CMS DAQ for High Luminosity - LHC (run4)

DAQ@LHC workshop 2016,

Frans Meijers – CERN EP-CMD

On behalf of CMS DAQ group

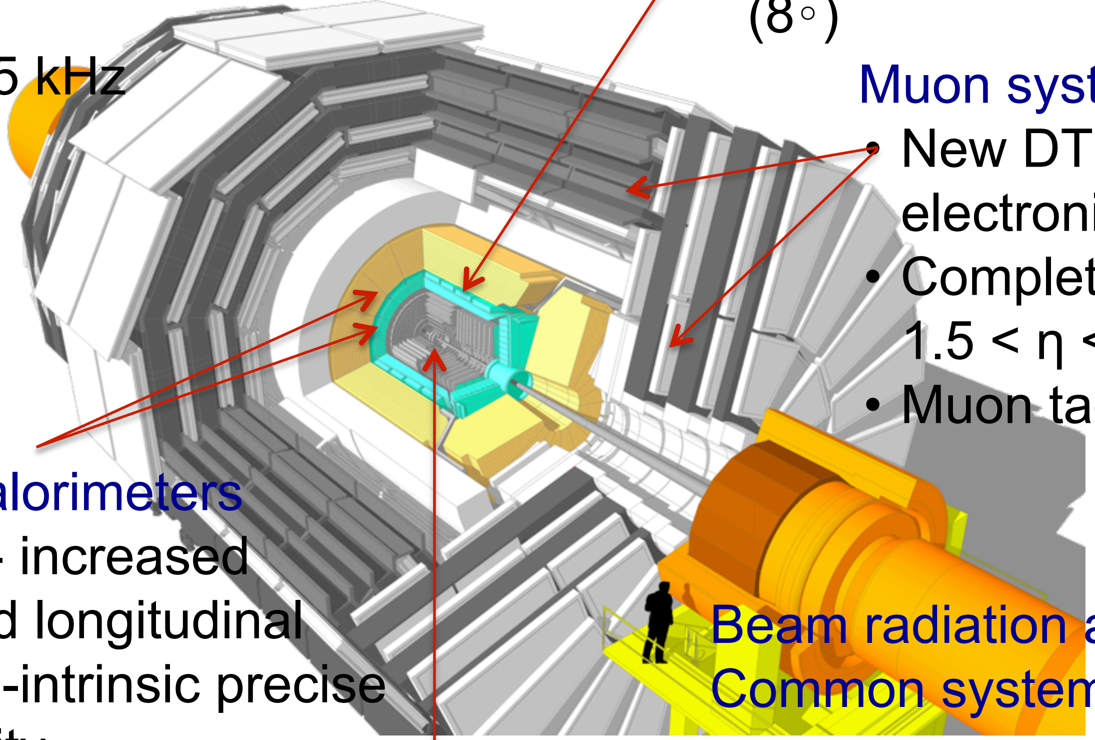
- CMS detector for HL-LHC (run 4 and beyond)
- Baseline Trigger / DAQ / HLT
- Idea for 40 MHz “DAQ scouting”

CMS
at HL-LHC

Schedule TDR

- CMS
  - 2014 Phase-II TP
  - 2015-Q3 Phase-II Upgrade scope
  - **2019 TDR for Trigger**
  - **2020 TDR for DAQ**

CMS “reference” upgrades for Phase-II (Technical Proposal)



Trigger/HLT/DAQ

- Track information in Trigger (hardware)
- Trigger latency 12.5 μs - output rate 750 kHz
- HLT output 7.5 kHz

Barrel EM calorimeter

- New FE/BE electronics
- Lower operating temperature (8◦)

Muon systems

- New DT & CSC FE/BE electronics
- Complete RPC coverage $1.5 < \eta < 2.4$
- Muon tagging $2.4 < \eta < 3$

New Endcap Calorimeters

- Rad. tolerant - increased transverse and longitudinal segmentation -intrinsic precise timing capability

Beam radiation and luminosity
Common systems and infrastructu

New Tracker

- Rad. tolerant - increased granularity - lighter
- 40 MHz selective readout (Pt≥2 GeV) in Outer Tracker for Trigger
- Extended coverage to $\eta \simeq 3.8$

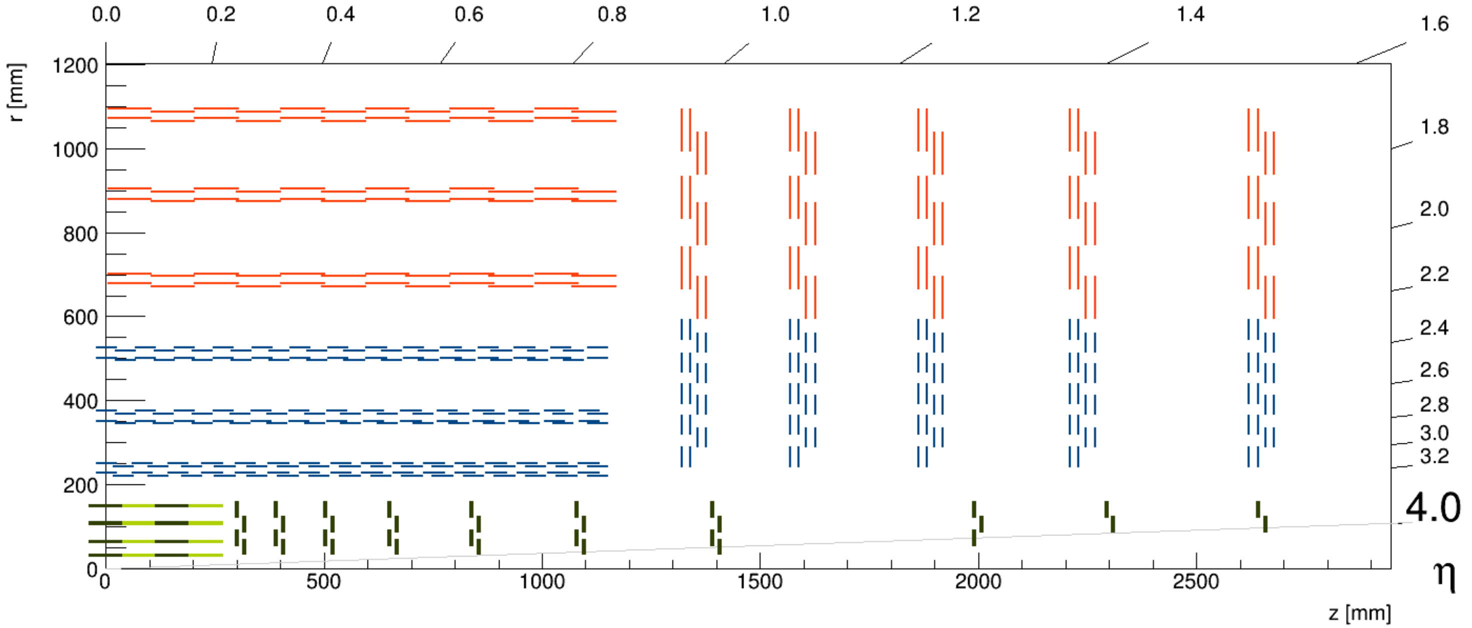
Tracker design

Pixel detector

- 4 layers at similar radii as in Phase-I - smaller pixel size & thin sensors for improved resolution
- Extended coverage to η = 3.8 to mitigate PU effect in forward region

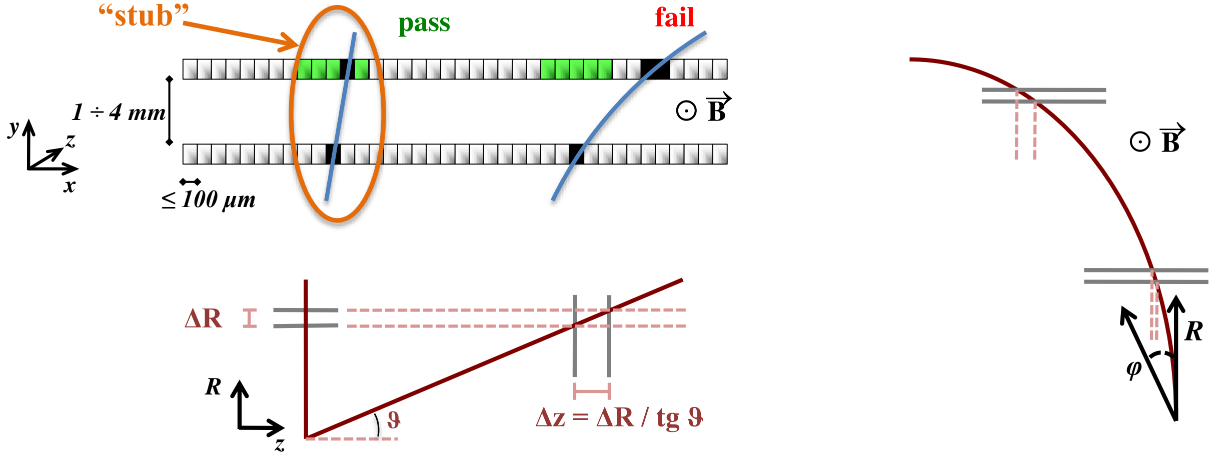
Outer Tracker

- 6(5) layers(disks) - strips ≃ 2.5-5 cm, 90-100 μm pitch optimized for cost
- 2 sensor modules for 40 MHz selective readout for L1-Trigger - macropixels in one sensor for z-measurement in 3 inner layers (PS modules)
- Light module design & mechanics - CO2 cooling (-30◦) - DC/DC powering



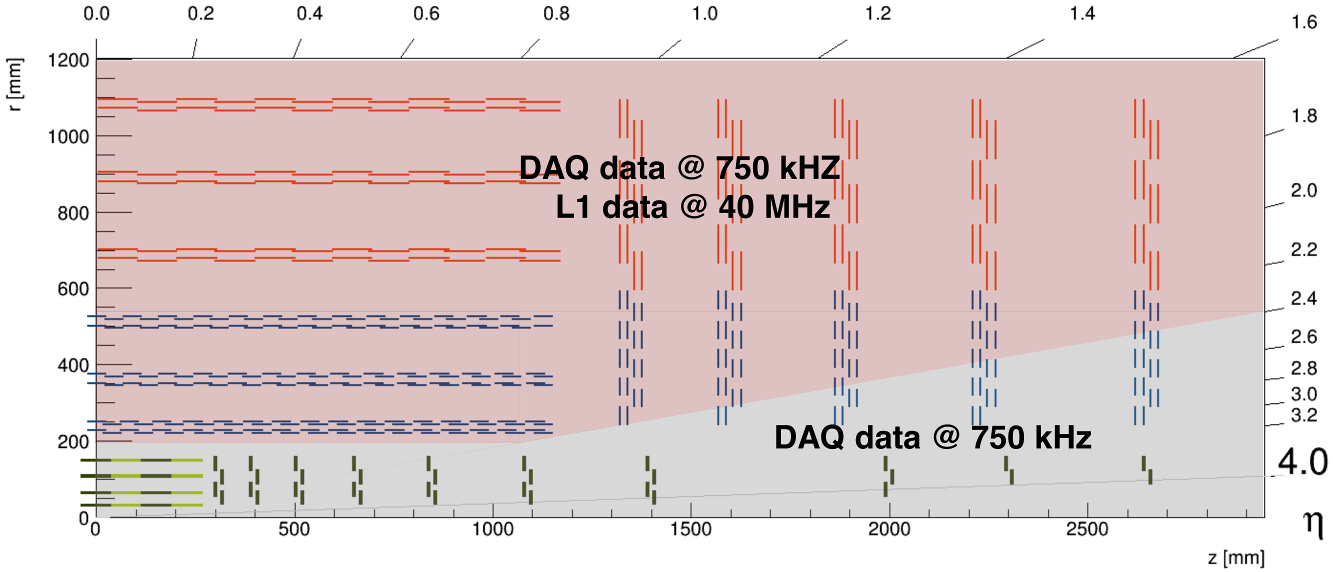
Tracker with trigger

Working principle of p_T modules



- Sensitivity to p_T from measurement of Δ(Rφ) over a given ΔR
  - For a given p_T, Δ(Rφ) increases with R
  - In the barrel, ΔR is given directly by the sensors spacing
  - In the end-cap, it depends on the location of the detector (tgϑ)
    - End-cap configuration typically requires wider spacing, and yields worse discrimination
- Optimize selection window and/or sensors spacing
  - To obtain, ideally, consistent p_T selection through the tracking volume
- The concept works down to a certain radius
  - 20÷25 cm with the CMS magnetic field and a realistic 100 μm pitch

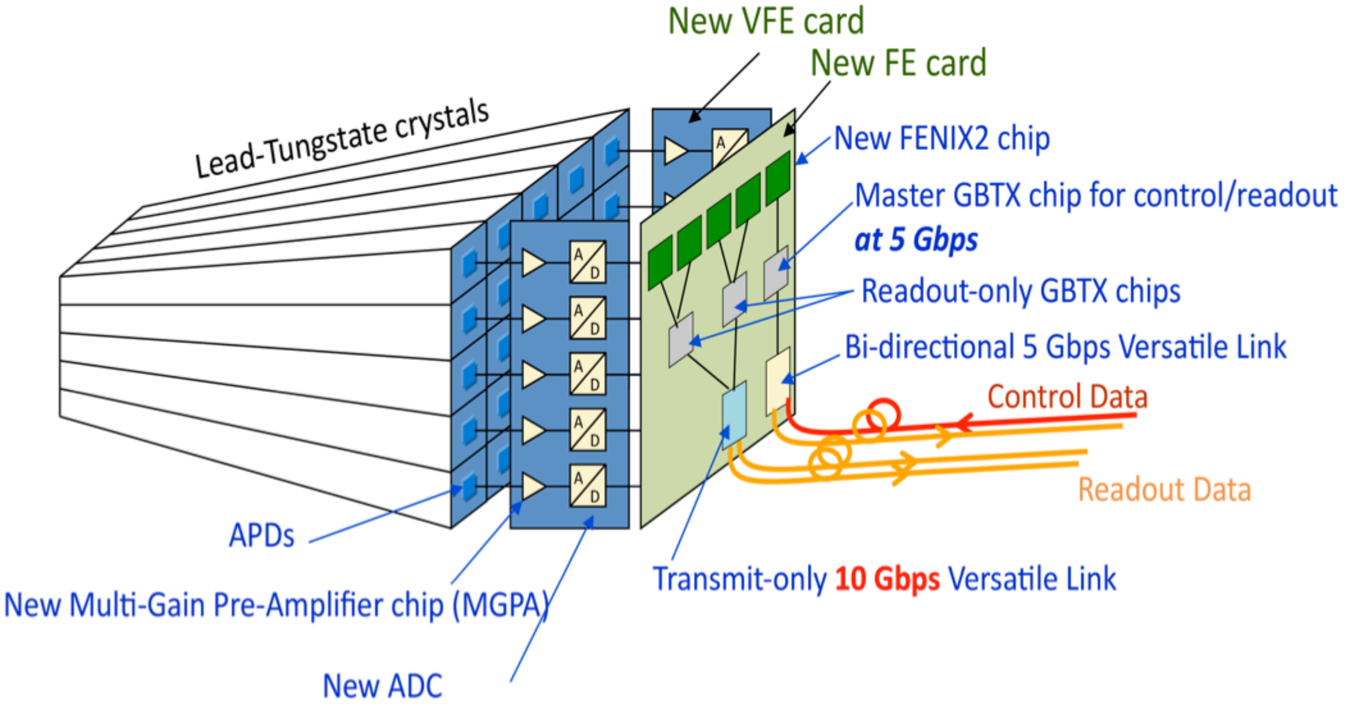
N.B. L1 tracking acceptance is limited at η~2.4



The L1 data output is disabled for modules located at low angle in the End Caps
(p_T discrimination insufficient to achieve reasonable bandwidth and stub purity)

Barrel electromagnetic calorimeter upgrade

- New very frontend - mitigate aging of APDs (noise) and spike background with shorter shaping time - opportunity to improve time resolution
- New frontend - L1-Trigger latency & rate - opportunity to provide crystal information at 40 MHz
- Operation at 8◦ - mitigate aging of APD (noise)


New VFE card
New FE card
Lead-Tungstate crystals
New FENIX2 chip
Master GBTX chip for control/readout at 5 Gbps
Readout-only GBTX chips
Bi-directional 5 Gbps Versatile Link
Control Data
Readout Data
A
D
APDs
New Multi-Gain Pre-Amplifier chip (MGPA)
New ADC
Transmit-only 10 Gbps Versatile Link

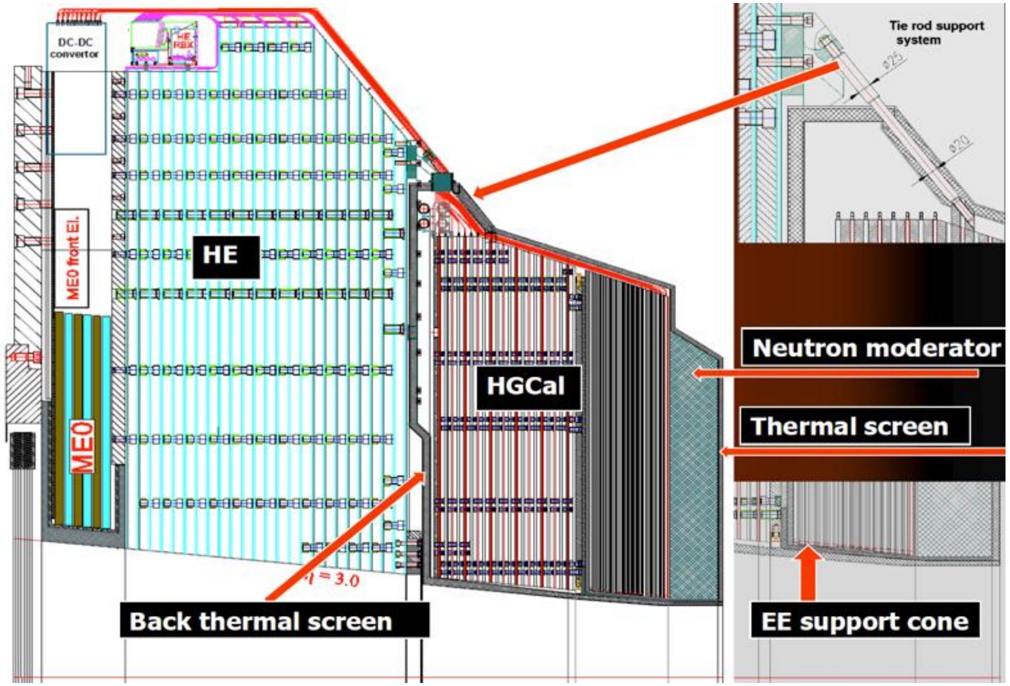
End cap Calorimeter design

High Granularity Calorimeter with silicon sensors - optimized on energy resolution - pad size selected for MIP S/N - high intrinsic precision timing ($\gtrsim$ 10 MIPs)

- Electromagnetic section (26 X_0, 1.5λ): 28 layers of Silicon-W/Cu absorber
- Front Hadronic section (3.5 λ): 12 layers of Silicon/Brass or Stainless Steel

Back Hadronic Calorimeter (HE) - similar design as present HE - radiation tolerant - increased granularity ($\simeq$ x4)

- BH (5 λ): 12 layers of Scintillator/Brass or Stainless Steel (2 depths readout)



Muon system reference design

New DT Minicrates - radiation tolerant - L1-Trigger rate - opportunity for full DT resolution at 40 MHz

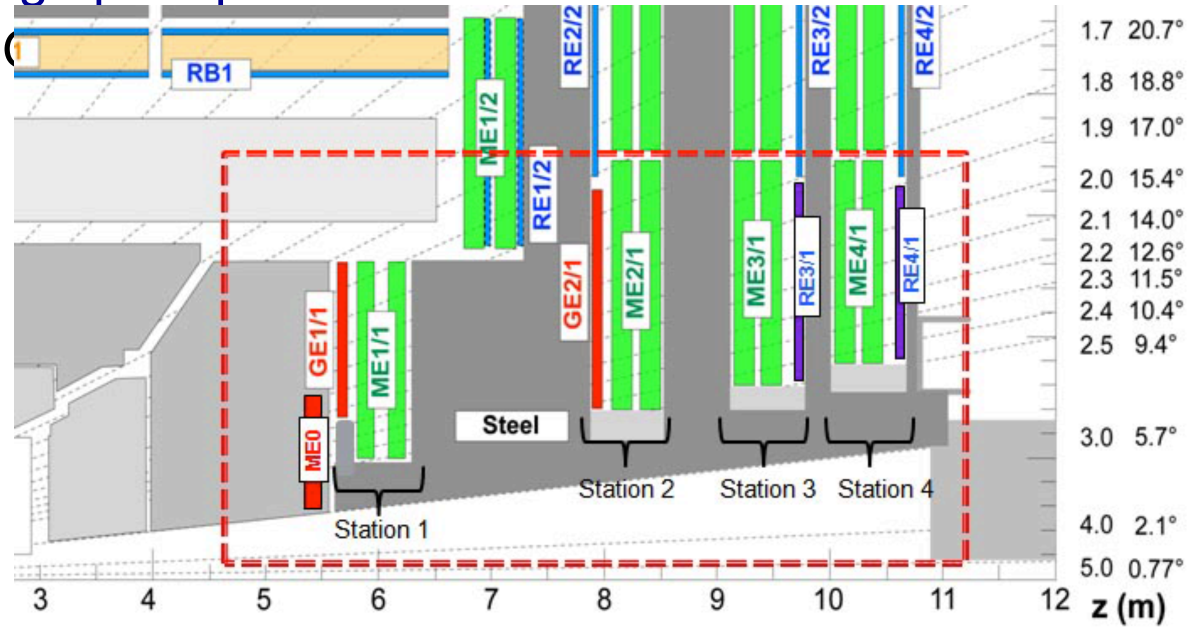
New CSC frontend in inner rings of station 2, 3 & 4 - L1-Trigger latency and rate

Complete RPC coverage in 1.5 < η < 2.4 - L1-Trigger performance and redundancy

- Pairs of triple GEM chambers in 2 first stations
- iRPC in stations 3 and 4

Muon tagging up to η = 3

- 6 triple G



L1-Trigger/HLT/DAQ upgrades

L1-Trigger - implement tracking information - latency of 12.5 µs - increase rate up to 750 kHz at 200 PU

- High BW and processing power boards in 2 layers - match detector information - produce Trigger objects

Trigger timing, throttling and control - allow trigger information to steer readout

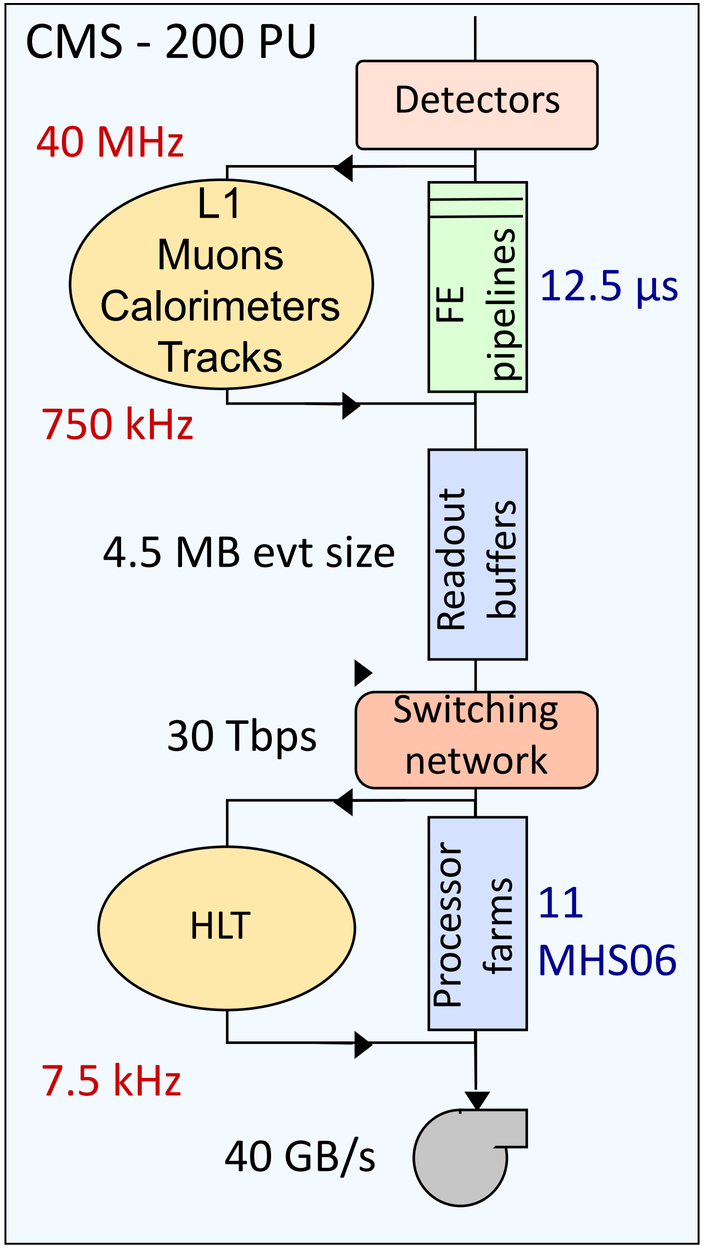
- High Band Width bi-directional links

DAQ - similar architecture as in current system (event builder, HLT and storage)

- Increased Band Width

High Level Trigger - similar reduction factor as present (1/100)

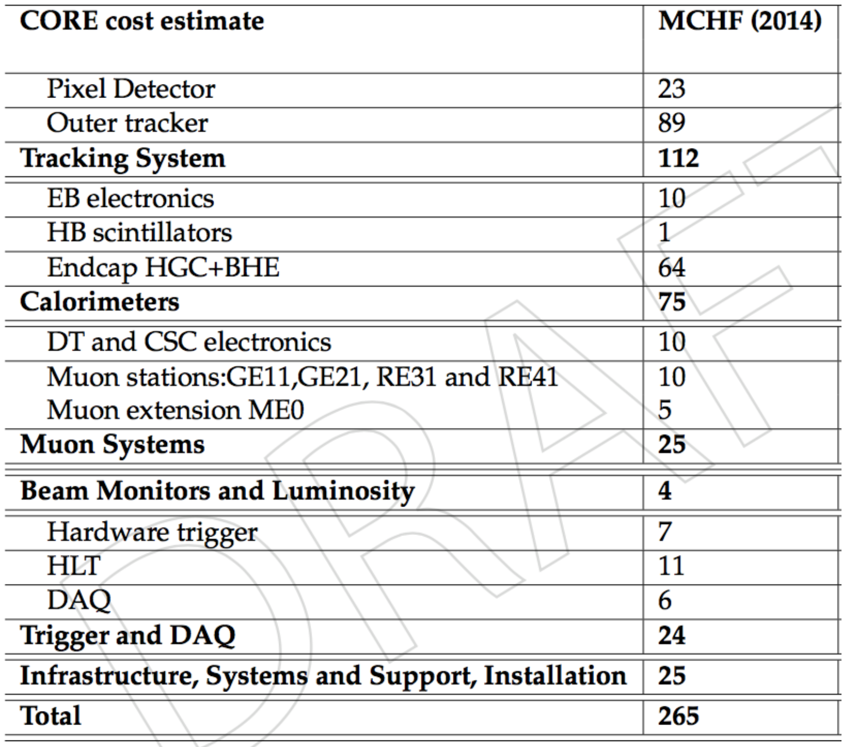
- Processing power scales as PU x L1-trigger rate



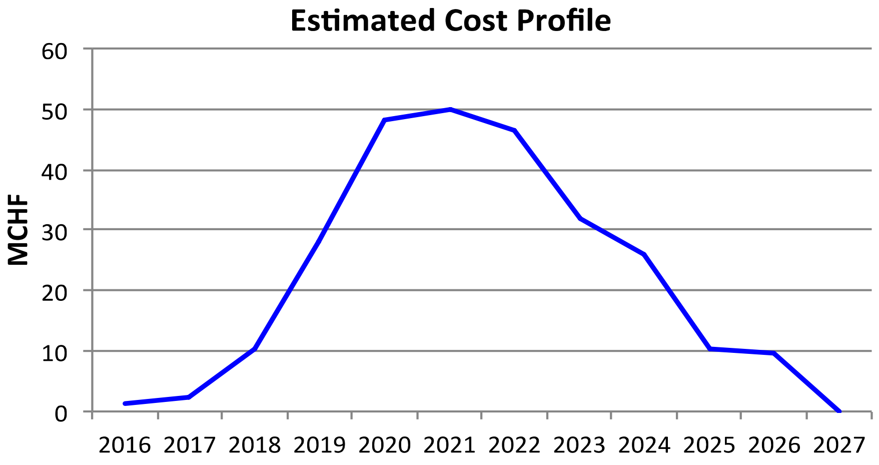
CMS “reference” upgrades cost estimates (Technical proposal)

The cost estimate for the CMS Phase II upgrade is 265 MCHF

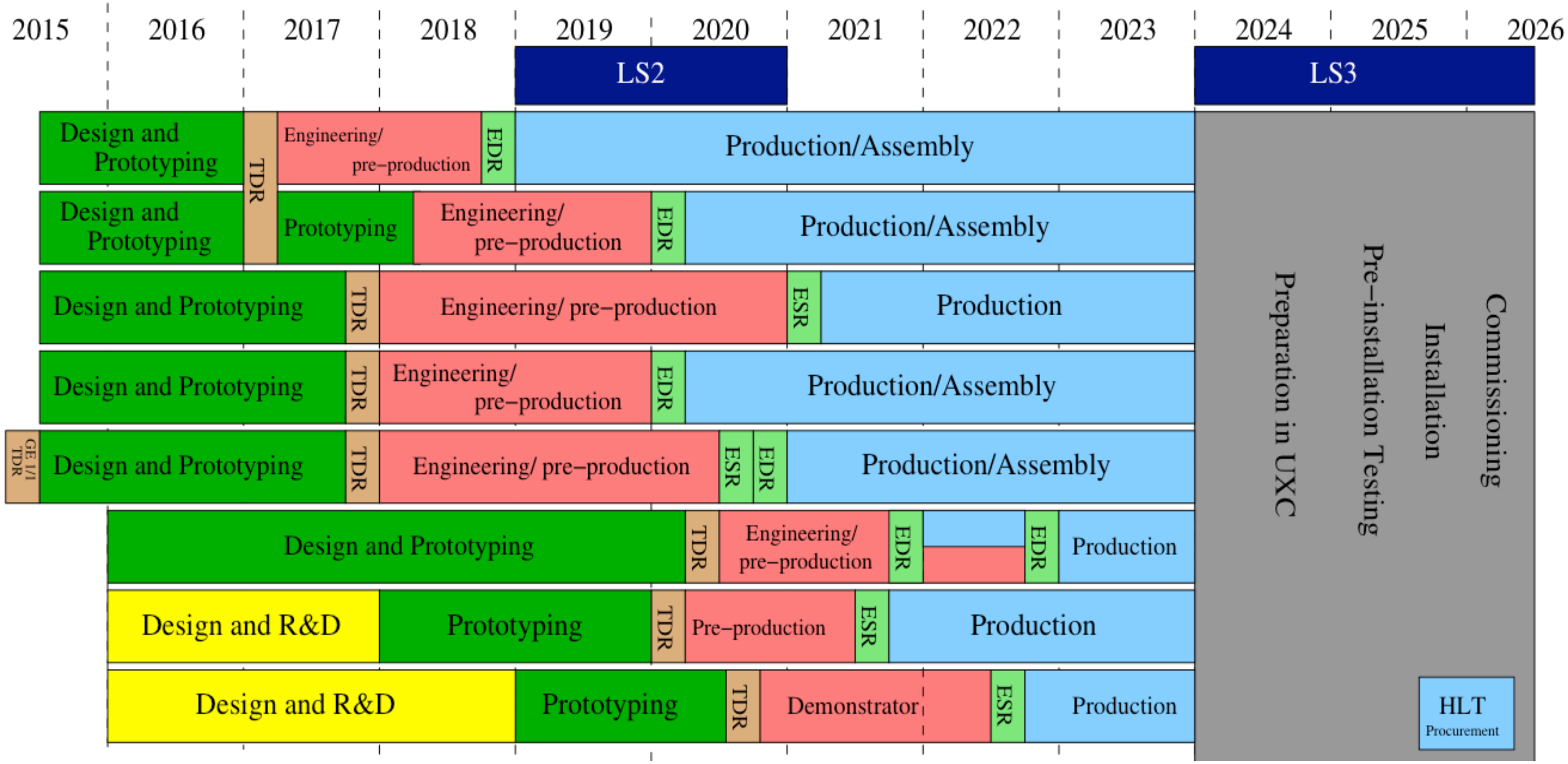
- It is established according to the CERN “CORE” rules
- For each upgrade project, estimates are made at the component or board level, mostly based on:
  - vendor information - this is the case for all major cost drivers
  - Scaling from original construction, LS1 work, or the ongoing Phase I upgrade
- Profile is based on major construction steps, updated to new LS3 schedule



CORE cost estimate	MCHF (2014)
Pixel Detector	23
Outer tracker	89
Tracking System	**112**
EB electronics	10
HB scintillators	1
Endcap HGC+BHE	64
Calorimeters	**75**
DT and CSC electronics	10
Muon stations:GE11,GE21, RE31 and RE41	10
Muon extension ME0	5
Muon Systems	**25**
Beam Monitors and Luminosity	**4**
Hardware trigger	7
HLT	11
DAQ	6
Trigger and DAQ	**24**
Infrastructure, Systems and Support, Installation	**25**
Total	**265**



Towards Technical Design Reports: calendar (i)



Additional avenues (not in scope document)

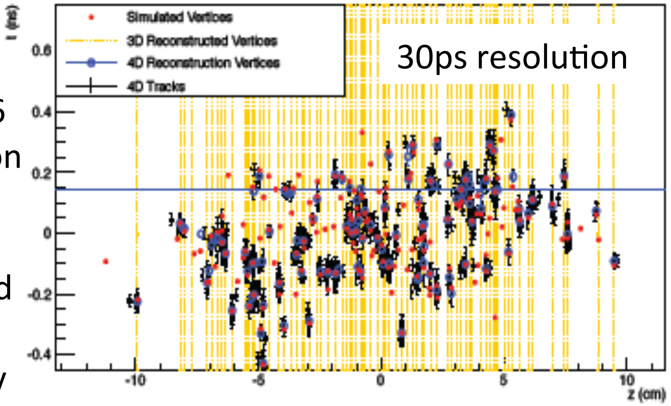
Towards Technical Design Reports

13

* Timing studies (not in TP scope)

https://indico.cern.ch/event/459736/contribution/3/attachments/1199316/1744449/FTRep_LindseyGray_03122015.pdf

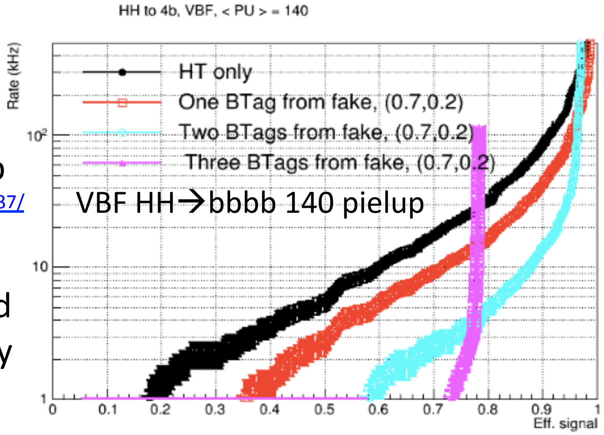
- Interim report expected next week, will be presented at the Physics week in February 2016
- Parameterization of EB and HGC timing precision and of representative MIP Timing Layer (TL)
- More work needed to evaluate performance individually and combined, for photons, jets and MET, target full studies by Sep. 2016
- Need careful evaluation of TL benefit, feasibility and implication for CMS integrity, target to review in fall 2016



30ps resolution

** Pixel in L1 trigger (not in TP scope)

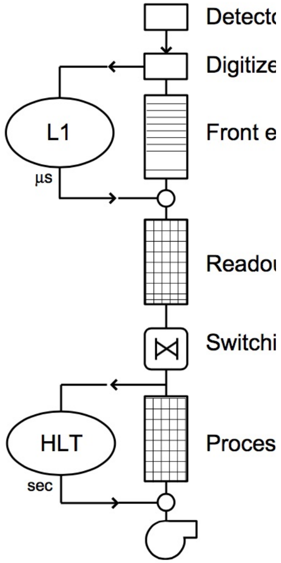
- Work in progress - presentation at Trigger workshop https://indico.cern.ch/event/455600/session/0/contribution/4/attachments/1185337/1718137/TTI_phase2trigger_11nov2015.pdf
- Need careful evaluation of benefit and feasibility since substantial implication on pixel chip design and trigger architecture - goal to converge on decision by fall 2016



VBF HH→bbbb 140 pielup

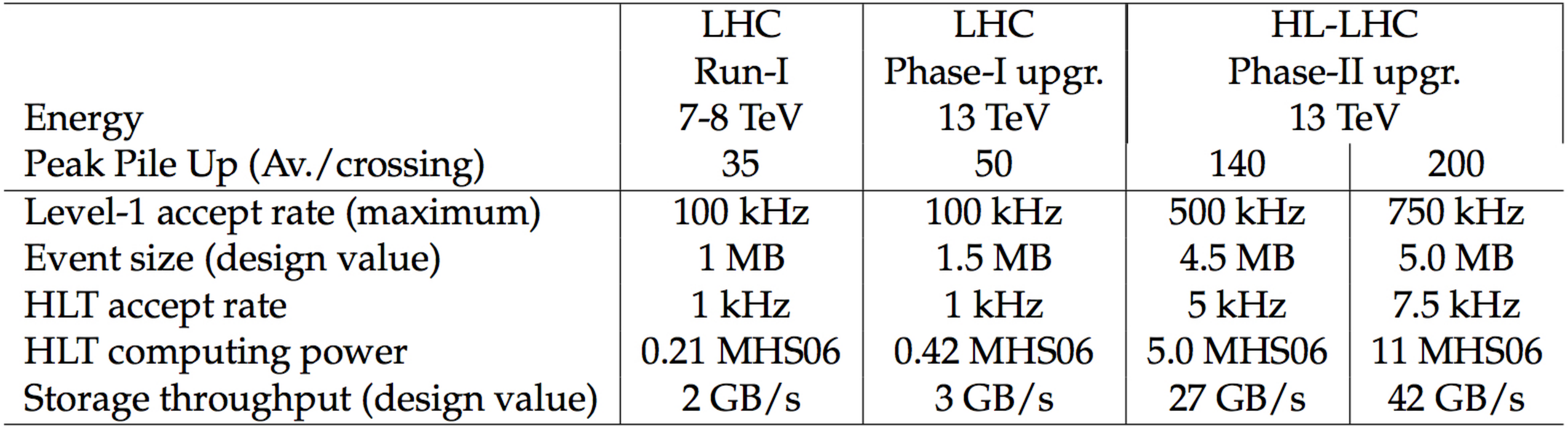
Trigger / DAQ / HLT

CMS L1 / DAQ / HLT



- Same two-level architecture as current system
  - L1 hardware trigger: 40 MHz clock driven, custom electronics
  - High Level Trigger (HLT): event driven, COTS computing nodes

Table 7.1: DAQ/HLT system parameters.



	LHC	LHC	HL-LHC	
	Run-I	Phase-I upgr.	Phase-II upgr.	
Energy	7-8 TeV	13 TeV	13 TeV	
Peak Pile Up (Av./crossing)	35	50	140	200
Level-1 accept rate (maximum)	100 kHz	100 kHz	500 kHz	750 kHz
Event size (design value)	1 MB	1.5 MB	4.5 MB	5.0 MB
HLT accept rate	1 kHz	1 kHz	5 kHz	7.5 kHz
HLT computing power	0.21 MHS06	0.42 MHS06	5.0 MHS06	11 MHS06
Storage throughput (design value)	2 GB/s	3 GB/s	27 GB/s	42 GB/s

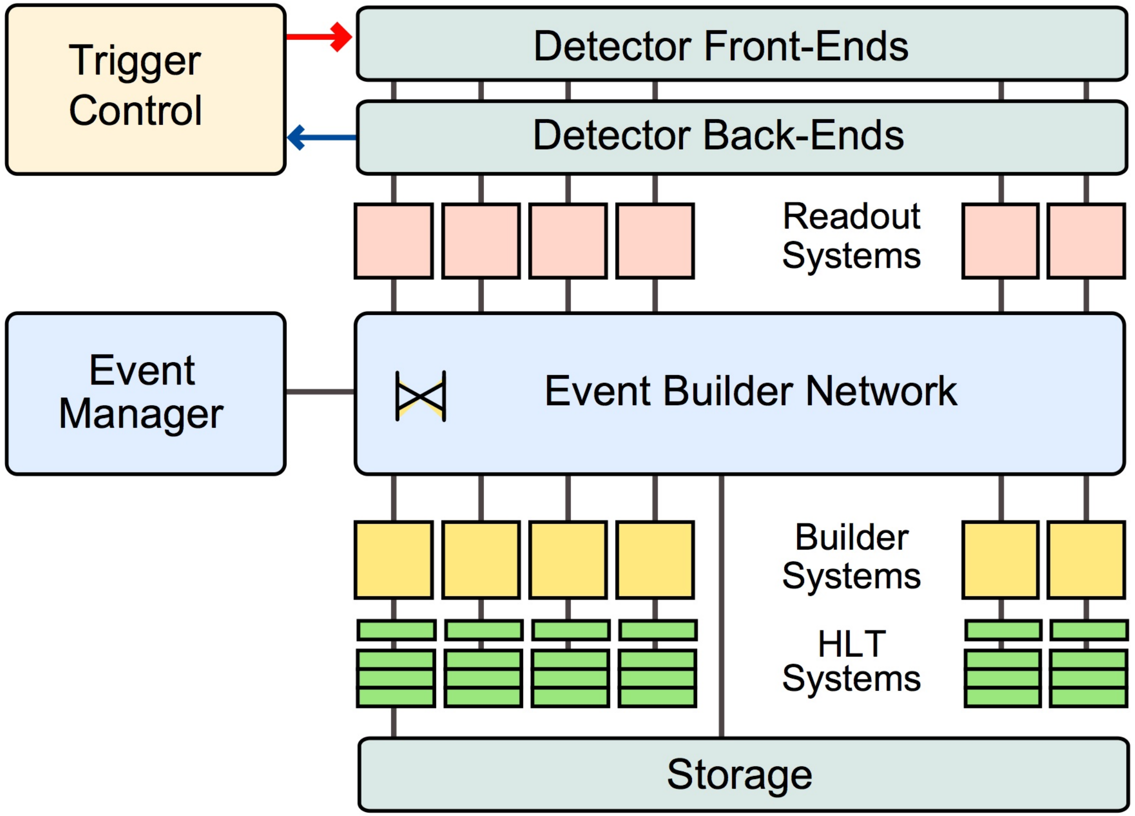
CMS Phase-II detector R/O parameters

Sub-det	# links on- 2 off-detector	Type (Gbps)	use	Data reduction	Event size (Mbyte)	#DAQ links (100 GBps)
TK-outer	13 k 2 k	GBT (4 G) GBT (9 G)	DAQ + Trig 20% + 80%	On-det	05. – 0.6	100
TK-pixel	1 k	lpGBT (9 G)	DAQ	On-det	0.7 – 1.0	200
ECAL-barrel	12 k	GBT (3 G)	streaming	Off-det	1.2	200
HCAL	2 k	GBT (3 G)	streaming	Off-det	0.2	40
HGCAL	9 k	lpGBT(9 G)	Streaming?	On-det?	1.2	200
Muons DT	6 k	GBT (3 G)	streaming	Off-det	0.1	20
Muons CSC	1 k	GBT (3 G)	DAQ+Trig 50%+50%	Off-det	0.1	20
Trigger						20
EVB					4.2-4.6	800

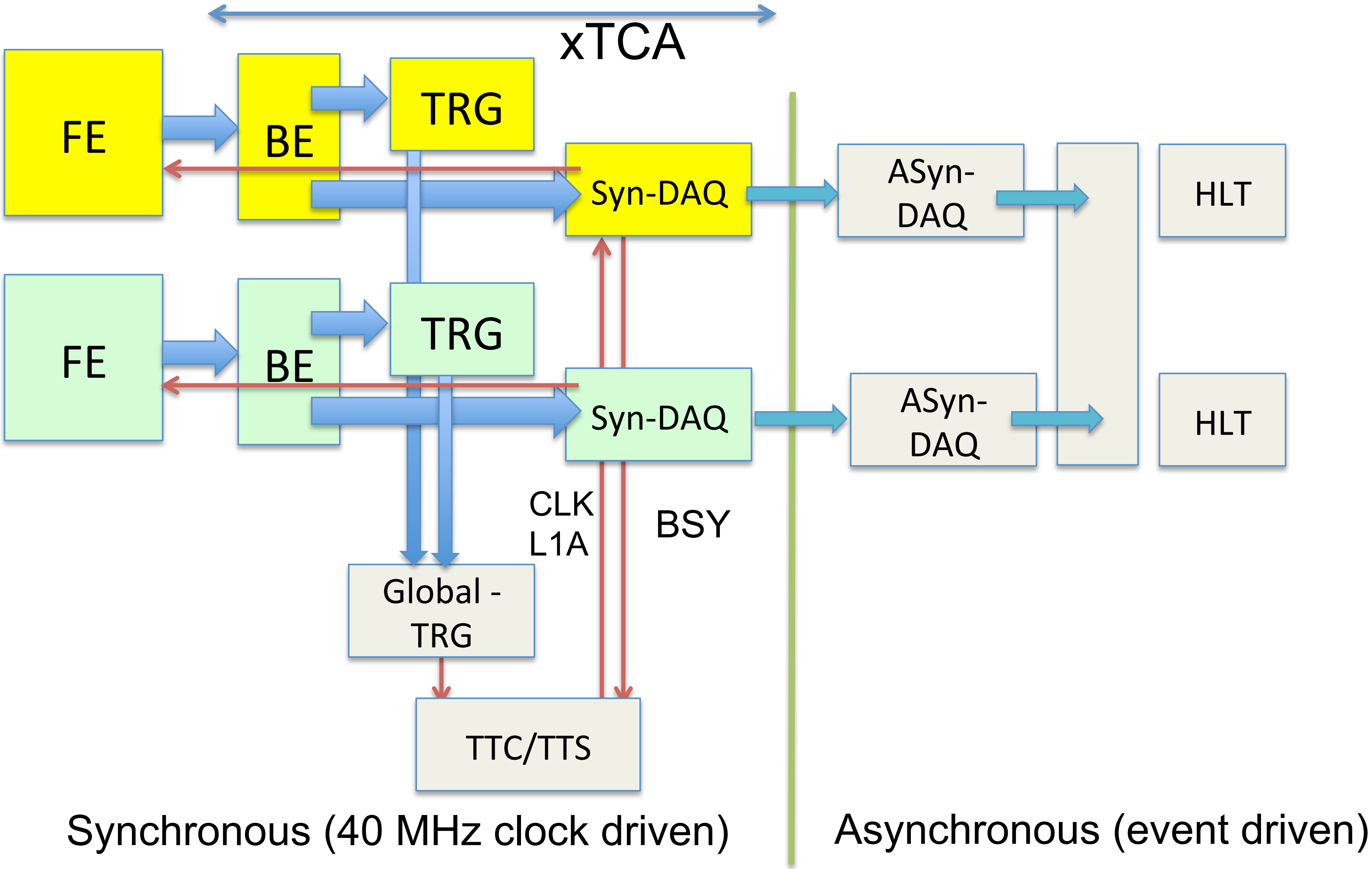
About 50 k “GBT” links

DAQ to 1st order

- All similar



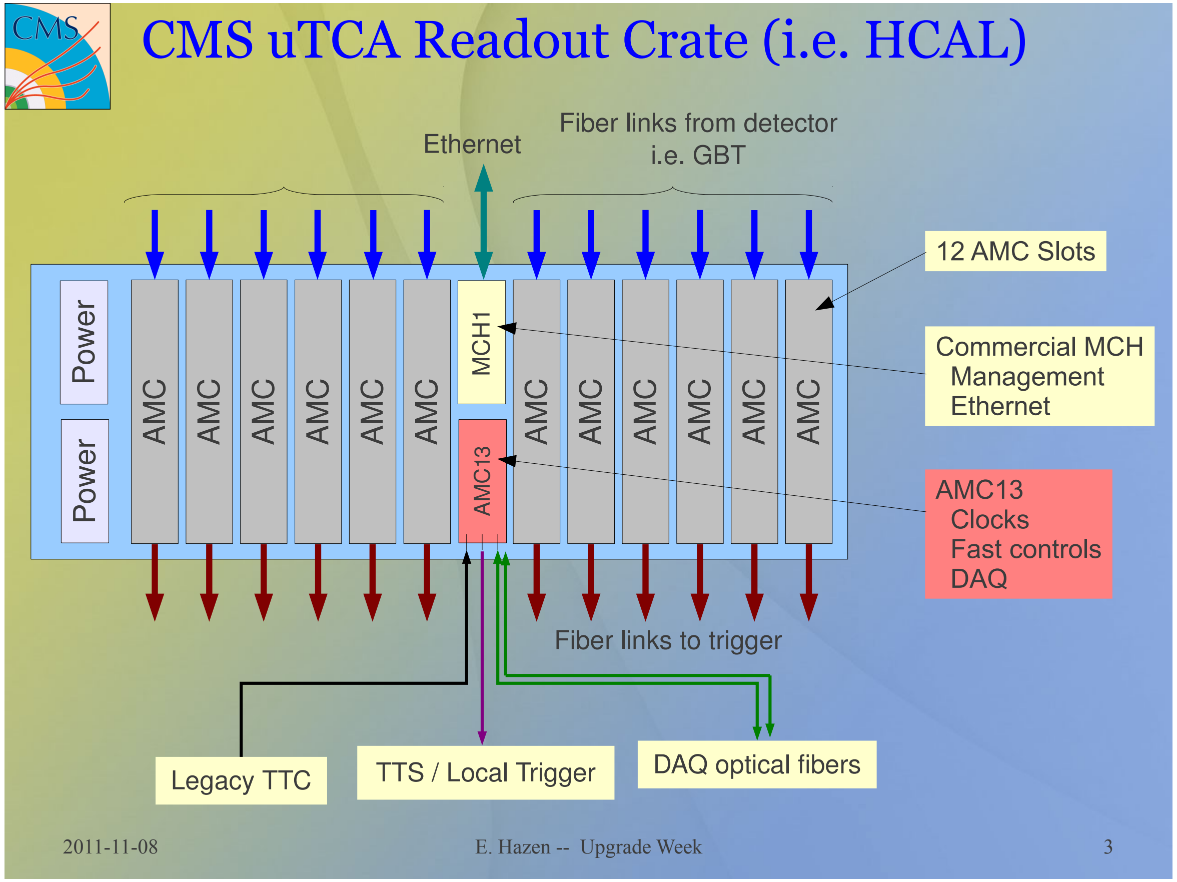
CMS Trigger DAQ



High level design – readout unit

- Synchronous domain (40 MHz clock driven)
  - Towards front-end electronics via serial point-to-point links 5-10 Gbps (mostly GBT)
  - Receive DAQ data
  - TTC and throttle
  - Optional Send / receive "DCS"
  - Optional Transmit to Trigger electronics (same or separate links as DAQ)
  - forward trigger selected data in case of "streaming" DAQ
- Asynchronous domain (event driven)
  - Data aggregation (concentrator)
  - Buffer
  - Sub-detector specific processing (data reduction, processing, especially suited for channel-by-channel processing)
  - Protocol conversion, transmit / receive standard commercial network

CMS phase-I upgrade



CMS uTCA Readout Crate (i.e. HCAL)

Ethernet

Fiber links from detector
i.e. GBT

12 AMC Slots

Commercial MCH
Management
Ethernet

AMC13
Clocks
Fast controls
DAQ

Power

Power

AMC AMC AMC AMC AMC AMC

MCH1

AMC13

AMC AMC AMC AMC AMC AMC

Fiber links to trigger

Legacy TTC

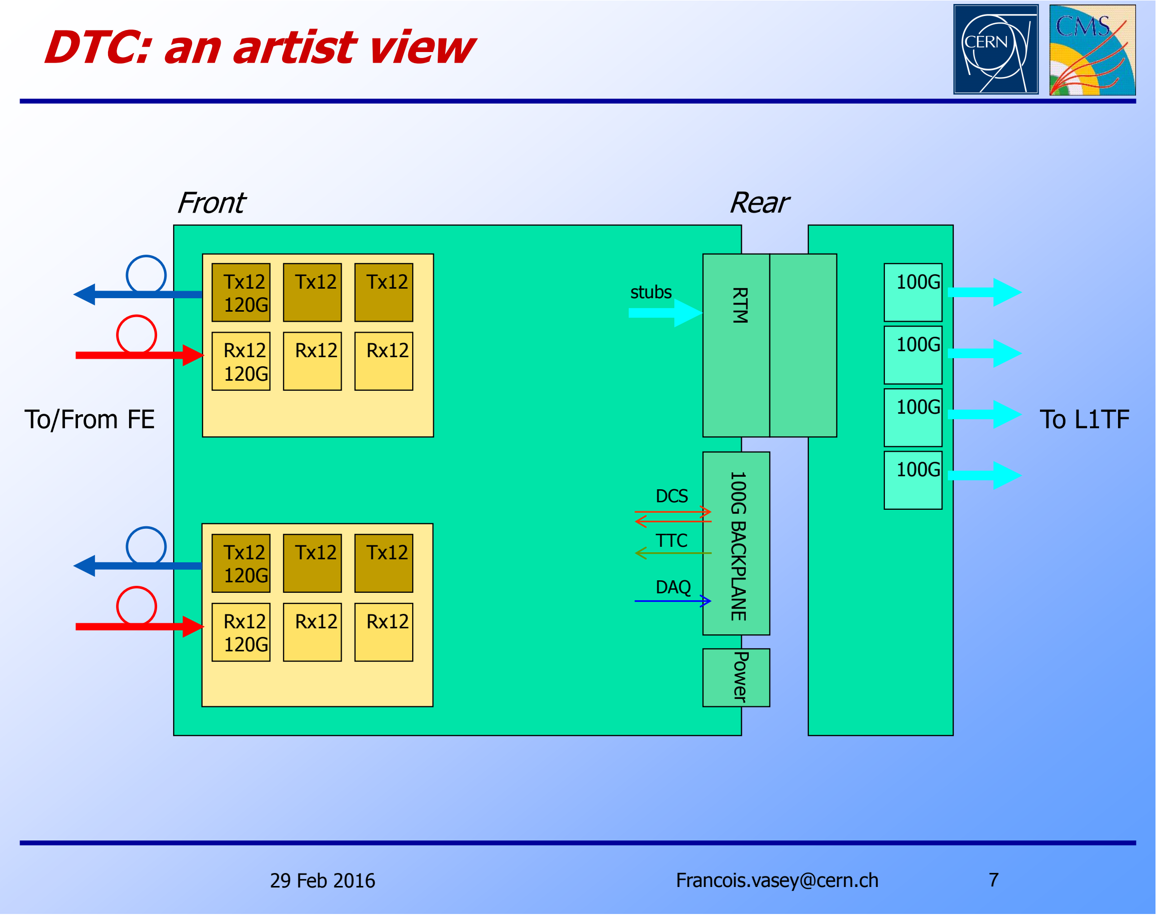
TTS / Local Trigger

DAQ optical fibers

2011-11-08 E. Hazen -- Upgrade Week 3

- Bifurcate stream coming from FE to Trigger
- Common Control+Timing HUB and DAQ concentrator (AMC13)
- In L1 upgrade re-use of various AMC cards (plug-in FW)

CMS Phase-II read-out example (TK)



DTC: an artist view

29 Feb 2016 Francois.vasey@cern.ch 7

- “DTC” has 72 GBT links (up to 10 Gbps)
- Full tracker will require 256 “DTC” cards
- Note: BW for Trigger/DAQ is ~80/20 %

CMS ideas for DAQ - RU

- Synchronous part done in xTCA
  - Located in Underground
- Asynchronous part of DAQ, some options
  - Point-to-point link to intermediate cDAQ “Hub” card
    - Mezzanine on leaf card
    - From leaf card via xTCA backplane to other slot
      - when concentration useful
  - Link and Protocol
    - Custom or Ethernet L2 or TCP when sufficient memory resources
    - Back pressure
  - cDAQ RU (preferably on surface)
    - PC with a commercial NIC (Ethernet L2), or custom card
    - Concentrator, buffer
    - Protocol converter to EVB network (Ethernet or HPC)
  - With 10 Gbps lanes (~200 m over MM) need 4000 fibre-pairs USC-SCX

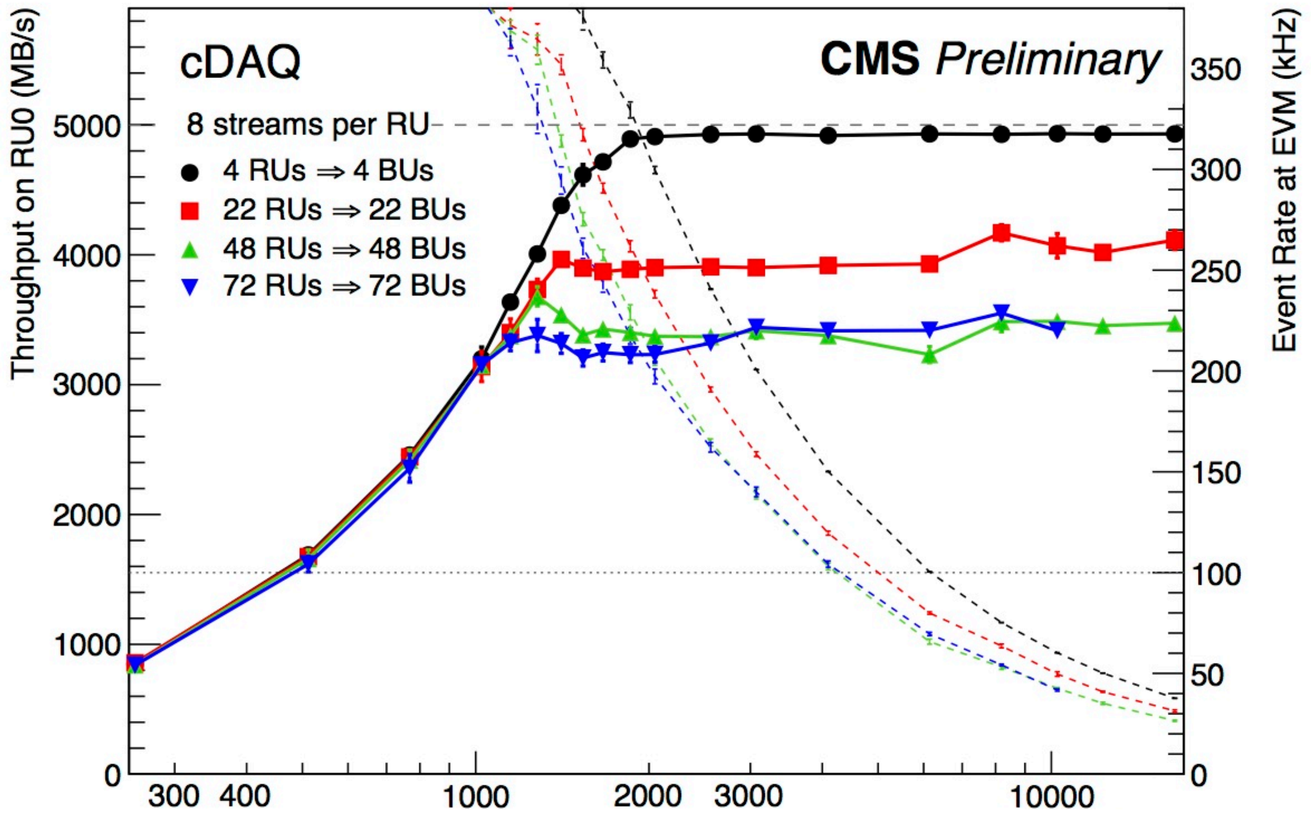
EVB design / implementation choices

- ROI based or **full event building**
  - Large implications for EVB – HLT interface and HLT framework
- Lossy (incomplete events) or **Lossless EVB**
  - Lossless
- Network technology
- Protocols
- Event Manager to control EVB process and optionally throttle trigger
- Effective throughput vs bi-section bandwidth
  - "folded" EVB to use bi-directional links
  - Traffic shaping ?

EVB network technology

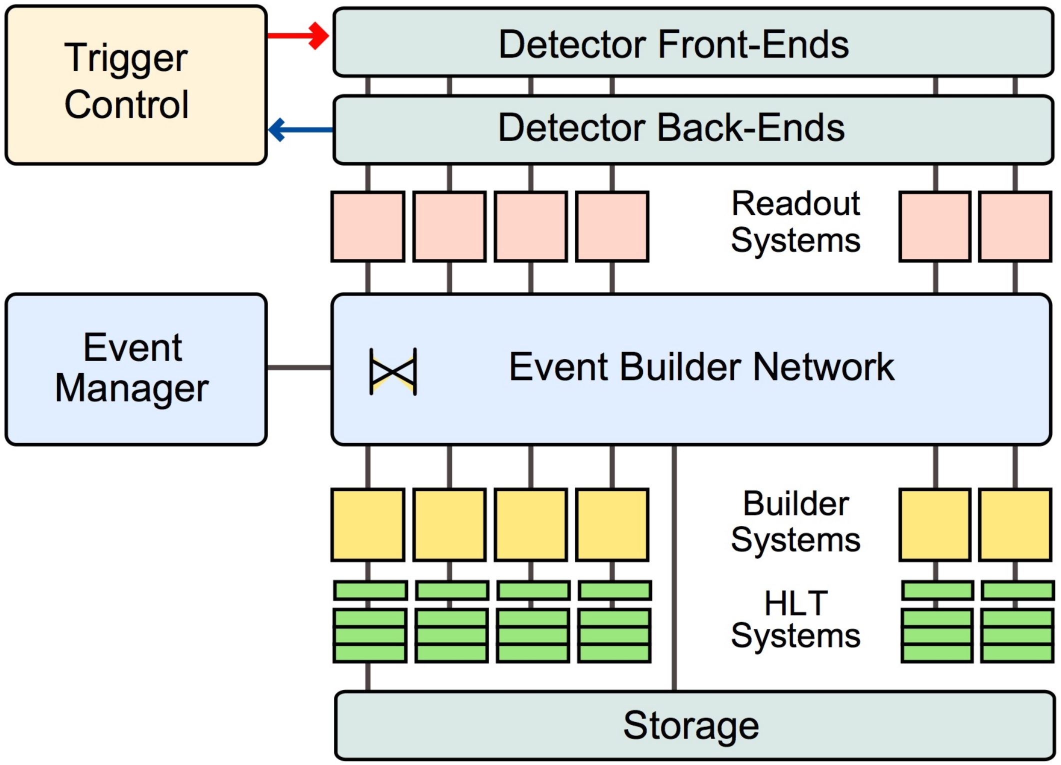
- Ethernet
  - Switches with or without “deep buffers”
  - L2 or TCP/IP ?
- HPC: Infiniband, Omnipath, ..
  - Trend for closer integration with CPU (Omnipath)
- 100 Gbps now (10x10 Gbps lanes or 4x25 Gbps lanes)
  - Likely 200 Gbps (4x50 Gbps lanes) at LS3
  - Severe length restriction on MM fibres (100 m for 25 Gbps lanes)
- Interfacing to FPGA based Read-Out Unit
  - Ethernet:
    - Layer2 ok, but transmission unreliable
    - (reduced) TCP/IP possible in FPGA, but needs memory for buffering
  - HPC: difficult

EVB scaling and switch technology



- CMS today Infiniband FDR (56 Gbps) EVB (switches w/o buffers)
- Scaling NxN: Efficiency is about 50% of line BW

CMS EVB baseline

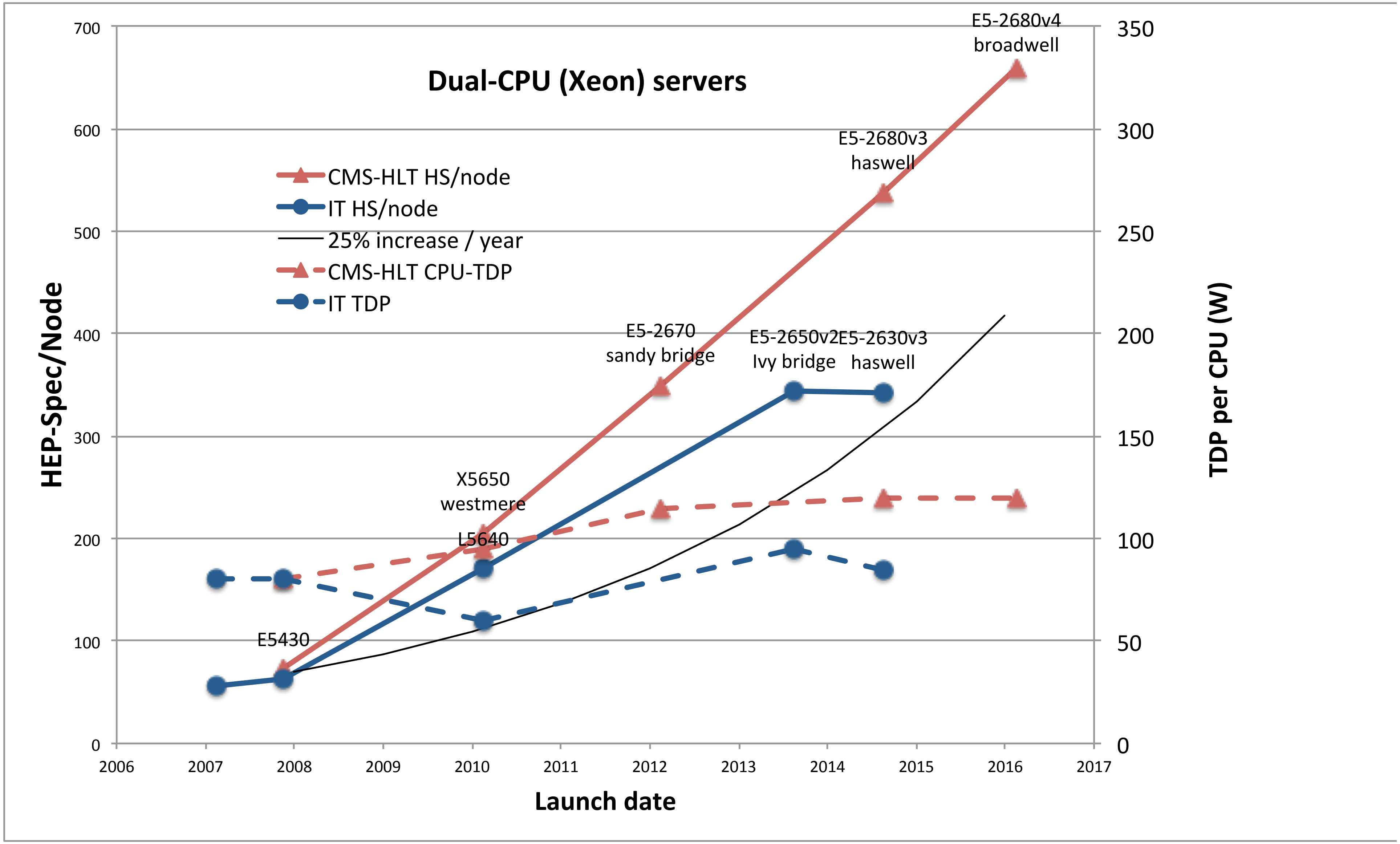


- 4 TB/s throughput
- Full event building
- Assume 400 servers, 200 Gbps links, ~30% efficiency

Extrapolation of HLT farm needs

- Guestimate (TP and scoping doc)
  - based on
    - Current detector,
    - PU dependence,
    - possible gain with using L1 Track trigger
  - Result:
    - PU = 140 / 200 and HLT input 500 / 750 kHz: need ~ 5.0 / 11.0 MHS06
- To fix ideas:
  - Need 11 MHS06 in 2026 (Atlas/CMS has ~ 0.7/0.5 MHS now)
    - Ignore that likely LHC-HL will not start with PU=200
  - Compare LHCb: need 3.3 MHS06 in 2021
- Extrapolated cost
  - WLCG [2014] 25% improvement per year ~1 $/HS06

Dual-Xeon server evolution

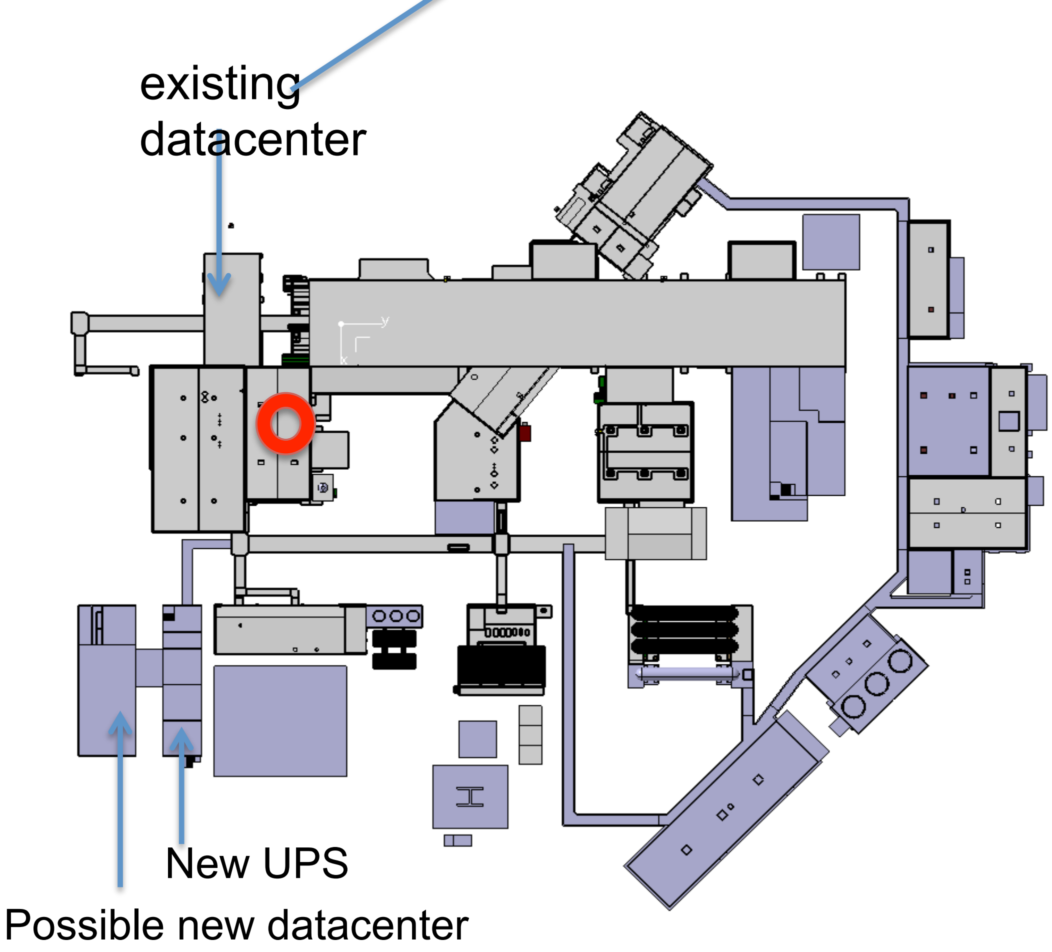


Extrapolation for HLT farm (dual Xeon based)

- To fix ideas: Need ~11 MHS for HLT in 2027
- Observed (~ 2007 – 2015) dual-Xeon servers
  - Start point Q1-2016: server with 0.66 kHS and 0.35 kW and X kCHF

Assumed perf. Increase of server	Exponential 25% / year [WLCG 2014]	Exponential 12.5 % / year	Linear 73 HS/year
11 years	12	3.7	2.2
#servers in Q1-27	1431	4562	7860
Total power Q1-27	0.5 MW	1.6 MW	3 MW

- Comments
  - Ignore installed base from run-3
  - Current CMS datacenter ~1 MW cooling looks insufficient

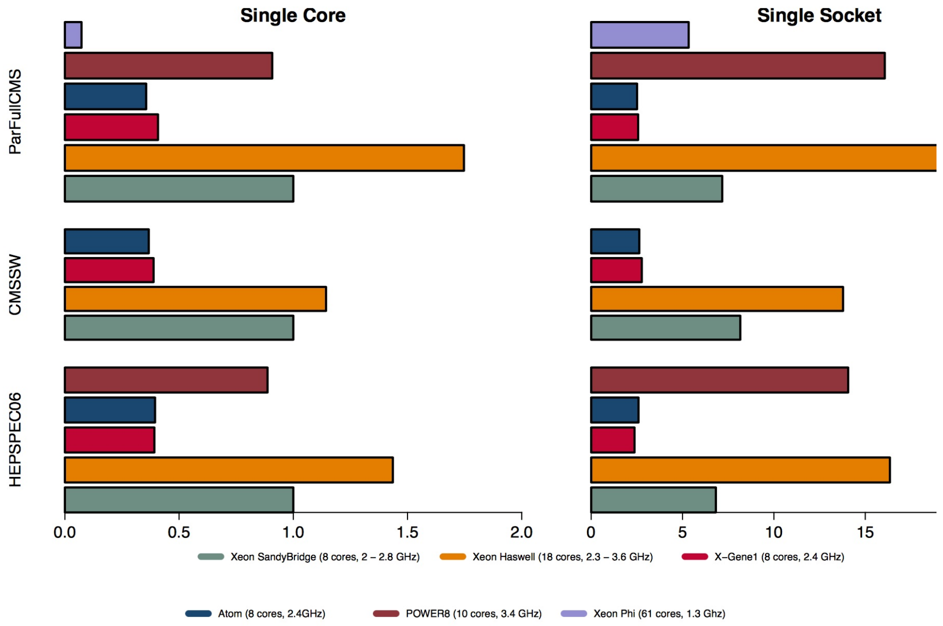


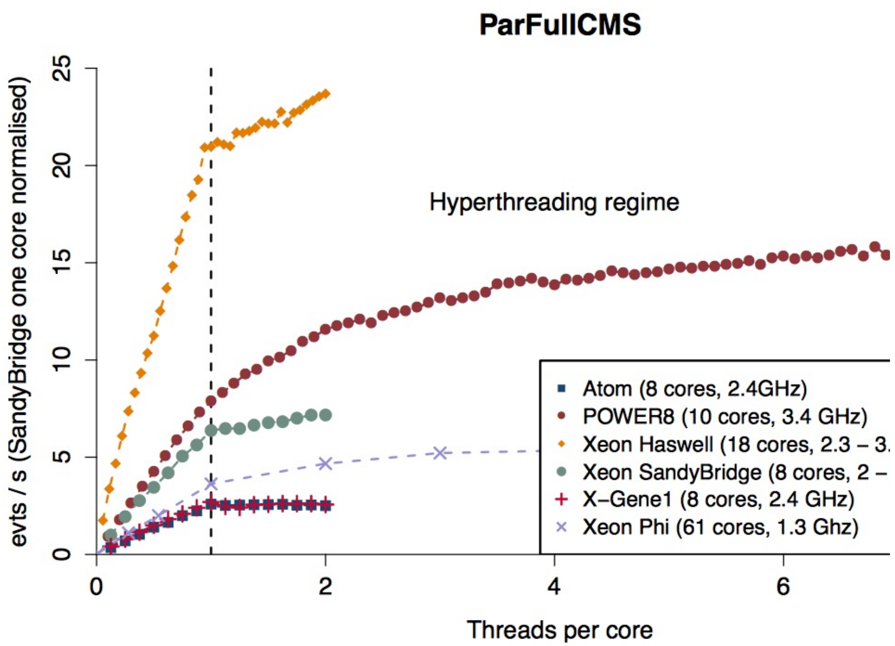
- Existing datacenter: racks, cooling to be redone if >1 MW

Specialized (co-) processors

- CMS experience, so far
  - CMS “standard” code, alternative platforms not competitive (price/perf) compared to Xeon (See http://arxiv.org/pdf/1510.03676v1.pdf)
- Very active area of research
  - Eg “Connecting the DOTS 2016” for tracking
- Concern
  - (fine grain) parallelization and vectorization
    - Requires extensive re-engineering of the code
  - Portability and long-term maintainability
- HLT farm versus worldwide Tier-n centers
  - HLT farm is under control of experiment, so easier to deploy alternative platforms

CMS standard offline code on specialized processors



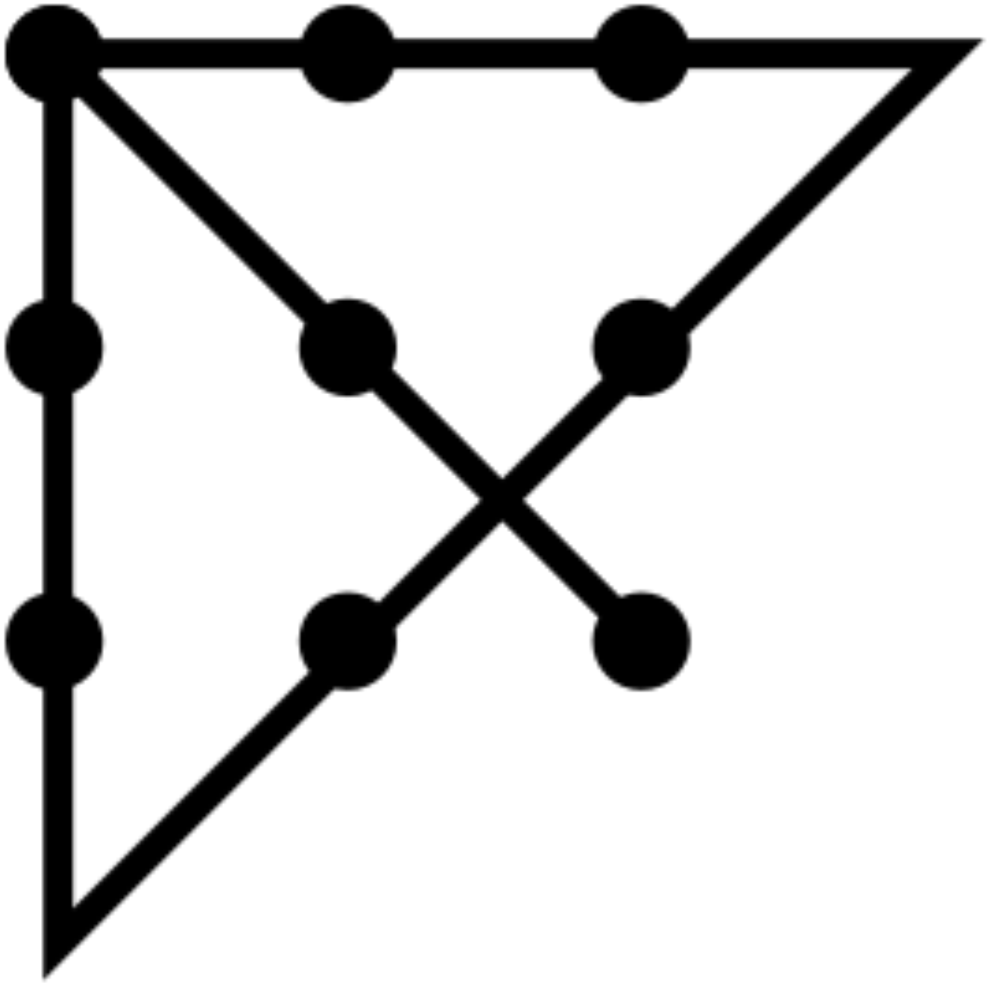


- Not competitive compared to dual-xeon (2015)
  - See http://arxiv.org/pdf/1510.03676v1.pdf

Storage

- CMS
  - 40 GB/s
  - Could use Cluster File System as now, or something else

The boss wants some new ideas
*it's time to think **outside the box**.*

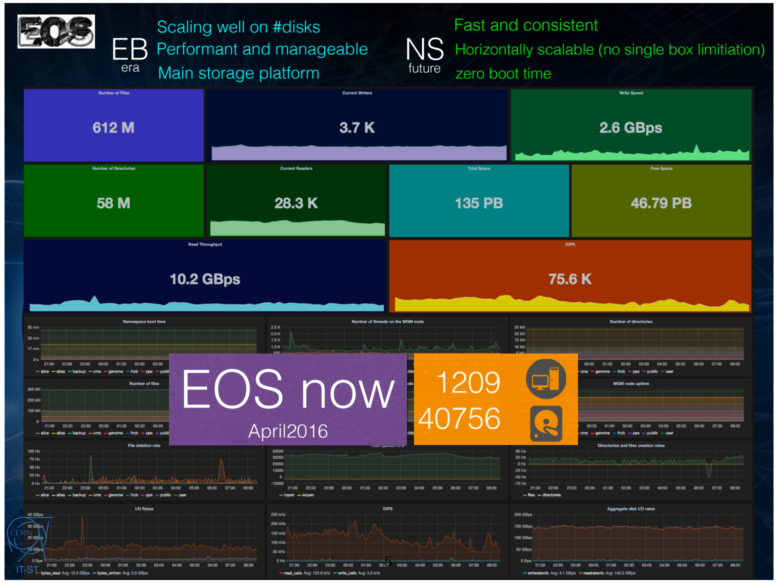


Remote data centre for HLT (I)

- Observation
  - Need new infrastructure on site for data centre >1 MW
  - 25 Gbps lanes over MM fibre go only up to 100 m, next hop is 10 km
- Suppose CERN provided a high bandwidth link from Pt5 to a central data-center (eg Prevessin)
  - 5 MB@750kHz = 4 TB/s, need 50 Tbps
- EVB at Pt5
  - Built full events, for convenience built lumi-sections (~10 s)
  - Store EVB output in filesystem (as today done on ram-disk)
  - Plenty of Space and BW with 3D-xpoint or SSDs
- Transfer files to remote datacenter to run HLT

Remote data centre for HLT (II)

- HLT run by offline leveraging IT services



EOS

EB era
Scaling well on #disks
Performant and manageable
Main storage platform

NS future
Fast and consistent
Horizontally scalable (no single box limitiation)
zero boot time

EOS now
April2016

  - “EOS++” on SDD for buffering
  - Condor batch for running jobs
  - Clean separation online / offline
  - Flexibility for waiting for calibration, multi-pass, etc ..

- Note:
  - For run 2,3: need 200 GB/s so 40x40 Gbps links (have 4x40 Gbps now)

Idea for 40 MHz “DAQ scouting”

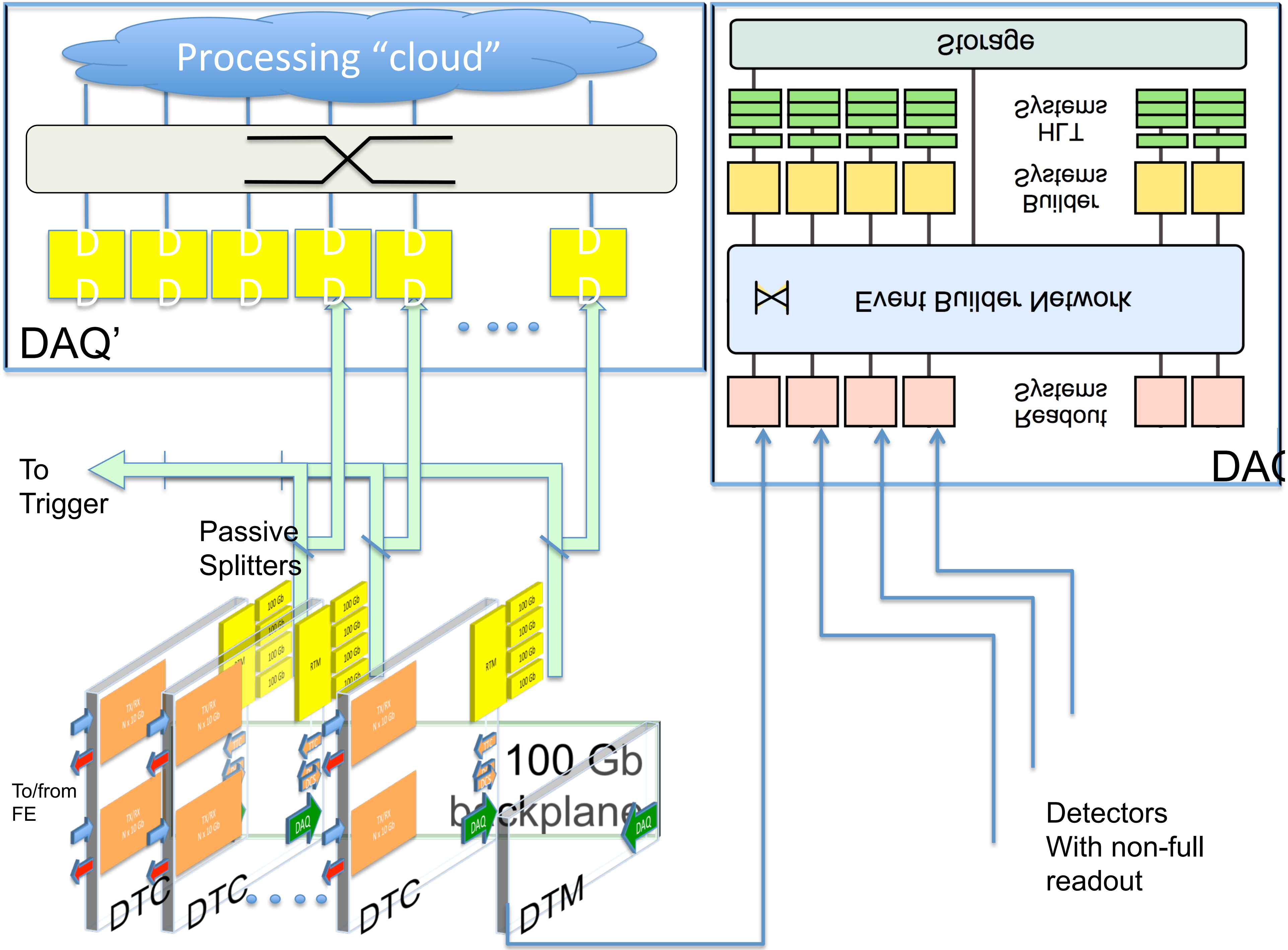
(E.Meschi)

HLT scouting in CMS **Today**

- At HLT look for processes
  - Characteristics
    - Difficult to select by L1 HW trigger
    - High rate
    - Common signatures (will not find it with pre-scaled “randoms”)
  - Too high rate to record at HLT output for offline processing
  - Do “analysis” inside HLT
    - Produce histogram and/or save minimal reconstructed info
- Once you observe a signal in a window (for example bump in invariant mass of jets) select the events in that window at HLT in the standard way
- Discovery !

Parasitic DAQ for scouting at 40 MHz at HL-LHC

- Look for processes
  - Characteristics
    - Difficult to select by L1 HW trigger
    - High rate
    - Common signatures
  - Do “analysis” inside parasitic DAQ
- Once you observe a signal with a certain signature select the events with that signature with standard L1/HLT
- Caveat
  - No full detector: Only tracker stubs, no pixel
  - Limited overall CPU available


Processing "cloud"
D D
D D
D D
D D
D D
D D
DAQ'
To Trigger
Passive Splitters
100 Gb
RTM
TX/RX N x 10 Gb
DAQ
100 Gb backplane
To/from FE
DTC
DTC
DTC
DTM
Storage
HLT Systems
Builder Systems
Event Builder Network
Readout Systems
DAQ
Detectors With non-full readout

Fantasy Server PC in 2027 and beyond

- I/O
  - PCIe gen4 15.7 Gbps per lane
  - A 16-lane gen-4 board can do 250 Gbps
- Memory
  - 3D Xpoint with high BW and capacity (~TB)
- Processing
  - Dual socket many core
  - Co-processing or embedded FPGA, GPU
- Config
  - 3 cards in a PC for input (receiving 40 MHz data stream out of ATCA)
  - 1 card for connection to processing cloud (ROI based)

Some numbers ..

- Config fantasy server PC of 2027+
  - 3 cards in a PC for input (receiving 40 MHz data stream out of ATCA)
  - 750 Gbps
  - 1 card for connection to processing cloud (ROI based)
- Size of system
  - Tracker: 128 servers for 256 DTC cards
  - Calorimeters: 500 servers to capture 2.5 MB @ 40 MHz
- Processing cloud is ROI based
  - No excessive BW requirements

CONCLUSION

CMS DAQ for LH-HLC

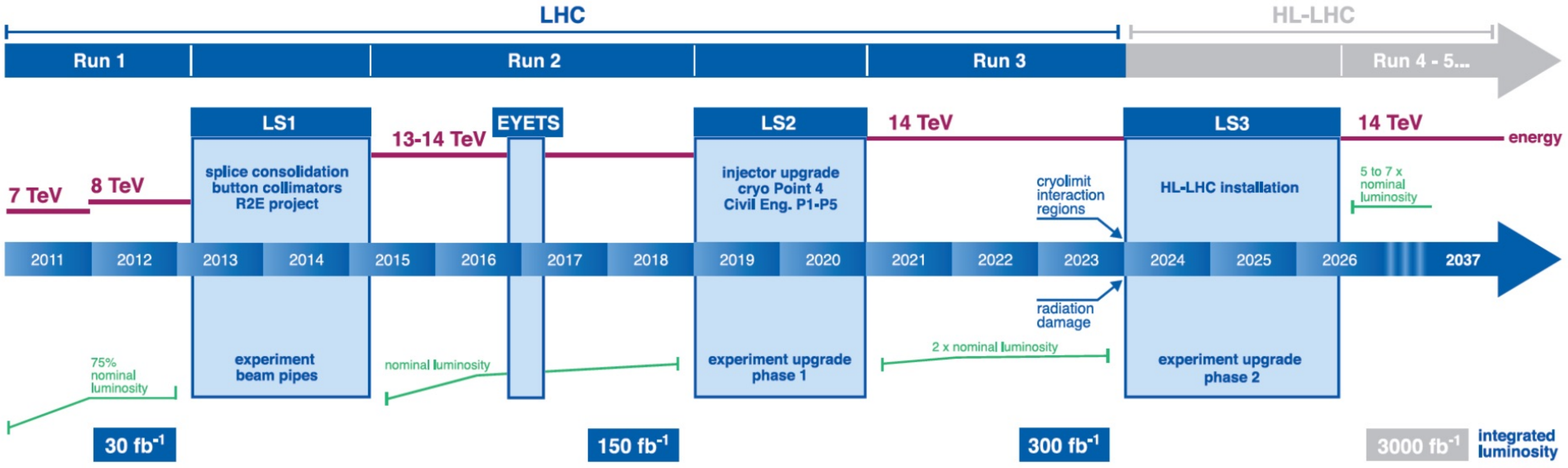
- Requirements for DAQ have been identified
- For PU=200
  - 750 kHz L1 rate and 5 MB event size
  - HLT reduction of 1 in 1000, so 7.5 kHz output
- Baseline design
  - Trigger and synchronous DAQ in xTCA, asynchronous DAQ (EVB, HLT) in commodity
  - Full event building
  - Looks entirely feasible for 2027
- HLT requirements in terms of CPU and costs not well understood

- There is time to think about new approaches

EXTRA MATERIAL

LHC / HL-LHC Plan

High Luminosity LHC



LHC

HL-LHC

Run 1 | Run 2 | Run 3 | Run 4 - 5...

LS1: splice consolidation, button collimators, R2E project; experiment beam pipes

EYETS

LS2: injector upgrade, cryo Point 4, Civil Eng. P1-P5; experiment upgrade phase 1

LS3: HL-LHC installation; experiment upgrade phase 2

7 TeV, 8 TeV, 13-14 TeV, 14 TeV, 14 TeV energy

2011, 2012, 2013, 2014, 2015, 2016, 2017, 2018, 2019, 2020, 2021, 2022, 2023, 2024, 2025, 2026, 2037

cryolimit interaction regions

radiation damage

75% nominal luminosity

nominal luminosity

2 x nominal luminosity

5 to 7 x nominal luminosity

30 fb^{-1}, 150 fb^{-1}, 300 fb^{-1}, 3000 fb^{-1} integrated luminosity

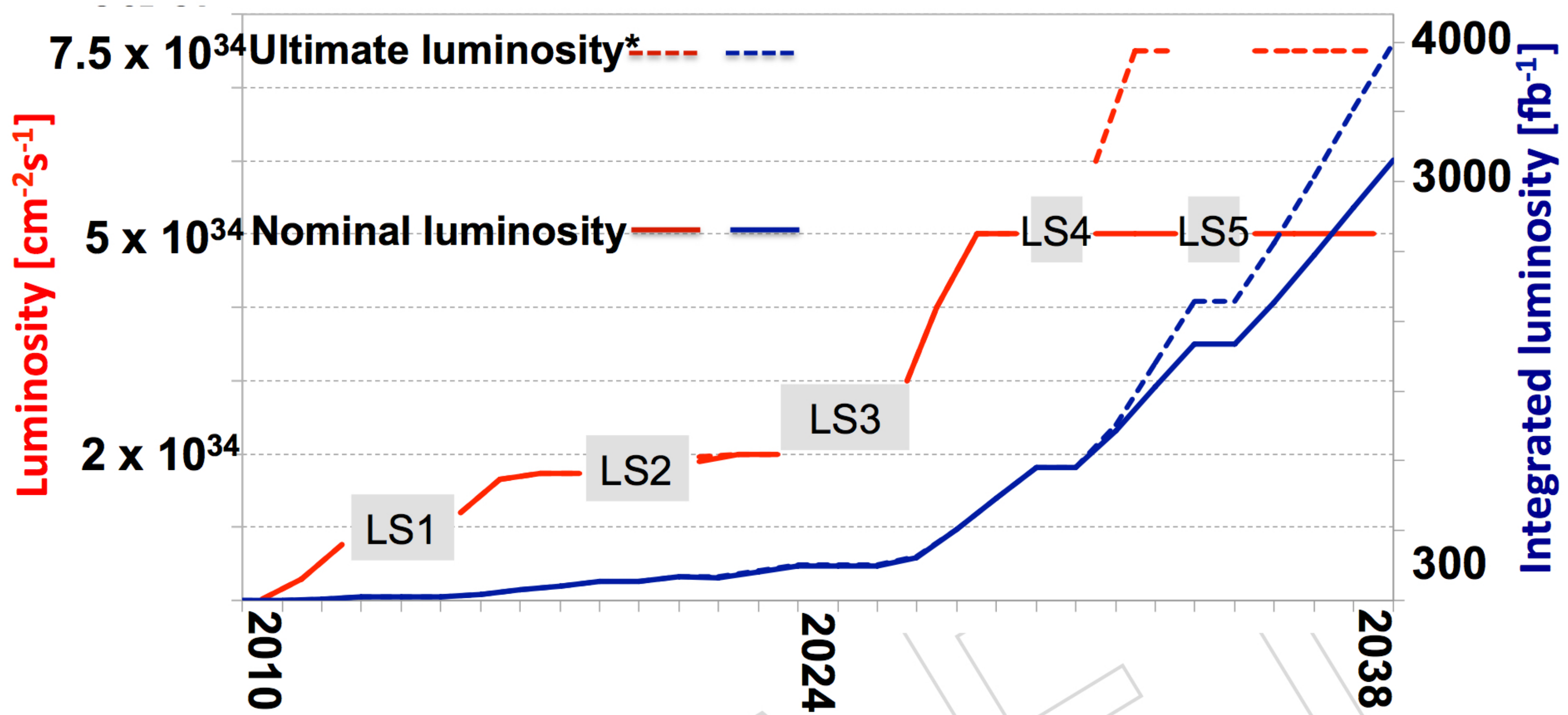


Figure 7: Projected LHC performance through 2040, showing present schedule for long shutdowns of LHC and projected luminosities.

DAQ Parameters

	Run	# HW/SW trigger levels	Level-x accept rate	Event size	EVB Effective thru	storage
Alice (Pb-Pb)	3	1 / 0	50 kHz	60 MB	0.5 TB/s	80 GB/s
LHCb	3	0 / 1	40 (30) MHz HLT 20 kHz	0.1 MB	4 TB/s	2 GB/s
Atlas	4	1 or 2 / 1	L0/L1 400 kHz or L0 1 MHz HLT 10 kHz	~ 5 MB	5 TB/s	50 GB/s
CMS	4	1 / 1	L1 750 kHz HLT 7.5 kHz	~ 5 MB	4 TB/s	40 GB/s

- Note:
  - Alice: HLT does extensive data reduction (factor 6) before EVB
  - Atlas: HW trigger L0/L1 or L0 under discussion