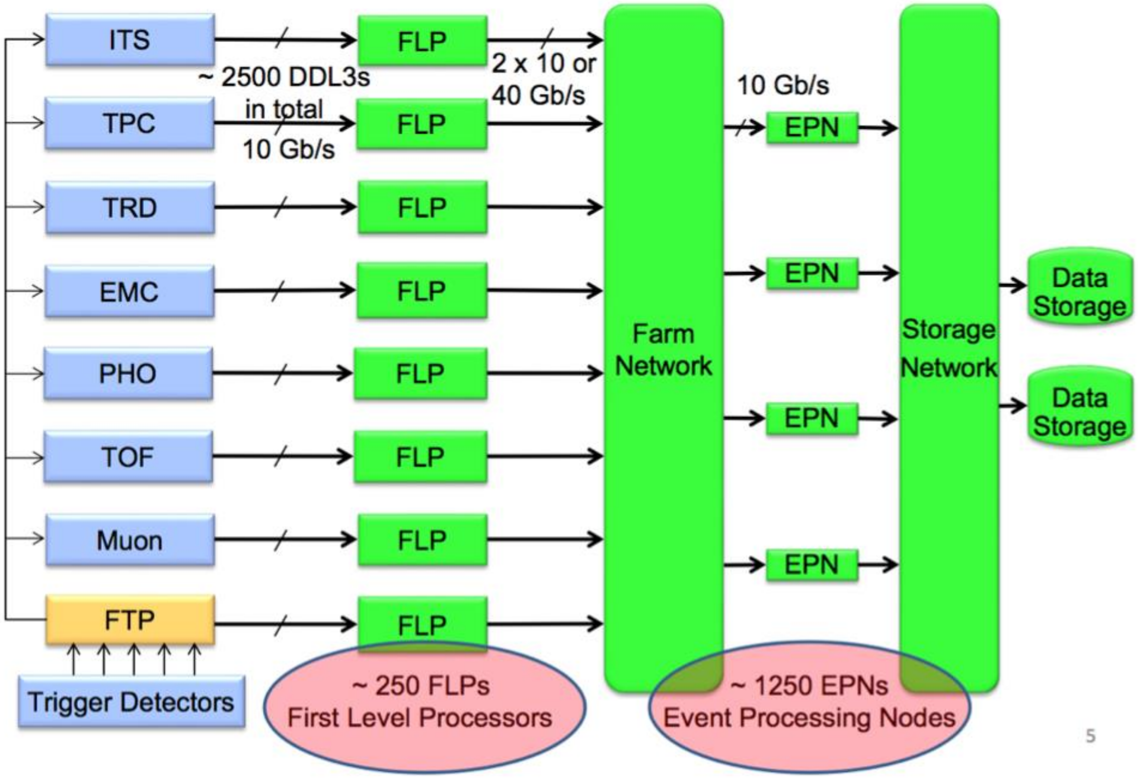
ALICE DAQ


Detectors electronics
TPC
TRD
Trigger and clock
Data Sample n+1
Data Sample n
Data Sample n-1
Raw Data Input
Local Processing
Frame dispatch
Global Processing
Storage
Event Building (asynchronous)
Offloading peaks
FLPs
Buffering
Raw data sample
Raw and Processed data
Data Reduction 0
- e.g. TPC clusters
Tagging
Calibration 0 (on local data, ie. partial detector)
Raw data sample
Time slicing
Time slicing could actually occur before, e.g. on the front-end or receiving part
Sub Time Frame
EPNs
Data aggregation
Full Time Frame
Full Time Frame
- Global reco
- Tracks/Tracklets
- Drift time
Data Reduction 1
Calibration 1 (on full detectors)
Condition & Calibration Databases
Raw data
QA/DQM
QA Db
Full Time Frame
Event
Information from CCDB
Analysis on Grid/Cloud
Compressed Time Frame
Data storage
Local Storage (active raw data) + latest AOD)
Up to date AOD
Fully Compressed Time Frame
Permanent Storage (archive)
Compressed Time Frame
Fully Compressed Time Frame, AOD
EPNs
Final reconstruction
Event building
Tagging
Data Reduction 2
Calibration 2 (on all events so far)
Well connected T1/HPC sites
Permanent Storage (archive)
TPC Data 1 TB/s
250 GB/s
125 GB/s
80 GB/s

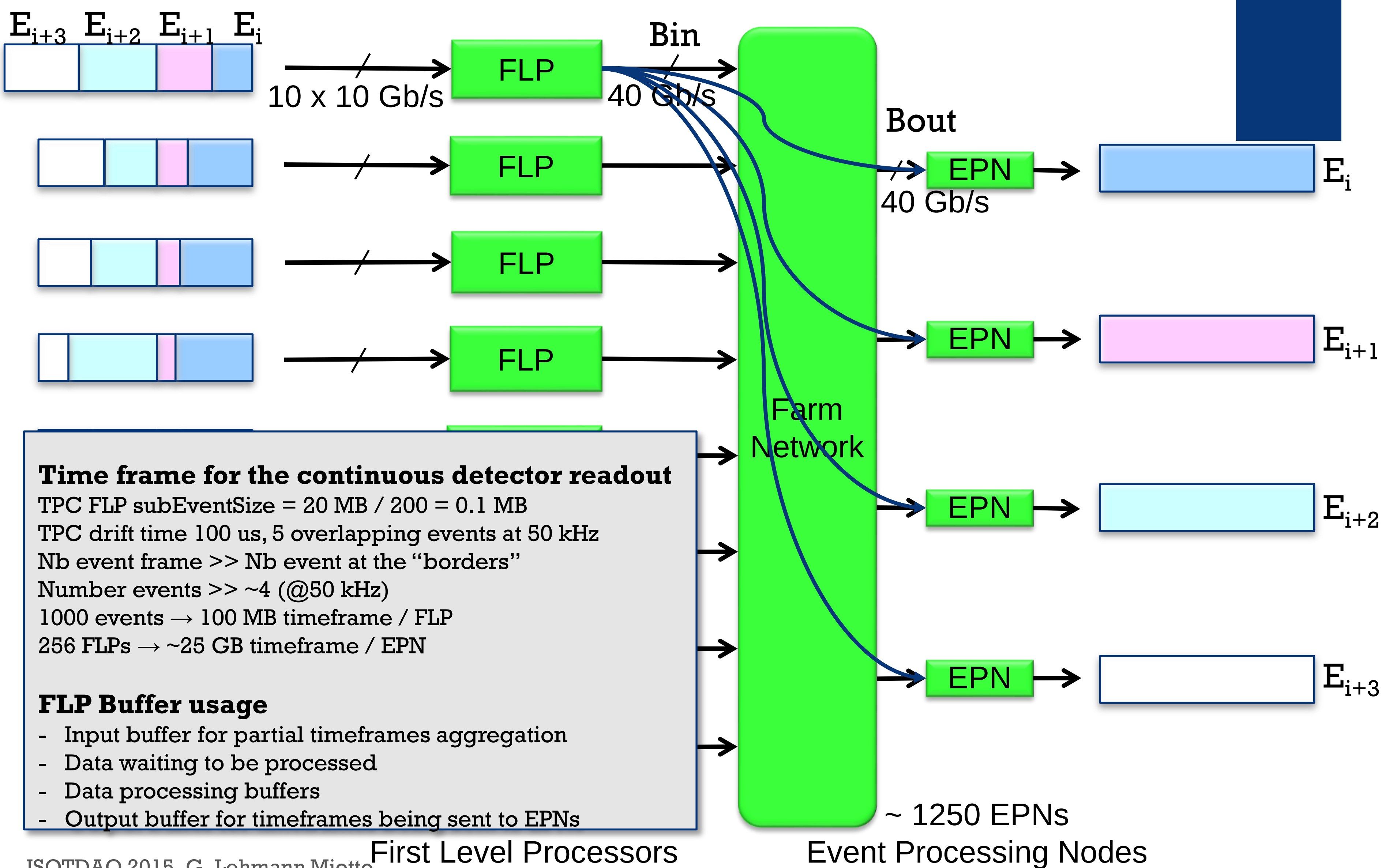
ALICE DAQ

- Event input: 1 TB/s
- Aim at x100 compression
  - Partial event building
- Compression start at each FLP and continues once event in EPN
- Later compression stages perform calibration that is fed in into earlier stages
- Compression preserves ability to re-calibrate offline

Detector	Input to Online System (GByte/s)	Peak Output to Local Data Storage (GByte/s)	Avg. Output to Computing Center (GByte/s)
TPC	1000	50.0	8.0
TRD	81.5	10.0	1.6
ITS	40	10.0	1.6
Others	25	12.5	2.0
Total	**1146.5**	**82.5**	**13.2**



Event building & Continuous Readout



Time frame for the continuous detector readout

TPC FLP subEventSize = 20 MB / 200 = 0.1 MB

TPC drift time 100 us, 5 overlapping events at 50 kHz

Nb event frame >> Nb event at the "borders"

Number events >> ~4 (@50 kHz)

1000 events → 100 MB timeframe / FLP

256 FLPs → ~25 GB timeframe / EPN

FLP Buffer usage

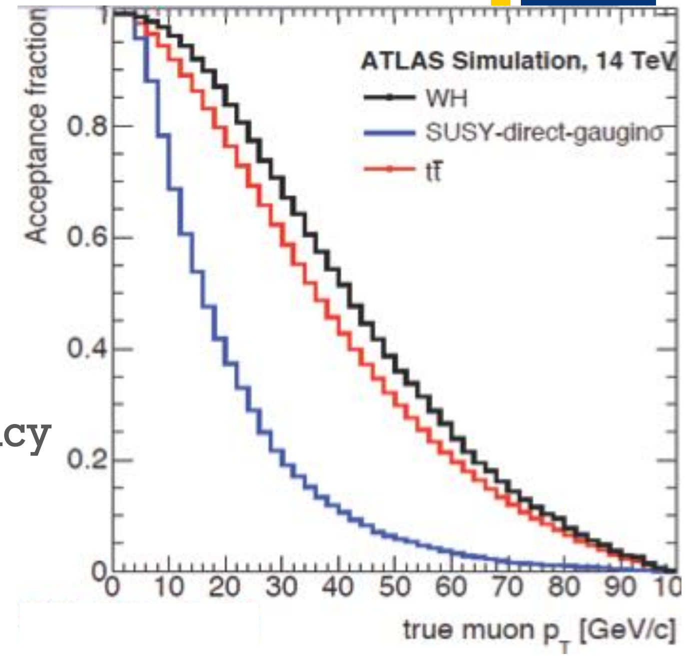
- Input buffer for partial timeframes aggregation
- Data waiting to be processed
- Data processing buffers
- Output buffer for timeframes being sent to EPNs

ALICE: Summary

- Abandon HW trigger in classical sense
- Varied latencies, busy and readout policies for different detectors
- DAQ/HLT will compress data, not select them
  - Goal is to achieve a x100 compression
  - Option of recording only results of reconstruction
- ALICE online and offline integrated into a single workflow
- A lot of research on viable computing platforms, algorithms and data structures optimizations

ATLAS & CMS for Run 4

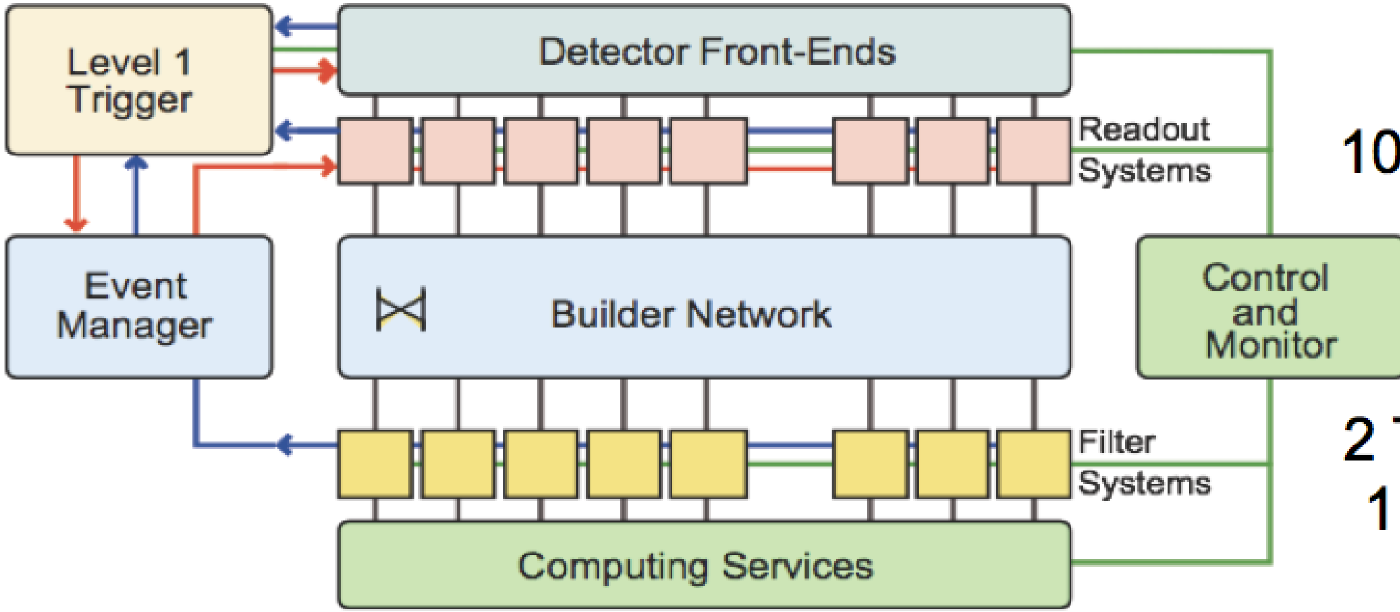
- Maintaining current physics sensitivity at HL-LHC challenging for trigger
  - EWK, top (and Higgs) scale physics remain critical at HL-LHC
  - 100kHz L1 bandwidth cannot fit interesting physics events at 13-14 TeV, $5x10^{34}cm^{-2}s^{-1}$
  - Increasing p_T thresholds reduces signal efficiency
    - Trigger on lepton daughters from H->ZZ at p_T ~ 10-20 GeV
    - Thresholds risk to increase beyond energy scale of interesting processes
  - Backgrounds from HL-LHC pileup reduces the ability to trigger on rare decay products
    - Leptons, photons no longer appear isolated and are lost in QCD backgrounds
    - Increased hadronic activity from pileup impacts jet p_T and MET measurements



+ ATLAS & CMS L1 Tracking Trigger

- Reduces leptonic trigger rate
  - Validate calorimeter or muon trigger object, e.g. discriminating electrons from hadronic ($\pi_0 \rightarrow \gamma\gamma$) backgrounds in jets
  - Addition of precise tracks to improve precision on p_T measurement, sharpening thresholds in muon trigger
  - Degree of isolation of e, γ,μ or τ candidate
  - Requires calorimeter trigger to work at finest granularity to reduce electron trigger rate
- Other triggers
  - Primary z-vertex location within 30 cm luminous region derived from projecting tracks found in trigger layers
  - Provide discrimination against pileup events in multiple object triggers, e.g. in lepton + jet triggers

CMS

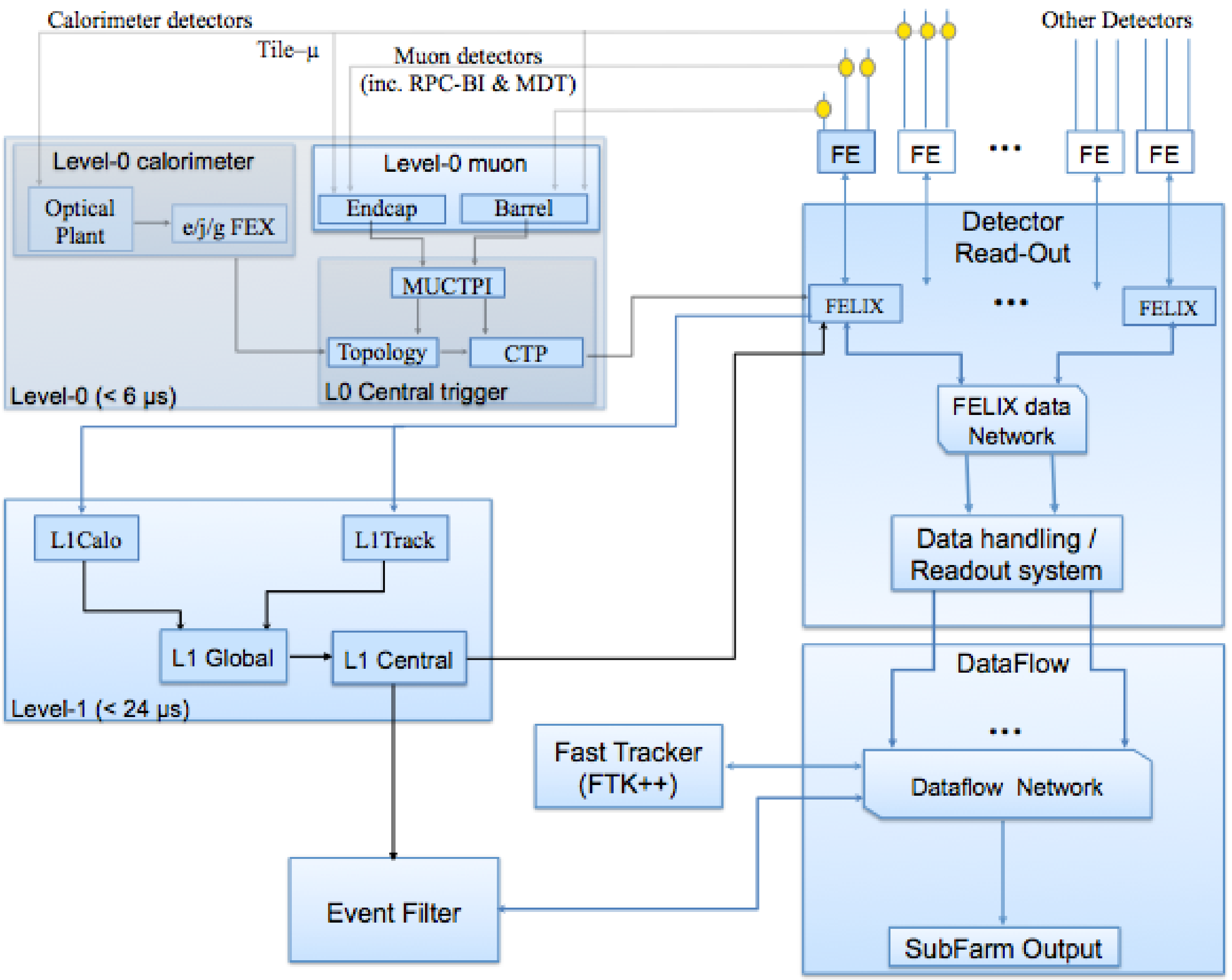


40 MHz
100 kHz → 750 kHz

2 Tbps → ~50 Tbps
1 kHz → ~7.5 kHz

- L1 tracking trigger calculated stand-alone, combined with calorimenter & muon trigger data regionally
- After regional correlation stage, physics objects transmitted to global trigger
- L1 trigger latency = 12.5 μs

ATLAS



- Divide L1 Trigger into L0/L1 of latency 6/30 μsec; rate ≤ 1 MHz/400 kHz
- HLT output 5-10 kHz
- L0 uses cal. & μ Triggers, which generate track trigger seeds
- L1 uses Track Trigger and more fine-grained calorimeter trigger information.

ATLAS & CMS: Summary

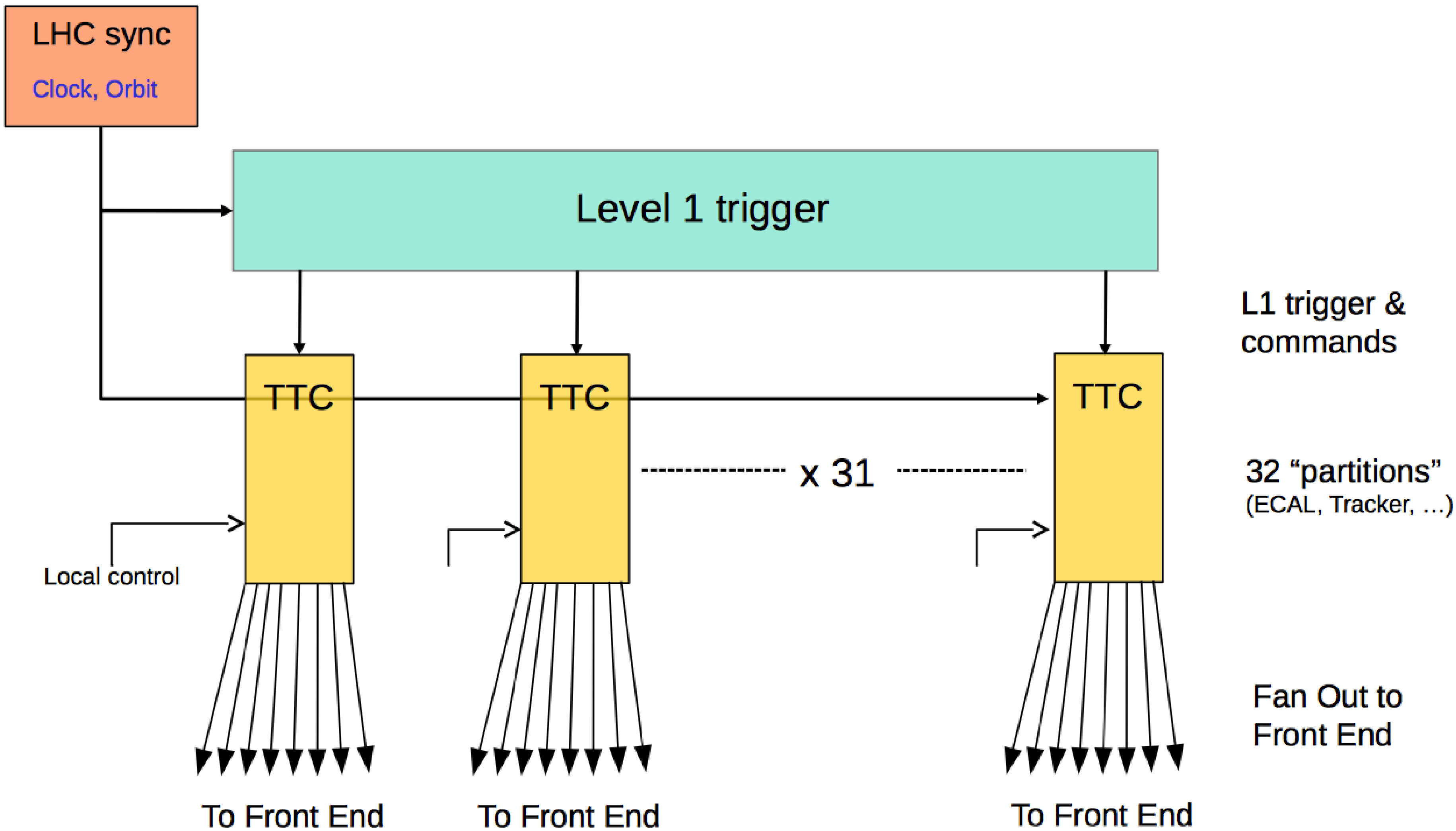
- ATLAS & CMS still need a hardware trigger
  - Ultra low mass, low power and high speed optical links could change this -> R&D
- L1 tracking triggers enable "Run 1" thresholds
  - Technically challenging and strongly coupled to tracker design -> R&D
- L1 global, calorimeter and muon triggers need upgrade to be able to exploit this
  - Also challenging due to the large data rates -> R&D
- Evolution of processing power of processors and co-processors critical for HLT
  - If we do not find clever solutions processing times in the HL-LHC era will explode -> R&D

Summary & Outlook

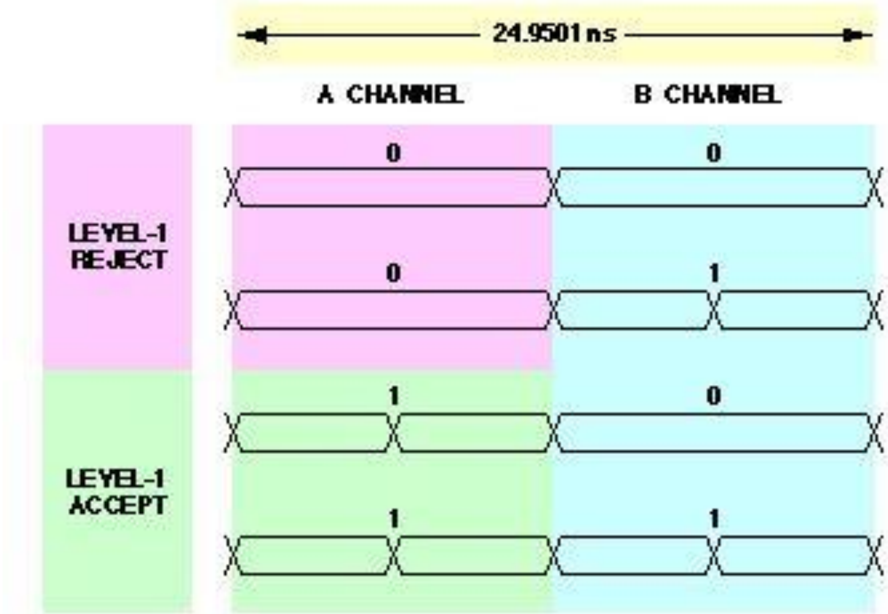
- The TDAQ systems of all four large experiments have a fascinating upgrade programme
- ALICE & LHCb are already designing the new systems in order to use them from 2019 onwards
  - New physics reach
  - Elimination of classic HW trigger stage
- CMS & ATLAS are in a phase of R&D in order to understand how to cope with HL-LHC and preserve the interesting physics
- In all cases, we rely on **your** clever ideas to find the best solutions within the constraints!

Backups

Timing, Trigger & Control at LHC



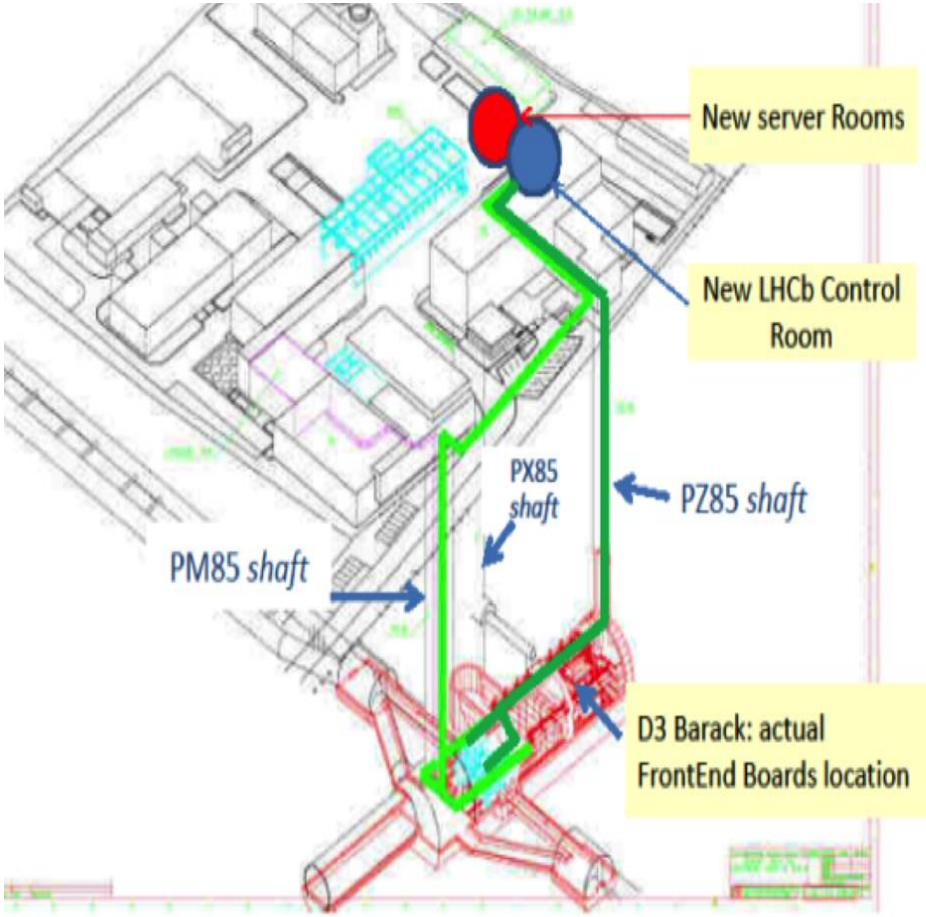
+ TTC Encoding: 2 Channels



- Channel A:
  - One bit every 25ns
  - constant latency required
    - Used to read out pipelines
  - For distribution of LVl1-accept
- Channel B:
  - One Bit every 25 ns
  - Synchronous commands
    - Arrive in fixed relation to LHC Orbit signal
  - Asynchronous commands
    - No guaranteed latency or time relation
  - "Short" broadcast-commands (Bunch Counter Reset, LHC-Orbit)
  - "Long" commands with addressing scheme
    - Serves special sub-system purposes

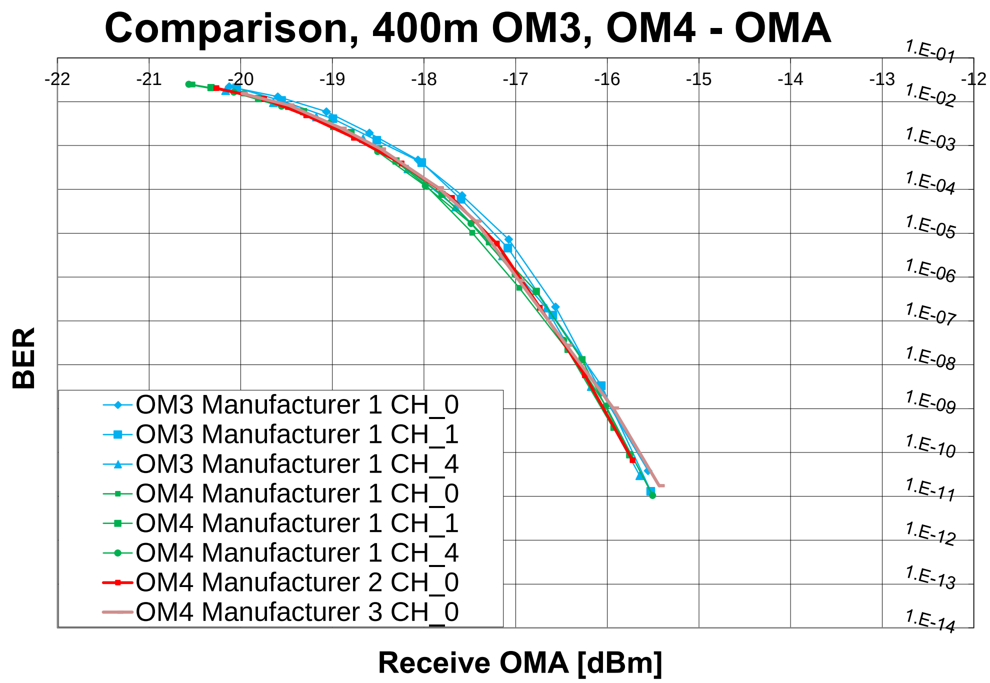
+Long-distance optical fibres

- Most compact system achieved by locating all Online components in a single location
- Power, space and cooling constraints allow such an arrangement only on the surface: containerized data-centre
- Versatile links connecting detector to readout-boards need to cover 300 m



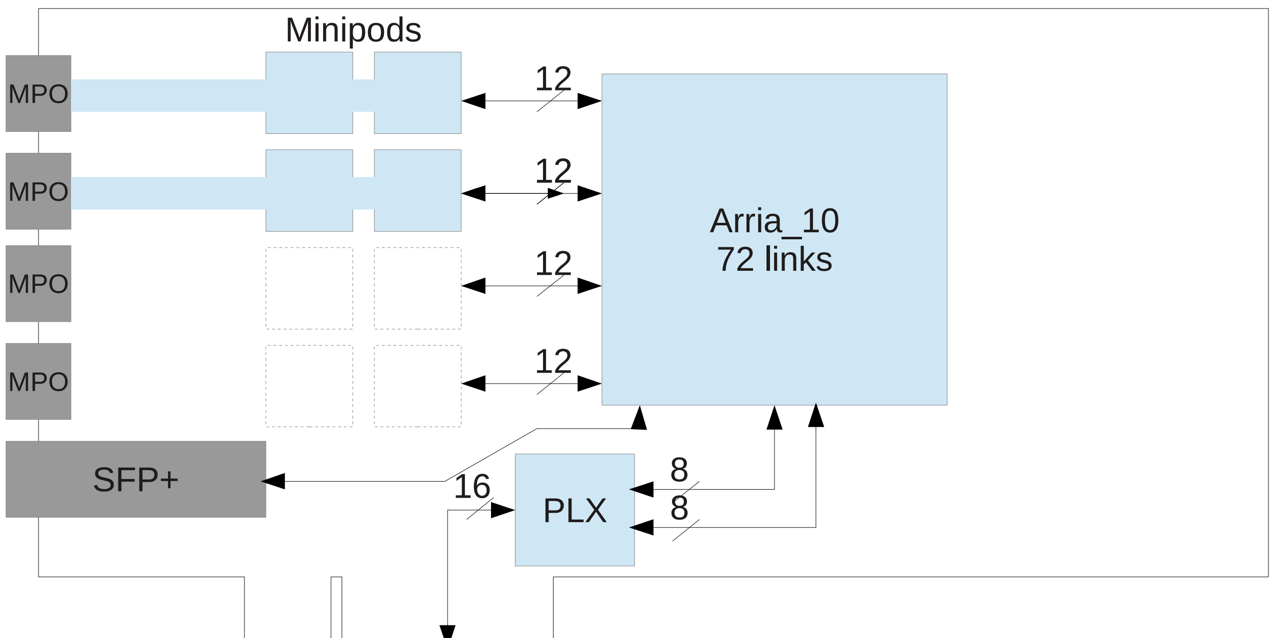
+Long distance versatile link lab tests

- Various optical fibres tested show good optical power margin and very low bit error rates
- For critical ECS and TFC signals Forward Error Correction (standard option in GBT) gives additional margin
- On DAQ links expect < 0.25 bit errors / day / link in 24/7 operation



PCIe40

- Up to 48 bi-directional optical I/Os (VL)
- Up to 100 Gbit/s I/O to the PC (PCIe Gen3 x 16 card)
- Designed by CPP Marseille. Firmware and production support by INFN Bologna, LAPP and CERN
- Universal building block for DAQ, ECS and TFC

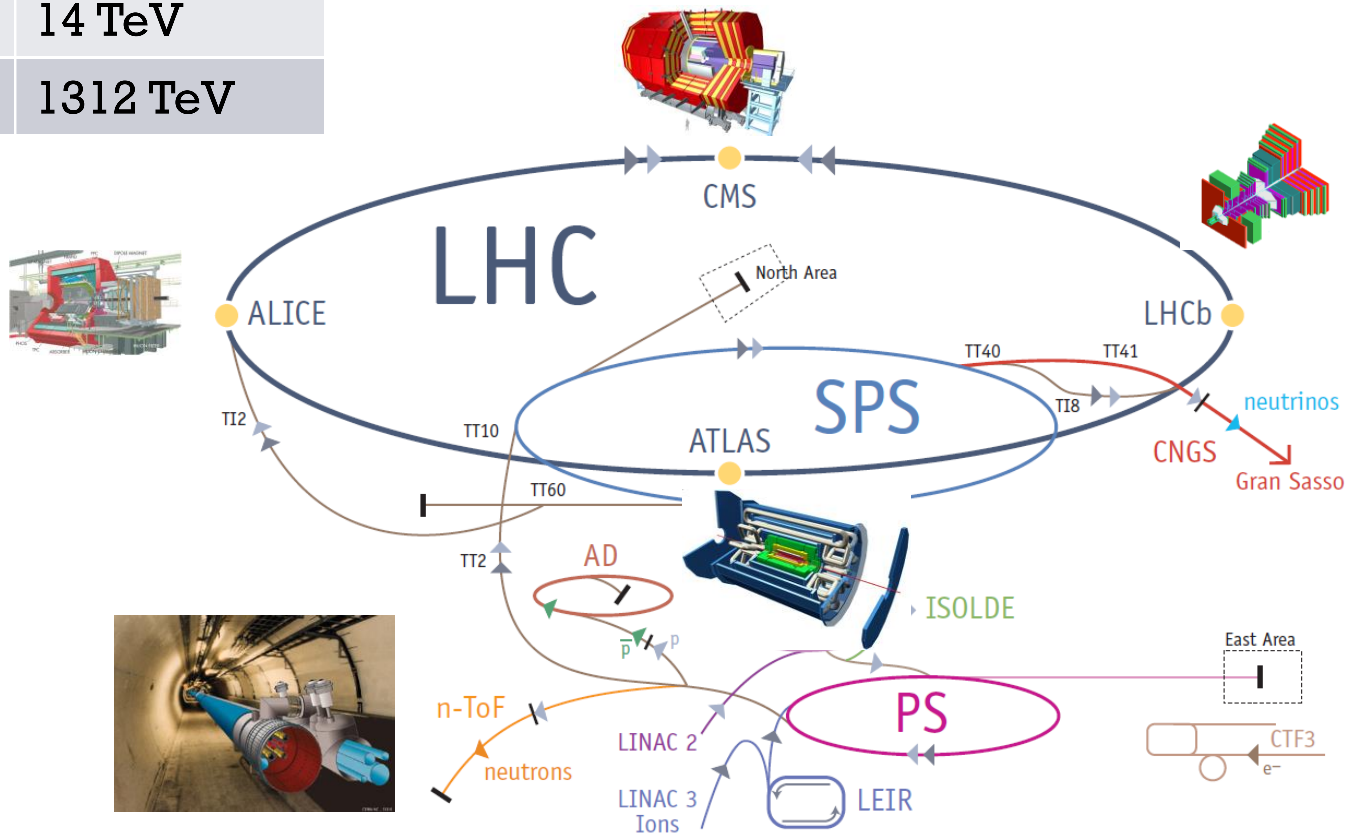


+ Network building & testing

- Core network will require a 500 port 100 Gbit/s device → this will be available
  - Internally probably a Clos (like) topology → need to carefully verify blocking factors and protocol
- Large scale tests require large system
  - Can test opportunistically in HPC sites

+ LHC: A Discovery Machine

	Beams	Energy
LEP	$e^+ e^-$	200 GeV
LHC	p p	14 TeV
LHC	Pb Pb	1312 TeV



+Current and future DAQ

	LHCb Run1 & 2	LHCb Run 3
Max. inst. luminosity	4 x 10^32	2 x 10^33
Event-size (mean – zero-suppressed) [kB]	~ 60 (L0 accepted)	~ 100
Event-building rate [MHz]	1	40
# read-out boards	~ 330	400 - 500
link speed from detector [Gbit/s]	1.6	4.5
output data-rate / read-out board [Gbit/s]	4	100
# detector-links / readout-board	up to 24	up to 48
# farm-nodes	~ 1000	1000 - 4000
# links 100 Gbit/s (from event-builder PCs)	n/a	400 - 500
final output rate to tape [kHz]	5	20 - 100