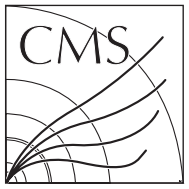
Available on CMS information server

CMS CR -2017/030



The Compact Muon Solenoid Experiment

Conference Report

Mailing address: CMS CERN, CH-1211 GENEVA 23, Switzerland

10 February 2017 (v3, 20 February 2017)

The CMS Data Acquisition - Architectures for the Phase-2 Upgrade

Jean-Marc Andre[5)], Ulf Behrens[1)], James Branson[4)], P. Brummer[2)], Olivier Chaze[2)], Sergio Cittolin[4)], Cristian Contescu[5)], B.G. Craigs[2)], Georgiana-Lavinia Darlea[6)], Christian Deldicque[2)], Zeynep Demiragli[6)], Marc Dobson[2)], Nicolas Doualot[5)], Samim Erhan[3)], Jonathan F. Fulcheri[2)], Dominique Gigi[2)], Michail Gładki[2)], Frank Glege[2)], Guillelmo Gomez-Ceballos[6)], Jeroen Hegeman[2)], Andre Holzner[4)], Mindaugas Janulis[2)], Raúl Jimenez-Estupiñá[2)], Lorenzo Masetti[2)], Frans Meijers[2)], Emilio Meschi[2)], Remigius K. Mommsen[5)], Srecko Morovic[2)], Vivian O'Dell[5)], Luciano Orsini[2)], Christoph Paus[6)], Petia Petrova[2)], Marco Pieri[4)], Attila Racz[2)], Thomas Reis[2)], Hannes Sakulin[2)], Christoph Schwick[2)], Dainius Simelevicius[2)], Petr Zejdl[5)]

Abstract

The upgraded High Luminosity LHC, after the third Long Shutdown (LS3), will provide an instantaneous luminosity of 7.5×10^{34} cm$^{-2}s^{-1}$ (levelled), at the price of extreme pileup of up to 200 interactions per crossing. In LS3, the CMS Detector will also undergo a major upgrade to prepare for the phase-2 of the LHC physics program, starting around 2025. The upgraded detector will be read out at an unprecedented data rate of up to 50 Tb/s and an event rate of 750 kHz. Complete events will be analysed by software algorithms running on standard processing nodes, and selected events will be stored permanently at a rate of up to 10 kHz for offline processing and analysis. In this paper we discuss the baseline design of the DAQ and HLT systems for the phase-2, taking into account the projected evolution of high speed network fabrics for event building and distribution, and the anticipated performance of general purpose CPU. Implications on hardware and infrastructure requirements for the DAQ "data center" are analysed. Emerging technologies for data reduction are considered. Novel possible approaches to event building and online processing, inspired by trending developments in other areas of computing dealing with large masses of data, are also examined. We conclude by discussing the opportunities offered by reading out and processing parts of the detector, wherever the front-end electronics allows, at the machine clock rate (40 MHz).

Presented at *CHEP 2016 22nd International Conference on Computing in High Energy and Nuclear Physics*

[1)] DESY, Hamburg, Germany

[2)] CERN, Geneva, Switzerland

[3)] University of California, Los Angeles, Los Angeles, California, USA

[4)] University of California, San Diego, San Diego, California, USA

[5)] FNAL, Chicago, Illinois, USA

[6)] Massachusetts Institute of Technology, Cambridge, Massachusetts, USA

The CMS Data Acquisition - Architectures for the Phase-2 Upgrade

J-M Andre[5], U Behrens[1], J Branson[4], P Brummer[2], O Chaze[2], S Cittolin[4], C Contescu[5], B G Craigs[2], G-L Darlea[6], C Deldicque[2], Z Demiragli[6], M Dobson[2], N Doualot[5], S Erhan[3], J F Fulcher[2], D Gigi[2], M Gładki[2], F Glege[2], G Gomez-Ceballos[6], J Hegeman[2], A Holzner[4], M Janulis[2a], R Jimenez-Estupiñán[2], L Masetti[2], F Meijers[2], E Meschi[2], R K Mommsen[5], S Morovic[2], V O'Dell[5], L Orsini[2], C Paus[6], P Petrova[2], M Pieri[4], A Racz[2], T Reis[2], H Sakulin[2], C Schwick[2], D Simelevicius[2a], and P Zejdl[5b]

[1] DESY, Hamburg, Germany
[2] CERN, Geneva, Switzerland
[3] University of California, Los Angeles, California, USA
[4] University of California, San Diego, California, USA
[5] FNAL, Batavia, Illinois, USA
[6] Massachusetts Institute of Technology, Cambridge, Massachusetts, USA
[a] also at Vilnius University, Vilnius, Lithuania
[b] also at CERN, Geneva, Switzerland

E-mail: emilio.meschi@cern.ch

Abstract. The upgraded High Luminosity LHC, after the third Long Shutdown (LS3), will provide an instantaneous luminosity of 7.5×10^{34} cm$^{-2}s^{-1}$ (levelled), at the price of extreme pileup of up to 200 interactions per crossing. In LS3, the CMS Detector will also undergo a major upgrade to prepare for the phase-2 of the LHC physics program, starting around 2025. The upgraded detector will be read out at an unprecedented data rate of up to 50 Tb/s and an event rate of 750 kHz. Complete events will be analysed by software algorithms running on standard processing nodes, and selected events will be stored permanently at a rate of up to 10 kHz for offline processing and analysis.

In this paper we discuss the baseline design of the DAQ and HLT systems for the phase-2, taking into account the projected evolution of high speed network fabrics for event building and distribution, and the anticipated performance of general purpose CPU.

Implications on hardware and infrastructure requirements for the DAQ "data center" are analysed. Emerging technologies for data reduction are considered. Novel possible approaches to event building and online processing, inspired by trending developments in other areas of computing dealing with large masses of data, are also examined.

We conclude by discussing the opportunities offered by reading out and processing parts of the detector, wherever the front-end electronics allows, at the machine clock rate (40 MHz). This idea presents interesting challenges and its physics potential should be studied.

1. Introduction: CMS Detector Upgrades for HL-LHC

The HL-LHC upgrade aims at increasing the integrated luminosity delivered to the experiments by approximately an order of magnitude with respect to the original design. This will be achieved

by a combination of increased beam intensity and improved focusing, and will require a major upgrade of several components of the accelerator and the injection chain. The resulting HL-LHC beam parameters are summarised in table 1. In order to keep under control the number

Table 1. HL-LHC projected running parameters. The instantaneous luminosity and average pileup figures are for levelled operation.

	$\mathcal{L}$	$\langle$ NPU $\rangle$	Vtx Density	$\int \mathcal{L}$ / year
Baseline	$5 \cdot 10^{34}$ cm^{-2} s^{-1}	140	0.8 /mm	250
Ultimate	$7.5 \cdot 10^{34}$ cm^{-2} s^{-1}	200	1.2 /mm	>300

of interactions per crossing (pileup), which is one of the main limitations to detector operation, the large integrated luminosity will be achieved by levelling, i.e. controlling the instantaneous luminosity by continuous variation of the beam focusing and/or overlap. In spite of this measure, the amount of overlapping interactions and their density, as well as the level of radiation produced in the forward region, dictate an almost complete redesign of the tracker and of the endcap calorimeter. One of the main physics focus of the HL-LHC physics program, the precision study of the Higgs boson properties, also requires extended coverage in the forward region to enable tagging of VBF Higgs production [1]. A new silicon tracker with trigger capabilities and extended coverage (see [1] chapter 2), high granularity endcap calorimetry largely based on silicon sensors (see [1] section 3.2), and extended muon detector coverage, will enhance the CMS acceptance and selection power. To help distinguish particles originating from the interesting vertex, timing information will be recorded in most detectors. Coping with beam conditions will also require replacing front-end electronics of legacy detector components to improve radiation hardness and readout speed, a higher selectivity of the trigger system, and a higher readout rate and bandwidth, which will also entail new strategies for the back-end electronics. The upgraded CMS detector will be read out at a rate close to 1 MHz by some 50k high-speed optical links, for an average expected event size of 5MB. The phase-2 requirements of the CMS Trigger and DAQ are summarised in table 2.

Table 2. CMS phase-2 trigger and DAQ projected running parameters.

Peak $\langle$ NPU $\rangle$	140	200
Level-1 accept rate (maximum)	500 kHz	750 kHz
Event Size (estimate [1])	4.5 MB	5.0 MB
HLT accept rate	5 kHz	7.5 kHz
HLT computing power (estimate [1])	5.0 MHS06	11.0 MHS06
Storage throughput	27 GB/s	42 GB/s

2. Baseline DAQ Design

The main goal of the DAQ is to provide the data pathway and time decoupling between synchronous data readout and processing (detector back-end and hardware trigger level) and asynchronous selection and permanent storage (figure 1).

Following the tradition [2, 3], the baseline CMS Trigger and Data Acquisition System will continue to feature two trigger levels, with only one synchronous first level (Level-1) consisting of

custom electronic boards and a second level (High Level Trigger, HLT) selecting on full detector event data with software algorithms running asynchronously on "standard" processors. The

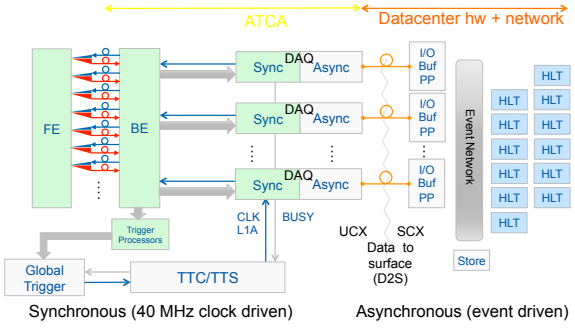


Figure 1. Synchronous and asynchronous DAQ.

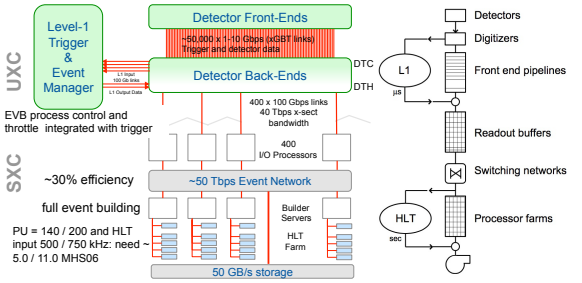


Figure 2. Principle schematics of the baseline phase-2 CMS DAQ

baseline architecture of the DAQ is sketched in figure 2. The optical links from detector front-ends are aggregated in detector-dependent back-end boards. A DAQ Hub (DTH), described in more detail in section 3, aggregates data from multiple back-ends and combines streams to feed high speed commercial optical links (D2S links) with 100 Gb/s or larger bandwidth, providing the necessary buffering for time decoupling and transmission using a reliable high-level protocol. The D2S links carry the data to surface where they are buffered in a set of I/O processors prior to event building. The DTH is also responsible for distributing trigger accept and timing signals, as well as trigger control codes for calibration and synchronisation, to the back-end electronics, from where they are usually redistributed to front-ends.

Full event building is achieved using a high speed switching network with a total effective cross-sectional throughput of 50 Tb/s. Events are assembled in a set of builder servers and stored in files on a fast access block device (based on DRAM or some other form of low-latency storage) where they can be accessed by one of the HLT computers, connected to the builder server via inexpensive small switches for processing, in a way similar to the current scheme [4], and finally transferred to a cluster filesystem for storage as in [5]. The overall design of the baseline event builder and the HLT infrastructure remains very similar to the one described in [3].

In the following sections we examine the implications of the baseline design for the various components and discuss possible new directions which can help building a better, cheaper DAQ system.

3. Data To Surface

As outlined in figure 1 a set of back-end boards, which are sub-detector-dependent, will work synchronously with the machine clock and receive data from the front-end, distribute Trigger, Timing and Control (TTC) signals, and generate trigger throttle requests based on the internal buffer occupancy (TTS). In general, Level-1 trigger input data will be received by separate back-end electronics for distribution to the Level-1 trigger processors, but in some cases (e.g the silicon tracker) the same back-end board will be responsible of routing trigger and DAQ data (received over the same links as DAQ) to the different destinations. The front-end uplinks will mostly be serial point-to-point bidirectional optical link running at 5-10 Gb/s (mostly lpGBT [6]). Some detectors will "stream" the entirety of their readout directly to the backend at the crossing rate, rather than buffering them at the front-end waiting for a Level-1 trigger. In this

case, data will be forwarded to the Level-1 processors and buffered at the back-end. Upon Level-1 decision, the relevant subset will be forwarded to the DAQ.

Table 3. CMS Phase-2 detector projected data links and event size summary.

subdet	FE daq links	FE trigger links	sub-event size	BE leaf cards	BE crates	DAQ links	notes
PxTk	1000	0	1.6	100	12	48	
SiTk	1500	13500	0.5	300	33	132	([7])
ECAL	8000	2000	1.2	100	12	48	([8])
HCAL	1000	1000	0.2	100	12	48	([8])
HGCAL	8000	8000	1.2	160	18	72	
DT	0	3000	0.1	60	8	16	([9])
CSC	0	600	0.2	12	2	8	([9])
GEM	500	1000	0.05	30	12	12	
RPC	1000	1000	0.06	40	5	20	
Other	1000	1000	0.1	20	3	12	
Total	25612	27500	5.0	922	117	416	

[7] SiTk uses the same links to multiplex trigger and DAQ data - here we use the known proportion of trigger data (90%) to split the number of links

[8] both legacy calorimeter systems will stream data at 40 MHz

[9] data are streamed at 40 MHz for both DT and CSC and trigger input is generated in the back-end

The modular electronics form factor of choice for the phase-2 upgrade back-end electronics is the ATCA standard [7]. The DAQ and TTC Hub ATCA board (DTH) will aggregate several input streams from individual back-end boards in an ATCA crate via the backplane or dedicated links, and route them to the surface (figure 3). The asynchronous part of the DAQ will be fully data driven. Sufficient local buffering will be necessary to enable conversion to a standard commercial protocol. A large FPGA will take care of the stream aggregation and protocol conversion, with the option to perform some sub-detector specific post-processing. To match the needs of the different sub-detectors in terms of output bandwidth, as outlined in table 3, the DTH will have to be capable of up to 1.2 Tb/s output bandwidth over twelve 100 Gb links, which could be internally bundled. The DTH will also receive and distribute, over the ATCA backplane, the clock and trigger signals from the trigger distribution system, which will be an evolution of the current TCDS system [8].

The current CMS D2S board (FeROL) uses a simplified TCP/IP implementation in FPGA and can drive up to 4 10 Gb/s links [9]. The DTH prototyping work will aim at reusing this implementation and proceed in two phases: a first, with the aim of combining the FeROL functionality with that of the current μTCA board used for data aggregation and trigger distribution [10], and a second, aimed at scaling the total cross-sectional bandwidth of the device up to the phase-2 specifications. In case the memory and/or bandwidth requirements could not be met, ethernet layer-2 could be adopted at the price of having to deal with unreliable transmission. The implementation of a more complex HPC protocol, like Infiniband, in the DTH seems difficult in perspective.

The D2S network will be organised as illustrated in figure 4, with a minimal amount of redundancy built in by using small switches allowing the redirection of links to different I/O processors.

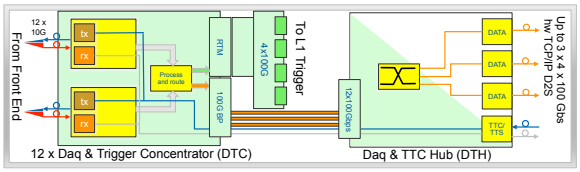


Figure 3. The DTH and a detector back-end board in an ATCA crate.

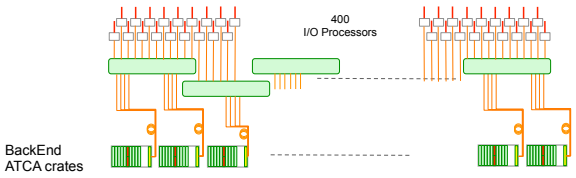


Figure 4. Principle schematics of the D2S system.

4. Event Building

The current CMS DAQ Event Builder [11] uses InfiniBand FDR switches to build events of up to 2 MB size at a Level-1 rate of 100 kHz. Considering the parameters of the phase-2 DAQ, assuming the same structure of the event builder as the current one, and using readily available 100 Gb/s technology (InfiniBand HDR or OmniPath) would require 400 I/O processors receiving data from the D2S system and 800 switch ports to connect them to the Builder Servers distributing complete events to the HLT (figure 2). Clearly, some further optimisation, considering the foreseeable evolution of hardware over the phase-2 timescale, is possible.

On the cluster interconnect front, multi-100 Gb/s is starting to hit the market and can be expected to become affordable over the next 5 years, allowing the reduction of the number of switch ports and EVB nodes. A corresponding evolution of I/O on-board interconnects can be expected over the same time scale. PCIe 4.0, with a speed of 16 GT/s per lane, is also expected to hit the server market over the next five years. On the other hand, server memory bandwidth increase is driven by the needs of many-core CPU and could well exceed that required to fully exploit top-of-the-line single bidirectional links at the relevant timescale for the CMS phase-2 upgrade. An hypothetical 2025 I/O server, like the one sketched in figure 5, could handle multiple input D2S links and perform event building in a folded architecture (as is planned e.g. for the LHCb upgrade [12]), whereby the same processor receives and organises input data, performs event building, and buffers the resulting complete events. This could bring the number of event builder nodes down by up to an order of magnitude.

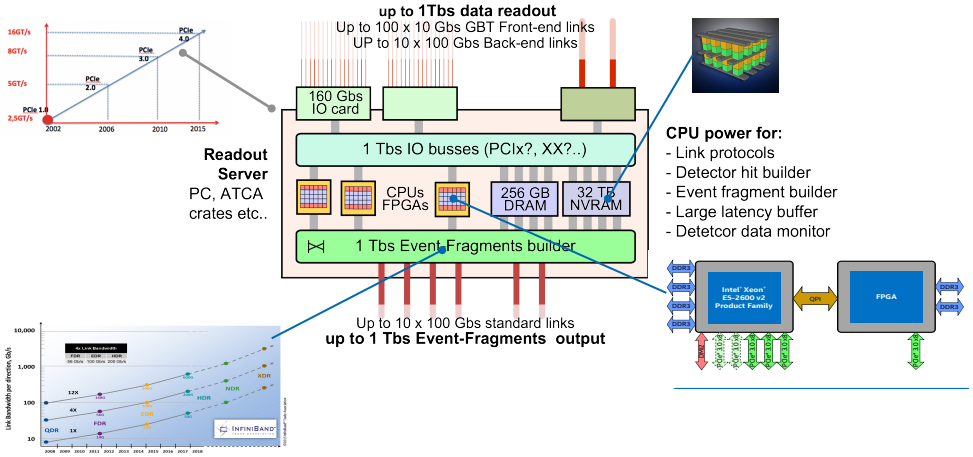


Figure 5. Possible features of a 2025 I/O processor.

Leveraging the computing power potentially available in the I/O processor, as well as ancillary engines, such as FPGAs, to perform pre-processing and organise data, one could further imagine

an alternative approach where event fragments are logically "federated" in large amounts of non-volatile higher latency memory in the I/O processors, and then analysed in place through the cluster interconnect [13].

5. High Level Trigger

One of the main strategic early choices of the CMS DAQ design was to use for the HLT the exact same framework and algorithms used offline [2]. This choice presents undoubtedly many advantages, in terms of rapidity of deployment of new triggers adapted to the physics needs, reproducibility of results, and control over the trigger efficiency, which are all paramount for the accuracy of the subsequent analysis. On the other hand, optimisation of the algorithms for specific processor architectures, as well as rapid adoption of new and more efficient programming paradigms, are harder with software that is developed by and has to cater to a vast community. As a result, as for most LHC experiments, the CMS HLT has been limited to use general purpose CPU and has had limited success in optimisation campaigns meant to exploit the most modern features like vectorisation and large-scale parallelism.

HepSpec06 [14] is a widely adopted benchmark for generic HEP applications, believed to well reflect the performance of different processors on HEP software. The evolution of the benchmark performance and TDP of server computers used in the CMS HLT, compared to the ones used in CERN IT services, are illustrated in figure 6. Assuming the current reconstruction and

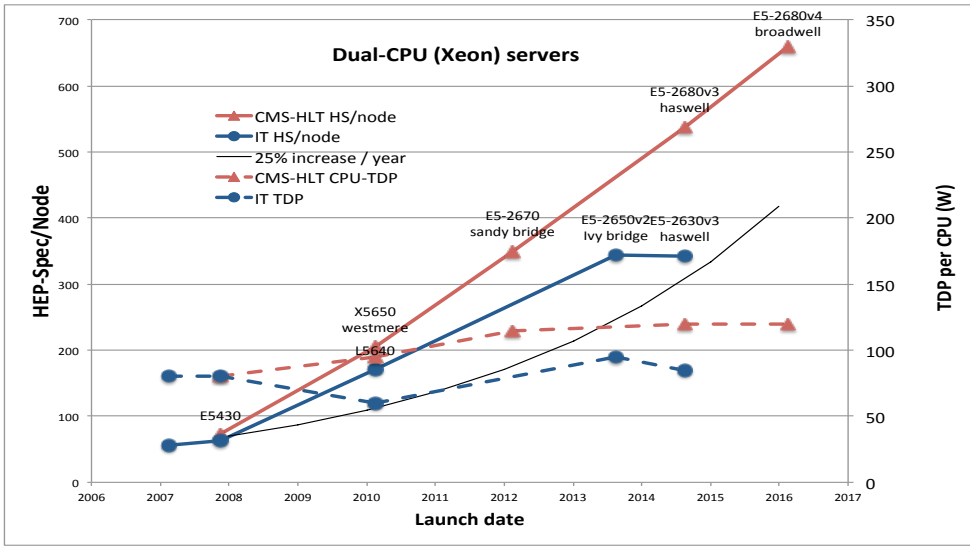


Figure 6. Evolution of the CPU HS06 benchmark and TDP of CMS HLT servers (red). The black solid curve shows the model whereby CPU power grows exponentially by 25% per year at constant TDP.

selection algorithms will scale with the number of channels and the detector occupancy, taking into account the increased Level-1 accept rate, and folding in the CPU savings anticipated from the use of Level-1 tracks in the HLT, we estimate the total required CPU power for the HL-LHC operation of CMS to be 5.0 and 11.0 MHS06 respectively for baseline (140 PU) and ultimate (200 PU) instantaneous luminosity (table 2). The number of HLT servers and the total cooling power required for the ultimate HLT farm of phase-2 are illustrated in table 4. The current CMS data center is sized to approximately match the most optimistic scenario of the first column in the table. Deploying a larger number of servers or a much larger cooling power would require a new data center and a major overhaul of the existing infrastructure. It is therefore paramount

Table 4. phase-2 HLT datacenter size and total power projections.

	exponential, 25%	exponential, 12.5%	linear 75HS06/year
benchmark ratio 2027/2016	12	3.7	2.2
number of servers	1400	4500	7800
total power (MW)	0.5	1.6	3

to investigate ways to reduce the TDP of the final system without compromising the physics program.

A fair amount of R&D work has been already directed at exploiting coprocessors and GPGPU as offload engines for specific reconstruction algorithms (see e.g. [15, 16, 17]). On the other hand, it is of particular importance in the context of the HLT to be able to rapidly deploy additional CPU power when conditions warrant, without the need for complex porting and revalidation of algorithms on new hardware. Instrumenting individual HLT compute nodes with coprocessor seems therefore a bad idea. Offloading highly parallelizable tasks to a network service, provided by a farm of specialized nodes, may seem more attractive. However, one must take into account that the additional latency introduced by the service may lead to livelock situations inside the reconstruction code running in many-core systems, thereby defeating the purpose of the offload. An attractive solution, explored for one particular case in [18], is to run algorithms preemptively on a dedicated farm receiving input directly from the event network, effectively making the results available as part of the input data for the HLT. The latter approach could be particularly suitable when combined with detector data federation and indexing, as discussed in 4, making best use of large amounts of low-latency, addressable non-volatile memory.

6. Partial Acquisition at Crossing Rate

As we have seen in section 3, in the upgraded CMS several detectors will be read out in streaming mode, i.e. for every crossing. On top of that, more precise trigger input information will be available at the full crossing rate in the detector back-end electronics, including tracks down to relatively low-pt and large pseudorapidity. It is therefore natural to ask whether collecting and analysing all these detector/trigger data available at 40 MHz may provide interesting additional functionality, for example to enable the study of channels lacking a well defined signature for hardware triggering, but not needing the full detector acceptance and/or full detector resolution.

While the full detector is being read out and processed at the Level-1 rate, a second, parallel system would run as an "opportunistic experiment" processing detector streams at 40 MHz. We call this system "Level-1 scouting" in analogy to [19]. For trigger input, the data could be extracted directly from the Level-1 links using passive splitters and fed to this system, which would pre-process them, organise them, and perform a fast one-pass analysis to produce ranked indices of physics objects to support specific query-based analyses [13]. Such a system, besides providing the capability to search for non trivial signatures over the full statistics, would naturally constitute an extra-fast-track calibration loop which could be used to assist "standard" scouting in the HLT, which is usually limited by the inferior quality of calibrations and alignment typically available online.

7. Conclusions

The CMS baseline DAQ architecture for the Phase-2 upgrade is feasible with readily available technology. The baseline event builder will require about eight hundred 100 Gb/s link ports, connecting 400 servers in the folded scheme. Scaling the current HLT CPU needs for increased event size, pileup, and rate, gives a total of about 11 MHS06 which, assuming a modest

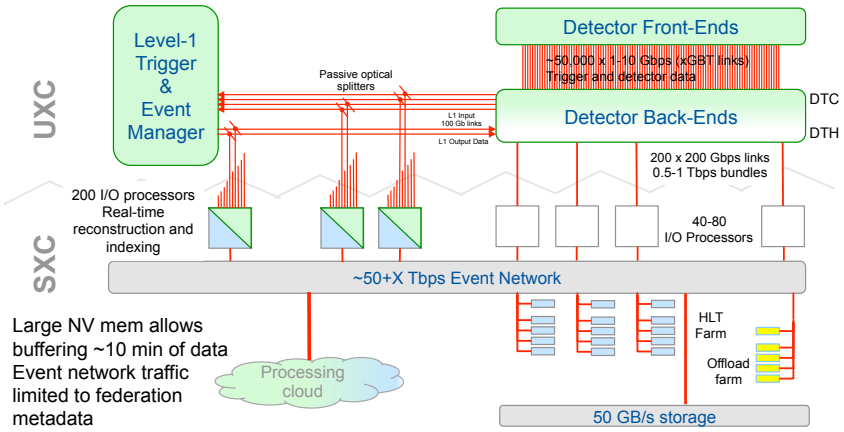


Figure 7. A cartoon of the phase-2 CMS DAQ including the 40 MHz scouting system

improvement of the processing power of 25% per year would require 1400 HLT servers. The evolution of server capacity remains however uncertain. Reduction of complexity and cost may result from tracking and late adoption of maturing technologies, in particular multi-100 Gb links and high bandwidth I/O servers with large amounts of addressable NV memory.

Prototyping and exploratory work will aim at developing efficient D2S solutions and studying different alternatives for the Event Network, including a "non-building" option. The single most challenging task will be keeping the size and cost of the HLT farm under control. Understanding the actual evolution of hardware in the coming decade, as well as studying the application of heterogeneous architectures to offload suitable algorithms, may help reducing the TCO. The adoption of new programming styles more suitable for truly distributed systems with large NV memory should also be considered and studied in more detail. A 40 MHz "scouting" system exploiting the Level-1 track primitives as well as data from streaming detectors may be an attractive addition to the CMS DAQ, with the potential to enable the study of particular physics channels where the lack of a well defined signature makes hardware triggering difficult. The feasibility and actual physics potential of such a system should be studied in more detail.

References

[1] Contardo D *et al.* (CMS Collaboration) 2015 *CERN-LHCC-2015-010*
[2] Cittolin S *et al.* (CMS Collaboration) 2002 *CERN-LHCC-2002-026*
[3] Hegeman J *et al.* 2016 *to be published in proceedings 2016 IEEE-NPSS Real Time Conference (RT)*
[4] Meschi E *et al.* 2015 *Journal of Physics: Conference Series* **664** 082033
[5] Darlea G L *et al.* 2015 *Journal of Physics: Conference Series* **664** 082009
[6] Felici D *et al.* 2014 *Journal of Instrumentation* **9** C01004
[7] URL `https://indico.cern.ch/event/471309`
[8] Hegeman J *et al.* 2015 *2015 IEEE NSS/MIC Conference Book* pp 1–3
[9] Bauer G *et al.* 2013 *Journal of Instrumentation* **8** C12039
[10] Hazen E *et al.* 2013 *Journal of Instrumentation* **8** C12036
[11] Mommsen R K *et al. these proceedings*
[12] Campora Perez D H *et al.* 2016 *Nucl. Instrum. Meth.* **A824** 280–283
[13] Meschi E 2015 *Journal of Physics: Conference Series* **664** 082032
[14] URL `http://w3.hepix.org/benchmarks/doku.php`
[15] Nowak A *et al.* 2014 *Journal of Physics: Conference Series* **513** 052024
[16] Halyo V *et al.* 2014 *Journal of Instrumentation* **9** P04005
[17] Funke D *et al.* 2014 *Journal of Physics: Conference Series* **513** 052010
[18] Pantaleo F *et al. these proceedings*
[19] Khachatryan V *et al.* (CMS Collaboration) *arXiv preprint arXiv:1604.08907*