CMS DAQ in Phase II

Requirements

Strategies

Status of hardware development

(Christoph Schwick for the CMS DAQ group

during the ACES 2020 workshop)

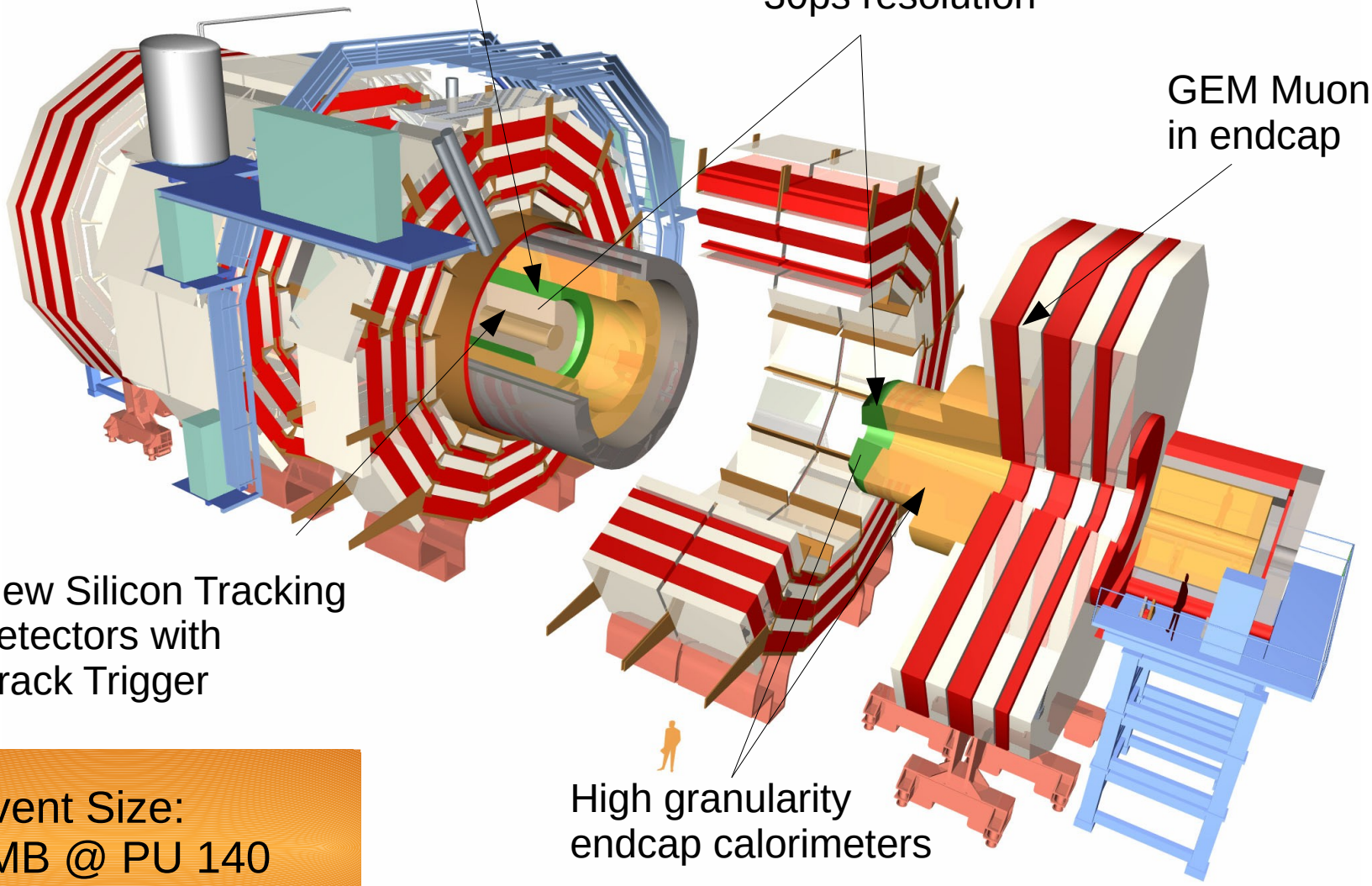
Introduction

Requirements

Setting the scene: the LHC

- LHC parameters foreseen for Phase II
  - HL-LHC Design Parameters
    - Energy 7TeV (possibly 7.5TeV at a later stage)
    - Baseline : $L = 5 \times 10^{34}\ s^{-1}cm^{-2}$ with pile up ~ 140
      5h levelling time, 9.2h optimal Fill length
    - Ultimate : $L = 7.5 \times 10^{34}\ s^{-1}cm^{-2}$ with pile up 200
      Using the margins of the baseline design
  - New for the DAQ: **long periods with peak luminosity due to levelling**
    - Expect this already in run 3
  - **More homogeneous use of resources**
    - Compare to Run 2: Design had to be adequate for peak lumi at the beginning of the fill (or a short levelling time), afterwards DAQ resources were only partially used

Setting the scene: upgraded CMS



Event Size:
6MB @ PU 140
8MB @ PU 200

Setting the scene: upgraded CMS

- Trigger strategy to achieve the required physics performance
  - 2 level trigger system as in Phase 1
  - New Track Trigger at 40MHz in Lvl1
  - Considering accelerators for the HLT (see below)

- Trigger DAQ Performance Parameters

CMS detector	LHC Run-2	HL-LHC Phase-2		
Peak ⟨PU⟩	60	140	200	
L1 accept rate (maximum)	100 kHz	500 kHz	750 kHz	x7
Event Size	2.0 MB [a]	6 MB	8 MB	x4
Event Network throughput	1.6 Tb/s	23 Tb/s	44 Tb/s	x7
Event Network buffer (60 seconds)	12 TB	171 TB	333 TB	x30
HLT accept rate	1 kHz	5 kHz	7.5 kHz	x7
HLT computing power [c]	0.5 MHS06	4.5 MHS06	9.2 MHS06	x18
Storage throughput	2.5 GB/s	31 GB/s	61 GB/s	x24
Storage capacity needed (1 day)	0.2 PB	2.7 PB	5.3 PB	x27

Interim TDR updated for 0.6MB MIP timing detector event size

DAQ Phase II upgrade strategy

- Constraints and observations for Phase II upgrade
  - Available person-power stays constant (optimistic assumption...)
  - During Run 3 the running DAQ systems needs to be maintained
    - Software updates for feature requests and bug fixes
    - Coordination of online operation
      - Shifter training
      - Experts shifts
      - Trouble shooting
    - Hardware replacements and repair
  - Experience with current system during Run 1 & 2 mainly positive
- Strategy
  - Use up to date technologies
  - Improve where lessons are to be learned
  - Adapt to new requirements
  - Profit from experience where possible

No revolution but evolution for DAQ in Phase II (hardware and software)

Status of the Phase II upgrade

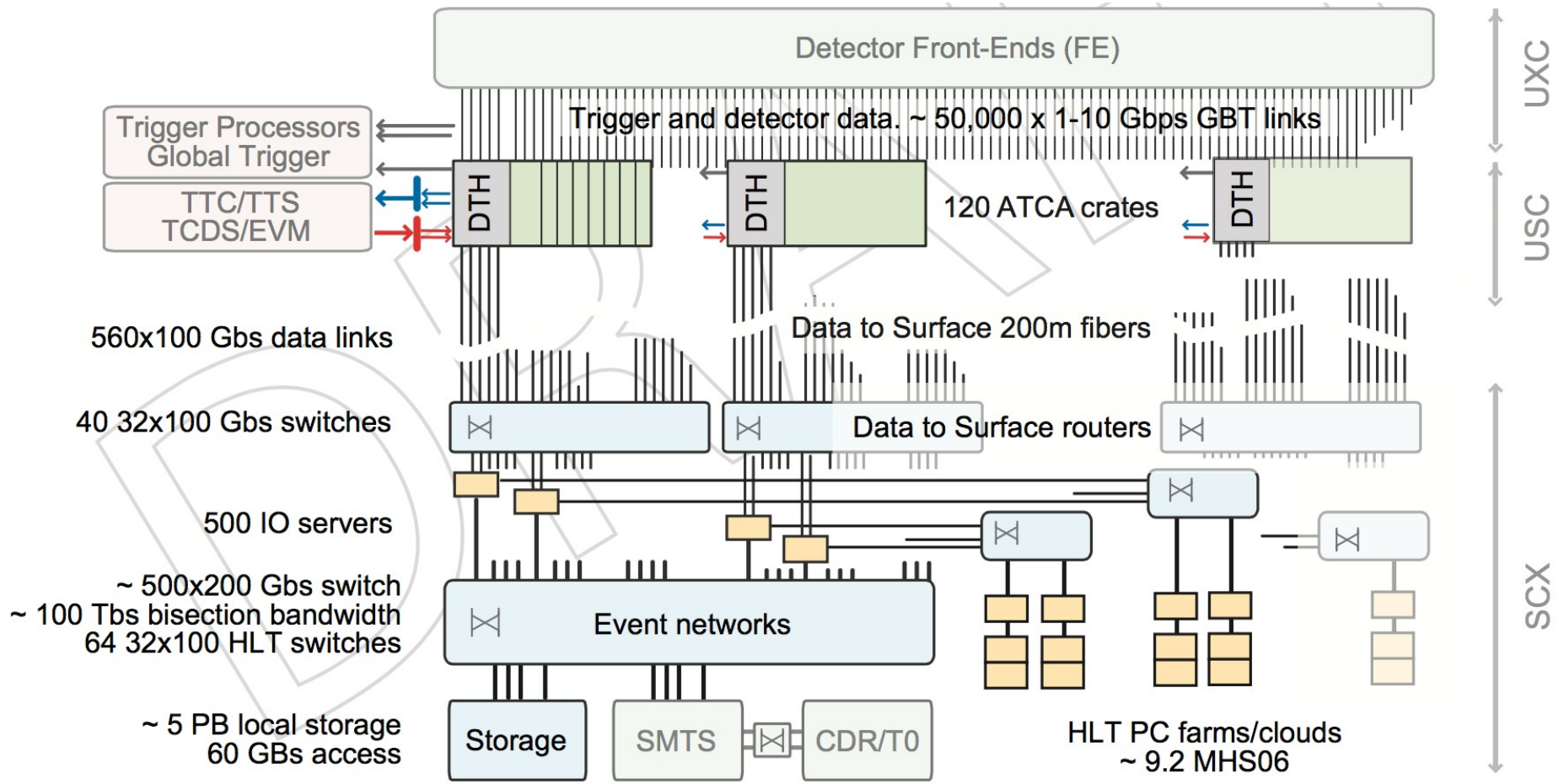
Overall system architecture

Backend readout hardware

Event Builder

HLT farm

Phase II DAQ: Overview

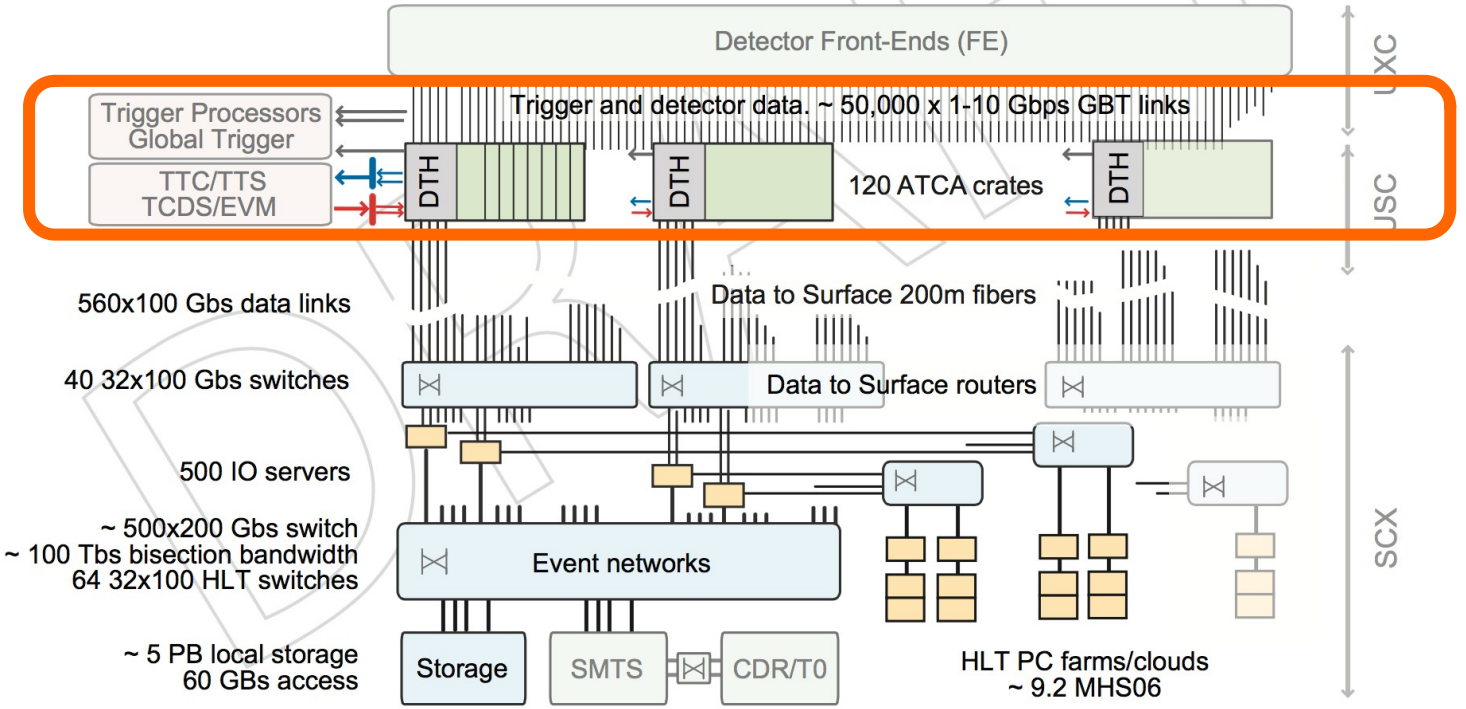


DAQ custom hardware: DTH

Backend readout

Trigger Control and Timing

Subdetector synchronisation



DAQ Custom Hardware: DTH (Daq and Timing Hub)

- Merging of several functionalities in one board
  - **DAQ**:
    - Interface of Backend Boards to Event builder: Receiving end of custom readout links (Slink Rocket)
      - On the sender side Sub-Systems are provided with an FPGA IP block
      - Interface conceptually similar to “traditional” SLINK
      - Output: TCP/IP to event builder network (Readout Unit: RU)
  - **Timing**
    - Precision timing distribution
      - Clock recovery with fixed and reproducible latency from
        - Jitter requirement < 15 ps (met in prototype : 10.6ps. Further improvements possible)
        - Clock distribution over backplane in ATCA crate
  - **Trigger Control System and Trigger distribution**
    - Assumption for Phase II: No TTCrx chips in CMS anymore: Can use new high speed protocol
    - New: Trigger type distribution with every trigger (over backplane)
      - Allows for new features: Luminosity triggers, debugging triggers, …
    - Implementation of trigger rules
  - **Synchronisation**
    - TTC like synchronisation messages
      - Concept TTC b-channel commands
    - Trigger Throttling system
      - Messaging system to signal readiness or synchronisation status of sub-system to trigger control system
  - See talk of **J. Hegeman : The Evolution of the CMS Timing Distribution System (Tue 17h30)**

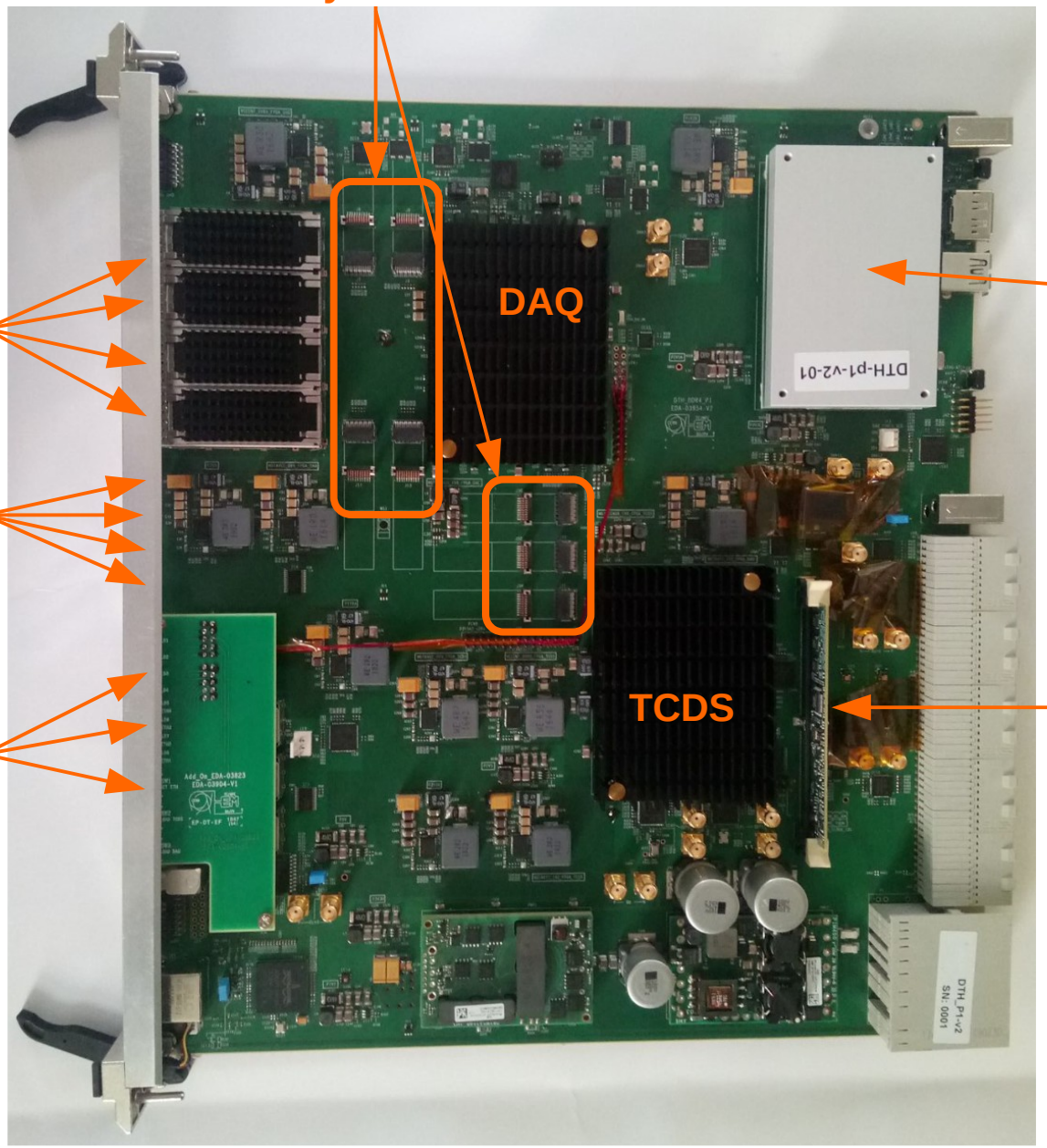
More functionalities of the DTH

- The DTH sits in the ATCA HUB slot
  - Needs to implement a Ethernet switch for the control network of the ATCA cards in the crate
    - Second HUB slot is used by sub-systems and cannot be used for insertion of a commercial card
    - An external switch would be too challenging for cabling and cooling in the crate and would not relieve the challenges of the PCB layout
  - Requirements
    - 1Gbps for the leaf cards
    - 10Gbps uplink in order not to create a bottle neck in the crate

DTH: DAQ part

- Challenge: Varying sub-system requirements for data throughput
  - Solution: Modular design with 2 different board versions
    - DAQ module for 400Gbps output streams
    - TCDS module for Trigger control, Timing and Synchronisation functionality
  - DAQ400 board:
    - ATCA board for HUB slot with one DAQ and one TCDS module
    - Needs to implement Ethernet switch for Control Network into the Backend boards.
  - DAQ800 board:
    - ATCA card with 2 DAQ 400 units
  - Subsystems can combine the DAQ400 board with a number of additional DAQ800 boards to achieve the desired bandwidth.

2nd DTH Prototype (p1v2)



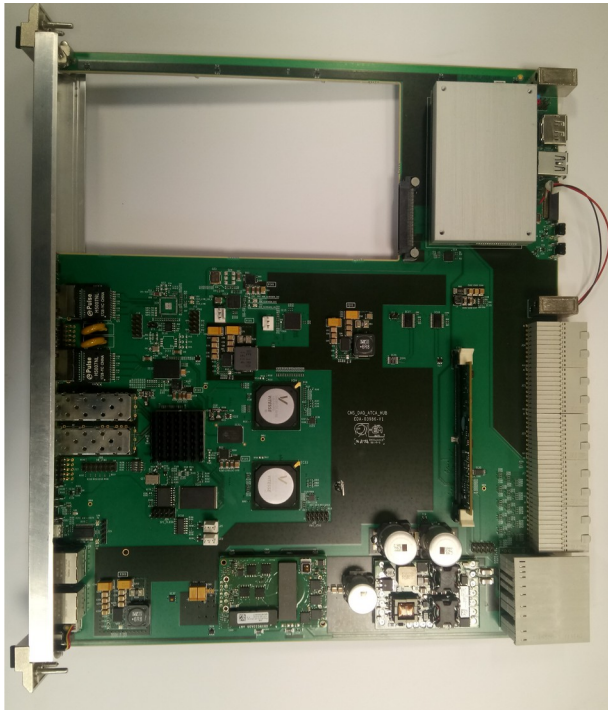
Status of the DTH

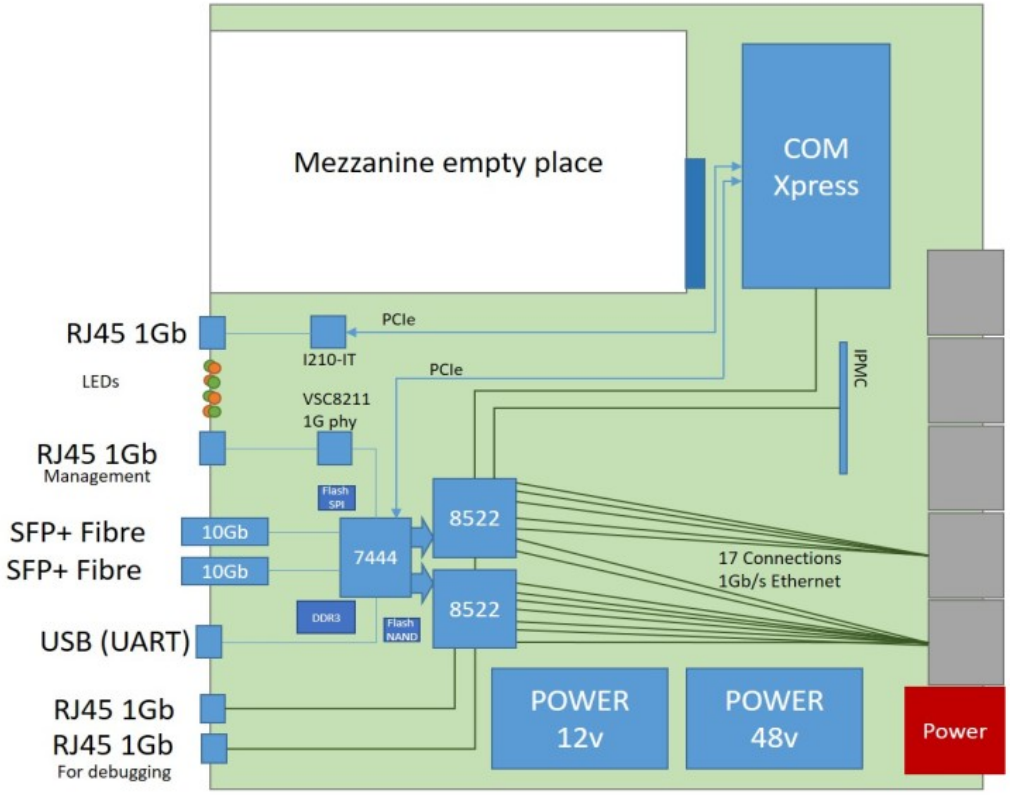
- Second Prototype has been produced (similar to DTH400 board)
  - **Available Features:**
    - Input via MPIO → Firefly adapters. Two alternative Slink Rocket speeds:
      - 4 x (4x25Gbps) = 16 x 25Gbps Slink Rocket input
      - 4 x (6x15.6Gbps) = 24 x 15.6 Slink Rocket inputs
    - Contains a DAQ and a TCDS FPGA
    - System Controller is a ComExpress PC card
    - Running CentOS Linux from CERN
    - Firmware currently supports
      - 2 input channels SlinkRocket
      - 2 100Gbps output streams TCP/IP over 1 QSFP transceiver
  - **Missing Features:**
    - Large buffer memory for TCP/IP buffers
      - Final version will use HBM blocks (High Bandwidth Memory Blocks) in new Xilinx FPGAs
    - Network Switch for control network

- DTH Kit for sub-detector tests
  - The current prototype will be prepared as a Kit for sub-systems to perform readout tests.
  - With a 100Gbps NIC card they can read out data from their backend cards to a PC.
  - Slink Rocket sender IPs are ready to be delivered to sub-systems to implement the SlinkRocket sender

Status of the DTH

- Switch prototype on independent ATCA card
  - Requirement: 1Gbps to all ATCA slots, 10Gbps uplink





  - Based on Vitesse chipset
  - Two 10Gbps uplinks
    - Could be used for redundancy
  - Managed switch
    - Allows to build virtual LANs for each card
    - QoS to guarantee bandwidth for IPMC

DTH: system aspects

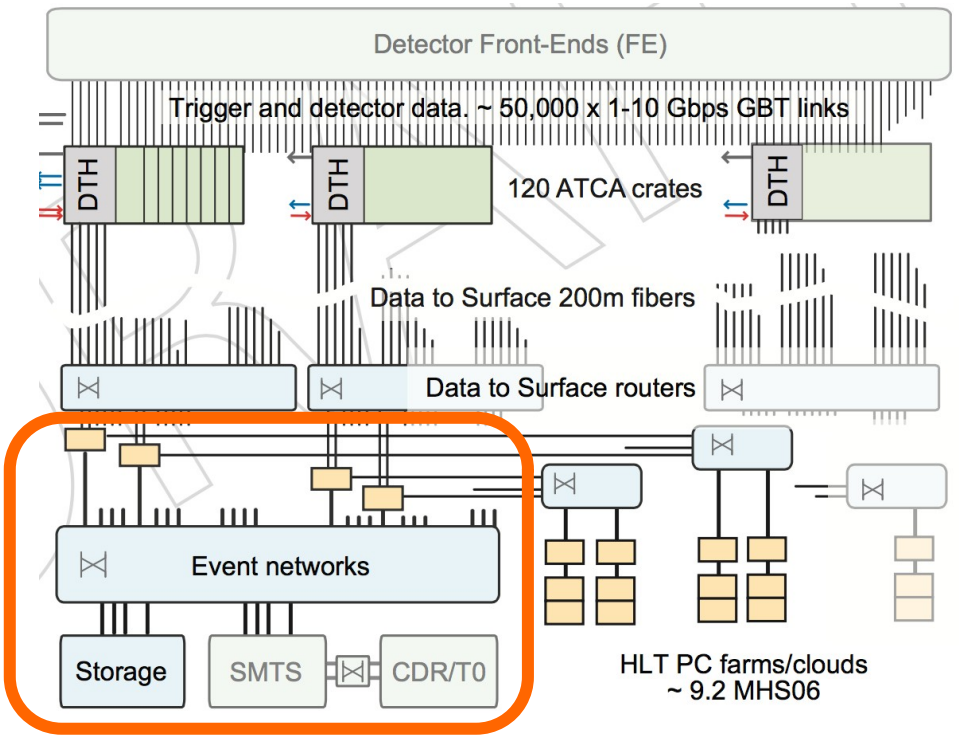
- System aspect: embedded Linux systems on ATCA cards
  - CMS so far did not have many embedded systems running on custom electronics boards
  - ATCA boards in CMS will all have a system controller running linux
    - In addition there will be the IPMC controller
  - The idea is to have a common cern wide standard linux system running on these controllers
    - SoC working group is developing this system
    - See R.Spiwoks: “SoC at CERN Overview and Outlook” in this workshop
  - Under discussion in CMS: how to handle the user software on these boards
    - Default solution: RPM installation as today
    - Alternative being investigated : Docker images
      - Who provides the image ?
      - Policies for what can be installed ?

DTH: open questions

- Data aggregation:
  - How to aggregate data in the DTH to reduce the rate of fragments to handle in the Event Builder
    - Aggregate various Slink Rocket inputs to super-fragments per event
      - Similar to the current operation mode
    - Aggregate fragments from one orbit and send the chunks directly to the Filter Units
      - Not much work left for the Builder Units
      - Requires more work and memory in the Filter Units

- Space on the board
  - Challenging (Impossible?) to implement all functionalities on one board
    - TCDS, DAQ, Switch, Controller PC
  - Consider to “export” either the switch or the Controller on an RTM card

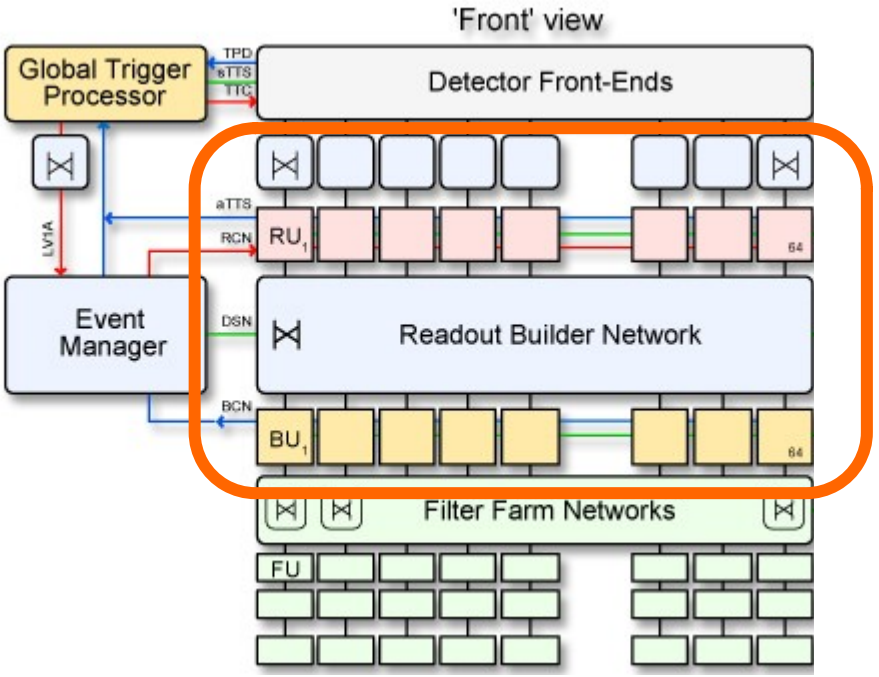
Event Builder



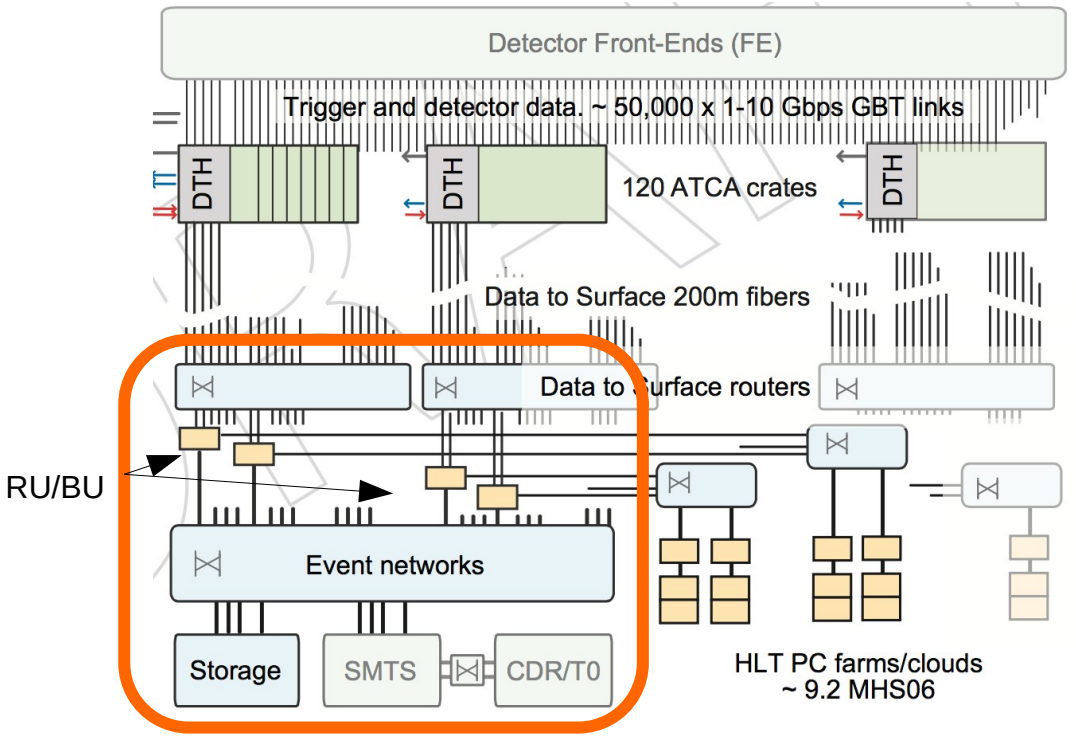
DAQ Upgrade Event Builder

- Important change of architecture (already foreseen for Run-3)
  - Folded Event Builder
  - Nodes contains Readout Unit and Builder Unit
  - Better use of purchased Network Bandwidth (Links are used in both directions)
  - Less nodes to be purchased

Run I & II



Run III & Phase 2

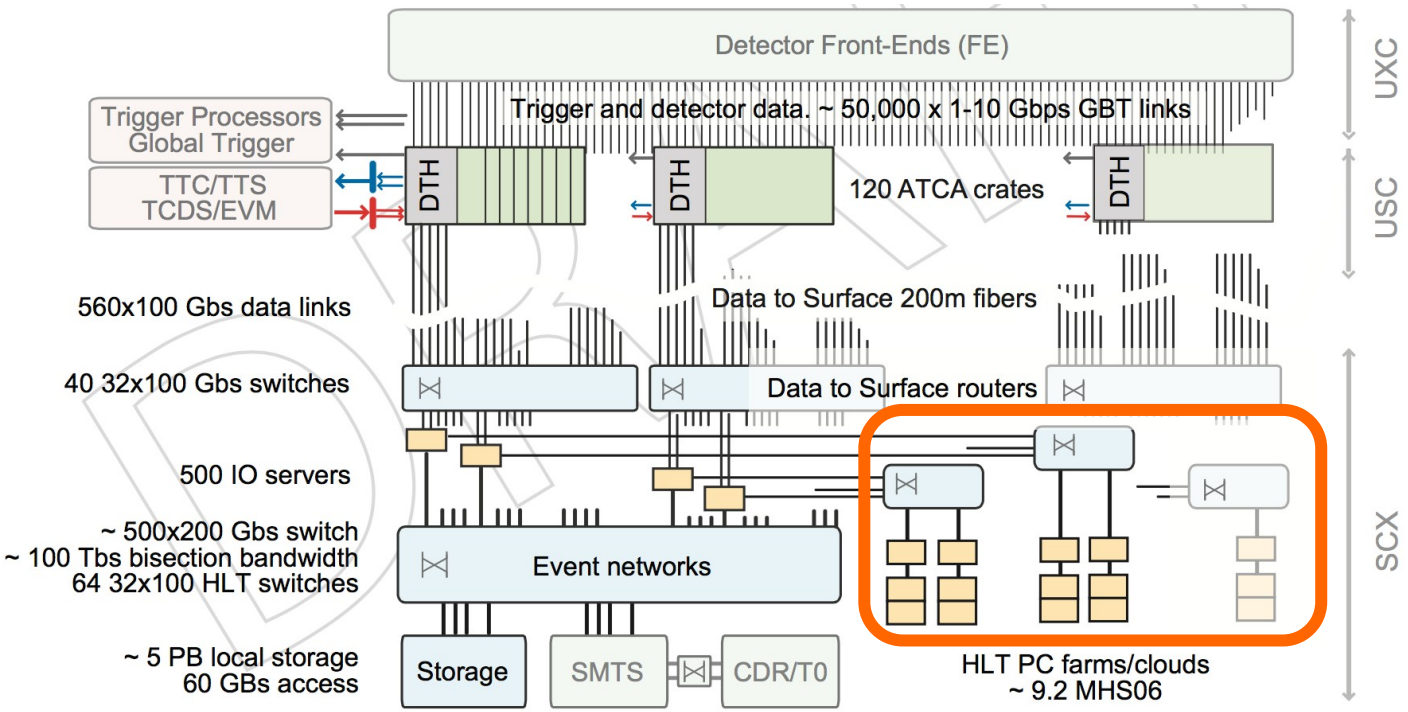


DAQ Upgrade Event Builder

- Data alignment checks will move to hardware (DTH)
  - Previously done in software (at max 100kHz)
  - Significant load on Processing nodes at high rate (L1 rate 750kHz)

- Technology tracking
  - Folded Event Builder is an option due to the performance of affordable server nodes
  - Servers with PCIe Gen4 16Gbps/lane → 16 lanes match 200Gbps NIC
    - 400Gbps is coming for routers now
  - Current status: RDMA (RoCE / Infiniband)
    - Infiniband was used Run 2
    - RoCE is baseline for Run 3

HLT Farm

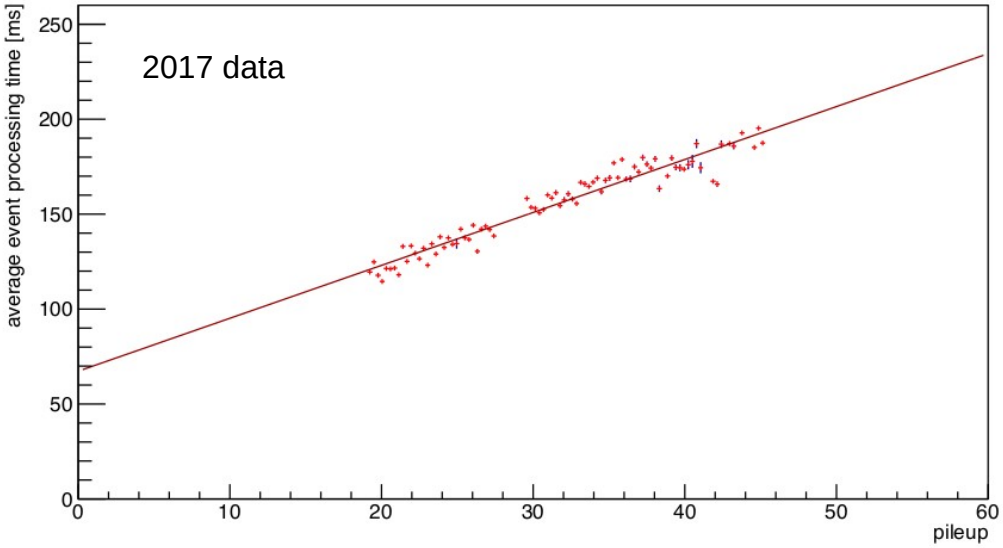

Detector Front-Ends (FE)
Trigger and detector data. ~ 50,000 x 1-10 Gbps GBT links
Trigger Processors
Global Trigger
TTC/TTS
TCDS/EVM
DTH
DTH
120 ATCA crates
DTH
UXC
USC
560x100 Gbs data links
Data to Surface 200m fibers
40 32x100 Gbs switches
Data to Surface routers
500 IO servers
~ 500x200 Gbs switch
~ 100 Tbs bisection bandwidth
64 32x100 HLT switches
Event networks
SCX
~ 5 PB local storage
60 GBs access
Storage
SMTS
CDR/T0
HLT PC farms/clouds
~ 9.2 MHS06

The HLT at Phase II

- Naive considerations for the required Processing Power
  - During Run 2 we observed the average HLT processing time per event as a function of the pileup (μ).
    - We see that the processing time increases approximately linearly in the pileup range from ~20 to ~50
    - If we naively extrapolate this to the regime of HL-LHC we would end up with a processing time of

      ~ 0.5s for μ ~ 140

      ~ 0.6s for μ ~ 200



  - During 2018 we performed a high pileup fill and measured processing time with a reduced HLT menu
    - This menu was not at all optimised for processing high pileup events
    - This test resulted in much higher processing times per event than the above naive extrapolation
      - Extrapolating naively the observations of this special fill to a pileup of μ=200 would result in a processing time of 4 seconds
  - **None of the above naive considerations can be assumed to be correct** but it shows that there are large unknowns in the required processing time and we need to have methods to reduce the processing time.

HLT in Run 2

- Approaches to minimise necessary CPU power
  - Software optimisation
  - Consider Accelerator cards in the HLT nodes
    - Better performance for the same price
    - CMS considers to use this approach in Run 3. In case CMS decides positively:
      - Each HLT node will be equipped with an NVIDIA accelerator card
    - Requires substantial changes in the software to make compatible to run on accelerator cards (ongoing work)
      - 24% of the HLT code can be off-loaded today
      - CMS holds tutorials for writing compatible code
      - Currently testing on NVIDIA accelerators
    - For Phase II it is hoped that 80% of the HLT code can be off-loaded to accelerators (in case this strategy will be pursued)

HLT reconstruction

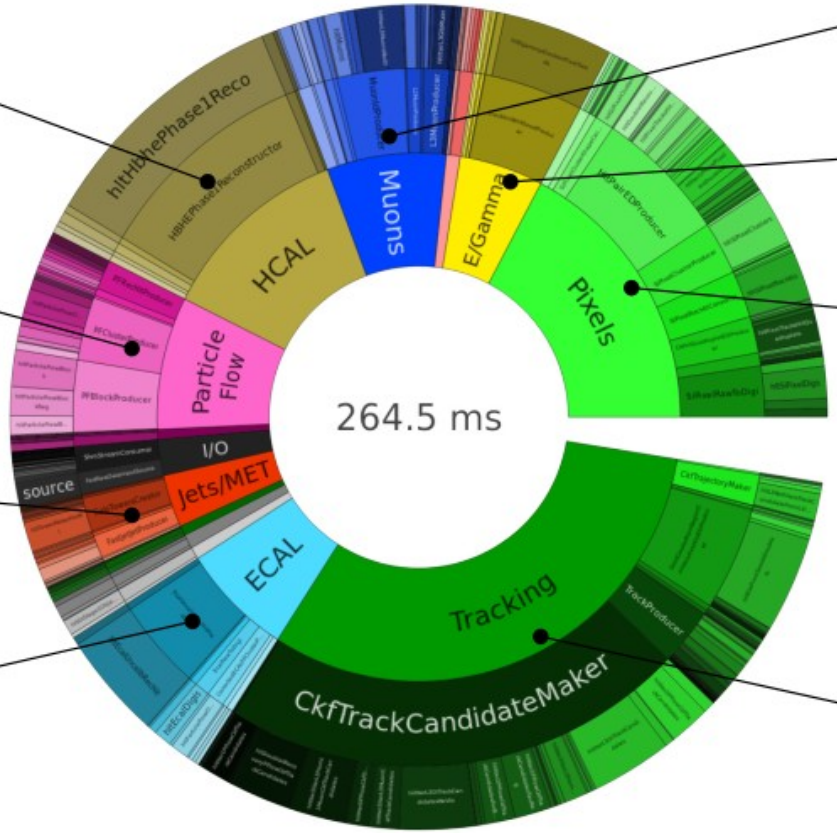
Time needed by various reconstruction steps

HCAL: local reconstruction and calibrations

particle flow reconstruction: regional inside jets and around leptons

jet and MET: reconstruction based on calorimetric and particle flow objects

ECAL: local reconstruction and calibrations



muons: detector local and regional global reconstruction

dedicated reconstruction for **electron** tracks

pixel tracking: global and regional reconstruction

2018 data
- 50 average pileup
- 2018 L1T and HLT

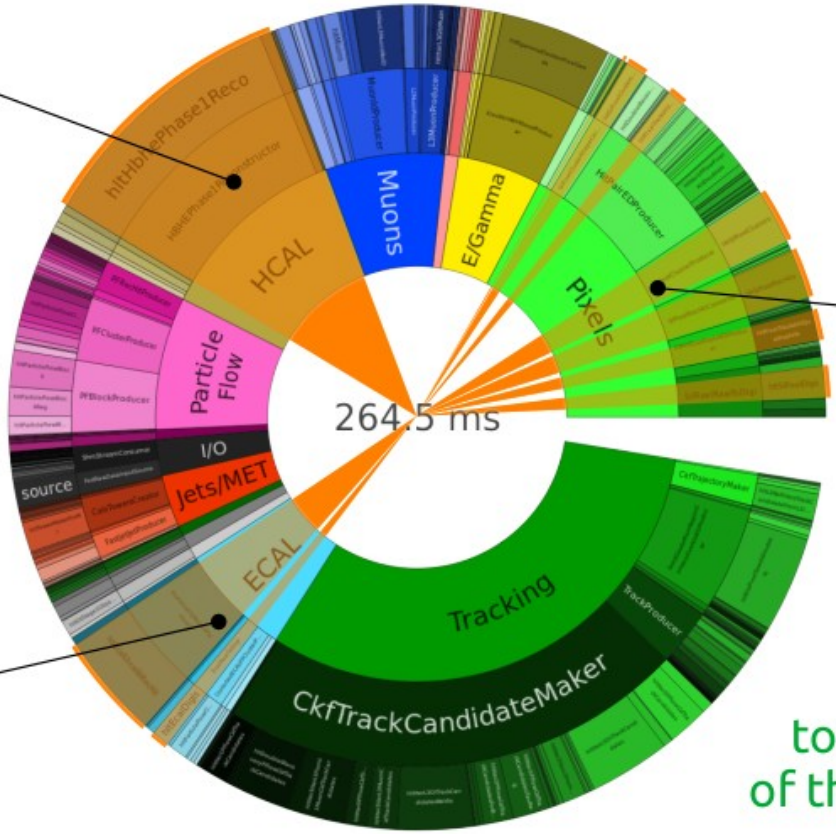
full tracking: regional, partial reconstruction

interactive version

link to circles

24% of HLT reconstruction can be off-loaded today

HCAL: local reconstruction and calibrations



pixel tracking: global reconstruction details on the next slides

ECAL: local reconstruction and calibrations

today we can offload ~24% of the online reconstruction !

Conclusion

- Phase II development of the DAQ system has started
  - Baseline design will be documented in the upcoming TDR

- The custom hardware development is the most advanced part
  - Merging various functionalities in one board
    - Data readout from Backend boards
    - Precision Clock distribution
    - Trigger distribution
    - Sub-detector Synchronisation
  - New Possibilities
    - Trigger types : selective or special triggers
    - Hardware checks of readout data alignment
  - The DTH board needs to be available to subdetectors for testing essentially now

- Some new strategies for Event Builder and Filter Farm will be used in Run 3 already
  - Folded event builder
  - Filter farm with accelerators (still to be signed off by collaboration)

- DAQ TDR will be delivered to the LHCC in June 2021