EPJ Web of Conferences **251**, 04023 (2021) https://doi.org/10.1051/epjconf/202125104023
CHEP 2021

The Phase-2 Upgrade of the CMS Data Acquisition

Gilbert Badaro[7], *Ulf* Behrens[6], *Andrea* Bocci[1], *James* Branson[2], *Philipp* Brummer[1,10], *Sergio* Cittolin[2], *Diego* Da Silva-Gomes[3,9], *Georgiana-Lavinia* Darlea[4], *Christian* Deldicque[1], *Marc* Dobson[1], *Dominique* Gigi[1], *Nekija* Dzemaili[1], *Maciej* Gladki[1], *Frank* Glege[1], *Guillelmo* Gomez-Ceballos[4], *Jeroen* Hegeman[1], *Thomas Owen* James[1], *Wei* Li[6], *Frans* Meijers[1], *Emilio* Meschi[1,*], *Remigius K.* Mommsen[3,11], *Srecko* Morovic[2], *Luciano* Orsini[1], *Ioannis* Papakrivopoulos[5,9], *Christoph* Paus[4], *Andrea* Petrucci[2], *Marco* Pieri[2], *Ema* Puljak[1], *Dinyar* Rabady[1], *Kolyo* Raychinov[1], *Attila* Racz[1], *Hannes* Sakulin[1], *Christoph* Schwick[1], *Dainius* Simelevicius[1,8], *Panagiotis* Soursos[1], *Andre* Stahl[6], *Uthayanath* Suthakar[1], *Cristina* Vazquez-Velez[1], and *Petr* Zejdl[1]

[1]CERN, Geneva, Switzerland
[2]University of California, San Diego, San Diego, California, USA
[3]FNAL, Chicago, Illinois, USA
[4]Massachusetts Institute of Technology, Cambridge, Massachusetts, USA
[5]Technical University of Athens, Athens, Greece
[6]Rice University, Houston, Texas, USA
[7]American University of Beirut, Beirut, Lebanon
[8]Also at Vilnius University, Vilnius, Lithuania
[9]Also at CERN, Geneva, Switzerland
[10]Also at Karlsruhe Institute of Technology, Karlsruhe, Germany
[11]Now at Zurich Instruments, Zurich, Switzerland

Abstract. The High Luminosity LHC (HL-LHC) will start operating in 2027 after the third Long Shutdown (LS3), and is designed to provide an ultimate instantaneous luminosity of 7.5×10^{34} cm^{-2} s^{-1}, at the price of extreme pileup of up to 200 interactions per crossing. The number of overlapping interactions in HL-LHC collisions, their density, and the resulting intense radiation environment, warrant an almost complete upgrade of the CMS detector. The upgraded CMS detector will be read out by approximately fifty thousand high-speed front-end optical links at an unprecedented data rate of up to 80 Tb/s, for an average expected total event size of approximately 8 – 10 MB. Following the present established design, the CMS trigger and data acquisition system will continue to feature two trigger levels, with only one synchronous hardware-based Level-1 Trigger (L1), consisting of custom electronic boards and operating on dedicated data streams, and a second level, the High Level Trigger (HLT), using software algorithms running asynchronously on standard processors and making use of the full detector data to select events for offline storage and analysis. The upgraded CMS data acquisition system will collect data fragments for Level-1 accepted events from the detector back-end modules at a rate up to 750 kHz, aggregate fragments corresponding to individual Level-1 accepts into events, and distribute them to the HLT processors where they will be filtered further. Events accepted by the HLT will be stored permanently at a

*corresponding author - e-mail: emilio.meschi@cern.ch

rate of up to 7.5 kHz. This paper describes the baseline design of the DAQ and HLT systems for the Phase-2 of CMS.

1 Introduction

The main purpose of the Data Acquisition system (DAQ) of a collider experiment is to provide the data pathway and time decoupling between the synchronous detector readout and data reduction, the asynchronous selection of interesting events in the software trigger level, and their permanent storage for offline analysis.

A conceptual scheme of the current CMS DAQ is illustrated in Fig. 1.

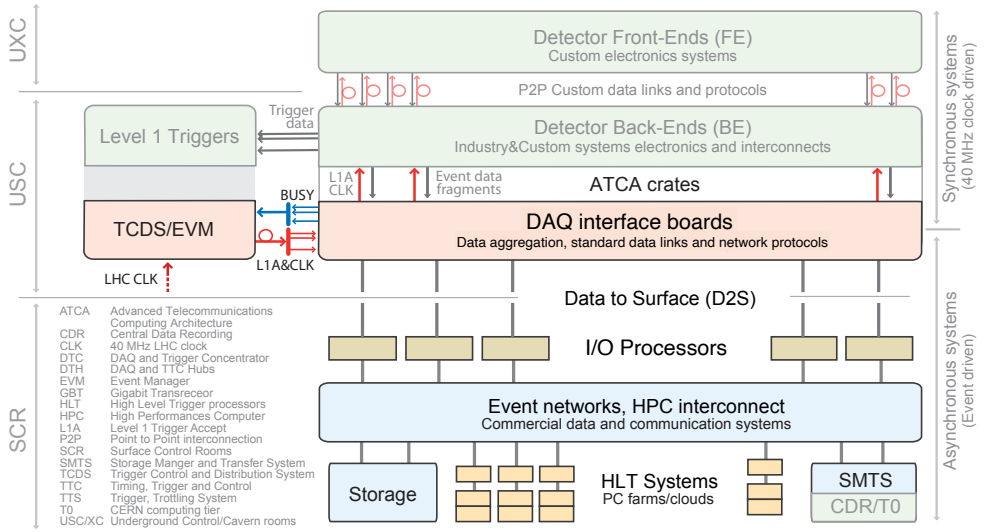

Figure 1. Schematic view of the current CMS DAQ. The D2S links bring detector data to a set of I/O servers that perform event building using an HPC interconnect fabric and serve data to the high level trigger system through the event network. Accepted events are stored locally in the HLT nodes and successively aggregated into a distributed storage system to be transferred to central computing resources in the CERN computing center.

The detector front-end (FE) and back-end (BE) are connected by bidirectional links. The downlinks are used by the BE to distribute the LHC bunch clock, the accept signal generated by the trigger system, and fast signals to control the FE electronics (e.g. counter resets). The uplinks transport digitized detector data from the FE to the BE modular electronics used to pre-process them, route the relevant portion to the hardware trigger processors, and the full resolution data to the DAQ system. This system is deadtime-less, in the sense that pipelines at every stage of the synchronous process are sized in such a way as to store data for the maximum latency required to decide whether to pass them on or drop them, under normal conditions, without losses. Normal conditions refer to the characteristic size of the data fragments to be buffered and transferred, and to the maximum rate that the links can sustain. Some constraints may also be imposed on the time between two subsequent accept signals due to specific design choices of the FE electronics. These constraints are referred to as trigger rules. Even if deadtime-less by design, the system must be protected from buffer

overruns, since the buffers themselves have finite size. These can arise from problematic or noisy detector readout channels, or from accelerator and/or beam conditions.

The synchronous portion of the DAQ system distributes the LHC bunch clock, trigger accept and fast control signals (Trigger, Timing and Control, TTC) to the BE, and collects the buffer status of the individual BE leaf boards in order to control the issuing of accept signals (Trigger Throttling System, TTS) and prevent buffer overruns. The main goals of the synchronous portion of the DAQ are to guarantee the collection of all data for events selected by the hardware trigger, and to keep the effective deadtime within a certain (small) limit. The timescales involved in the synchronous part of the system are related to the latency of the hardware that processes the trigger decision (4 μs for the current CMS detector).

Each BE module uses standardized DAQ firmware to transfer its accepted event data to a DAQ interface board over an asynchronous point-to-point link, using a protocol with flow control. There, data are aggregated into larger fragments for efficient use of the network bandwidth and then transported to the surface (Data To Surface, D2S) over high-speed links, using a standard protocol, into the memory of commercial CPU servers used as I/O processors. To guarantee protection against congestion at the destination, a lossless protocol is preferred for the D2S, even though this requires additional buffer space at the source to allow re-sending lost packets. The inherent timescale of this step is of the same order as the allowed time window to resend lost packets, of order 1 ms.

A high performance switched network (Event Network) interconnects the I/O processors to enable the assembly of fragments corresponding to individual accepted events in the memory of a single computer. This process is called event building. After events are built in one of the I/O processors, they are stored locally until one of the HLT processors can pick them up and analyse them. The average processing time for the HLT is of order 200–300 ms per event (3–4 s per event for the detector and conditions of HL-LHC, as measured on today's CPUs), with the most complex events taking up to ten times as much. The buffer for built events is designed for a maximum latency of up to 1–2 minutes, to allow for the startup time and large fluctuations typical of a purely software selection algorithm.

Accepted events are stored locally before being assembled into larger data-set files for efficient long-term storage. The local storage decouples the data acquisition from the transfer process and must provide enough buffer to absorb fluctuations in the transfer speed and to enable uninterrupted data taking in case of outage of the transfer link. This is realized using distributed or network storage attached to the same switched network used for event building. Finally, entire data-set files are transferred to central computing resources (Tier-0, located in the CERN computing center, some 10 km away) over long-distance links, to be stored for the subsequent offline reconstruction, which typically happens within a few days of the actual data-taking.

2 The Phase-2 CMS Upgrade

The number of overlapping interactions in HL-LHC collisions, their density, and the resulting intense radiation environment warrant an almost complete upgrade of the CMS detector for Phase-2. Some of the main physics foci of the CMS Phase-2 physics program, including the precision study of the Higgs boson properties, also require extended coverage in the forward region. New tracking detectors will be installed, with an Inner Tracker featuring small size pixel sensors and an Outer Tracker equipped with strip and macro pixel sensors, extending the coverage to $|\eta| = 3.8$ and providing tracking information to the Level-1 trigger [1]. An extended coverage, high granularity endcap calorimeter [2], largely based on silicon sensors, with over 6 million readout channels will also be installed. This sampling calorimeter will provide shower separation and identification adapted to harsher conditions

in the forward region of the detector. New muon detectors will complement the existing ones and extend the muon acceptance, redundancy, and selection power [3]. To help distinguish particles originating from the interesting vertex, time of flight information will be recorded by a dedicated minimum ionizing particle (MIP) timing detector (MTD) [4], by the new endcap calorimeter, as well as by the electromagnetic portion of the barrel calorimeter. A new luminometer [5], based on forward pixel rings, will provide a redundant and improved luminosity measurement. Coping with beam conditions, increased multiplicity, and ageing effects in the presence of a higher particle flux will require the replacement of the front-end electronics of legacy detector components to improve radiation hardness and readout speed [3, 6].

The increased complexity of the detector and the increased multiplicity and pileup density will require a redesign of the Level-1 trigger, which will accept events at a maximum rate of 750 kHz. Events of average size 10 MB will have to be assembled at this rate and processed by the HLT to select only one event in one hundred for permanent storage. The Phase-2 parameters are summarized in Table 1 and compared to the current (Phase-1) DAQ. The total required DAQ throughput for Phase-2 is up to a factor 40 larger with respect to Phase-1. Estimates based on simulation and the current reconstruction code indicate that the computing power required for the HLT will be 21 to 46 times larger than the Phase-1 for the design pileup level of 140 interactions per crossing, and 200 interactions per crossing scenarios respectively.

Table 1. CMS Phase-2 trigger and DAQ projected running parameters for the two pileup scenarios of 140 and 200, compared to the design values of the current (Phase-1) system.

	LHC	HL-LHC	
CMS detector	Phase-1	Phase-2	
Peak average pileup	60	140	200
L1 accept rate (maximum)	100 kHz	500 kHz	750 kHz
Event Size at HLT input	2.0 MB [a]	7.8 MB	9.9 MB
Event Network throughput	1.6 Tb/s	31 Tb/s	60 Tb/s
Event Network buffer (60 s)	12 TB	234 TB	445 TB
HLT accept rate	1 kHz	5 kHz	7.5 kHz
HLT computing power [b]	0.8 MHS06	17 MHS06	37 MHS06
Storage throughput [c]	2 GB/s	31 GB/s	61 GB/s
Storage throughput (Heavy-Ion)	12 GB/s	61 GB/s	61 GB/s
Storage capacity needed (1 day [d])	0.2 PB	2.0 PB	3.9 PB

[a] Design value.
[b] Does not include Data Quality Monitoring.
[c] The storage throughput is defined as the effective throughput with concurrent recording and transfer. The throughput required is determined by the (raw) input event size, the addition of HLT products and (zlib level-1) compression (combined factor 0.7 observed in Phase-1 and assumed the same for Phase-2) and the additional output streams for calibration and monitoring purposes (factor 1.4 for Phase-1 and factor 1.1 assumed for Phase-2).
[d] Assuming an accelerator duty cycle, i.e. the fraction of time spent in stable colliding beams, of 75 %.

3 Baseline Architecture of the DAQ Upgrade

The baseline architecture of the Phase-2 CMS DAQ, illustrated in Fig. 2, leverages the experience accumulated during Run-2 and in the preparation for Run-3. It makes use of a well-established design and, in most cases, of technologies that have been successfully deployed or are being deployed for Run-3.

The structure of the CMS Phase-2 unified data readout is largely dictated by the choice of the Advanced Telecommunications Computing Architecture (ATCA) standard form-factor

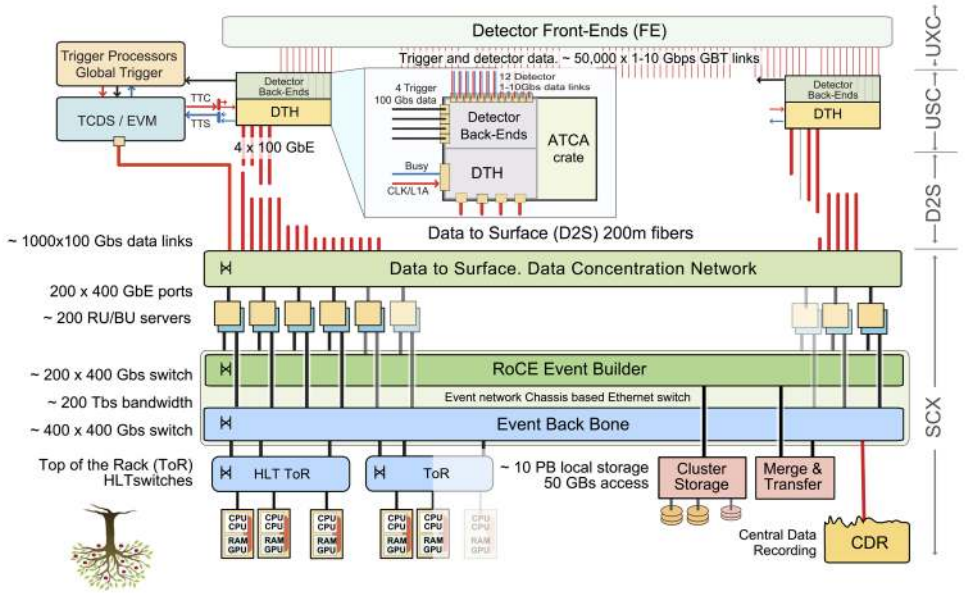

Figure 2. Layout of the CMS Phase-2 DAQ. The experiment and front-end electronics are located in the underground experimental cavern (UXC). The back-end electronics crates including the DAQ equipment are located in the underground service cavern (USC). The DAQ links (D2S) connect the USC to the surface complex (SCX) where the DAQ data center is located.

for the detector back-end electronics. The ATCA standard, with more available real estate on a single board, together with the availability of powerful new FPGA families, allow the combination of the DAQ data aggregation from back-end boards and the TTC/TTS functionality in a single custom board, the DAQ, Trigger and Timing Hub (DTH), to be installed in the sub-detector back-end crates. A prototype of the DTH board, with full functionality but reduced performance, has been produced (Fig. 3). It can accept up to 16 input optical links at 25 Gb/s, or 24 links at 16 Gb/s (hence its full name, DTH-400) using a serial point-to-point protocol with flow control, evolved from the Run-2 S-Link Express. In order to support sub-detectors with larger throughput per crate, another ATCA board, the DAQ-800, without the timing functionality, but with double the input connectivity, will also be developed. The DTH will connect to the surface data concentration network switch over up to five 100-GBASE-CWDM4 links using single-mode optical fibers, providing the necessary bandwidth and the signal range needed to cover the ~200 m distance between the underground service cavern (USC) and the surface data center (SCX).

Based on the most up-to-date information, a total of over 1200 ATCA back-end boards in about 130 crates will be needed to receive front-end data from about 50000 front-end links of the different sub-detectors. Data for Level-1 accepted events will be collected, for each ATCA crate, by one or more DTH/DAQ boards connected to adjacent boards via dedicated front-panel high-speed optical connections.

3.1 Timing and Trigger Control and Distribution System

In the CMS Phase-1 upgrade, the trigger, timing, and control (TTC) was integrated with the trigger control (TCS) and the trigger throttling (TTS) into a single system: the TCDS [7]. While the revision of the TTC/TTS system was primarily dictated by the need to integrate

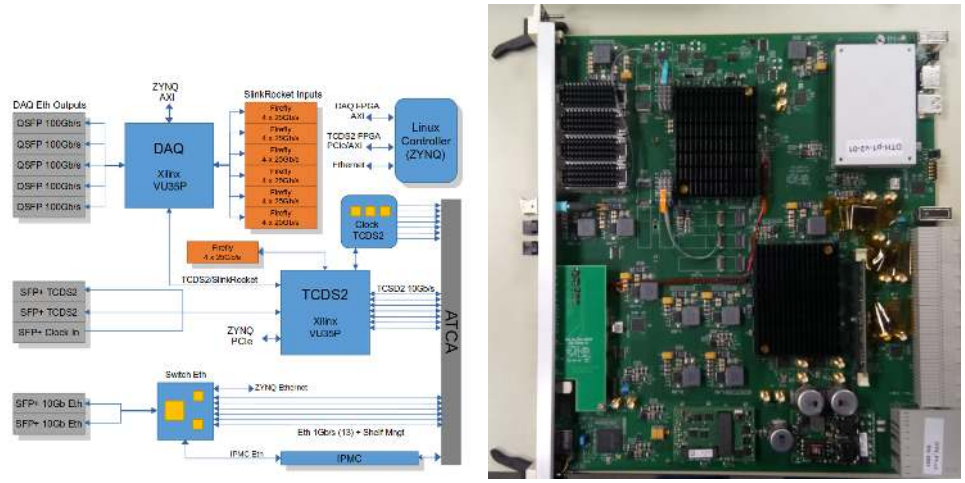

Figure 3. Left: block diagram of the DTH board. Right: DTH P1-v2 board prototype. This prototype uses two KU15P FPGA. The prototype 2 of the board, currently being designed, will use a VU35P FPGA with on-chip HBM.

the Phase-1 MicroTCA-based back-end electronics, the new system also allowed the streamlining of the trigger control functionality and a better support for the special requirements related to the operation of the luminosity detectors. Building on this success, the Phase-2 TCDS system (TCDS2) is designed to profit from the adoption by all sub-system back-ends of a single standard, and from the general availability of high-speed serial links in modern FPGAs. The clock distribution, embedded in the same high-speed data stream used to distribute trigger and synchronization commands, must fulfill stringent requirements, dictated by the timing detectors, in terms of clock quality and phase stability. A new streamlined distribution of calibration and special triggers, as well as support for multiple physics trigger types, are also envisaged. The TCDS2 architecture (Fig. 4) is considerably simplified by the physical integration of most of its functionality in a single board type, also comprising the DAQ data collection functionality (Fig. 3). A separate TCDS2 FPGA will handle the distribution of a precision clock and of the TTC stream, and the collection of TTS status, over the ATCA crate backplane. The TTC will use a special time-compensated link implementation (TCLink [8]) overlaid on top of the lpGBT protocol also used for communication with the on-detector front-ends [9].

A demonstrator clock distribution chain has been built based on prototype ASICs and prototype DTH hardware, spanning an emulated captain, a basic but real DTH, a basic subsystem back-end, and an lpGBT evaluation board as surrogate front-end. Preliminary clock quality measurements (Fig. 5) indicate that the achieved clock quality (in terms of random jitter) is well within the 10 ps RMS requirement of the Phase-2 CMS timing detector [4].

3.2 Data To Surface

The use of a reduced TCP/IP firmware implementation for the transfer of data from the detector back-end electronics to the surface data center (D2S) has been proven a dependable approach during CMS Run-2 data-taking, using optical links with speed up to 10 Gb/s [10]. A reliable standard protocol enables the direct connection of the readout links to a commercial switched network. Thanks to the large number of high-speed serial I/O available on modern

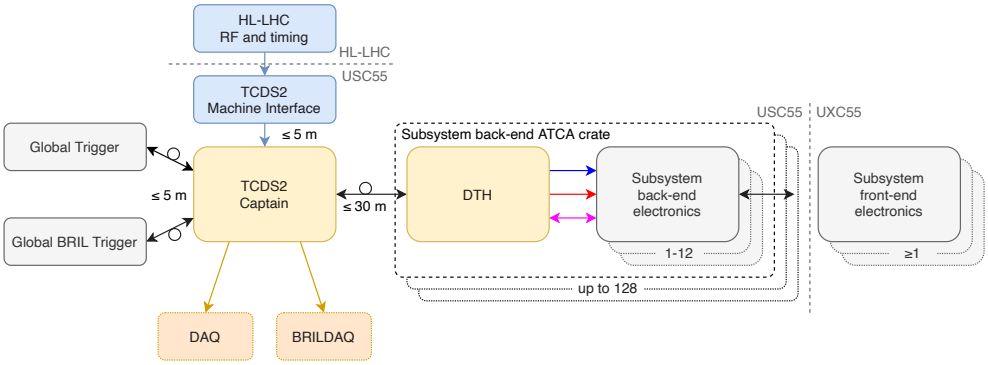

Figure 4. Simplified architectural overview of the TCDS2. The Machine Interface connects CMS to the HL-LHC RF systems. CMS data-taking is governed by the TCDS2 captain. The captain receives physics L1A candidates from the Global Trigger, BRIL triggers from the Global BRIL Trigger, and bunch clock and orbit signals from the Machine Interface. The captain distributes clocks, triggers, and timing and book-keeping information to all subsystems via the DTH. In opposite direction, the captain collects data-taking readiness information from all subsystems via the DTH. The TCDS2 captain delivers an event record with data provenance identifiers to the CMS DAQ, as well as a non-event-based 'BRILDAQ' record containing experiment live-time information to the luminosity systems.

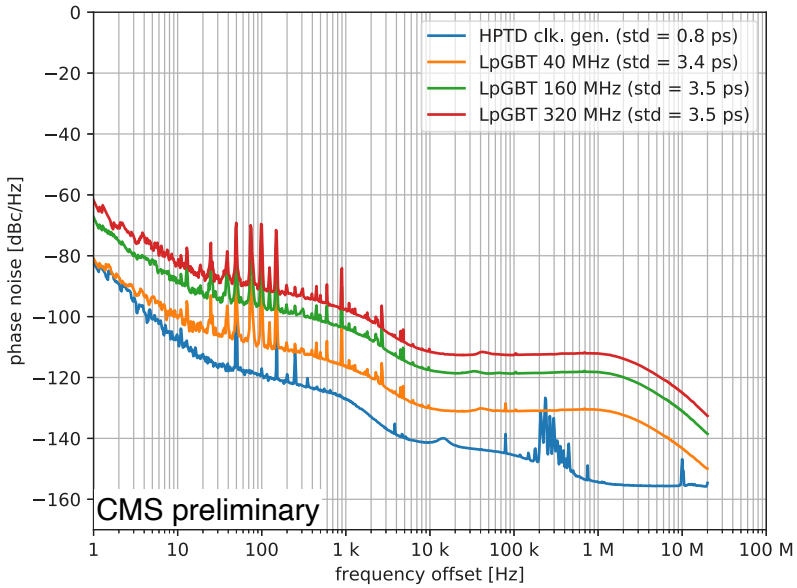

Figure 5. Phase noise and random jitter of the front-end output clock measured at the output of the lpGBT ASIC. The observed jitter comfortably meets the requirements of the Phase-2 CMS timing detector. Note: preliminary results based on prototype hardware.

FPGAs, it is today possible to overhaul this design to operate at 100 Gb/s, meeting the additional bandwidth demand arising from the increased granularity of upgraded sub-detectors and the higher Level-1 accept rate envisaged for Phase-2. In addition, large amounts[1] of on-chip high-bandwidth memory (HBM) not only enable an optimal management of link congestion at the destination port by providing large buffers at the sender side, but can also be used to handle adverse situations in the TCP/IP stack of the receiving host.

Slightly under one thousand D2S links are necessary. This is estimated from the number of DTH-400 and DAQ-800 boards required per back-end crate, the throughput per crate based on projections of the event fragment sizes by sub-detectors, and the total number of crates per sub-detector.

The bandwidth utilization of the input links to the Data Concentration Network switch complex (about one thousand) is expected to vary from 10 to 90%. The purpose of the Data Concentration Network, besides providing flexible routing, is to aggregate this traffic efficiently into higher-speed links. The choice of one or the other link speed for the output ports will be dictated by the cost and the actual availability of network interfaces and servers capable of handling the corresponding concurrent I/O, and is discussed in the following section.

3.3 Event Building

In Run-2, CMS adopted an InfiniBand switched fabric [11] for its event builder. The InfiniBand standard, widely used in HPC, lifts a large part of the load of handling network traffic from the host CPU by implementing Remote Direct Memory Access (RDMA). This enables full exploitation of the physical link available bandwidth, with low latency and reduced CPU load at the endpoints. The use of RDMA, together with an order of magnitude increase in line-speed, allowed a radical reduction of the size and cost of the Run-2 event builder with respect to Run-1.

More recently, network interfaces (NICs) supporting the RoCE protocol (RDMA over Converged Ethernet, encapsulating InfiniBand packets into Ethernet packets) have become available. These enable RDMA over a standard Ethernet switch fabric, which is usually less expensive and easier to manage. The Run-3 CMS event builder, currently being deployed, will receive data and run event building at 2 Tb/s over a unique set of 60 single-socket "readout unit/builder unit" (RU/BU) servers, connected to a Data to Surface (D2S) network receiving over five hundred 10 Gb/s links, concentrated into sixty 100 Gb/s links. The servers are connected to a 100 Gb/s RoCE-enabled network used to perform the event building, while complete events are served to HLT nodes over a second backbone network at the same link speed. The advantage of using a single-socket system lies in the simplified memory access architecture, resulting in more efficient use of memory bandwidth and simplified application configuration with respect to NUMA-based systems. Initial measurements of event builder traffic on nodes of the Run-3 event builder[2] indicate that throughput close to the line-speed can be obtained for message sizes larger than 10 KB.

Higher Ethernet link speeds of 200 and 400 Gb/s are already available for backbone switches. As with previous generations of link speeds, NICs are expected to follow suit and to become affordable on the timescale required for the installation of the Phase-2 system (beginning of year 2025). A 400 Gb/s RoCE-based event builder will in practice require servers that support the next generation PCIe (Gen5), which some vendors announce as imminent at the time of writing. The increased memory bandwidth of server-grade processors, coupled with a large number of Gen-5 PCIe lanes, will enable even a single-socket server to handle

[1] For example, the Xilinx VU35P FPGA has 8 GB of HBM memory partitioned over two stacks of 8 blocks.

[2] The Run-3 event builder uses single-socket nodes equipped with AMD EPYC 7502P processor and Mellanox ConnectX-6 100 GbE NICs interconnected via a Juniper Networks QFX100016 deep buffered switch.

concurrently the input from a 400 Gb/s NIC with the output required for both event building traffic and to serve the resulting complete event data to the HLT processors. For Phase-2, the use of a single set of servers for D2S readout and event building, and the concentration of the 100 Gb/s D2S links to 400 Gb/s, allows the size, cost, and complexity of operation of the D2S-event builder complex to remain similar to the ones of the current system. This is all the more remarkable if one considers the increase in required throughput of more than a factor 40. The choice of the event builder (RU/BU) server architecture and form factor depends mostly on this aspect and the available aggregated memory bandwidth. Should such architectures not be available (or affordable) a 200 Gb/s event builder switch could be used. The number of "RU/BU" servers will have to be doubled, but this does not present a major technical obstacle.

3.4 HLT Data Distribution and Collection

Profiting from the large amounts of physical memory that can be installed in modern servers at a relatively modest cost, the file-based HLT infrastructure (F3), deployed in Run-2 [12], enabled CMS to achieve a high level of decoupling between the event builder, the HLT algorithms, running inside the same reconstruction framework (CMSSW) also used for offline reconstruction and analysis, and the storage and transfer system responsible for the aggregation of HLT-accepted events for subsequent transfer to Tier-0. This is achieved by storing raw data of complete events into files in a memory-based file system on the event builder nodes[3].

The actual ability of the current file-based HLT I/O system, relying on nfs over TCP/IP, to meet the bandwidth, latency and reliability demands of the Phase-2 system, needs to be verified. Alternatives include the use of an RDMA-enabled network file systems, or the transition to a loosely coupled request-response protocol similar to the one currently used to uniquely distribute files. The baseline configuration assumes the use of RAM for the memory-based file system, an event backbone switch and a top of the rack switch matching the event builder link speed to the maximum input bandwidth corresponding to the HLT unit's processing power, and minimal modifications to the Run-2 software infrastructure used to serve input data to the HLT processors and to collect and combine output data.

3.5 Co-processors for the HLT

Following the general trend of HPC towards heterogeneous platforms, and in a way similar to other HEP experiments like ALICE [13] and LHCb [14], the Run-3 CMS HLT processing nodes will incorporate a GPU co-processor. A filter farm consisting of heterogeneous platforms offers the opportunity to optimize the compute power by choosing the architecture most appropriate for each task. Currently some of the most time consuming components of the online reconstruction (Fig. 6) can be offloaded to a GPU, achieving an overall speedup[4]

[3]In the file-based filter farm, each HLT processor gets access to the raw data files through a network file system (nfs) remote mount. A simple arbitration server ensures that each raw file is assigned to one and only one HLT process. The size of the file system is tailored to provide enough buffer space to allow for fluctuations in the HLT execution time, as well as the typical startup time required by the offline framework to e.g. load geometry and calibration constants. The number of events per file can be tailored to the number of individual servers attached to each event builder node to avoid starvation. Individual HLT processes store accepted events on the local disk of the server they execute on, and the multiple output files are merged back into the remote memory based file-system via the network mount. The operation of the HLT processes, the distribution of input files, and the aggregation of the output, are controlled and monitored by a set of services that use a standard kernel subsystem to detect new files. This approach also enables the HLT processes to continue after a run is stopped and until completion, which again decouples the start of a new run from potential tails in processing the last events of the previous run.

[4]These results were obtained running the HLT configuration being developed for Run-3 on a dual-socket machine equipped with two AMD EPYC "Rome" 7502 processors and one NVIDIA Tesla T4 GPU.

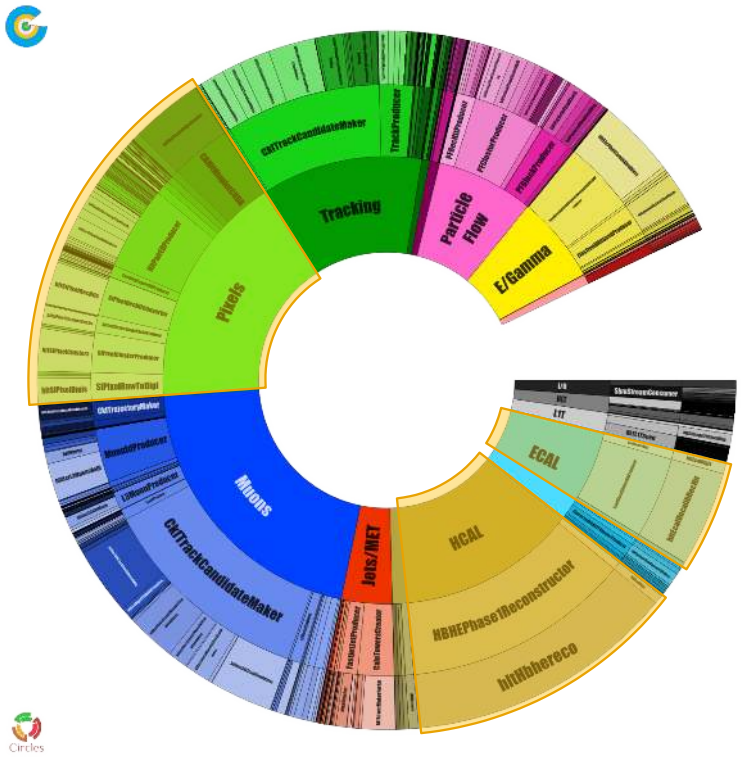

Figure 6. Fraction of the time spent in the major components of the High Level Trigger online reconstruction. The highlighted slices indicate the parts that can be offloaded to run on GPUs: the Pixel local, track and vertex reconstruction; the HCAL local reconstruction; and the ECAL local reconstruction.

of about 25%. The HLT farm foreseen for Run-3 requires roughly 800 kHS06 – 600 kHS06 coming from traditional CPUs and the equivalent of 200 kHS06 provided by GPUs. This is roughly the *break-even* point, where the cost of adding the GPUs is offset by the increase in the event processing throughput, and leaves the GPUs partially underutilized. Other reconstruction algorithms targeting both Run-3 and Phase-2 are being ported to run on GPUs, with the goal to offload at least 50% and 80% of the HLT, respectively, during Run-4 and Run-5. As the fraction of the HLT being offloaded increases, the system can be expanded adding a second GPU on each node or – targeting Run-4 – redesigned to use more powerful GPUs.

The choice of NVIDIA GPUs over other types of accelerators was mainly dictated by the flexibility of the platform, the maturity of the development tools for algorithm developers, and the cost. Many of the necessary modifications required in the reconstruction framework to enable offloading parts of the reconstruction [15], however, will be readily usable for other types of co-processors. While we use GPU benchmarks and projections of GPU computing power evolution to define the structure and cost of the Phase-2 system, the design remains flexible and agnostic to the adoption of different types of co-processors.

In order to reduce the risk of vendor lock-in, and to be able to easily exploit GPUs from different vendors and eventually other kinds of accelerators, CMS is investigating different solutions for "performance portability", like the Alpaka [16] and Kokkos [17] libraries, and the HIP [18] and SYCL [19] frameworks. As the CMSSW software is shared by the HLT and

offline processing, this will also allow CMS to leverage offline heterogeneous resources, like HPC centers.

3.6 Storage System

In Run-2 CMS adopted a cluster file system, Lustre, for the local storage of data of events accepted by the HLT prior to their transfer to Tier-0. Storage hardware is connected to the InfiniBand event builder backbone, featuring a total capacity of 750 TB and capable of delivering an aggregate (read+write) throughput of approximately 12 GB/s, well matched to the most demanding use case, i.e. heavy ion runs[5]. The Run-2 system will be replaced for Run-3 by a similar, more powerful system still based on Lustre and connected to the event builder over 100 GbE using RoCE.

While it appears that a storage system based on the same technology as the Run-3, i.e. Lustre over RoCE-enabled ethernet, exploiting the same switch used for event building, will likely meet the requirements of Phase-2 (31–61 GB/s throughput and a total storage of 2–4 PB, respectively for an average pileup of 140 and 200) at a reasonable cost, a technology survey will be conducted, approximately two years ahead of the installation, with the aim of identifying and testing the most promising solutions and select the one with the best cost/performance ratio among those meeting the requirements.

4 Conclusions

The CMS baseline DAQ architecture for the Phase-2 upgrade, as outlined above, can be realised with readily available technologies. Reasonable estimates of the evolution of the cost/performance of these technologies indicate that the Phase-2 DAQ/HLT system requirements can be met at a target cost equal or less than 10% of the total cost of the upgrades. Further evolution could help reduce the cost, or enable the system to be updated during the Phase-2 operation to meet additional requirements arising from new physics goals of CMS.

References

[1] CMS Collaboration, Tech. Rep. CERN-LHCC-2017-009, CMS-TDR-17-001 (2017)
[2] CMS Collaboration, Tech. Rep. CERN-LHCC-2017-023. CMS-TDR-019 (2017)
[3] CMS Collaboration, Tech. Rep. CERN-LHCC-2017-012. CMS-TDR-016 (2017)
[4] CMS Collaboration, Tech. Rep. CERN-LHCC-2019-003. CMS-TDR-020 (2019)
[5] CMS Collaboration, CMS-NOTE **2019-008** (2020)
[6] CMS Collaboration, Tech. Rep. CERN-LHCC-2017-011. CMS-TDR-015 (2017)
[7] J. Hegeman et al., in *2015 IEEE NSS/MIC Conference Book* (2015), pp. 1–3
[8] E. Brandao De Souza Mendes et al., PoS **TWEPP2019**, 057. 5 p (2020)
[9] P. Moreira, *The lpgbt project, status and overview*, Presented at the 2016 ACES workshop at CERN (8-March-2016), `https://indico.cern.ch/event/468486/contributions/1144369/`
[10] P. Zejdl et al., Journal of Instrumentation **8**, C12039 (2013)
[11] R.K. Mommsen et al., J. Phys. Conf. Ser. **898**, 032020 (2017)

[5] A cluster file system provides reliable concurrent access to a central storage from multiple sources, removing the need for in-house development of special protocols and applications. A commercial storage system can more easily be optimized in terms of network connectivity and performance with respect to direct attach JBODs, for example. Built-in redundancy and automatic fail-over capabilities guarantee uninterrupted operation of the experiment and avoid dead-time.

[12] E. Meschi et al., J. Phys. Conf. Ser. **664**, 082033 (2015)

[13] Rohr, David, EPJ Web Conf. **245**, 10005 (2020)

[14] R. Aaij, J. Albrecht, M. Belous, P. Billoir, T. Boettcher, A. Brea Rodríguez, D. vom Bruch, D.H. Cámpora Pérez, A. Casais Vidal, D.C. Craik et al., Computing and Software for Big Science **4** (2020)

[15] A. Bocci et al., EPJ Web Conf. **245**, 05009 (2020)

[16] E. Zenker, B. Worpitz, R. Widera, A. Huebl, G. Juckeland, A. Knüpfer, W.E. Nagel, M. Bussmann, *Alpaka–An Abstraction Library for Parallel Kernel Acceleration*, in *2016 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW)* (IEEE, 2016), pp. 631–640

[17] H.C. Edwards, C.R. Trott, D. Sunderland, Journal of Parallel and Distributed Computing **74**, 3202 (2014), domain-Specific Languages and High-Level Frameworks for High-Performance Computing

[18] Advanced Micro Devices, *Heterogeneous-computing interface for portability (hip) programming guide* (2021), online; accessed: 2021-06-01, `https://rocmdocs.amd.com/en/latest/Programming_Guides/Programming-Guides.html`

[19] Khronos Group, *Sycl 2020 specification* (2021), accessed: 2021-06-01, `https://www.khronos.org/registry/SYCL/specs/sycl-2020/html/sycl-2020.html`