The Phase-2 Upgrade of the CMS DAQ
Interim Technical Design Report

CMS Collaboration
Editors
F. Meijers, E. Meschi

Chapter Editors

Contributors

Acknowledgements
The editors wish to thank A. Bocci and F. Pantaleo for useful input and discussions.
Feedback from all readers, the Phase-2 upgrade coordinators, and the chair of the CMS Phase-2 TDRs editorial board (C. Lourenço) helped improve the quality of this document.
A special thanks goes to Sergio Cittolin for invaluable advice and help.
## Contents

1  Overview .......................................................... 9
   1.1  Introduction: CMS Detector Upgrades for HL-LHC ...................... 10
   1.2  Conceptual DAQ Design ........................................... 11
   1.3  Detector Readout .................................................. 12
       1.3.1  Readout Strategy .......................................... 13
       1.3.2  Outer Tracker ............................................. 13
       1.3.3  Inner Tracker ............................................. 14
       1.3.4  MIP Timing Detector (MTD) ................................. 14
       1.3.5  ECAL Barrel Calorimeter (EB) ............................ 15
       1.3.6  HCAL Barrel Calorimeter (HB) ............................ 15
       1.3.7  Outer Barrel (HO) Hadronic Calorimeter ................. 16
       1.3.8  Forward (HF) Hadronic Calorimeter ..................... 16
       1.3.9  Endcap Calorimeter ...................................... 16
       1.3.10  Muon Drift Tubes (DT) system ........................... 17
       1.3.11  Muon Cathode Strip Chamber (CSC) system ............... 17
       1.3.12  Muon Gas Electron Multiplier (GEM) system ............. 17
       1.3.13  Muon Resistive Plate Chamber (RPC) system .......... 18
       1.3.14  Level-1 Trigger .......................................... 18
       1.3.15  Luminosity and Beam Background Detectors (BRIL) .......... 18
       1.3.16  Contingency for Readout of Additional Detectors ............ 19
       1.3.17  DAQ Link Count ........................................... 19
   1.4  Event Network and Intermediate Buffer ................................ 19
   1.5  Online Selection and Permanent Storage ................................ 21
       1.5.1  HLT Processing Requirements ................................ 21
       1.5.2  Storage Requirements ...................................... 22
       1.5.3  “Opportunistic Cloud” on the HLT Infrastructure ......... 23
   1.6  Heavy Ion Operation in Phase-2 .................................. 23
   1.7  Summary .......................................................... 23

2  Phase-2 DAQ Baseline Design ........................................ 25
   2.1  Data To Surface .................................................. 26
       2.1.1  DAQ and TCDS Hub ......................................... 27
       2.1.2  DTH implementation ........................................ 28
       2.1.3  Roadmap for DTH .......................................... 29
   2.2  Trigger Control and Distribution .................................. 30
       2.2.1  Phase-1 TCDS ............................................... 31
       2.2.2  Phase-2 TCDS ............................................... 32
       2.2.3  Trigger Rules and Deadtime ................................ 35
       2.2.4  Precision Clock Distribution .............................. 35
       2.2.5  TTC Machine Interface .................................... 36
       2.2.6  Roadmap for the TCDS .................................... 36
2.3 Event Builder .................................................. 37
  2.3.1 Roadmap for the Event Builder ..................... 37
2.4 High Level Trigger .......................................... 38
  2.4.1 Evolution of HLT Processing Nodes ............... 38
  2.4.2 Roadmap for HLT .................................. 39
2.5 Storage Manager and Transfer System .................. 40
  2.5.1 Roadmap for SMTS ................................ 40
2.6 Software Infrastructure, Control and Monitoring ....... 41
  2.6.1 Online Software Services ........................... 41
  2.6.2 Possible development directions for overall CMS online software .... 42
  2.6.3 Software Roadmap .................................. 42
2.7 Detector Control System ................................... 43
  2.7.1 DCS Roadmap ....................................... 43
2.8 Baseline Bill of Material ................................... 43

3 Design Optimization and Technology Survey ................. 47
  3.1 Data to Surface .......................................... 47
  3.2 Trigger Control and Distribution ....................... 48
  3.3 Event Builder ............................................ 48
  3.4 Heterogeneous Architectures for the HLT ............ 49
    3.4.1 Case Study of using GPUs ......................... 51
  3.5 Storage Manager And Transfer System ................ 51
    3.5.1 Data Format ....................................... 51
    3.5.2 Storage Technology ................................ 51
  3.6 Conclusions .............................................. 52

4 Online Infrastructure at Point 5 ............................. 53
  4.1 Introduction ............................................. 53
  4.2 DAQ infrastructure in USC .............................. 53
  4.3 Online Data Center ...................................... 53
    4.3.1 Technical Infrastructure Roadmap up to LS3 ......... 54
    4.3.2 Phase-2 Online Data Center Options ............... 54
  4.4 IT Infrastructure and Systems ........................ 56
    4.4.1 Control Network .................................... 56
    4.4.2 Operating Systems .................................. 57
    4.4.3 Sub-detector Hardware Support Infrastructure .... 57
    4.4.4 Core Computing Services ............................ 58
  4.5 Conclusions .............................................. 59

5 Development Directions ...................................... 61
  5.1 Event Building over D2S ................................. 61
  5.2 TCDS Support for Prioritized Throttling ............... 61
  5.3 TCDS Support for Latency Measurement ................ 62
  5.4 Federated Event Data Indexing .......................... 62
Contents

5.5 Partial Acquisition at Bunch Crossing Rate .......................... 63

6 Organization, Schedule and Costs ............................... 65
  6.1 Organization .................................................. 65
  6.2 Schedule ....................................................... 65
  6.3 Cost Estimate .................................................. 68

A Optical Link Technologies for LHC Upgrades .................. 71
  A.1 Phase-1 and Phase-2 Upgrades .................................. 71
  A.2 Optical Link Systems ........................................... 71

B Modular Electronics ................................................. 73
  B.1 MicroTCA ...................................................... 73
    B.1.1 Phase-1 CMS MicroTCA systems ......................... 74
  B.2 Phase-2 Modular Electronics .................................. 74
  B.3 ATCA .......................................................... 74

C Prototype P1 for the DAQ and TCDS Hub ....................... 77

D Technical Infrastructure and Cooling .......................... 81
  D.1 Current Infrastructure in the Service Cavern (USC) ............. 81
  D.2 Parameters of the Current Surface Counting Room ............... 81
  D.3 Free Cooling for a Data Center in the Geneva Area ............. 82

Glossary of Special Terms and Acronyms ......................... 90

References .......................................................... 91
Chapter 1

Overview

The High Luminosity LHC (HL-LHC), after the third Long Shutdown (LS3), will provide an ultimate instantaneous luminosity of $7.5 \times 10^{34} \text{ cm}^{-2}\text{s}^{-1}$, 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 in the second half of 2026 (Fig. 1.1).

The upgraded detector will be read out at an unprecedented data rate of up to 50 Tb/s and an event rate of up to 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 7.5 kHz for offline processing and analysis. For comparison, the current running parameters of CMS are 2 Tb/s data rate, 100 kHz event rate, and a total rate to permanent storage of about 1 kHz.

This Interim Document describes the requirements and the principle baseline design of the Data Acquisition (DAQ) and High Level Trigger (HLT) systems for the Phase-2 CMS. The aim of the document is to show that the requirements can be met using readily available technology and at an affordable cost, discuss the timeline up to the Phase-2 DAQ and HLT TDR in Q2-2021, and indicate the areas of research and development being pursued to achieve an optimal cost/performance ratio for the final system.

Chapter 1 presents the requirements for the readout, processing, and storage of the Phase-2 detector data, taking into account the current status of the sub-detector upgrade design and development, as well as the current understanding of the underlying physics requirements.
Chapter 1. Overview

Chapter 2 discusses the baseline DAQ and HLT structure for 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 CPUs.

In Chapter 3 we examine the implications of the baseline design for the various components and discuss relevant emerging technologies that can help build a better, cheaper DAQ/HLT system.

In Chapter 4 the implications of the baseline design and the different optimization options with respect to hardware and infrastructure requirements for the online data center are analysed.

In Chapter 5 we discuss development directions that can either provide new interesting features, or simplify the DAQ/HLT design. Possible novel approaches to event building and online processing, inspired by technology trends in other areas of computing that deal with large volumes of data, are discussed. Finally, an addition to the baseline DAQ, called the “40 MHz scouting system”, is presented.

Chapter 6 provides an outline of the organization and schedule of the project, as well as an initial estimate of the cost.

1.1 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 injection chain. The resulting HL-LHC beam parameters are summarised in Table 1.1. In order to keep the number

Table 1.1: HL-LHC projected running parameters. The instantaneous luminosity and average pileup figures, expressed as the average number of interactions per crossing, ⟨PU⟩, are for levelled operation.

<table>
<thead>
<tr>
<th></th>
<th>⟨L⟩</th>
<th>⟨PU⟩</th>
<th>Vertex Density</th>
<th>∫L/year</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>5·10^{34} cm^{-2}s^{-1}</td>
<td>140</td>
<td>0.8/mm</td>
<td>250 fb^{-1}</td>
</tr>
<tr>
<td>Ultimate</td>
<td>7.5·10^{34} cm^{-2}s^{-1}</td>
<td>200</td>
<td>1.2/mm</td>
<td>&gt; 300 fb^{-1}</td>
</tr>
</tbody>
</table>

of interactions per crossing (pileup) limited to a value which results in the best possible data quality of the upgraded detector, the instantaneous luminosity will be adjusted by varying the beam focusing and/or overlap (this process is called “levelling”). 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 calorimeters. Some of the main physics foci of the HL-LHC physics program, including the precision study of the Higgs boson properties, also require extended coverage in the forward region\(^1\). A new silicon tracker with trigger capabilities and extended coverage [2], high granularity endcap calorimetry largely based on silicon sensors (Ref. [1], Section 3.5), and extended muon detector coverage [3] will enhance the CMS acceptance and selection power. To help distinguish particles originating from the interesting vertex, timing information will be recorded by a dedicated timing detector, and by the new endcap calorimeter, as well as

\(^1\)To enable tagging of Vector Boson Fusion (VBF) Higgs production [1].
the electromagnetic portion of the barrel calorimeter. Coping with beam conditions will also require the replacement of the front-end electronics of legacy detector components to improve radiation hardness and readout speed [3, 4]. A higher selectivity of the trigger system [5], and a higher readout rate and throughput, will also entail new strategies for the back-end electronics. The upgraded CMS detector will be read out at a rate of up to 750 kHz by approximately 50k high-speed front-end optical links, for an average expected total event size of approximately 7.5 MB.

### 1.2 Conceptual DAQ Design

The main purpose of the DAQ system is to provide the data pathway and time decoupling between the synchronous detector readout and data reduction, and the asynchronous selection of interesting events and their permanent storage for offline analysis, as illustrated in the cartoon of Fig. 1.2. The detector front-end (FE) and back-end (BE) are connected by bidirectional links.

![Diagram](https://via.placeholder.com/150)

**Figure 1.2:** Conceptual design of the overall CMS DAQ.

The downlinks are used by the BE to distribute the master clock, the accept signal generated by the trigger system, and fast signals to control the FE electronics (typically 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.

Following the present established design [6, 7], the CMS trigger and data acquisition system will continue to feature two trigger levels, with only one synchronous hardware-based first level (Level-1 trigger, L1) consisting of custom electronic boards and a second level, the High Level Trigger (HLT), selecting on full detector event data with software algorithms running asynchronously on standard processors.

Some sub-detectors may separately transfer reduced input data for the hardware trigger, while buffering full-resolution information in the front-end, waiting for an accept, while others sim-
ply stream data for every machine bunch-crossing (BX). This choice is largely dictated by the number of links required for one or the other option, and the corresponding power distribution and cooling requirements. 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 normally 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 channels, or from machine conditions.

The synchronous portion of the DAQ system distributes the master 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 (12.5 µs for the Phase-2 CMS detector).

Each BE module uses standardized DAQ firmware to transfer its accepted event data to a DAQ concentrator over an asynchronous point-to-point link, using a protocol with flow control. There, data are concentrated and then transported to the surface (Data To Surface, D2S) over high-speed links, using a standard protocol, into the memory of commercial I/O processors. To guarantee protection against congestion at the destination, a lossless protocol is preferred for the D2S, even though it requires enough buffer space at the source to allow resending 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 of an individual accepted event 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 timescale inherent to the HLT processing is of order 1 s, but the allowed latency can and should be larger (up to 1–2 min) to allow for the large service time fluctuations typical of a purely software selection algorithm.

Accepted events are stored locally before being assembled into larger dataset files for efficient long-term storage. This is usually carried out using some form of distributed or network storage attached to the same switched network used for event building. Finally, entire dataset files are transferred to central computing resources (tier0) over long-distance links, to be stored for the subsequent offline reconstruction, which can happen within days of the actual data-taking.

### 1.3 Detector Readout

The custom electronics for the Phase-2 upgrade is based on commonly developed components where possible. In particular, the high speed optical data transmission between sub-detector front-end and back-end systems will use the low power Gigabit Transmission (lpGBT) chipset and Versalink+ optical links (see Appendix A), which can operate at 5 or 10 Gb/s. The number
of links is determined by the geometry of the detector and/or the bandwidth required.

The modular electronics form factor of choice for the Phase-2 upgrade back-end electronics is the Advanced Telecommunications Computing Architecture (ATCA) standard [8]. The ATCA standard specifies the board format, crate backplane and communication fabric. The rationale for upgrading from the Phase-1 MicroTCA detector readout back-end electronics to ATCA is discussed in some detail in Appendix B. A strategy for the integration of possible legacy systems has not been agreed yet. This must include adapters to interface to the new trigger and clock distribution (see Section 2.2.2). It is expected that these points will be clarified by the time of the TDR.

1.3.1 Readout Strategy

The DAQ concentrator board exposes a standardized interface to the detector back-end electronics, with a simple point-to-point protocol, for ease of integration in the subdetector boards. The DAQ concentrator will aggregate data from one back-end ATCA crate and convert from the custom point-to-point protocol to a standard one, apt to transport data to the surface counting room. Based on today’s technology, a typical link to be used for DAQ data transportation to the surface counting room will run at a speed of 100 or 200 Gb/s, resulting from the bundling of 25 Gb/s or 50 Gb/s physical links. The requirements and design considerations for the DAQ concentrator board will be discussed in Section 2.1.1. For now, we assume that the total maximum output bandwidth per concentrator board over the DAQ link will be of order 1.2 Tb/s. Because of the wide variety of throughput density for the different sub-detectors, it is to be expected that a modular design with units of, for example, 400 Gb/s will be necessary to optimize the cost of the D2S infrastructure.

In what follows, the readout needs of the individual sub-detectors after the Phase-2 upgrade are discussed, in view of establishing as accurate an account as possible of the throughputs involved, of the number of back-end units, and of the total bandwidth required for the DAQ. Wherever relevant, the ultimate LHC scenario with \(\langle PU\rangle = 200\) is assumed for the estimate of the data volume, which is based on the simulation of the relevant number of minimum bias events superimposed to a \(\bar{t}t\) “signal” event, without Level-1 selection. Note that for all legacy detectors, the effect of the Level-1 selection in biasing towards higher pile-up is irrelevant, either because all channels are read out for every crossing (as is the case for the barrel calorimeters), or because the occupancy is very low (as is the case for the muon detectors). In the remaining cases, this effect is expected to be small. On the one hand, the Phase-2 Level-1 is designed to produce a selection as independent of pile-up as possible. On the other hand, the event sizes are dominated by the large minimum bias multiplicity. The possible remaining small bias will be studied in the final TDR using the full simulation of the Level-1 trigger.

1.3.2 Outer Tracker

The new Outer Tracker (OT) (Ref. [2], Chapter 3) uses a total of 7680 silicon modules with strip-strip (2S) sensors and 5616 silicon modules with strip-macro-pixel (PS) sensors. The 2S module uses one lpGBT link running at 5 Gb/s, whereas the PS module uses one lpGBT link running at 5 or 10 Gb/s for the innermost modules. The same lpGBT links are used to transport both the trigger input data (stubs) at the bunch crossing rate, and the DAQ data (sparsified clusters) for the Level-1 accepted events. It is estimated that about 80% of the OT bandwidth will be dedicated to trigger data.

The OT back-end board, the Data, Trigger and Control (DTC)\(^2\) board, will merge up to 72

\(^2\)See Section 3.2.4.3 in Ref. [2].
lpGBT links. The exact way the merging of links from the different parts of the detector will be organized in order to balance the data throughput is still under development. A possible configuration organizes the readout in nine 40° sectors, each with one crate with twelve 2S DTCs and one crate with twelve PS DTCs. A preliminary simulation study estimated the output of the DTC board to the DAQ in the range of 10–90 Gb/s, depending on which part of the detector is read out by the DTC. The OT total sub-event size is estimated at 1.15 MB.

The Level-1 Track Trigger (TT) will also produce data for the DAQ. As this is an entirely new system, its actual architecture is not yet completely finalized. Three approaches are described in Section 3.5 in Ref. [2]. For the following, it is assumed that the system will be distributed over 18 crates, and that the total amount of data produced for the DAQ will be on average 10 kB per event, depending on the rate of readout of intermediate TT products, read out for diagnostic purposes. The final TT products are expected to contribute only 1 kB per event, corresponding to about 70 Level-1 tracks on average. The event record of the intermediate TT products will likely be larger than the remaining 9 kB, but the format and content are not yet firm. Since its readout will be pre-scaled according to needs, the total size should be regarded as an educated guess, good to the order of magnitude, but will in any case remain very small compared to the total event size.

### 1.3.3 Inner Tracker

The new Inner Tracker (IT) (Ref. [2], Chapter 4) will consist of approximately 10k pixel chips feeding 1260 lpGBT (10 Gb/s) links. As is the case with the present pixel detector, only the data of the Level-1 triggered events are transmitted to the back-end system. It is planned for the IT to accommodate 72 links per back-end board. The back-end board will, to the extent possible, be based on the same building blocks as the outer tracker DTC, discussed in the previous section. A crate with 6 pixel DTC boards can accommodate the readout of a quarter of the IT, for a total of 4 crates with 24 DTC boards. The total anticipated output bandwidth required per pixel DTC will be half the peak input bandwidth, as lpGBT readout links are not used at full bandwidth and some data compression can be implemented by modern high performance FPGAs. This yields an output throughput per DTC of $72 \times \frac{10}{2} = 360$ Gb/s. The exact compression factor achievable is not yet specified, and the cost of uncompressing data in the HLT must be evaluated. A preliminary estimate from simulation of the average sub-event size is approximately 1.4 MB.

### 1.3.4 MIP Timing Detector (MTD)

The new MIP Timing Detector (MTD) [9] will consist of a Barrel Timing Layer (BTL) and an Endcap Timing Layer (ETL). The BTL comprises 258,048 LYSO:Ce crystals with Silicon Photomultiplier (SiPM) photo-sensors. The two endcap ETL sections will consist of low gain avalanche silicon sensors with 2624 modules each containing 1536 pads. The architecture of the readout electronics is currently under study. One option is to implement the L1A data filtering in the front-end chip, and only transmit the Level-1 triggered events to the back-end electronics. With some level of data concentration at the front-end, 1000 lpGBT (10 Gb/s) links are required for the BTL, and a total of 438 lpGBT (10 Gb/s) links for the ETL. A possible ATCA

---

3These large variations are due predominantly to the different number of charged tracks per unit volume in the different layers of the detector. The relative spread does not depend on the exact value of the pileup.

4The OT event record includes the sparsified clusters and the stubs sent to the track trigger (about 10% compared to the cluster data).

5See Section 4.3 in Ref. [2].

6See Sections 2.1.3 and 2.2.3 in Ref. [9].
back-end system would use BE boards with 100 lpGBT input channels, and hence consist of one crate for the BTL with 11 boards and one crate for the ETL with 5 boards. The expected total sub-event size from simulation is about 24 kB for the BTL and about 36 kB for the ETL.

### 1.3.5 ECAL Barrel Calorimeter (EB)

The barrel electromagnetic calorimeter (ECAL) consists of 61 200 lead tungstate crystals with Avalanche Photo-Diodes (APD) photo-sensors. The EB front-end and back-end electronics will be replaced for the Phase-2 upgrade (Ref. [4], Section 3). The crystals are read out in groups of 25, resulting in approximately 2500 groups. The APDs will be read out at 160 MHz (oversampling is necessary to provide some rejection power against large pileup\(^7\)) and each readout group will require four lpGBT (10 Gb/s) links at about 90% bandwidth usage, or a total of 10 000 links for the whole detector\(^8\). The back-end board, called the Barrel Calorimeter Processor (BCP)\(^9\), will merge about 96 lpGBT links, for a total of 108 boards in twelve crates. The BCP will perform data buffering and trigger primitive generation. Zero suppression and selective readout functions could be implemented if necessary, but are not being considered in the baseline design. This will result in an EB sub-event\(^10\) (fixed) size of about 1.6 MB, independent of pileup.

### 1.3.6 HCAL Barrel Calorimeter (HB)

The barrel hadronic calorimeter (HCAL) is a sampling calorimeter that uses a brass absorber and plastic scintillator as the active material. Currently HPDs are used as photo-sensors.

The Phase-1 upgrade of the Hadron calorimeters [10] includes increasing the depth segmentation of the calorimeter, the replacement of the photo-sensor devices, and the replacement of the front-end readout, as well as the back-end electronics. A new optical link between FE and BE operates at 5 Gb/s and is based on the Versatile Link (VL) transceiver (see Appendix A) and a front-end radiation-tolerant FPGA.

As the first step of the Phase-1 upgrade of the HB, the back-end electronics has been replaced in 2015 by a MicroTCA based system, described in Section 6 in Ref. [10]. The Phase-1 upgrade will be completed during LS2 by the replacement of the HPDs with SiPMs, and the replacement of the front-end electronics.

The Phase-2 upgrade of the HCAL is described in Section 1.9 in Ref. [4]. Regarding the HB, the front-end electronics and the 1152 front-end optical links from the Phase-1 upgrade will remain unchanged for Phase-2. However, the Phase-2 HB back-end system\(^11\) will be upgraded for the higher Level-1 rates, and use the same BCP boards as the EB, described in the previous section. The ATCA back-end system will consist of 18 BCP boards arranged in two crates. Zero suppression is not being considered in the baseline design. This will result in an HB sub-event\(^12\) (fixed) size of 0.24 MB, independent of pileup.

---

\(^7\) And suppression of anomalous APD signals.

\(^8\) Including some channels used for transparency monitoring and calibration.

\(^9\) See Section 3.8 in Ref. [4].

\(^10\) The EB event record includes the readout of 5 bunch crossings for the data and 1 bunch crossing for the Level-1 primitives.

\(^11\) See Section 3.9 in Ref. [4].

\(^12\) The HB event record includes the readout of 8 bunch crossings for the data and 4 bunch crossings for the Level-1 primitives.
1.3.7 Outer Barrel (HO) Hadronic Calorimeter

The HPDs of the Outer Hadronic (HO) calorimeter were replaced with SiPMs during LS1. According to the current plan, the on-detector electronics will neither undergo a Phase-1, nor a Phase-2 upgrade. The 768 links from the legacy on-detector electronics are operated at a relatively low 1.6 Gb/s speed. The off-detector electronics, which is currently still the original VME-based electronics, is planned to be upgraded to MicroTCA before or during LS2. At this stage, there is no plan to replace the MicroTCA based back-end electronics for Phase-2. It is assumed that data from HO will be aggregated using an ATCA based adapter board. Zero suppression is not being considered in the baseline design. This will result in an HO sub-event (fixed) size of 30 kB, independent of pileup.

1.3.8 Forward (HF) Hadronic Calorimeter

The Forward Hadronic (HF) calorimeter is a Cherenkov calorimeter based on steel absorber and quartz fibers, using Photomultiplier Tubes (PMTs) as photo-sensors. The Phase-1 upgrade of the HF (Ref. [10], Chapter 4) has been completed in LS1. The HF detector will not be modified for Phase-2. At the time of writing, there is no plan to replace the MicroTCA based back-end electronics either, but to extend the system to cope with the higher Level-1 rates. It is assumed that data from HF will be aggregated using an ATCA based adapter board. Zero suppression is not being considered in the baseline design. This will result in an HF sub-event (fixed) size of 60 kB, independent of pileup.

1.3.9 Endcap Calorimeter

The new integrated endcap calorimeter (Ref. [1], Section 3.5) will encompass three sections:

- the electromagnetic section, EE, consisting of 28 layers of silicon sensors and brass absorber;
- the front hadronic section, FH, consisting of 12 layers of silicon sensors and steel absorber;
- the backing hadronic section, BH, consisting of an inner portion, with 12 layers of silicon sensors, and an outer portion with 12 scintillator layers read out by SiPMs.

The front-end is organized in around 70 'panels' per layer, each reading out 6 or 12 sensors. Each panel will have at least one lpGBT 10 Gb/s link or a total of around 8000 links for the 104 detector layers. The front-end electronics acquires the data with relatively low threshold in order to be sensitive for a MIP, and transmits the Level-1 accepted events to the back-end. The event size is expected to be approximately 2 MB, so the total DAQ readout throughput is about 16 Tb/s. This implies that the lpGBT DAQ links will be used at about 25%. The baseline back-end board could be able to concentrate 96 input links. Hence, the minimum number of BE boards is 84. A possible configuration uses 9 crates with 12 back-end boards each and two DAQ concentrator cards per crate.

One in two detector layers is used for triggering purposes. The trigger data are acquired with a higher threshold and transmitted over separate lpGBT (10 Gb/s) links to the Trigger Primitive Generation (TPG) back-end. The architecture of the TPG system is currently under study. A possible configuration consists of two stages. The first stage with 96 TPG boards in 8 crates receives data from the front-end and forms 2D-clusters. The second stage with 48 TPG boards

---

13Note that the Phase-0 HO on-detector electronics uses the TTCrx ASIC.
14The HO event record includes the readout of 8 bunch crossings for the data and no Level-1 primitives.
15The HF event record includes 3 bunch crossings for the data and 2 bunch crossings for the Level-1 primitives.
in 4 crates forms 3D-clusters, and transmits them to the Level-1 trigger system. A preliminary estimate of the EC-TPG average sub-event size is 0.2 MB, depending on the rate of full-readout events.

1.3.10 Muon Drift Tubes (DT) system

The Drift Tubes (DT) muon system consists of 250 DT gas wire chambers. The scope of the Phase-2 DT upgrade (Ref. [3], Chapter 3) includes the replacement of the on-detector electronics and its receiving back-end electronics, reorganizing the role of both parts in the trigger and readout chains. This is scheduled for LS3. The DT muon system will require 3840 lpGBT links, the number of links being dominated by requirements related to the geometry of the detector and not the bandwidth. The provisional back-end layout\(^\text{16}\) consists of 5 crates with 12 boards (48 lpGBT inputs) each, to cover all 60 DT sectors, and 2 crates with 12 BE boards (40 lpGBT inputs) each, to cover the 12 DT wedges. In addition, a single crate with 12 boards is used to provide the DT trigger primitives to the Level-1 trigger. This crate needs also to be read out by the DAQ. The total sub-event size is expected to be approximately 130 kB.

1.3.11 Muon Cathode Strip Chamber (CSC) system

The Cathode Strip Chamber (CSC) system consists of 540 trapezoidal modules arranged into four endcap disks. The scope of the Phase-2 CSC upgrade (Ref. [3], Chapter 4) includes the selective replacement of various modules in the chain comprising on-chamber electronics to off-chamber electronics to back-end electronics, in order to provide adequate throughput for the higher Level-1 rates. The module replacement is scheduled for LS2 and LS3. The CSC back-end electronics will handle 900 optical links from the off-chamber electronics at speeds up to 6.4 Gb/s. The back-end layout is not yet finalized. However, it is expected that the final concentrator modules implementing the DAQ interface will be arranged in two ATCA crates, one per endcap. The expected event size is approximately 0.2 MB\(^\text{17}\).

1.3.12 Muon Gas Electron Multiplier (GEM) system

The new Gas Electron Multiplier (GEM) muon system (Ref. [3], Chapter 6) will encompass three parts:

- the GE1/1 system \([11]\), consisting of 144 GE1 chambers, foreseen for installation in LS2
- the GE2/1 system, consisting of 72 GE2 chambers, foreseen for installation before or during LS3
- the ME0 system, consisting of 216 ME0 modules, foreseen for installation during LS3

The GE1/1 front-end is organized with one Opto-Hybrid (OH) board per chamber, with three GBT (4.8 Gb/s) DAQ links each. The trigger data are transmitted over separate optical links. The back-end electronics connects to 432 GBT links for the DAQ in total. The baseline plan is to use MicroTCA electronics, with 12 boards in total. We assume for now that the data from this MicroTCA system will be aggregated by a separate ATCA based adapter or concentrated by the ATCA based GE2/1 system. The total GE1/1 sub-event size is expected to be small (2 kB, using zero-suppression).

The GE2/1 front-end is organized with two OH boards per chamber, with one lpGBT DAQ link each. The trigger data are transmitted over separate optical links. The ATCA based back-end

\(^{16}\)See Section 3.5.3 in Ref. [3].

\(^{17}\)See Section 4.4.5 in Ref. [3].
electronics connects to 144 lpGBT links in total, and consist of 8 BE boards in a single crate. The total GE2/1 sub-event size is expected to be small (1 kB, using zero-suppression).

The ME0 front-end is organized with one OH boards per module, with one lpGBT DAQ link. The trigger data are transmitted over two separate lpGBT links. The ATCA based back-end electronics consist of 12 BE boards in a single crate. The total ME0 sub-event size is expected to be approximately 118 kB, using zero-suppression.

1.3.13 Muon Resistive Plate Chamber (RPC) system

The Phase-2 upgrade of the Resistive Plate Chamber (RPC) system (Ref. [3], Chapter 5) includes a replacement of the electronics. However, the RPC will only be connected to the Level-1 system, as the DAQ sub event record data would be identical to the RPC part of the Level-1 data record.

1.3.14 Level-1 Trigger

In order to estimate the total data size and number of links required for the readout of the Level-1 system, we considered the different components involved in processing the L1 decision, from the detector primitives to the final global decision. Because of the special treatment of primitive data by the different sub-detectors, the L1 tracks and the primitives from the endcap calorimeter are considered separately (see the corresponding detector sections). Only the Barrel Calorimeter and the Barrel and Endcap Muon Trigger systems, the Correlator, which combines calorimeter and muon information with tracks, and the global trigger, forming the final decision, are considered in the throughput estimate. The number of boards (120) and crates (14) are only intended for use in Chapter 2 to guide the estimate of bandwidth requirements per crate, and should not be regarded as final. In this estimate, it is assumed that the intermediate information used as input by each processor board will be stored for Level-1-accepted events. Hence, the estimated output throughput from the preceding level is scaled by a factor 0.75/40 for the purpose of DAQ readout (750 kHz L1 accept out of the 40 MHz crossing rate). A total event size of 150 kB is estimated for the L1 system from these assumptions.

1.3.15 Luminosity and Beam Background Detectors (BRIL)

The BRIL (Beam Radiation Instrumentation and Luminosity) project deploys currently three online luminosity measurement systems [12]: the Fast Beam Conditions Monitor (BCM1F) [13] and the Pixel Luminosity Telescope (PLT) [14], both installed inside the pixel detector volume, and the Forward Hadron calorimeter (HF-lumi). All these detectors perform bunch-by-bunch measurements, using the deadtime-free BRILDAQ readout architecture [15] that is independent of the Level-1 accept driven DAQ, discussed so far. In this architecture, data are captured every bunch crossing and accumulated in histograms in hardware. These histogram counters are accumulated over a configurable number ($2^{12} - 2^{14}$) of LHC orbits and read out at these intervals. The synchronisation of the three luminometers relies on the BRILDAQ synchronization scheme introduced in the Phase-1 TCDS (see Section 2.2.1).

It is currently planned to perform the luminosity measurement for Phase-2 with the following systems:

1. A new “BCM1F-like” detector at a different location, operating in the same way as the current BCM1F detector;

2. The current Forward Hadron calorimeter (HF-lumi), as it is currently operated;
3. The extension (TEPX) of the Inner Tracker (see Section 10.3 in Ref. [2]). Data will be read out from the front-end once triggered by either a regular Level-1 trigger or a dedicated luminosity trigger. Clusters are formed in the back-end electronics, and the data corresponding to a luminosity trigger are recognised. The number of pixel clusters and coincidences of pixels are then recorded separately per bunch crossing, and histogrammed in hardware, similar to the scheme used for systems 1–2.

The new scheme with luminosity triggers requires functionality beyond the current TCDS system and will be discussed in Section 2.2.2. During regular data taking at nominal pileup, the pixel luminosity triggers will be restricted to be not more than 10% over the nominal Level-1 trigger rate (750 kHz) in order to remain within the bandwidth of the Inner Tracker optical links.

The event (non-histogrammed) data will also be read out and recorded. This is necessary for offline monitoring and efficiency studies. The data volume from TEPX is approximately 0.6 MB/event at the luminosity trigger rate.

1.3.16 Contingency for Readout of Additional Detectors

At the time of writing, it is still unclear if additional sub-detectors will be included in the read-out of the Phase-2 CMS. One such example is the CMS/TOTEM Precision Proton Spectrometer (CT-PPS). Such additions are expected to contribute a small amount of data and are addressed as contingency in the baseline design of Chapter 2.

1.3.17 DAQ Link Count

The projected data throughput for each sub-detector and number of back-end boards and crates are summarized in Table 1.2. Both the total throughput and the total event size result from the simple addition of the individual estimated average contributions from each sub-detector. The total throughput can be used to estimate the minimum number of D2S links required in a perfectly optimized arrangement. Using 100 Gb/s as the unit link speed, 440 links would be required. Since it will not be possible to perfectly balance the throughput from areas of different occupancy, and because of the link speed quantization, the actual number of links will be larger. Chapter 2 contains a detailed discussion for the purpose of estimating the Bill of Material.

1.4 Event Network and Intermediate Buffer

The event network, as introduced in Section 1.2, provides the interconnection to perform event building by assembling fragments of an individual event in the memory of a single processor. From Table 1.2, the total average throughput requirement for the event network is 44 Tb/s. The total required bandwidth must allow for a switch efficiency that can be of 30 to 50% depending on the configuration. Hence, the total event network bi-sectional bandwidth required will be between approximately 90 and 150 Tb/s.

The DAQ system must also include a certain amount of intermediate buffering memory, necessary for the system to absorb fluctuations in the HLT service time, in particular when loading 18These figures are based on experience with HPC interconnects in the DAQ system of Run-1 (Myrinet) and Run-2 (Infiniband-FDR). In particular, with the current DAQ2 event builder, 30% corresponds to the efficiency achievable with the event builder traffic without any optimization. After extensive optimization, and exploiting the mostly unidirectional nature of the event builder traffic, efficiencies of order 60% were obtained. It is anticipated that optimisation in a folded event builder architecture will only partially match the DAQ2 results.
Table 1.2: CMS Phase-2 detector projected data links, ATCA back-end configuration and event size summary. Data are obtained from the technical design reports wherever possible. Average throughput estimated from event size assuming 750 kHz Level-1-accept rate.

<table>
<thead>
<tr>
<th>Subdetector</th>
<th>Front-end</th>
<th>Sub-event size (MB)</th>
<th>Back-end boards</th>
<th>Back-end crates</th>
<th>Average throughput (Tb/s)</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>lpGBT links$^a$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Outer Tracker</td>
<td>13 000</td>
<td>1.15</td>
<td>216</td>
<td>18</td>
<td>6.90</td>
<td>(6)</td>
</tr>
<tr>
<td>Track Trigger</td>
<td></td>
<td>0.01</td>
<td></td>
<td>18</td>
<td>0.06</td>
<td></td>
</tr>
<tr>
<td>Inner Tracker</td>
<td>1 260</td>
<td>1.44</td>
<td>24</td>
<td>4</td>
<td>8.64</td>
<td>(5)</td>
</tr>
<tr>
<td>MIP Timing Det. - BTL</td>
<td>1 000</td>
<td>0.02</td>
<td>11</td>
<td>1</td>
<td>0.14</td>
<td></td>
</tr>
<tr>
<td>MIP Timing Det. - ETL</td>
<td>438</td>
<td>0.04</td>
<td>5</td>
<td>1</td>
<td>0.22</td>
<td></td>
</tr>
<tr>
<td>ECAL Barrel</td>
<td>10 000</td>
<td>1.58</td>
<td>108</td>
<td>12</td>
<td>9.49</td>
<td></td>
</tr>
<tr>
<td>HCAL Barrel</td>
<td>other</td>
<td>0.24</td>
<td>18</td>
<td>2</td>
<td>1.45</td>
<td></td>
</tr>
<tr>
<td>HCAL - HO</td>
<td>legacy</td>
<td>0.03</td>
<td>-</td>
<td>1</td>
<td>0.18</td>
<td>(4)</td>
</tr>
<tr>
<td>HCAL - HF</td>
<td>other</td>
<td>0.06</td>
<td>-</td>
<td>1</td>
<td>0.36</td>
<td>(6)</td>
</tr>
<tr>
<td>Endcap CALO</td>
<td>8 000</td>
<td>2.00</td>
<td>108</td>
<td>9</td>
<td>12.00</td>
<td></td>
</tr>
<tr>
<td>Endcap CALO TPG</td>
<td>9 000</td>
<td>0.20</td>
<td>144</td>
<td>12</td>
<td>1.50</td>
<td></td>
</tr>
<tr>
<td>muon DT</td>
<td>3 840</td>
<td>0.13</td>
<td>84</td>
<td>8</td>
<td>0.78</td>
<td></td>
</tr>
<tr>
<td>muon CSC</td>
<td>other</td>
<td>0.20</td>
<td>-</td>
<td>2</td>
<td>1.20</td>
<td>(4)</td>
</tr>
<tr>
<td>muon GEM - GE1/1</td>
<td>other</td>
<td>0.002</td>
<td>-</td>
<td>1</td>
<td>0.01</td>
<td>(4)</td>
</tr>
<tr>
<td>muon GEM - GE2/1</td>
<td>144</td>
<td>0.001</td>
<td>8</td>
<td>1</td>
<td>0.01</td>
<td></td>
</tr>
<tr>
<td>muon GEM - ME0</td>
<td>216</td>
<td>0.12</td>
<td>12</td>
<td>1</td>
<td>0.71</td>
<td></td>
</tr>
<tr>
<td>muon RPC</td>
<td>other</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>(6)</td>
</tr>
<tr>
<td>Level1</td>
<td></td>
<td>0.15</td>
<td>120</td>
<td>14</td>
<td>0.90</td>
<td>(6)</td>
</tr>
</tbody>
</table>

Total: 7.4 >858 >106 44

$^a$Count of links based on lpGBT-VL+ (see Appendix A), “Other” links include the use of VL(+) components with radiation-tolerant FPGA at the front-end.

$^b$The Outer Tracker uses the same links to multiplex trigger and DAQ data — the proportion of trigger data is expected to be roughly 80%.

$^c$Assumes some compression factor is achievable in the back-end board, see text of Section 1.3.3.

$^d$Legacy MicroTCA back-end, assume ATCA adapter.

$^e$Only connected to Level-1 system.

$^f$The number of crates is estimated from the number of boards in the current Level-1 Trigger Interim Document draft, the throughput from the estimated output bandwidth from each stage scaled to 750 kHz.
1.5 Online Selection and Permanent Storage

The maximum accept rate to permanent storage must take into account the expected physics needs and the offline processing capabilities. These are discussed in some detail in the CMS Technical Proposal [1] and essentially call for a rejection factor of about 100 between the Level-1 accept rate and the HLT accept rate. The requirements on the HLT CPU power and the permanent storage throughput are derived in what follows assuming this rejection factor.

1.5.1 HLT Processing Requirements

The HLT selects and classifies events based on full detector data with software algorithms running on “standard” processors. The timing of the HLT has been analyzed since the start of data taking. The breakdown of the CPU time currently spent in the reconstruction code is approximately: 10% in ECAL calorimeter, 10% in HCAL calorimeter, 35% in tracking and 20% in particle flow.

HepSpec06 (HS06) [16] is a widely adopted benchmark for generic HEP applications, believed to reflect well the performance of different processors on HEP software. The performance of the HLT across different generations of HLT processors used in CMS is found to be proportional, within 10%, to the corresponding HepSpec06 (HS06) [16] benchmark figure. Since HS06 is a convenient metric to estimate the projected performance of general purpose servers within the Phase-2 timescale, and for ease of comparison with other experiments, CPU requirements are therefore quoted in HS06 in what follows. In case some of the HLT processing could be provided in the future by alternative architectures, e.g. by co-processor offloads, it is proposed to break down the required HLT resources into the HEPSpecs provided by general purpose CPUs, and the additional CPU power provided by co-processors, expressed in HEPSpec equivalent.

It is furthermore observed that the HLT processing time scales with pile-up for a fixed L1 conditions and calibrations. For the Phase-2 parameters, this requires 333 TB of distributed buffering, assuming a buffering capability of about 60 s of data taking. A more detailed discussion of event network buffering is deferred to Chapter 2.
selection, within the range currently accessible in Run-2 (see Fig. 1.3).

![Figure 1.3: Average HLT CPU time per event as a function of pile-up during 2017 data taking.](image)

The analysis of the HLT processing time in 2017, over a PU range up to 46, as shown in Fig. 1.3, can be used to produce a linear extrapolation of the CPU time per event required for the two PU scenarios of Phase-2. These translates into an estimate of the required CPU power of 4.5 and 9.2 MHS06 respectively for the $\langle PU \rangle = 140$, 500 kHz Level-1 and $\langle PU \rangle = 200$, 750 kHz Level-1 scenarios, as reported in Table 1.3. These figures should be considered valid with a substantial uncertainty. This is due partly to the uncertainty on the extrapolation and on the accuracy with which the benchmark reflects the actual HLT performance on a given processor. Furthermore, the Phase-2 detector and the Level-1 trigger are significantly different from the Phase-1. On the one hand the added complexity of the reconstruction for the endcap calorimeter has not been taken into account. On the other hand, the tracks found above $p_T = 2$ GeV by the Level-1 trigger, and possibly the results of particle flow at Level-1, will save processing time at the HLT. Finally, possible gains may also be expected from the continued improvement of the software.

Better estimates based on full simulation, Phase-2 reconstruction code, and a strawman HLT menu are expected to become available at the time of the TDR.

### 1.5.2 Storage Requirements

For the storage throughput estimate, the aforementioned accept rate and average event size are used. A factor 1.1 is applied to account for the additional data from triggers used for calibrations, monitoring and diagnostic.

---

24 On a Xeon E5-2680v4 CPU core.
25 It is to be noted that particle flow algorithms have been used in the HLT since Run-1. Particle flow was initially introduced in 2011 in a small subset of triggers, with its use expanded in 2012 and further expanded and optimized in Run-2. The extrapolation shown in Fig. 1.3 is based on a production HLT menu that uses it.
26 In the current CMS DAQ system, data from triggers used for calibrations, monitoring and diagnostic contribute an additional 25% to the nominal throughput. Some of these data streams require a constant rate, and their size...
1.5.3 “Opportunistic Cloud” on the HLT Infrastructure

Currently, the HLT farm is used outside of stable beams periods as a “cloud” resource to process standard CMS offline workflows (reconstruction, Monte Carlo) [17]. This is done in technical stop periods as well as during inter-fills, and significantly contributes to the CMS offline processing capacity. The use of HLT CPU for offline processing constitutes an important optimization of resources and is expected to be continued in both Run-3 and Phase-2.

1.6 Heavy Ion Operation in Phase-2

The HL-LHC will continue to be able to operate in PbPb collider mode. The PbPb interaction rate at the LHC center-of-mass energy is about 50 kHz, and the average PbPb interaction corresponds to about 200 pp pile-up, with central collisions peaking at 300. Since central collisions only represent some 10% of the total, and taking into account the Phase-2 readout scheme of the various sub-detectors, only the average size and the total bandwidth are relevant. Hence, the entire 50 kHz raw event rate can in principle be fed to the HLT, assuming just a simple minimum bias Level-1 trigger. As the installation of a storage system capable of higher bandwidth than required for pp operation is not currently foreseen, the HLT output rate will be limited to about 7 kHz. Detailed studies are needed, using the full Phase-2 simulation and reconstruction, to devise a Level-1 and HLT strategy to select heavy-ion events to be recorded. In this context, the use of Level-1 in tagging mode to seed the HLT is particularly relevant.

1.7 Summary

The Phase-2 requirements of the CMS Trigger and DAQ are summarised in Table 1.3 for two different pileup scenarios. For the rest of this document, only the peak $\langle PU \rangle = 200$ scenario will be considered.

Table 1.3: CMS Phase-2 trigger and DAQ projected running parameters, compared to the design values of the current Run-2 system.

<table>
<thead>
<tr>
<th>CMS detector</th>
<th>LHC Run-2</th>
<th>HL-LHC Phase-2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Peak $\langle PU \rangle$</td>
<td>60</td>
<td>140</td>
</tr>
<tr>
<td>L1 accept rate (maximum)</td>
<td>100 kHz</td>
<td>500 kHz</td>
</tr>
<tr>
<td>Event Size</td>
<td>2.0 MB $^a$</td>
<td>5.7 MB $^b$</td>
</tr>
<tr>
<td>Event Network throughput</td>
<td>1.6 Tb/s</td>
<td>23 Tb/s</td>
</tr>
<tr>
<td>Event Network buffer (60 seconds)</td>
<td>12 TB</td>
<td>171 TB</td>
</tr>
<tr>
<td>HLT accept rate</td>
<td>1 kHz</td>
<td>5 kHz</td>
</tr>
<tr>
<td>HLT computing power $^c$</td>
<td>0.5 MHS06</td>
<td>4.5 MHS06</td>
</tr>
<tr>
<td>Storage throughput</td>
<td>2.5 GB/s</td>
<td>31 GB/s</td>
</tr>
<tr>
<td>Storage capacity needed (1 day)</td>
<td>0.2 PB</td>
<td>2.7 PB</td>
</tr>
</tbody>
</table>

$^a$Design value.

$^b$Obtained by scaling the Event Size at $\langle PU \rangle = 200$ with pile-up (140/200), except for sub-detectors with fixed size readout.

$^c$Does not include Data Quality Monitoring.

is not expected to increase with pileup. Given the uncertainty on new requirements from upgraded detectors, however, a margin of 10% is maintained.
Chapter 2

Phase-2 DAQ Baseline Design

The baseline architecture of the DAQ is sketched in Fig. 2.1. The optical links from detector front-ends are aggregated in detector-dependent back-end boards. A DAQ and TCDS Hub (DTH), described in more detail in Section 2.1, 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, connecting the DTH output to the I/O servers responsible for 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 the I/O servers themselves, 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 servers via inexpensive small switches, for processing, in a way similar to the current scheme [18], and finally transferred to a cluster filesystem for storage, as described in Ref. [19]. The overall design of the baseline event builder and the HLT infrastructure remains very similar to the one
described in Ref. [7].

In the following sections we examine the implications of the baseline design for the various components.

### 2.1 Data To Surface

As outlined in Fig. 1.2 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 TTC signals, and generate a TTS state based on the internal buffer occupancy. 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 Outer Tracker) the same back-end board will be responsible for the routing of the trigger data (received over the same links as those for the DAQ) to the different destinations. The front-end uplinks will mostly be serial point-to-point bidirectional optical links running at 5–10 Gb/s (mostly lpGBT, see Appendix A). Some detectors will stream the entirety of their readout directly to the backend at the bunch crossing rate, rather than buffering them at the front-end waiting for a L1 accept. In this case, data will be forwarded to the L1 processors and buffered at the back-end. Upon a L1 accept, the relevant subset will be handed over to the DAQ.

The purpose of the Data To Surface system (D2S) is to aggregate several input streams from individual back-end boards, convert the aggregated streams into a standard switched network protocol, and route the data from the service cavern to the surface counting room and into the memory of the servers forming the event builder. Time decoupling from the synchronous readout in the back-end boards requires sufficient local buffering to enable the correct handling of congestion on the destination end. Irrespective of the exact implementation of the aggregation and conversion, the D2S network will be organised as illustrated in Fig. 2.1, with a minimal amount of redundancy obtained by using small Top Of the Rack (TOR) switches (D2S routers) allowing the redirection of D2S links to different I/O processors.

The baseline D2S links, with a bandwidth of 100 Gb/s and a range of 300 m, can be implemented using existing standard commercial fibres and transceivers with four 25 Gb/s lanes. Higher speed lanes, not necessarily reducing the total cost, may become available by the time of installation.

In the Phase-1 upgrade, MicroTCA based back-end systems were introduced (see Appendix B), where the crate-level data aggregation, the TTC distribution, and the collection of the TTS states are combined in a single board [20]. The board is connected to the Trigger Control and Distribution System (TCDS) using a dedicated optical link to a Partition Interface (PI) board, distributing TTC signals and collecting TTS status from multiple crates in a partition. The aggregated data from one crate are moved, using an ad-hoc protocol, over a point-to-point optical link to a separate board (Fed Readout Optical Link, FeROL [21]), where a simplified FPGA implementation of the TCP/IP protocol is used to transfer data to the surface over up to four 10 Gb/s links. The Phase-2 D2S system will aim to further integrate the two functions. An immediate

---

1These rad-hard links have been developed specifically for the readout of front-ends for the HL-LHC upgrade. They do not provide symmetric up- and down-links in terms of bandwidth, and correspond to the requirements of most of the detector upgrades. These links are not part of the “central” DAQ system of CMS, but belong to the sub-detector projects. They should not be confused with the “Data-to-Surface” links from the DAQ and TCDS Hub board (TCDS) to the surface data center, discussed later on.

2Specifically using 4 × 25 G Coarse Wave Division Multiplexing (CWDM4) transceivers and a single fibre pair (Single Mode Fibre, SMF). The use of CWDM potentially reduces the cost of the fibre installation between underground and the surface, with respect to the use of Multi Mode Fibers.
advantage of this integration will be that data aggregation and TTC/TTS will map to the same tree, thus easing testing and diagnostics: the full granularity information on the status of the individual back-end boards and the DAQ links will be immediately accessible for monitoring, and it will be possible to emulate triggered acquisition directly from the concentrator cards. The integration will also add further flexibility and functionality to the TTC/TTS system. The latter will be discussed in more detail in Section 2.2. Finally, a single DAQ and TCDS Hub with a standard interface to the detector back-end electronics will be easier to upgrade, e.g. to implement a more powerful D2S protocol, something which is likely to be necessary over the lifetime of the Phase-2 experiment.

The following sections discuss a possible implementation of the aggregation and conversion layer that matches the requirements of the different sub-detectors in terms of their expected throughput per board and per crate, as outlined in Table 1.2.

### 2.1.1 DAQ and TCDS Hub

In consideration of the structure of the ATCA crate, discussed in detail in Appendix B, it is natural to consider a standardized D2S board to be located in one or both of the hub slots.

The DAQ and TCDS Hub (DTH) will combine the following functionality:

- Data aggregation for optimal DAQ link usage;
- Translation to a standard protocol (TCP/IP);
- Distribution of Timing signals and Control to each individual back-end board in the crate;
- Collection and (pre)processing of the individual board status for fast monitoring and statistics.

The DTH aggregates one or multiple input data streams from each individual back-end board in an ATCA crate via dedicated links connected on the front panel, performs data buffering and protocol translation, and provides output links to be routed to the surface (Fig. 2.2).

The timing and control signals (i.e. LHC clock, L1 accept, resync, reset, etc.) are received by the DTH from the TCDS over an optical link, and distributed through the backplane to all node slots. Finally, each node slot sends its TTS status along with monitoring data to the DTH over the backplane. The DTH, in turn, elaborates a global status for use in higher level logic to generate trigger throttling as necessary. Monitoring data are also aggregated/(pre)processed if needed, and made available for debugging, diagnosis and statistical analysis.
2.1.2 DTH implementation

As seen in Table 1.2, the average data throughput per sub-system ranges from (less than) 0.1 Tb/s up to 12 Tb/s. In order to define the optimal parameters of a DTH board, Table 2.1 illustrates the requirements in terms of throughput per crate and per individual board.

<table>
<thead>
<tr>
<th>SubDet</th>
<th>Avg throughput (Tb/s) per BE crate</th>
<th>Avg throughput (Tb/s) per BE board</th>
<th>Avg throughput (Gb/s) per BE board</th>
</tr>
</thead>
<tbody>
<tr>
<td>Outer Tracker</td>
<td>6.90</td>
<td>0.38</td>
<td>32</td>
</tr>
<tr>
<td>Track trigger</td>
<td>0.06</td>
<td>0.003</td>
<td>&lt;10</td>
</tr>
<tr>
<td>Inner Tracker</td>
<td>8.64</td>
<td>2.16</td>
<td>360</td>
</tr>
<tr>
<td>MIP Timing Det. - BTL</td>
<td>0.14</td>
<td>0.14</td>
<td>13</td>
</tr>
<tr>
<td>MIP Timing Det. - ETL</td>
<td>0.22</td>
<td>0.22</td>
<td>43</td>
</tr>
<tr>
<td>ECAL Barrel</td>
<td>9.49</td>
<td>0.79</td>
<td>88</td>
</tr>
<tr>
<td>HCAL Barrel</td>
<td>1.45</td>
<td>0.72</td>
<td>80</td>
</tr>
<tr>
<td>HCAL HO</td>
<td>0.18</td>
<td>0.18</td>
<td>-</td>
</tr>
<tr>
<td>HCAL HF</td>
<td>0.36</td>
<td>0.36</td>
<td>-</td>
</tr>
<tr>
<td>Endcap CALO</td>
<td>12.00</td>
<td>1.33</td>
<td>111</td>
</tr>
<tr>
<td>Endcap CALO TPG</td>
<td>1.50</td>
<td>0.10</td>
<td>&lt;10</td>
</tr>
<tr>
<td>muon DT</td>
<td>0.78</td>
<td>0.10</td>
<td>&lt;10</td>
</tr>
<tr>
<td>muon CSC</td>
<td>1.20</td>
<td>0.60</td>
<td>-</td>
</tr>
<tr>
<td>muon GEM - GE1/1</td>
<td>0.01</td>
<td>0.01</td>
<td>-</td>
</tr>
<tr>
<td>muon GEM - GE2/1</td>
<td>0.01</td>
<td>0.01</td>
<td>&lt;10</td>
</tr>
<tr>
<td>muon GEM - ME0</td>
<td>0.71</td>
<td>0.71</td>
<td>59</td>
</tr>
<tr>
<td>muon RPC</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Level1</td>
<td>0.90</td>
<td>0.06</td>
<td>&lt;10</td>
</tr>
</tbody>
</table>

Because of the wide variety of required per-crate and per-board throughput, a unique, monolithic design of the DTH would introduce unnecessary inefficiency. Indeed, data from a subset of the leaf boards in a crate can be concentrated and processed independently, insofar as the output links are used efficiently. Hence, when the required throughput grants it, such a subset can be handled by one FPGA operating independently as a data module in a board providing multiple data modules. Assuming a unit D2S link to be capable of 100 Gb/s, and that a “data module” on a DTH can drive a certain number of such links, the following two options are considered:

- all node slots connected to a single-module board, with one FPGA driving as many 100 Gb/s TCP/IP output streams as possible;
- three modules on a single board, each reading out a fraction of the crate.

In certain cases, the required crate throughput may exceed the maximum bandwidth of a single DTH board. In this case, it will be possible to install one DTH per hub slot, each reading out and providing TTC/TTS functionality for a separate portion of the crate.

Preliminary analysis of vendor roadmaps indicates that suitable FPGAs, providing enough high-speed transceivers to fit a module with 400 Gb/s input and $4 \times 100$ Gb/s outputs in a single component, are available or will soon be available at an affordable price. Additional connectivity requirements must be taken into account for the buffer memory necessary to manage a reliable connection and are discussed further in the text. Since both the input and output of the DTH consist of multiple links, the modular approach provides sufficient aggregation. A
monolithic design, using a single FPGA, is therefore not considered further, since larger FPGAs, with more high-speed transceivers, are much less cost effective than the modular design.

On the input side, both options, transferring data over the backplane or from the front panel via (multiple) point to point links, have been considered. Assuming the Fabric Interface to be fully available for DAQ, this would impose a hard limit of 100 Gb/s for each node-to-hub connection. According to Table 2.1, two sub-systems exceed this limit. Considering that all numbers in the table are averages, significant contingencies will be necessary to take into account hotspots and fluctuations. For example, for the Outer Tracker, simulation studies indicate an average of approximately 30 Gb/s per back-end board, but with a large spread of 10–90 Gb/s (see Section 1.3.2). In addition, FI links could be used for clock and TTC distribution, reducing even further the available bandwidth. Finally, flexible input data balancing, required to ensure efficient usage of available bandwidth, would be significantly hampered by the use of the backplane, while the specifications for the maximum speed of 25 Gb/s over the backplane lanes are still in flux. For all these reasons, only input via the front panel will be considered in what follows.

Since board layout at high clock speed can be a serious challenge, technologies simplifying it with respect to signal integrity are being considered, such as embedded micro optical engines. Such devices, available today as proprietary technology, implement flyover high-speed multiple optical data links in a single package up to 28 Gb/s, effectively moving the electro-optical conversion process very close to the FPGA. Standardization efforts are ongoing (COBO [22]), which could lead to a family of compatible components from multiple vendors in the near future. The DTH prototyping program will explore these technologies.

To conclude, a modular design is chosen for the DTH, where, for an individual data module, the input streams are connected to 16 high-speed transceivers, each capable of up to 25 Gb/s, over front panel point-to-point optical links. Furthermore, the option of 24 high-speed 16 Gb/s transceivers, resulting in the same aggregate throughput, is being studied (see Appendix C). A basic, 16 (or 24) inputs and 4 × 100 Gb/s output DTH, based on a single module and called DTH400, and a second board with three modules, capable of up to 1200 Gb/s for high throughput, the DTH1200, are assumed in the following.

In Table 2.2, the number of DTH400 and DTH1200 boards required per sub-detector, as well as the corresponding number of D2S links, are summarized. The numbers include a small contingency to account for the possible readout of additional detectors (see Section 1.3.16).

### 2.1.3 Roadmap for DTH

Establishing a detailed roadmap with milestones for different prototypes, the desired features, and their testing and integration with back-end electronics, will require further coordination with sub-detectors and synchronization of the planning. The following is a first attempt at specifying the most important steps to achieve a production board with the necessary performance by the time of installation during LS3.

- **DTH Specifications by end 2017**, including indication of the technologies/components to be used for the first prototypes. Technology and components for prototypes may be different from production.
- **First prototype (P1) by mid-2018**: will include the ATCA base elements for a hub card, a generic design of the timing/monitoring unit and a functional data module with limited performance (e.g. limited bandwidth and/or number of input/output links). The target is to verify the ATCA base element compliance and provide a first
Table 2.2: Projected minimum number of DTH and DAQ D2S optical links (100 Gb/s) for each sub-detector.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Outer Tracker</td>
<td>18</td>
<td>1</td>
<td>-</td>
<td>18</td>
<td>-</td>
<td>4</td>
<td>72</td>
<td></td>
</tr>
<tr>
<td>Track Trigger</td>
<td>18</td>
<td>-</td>
<td>1</td>
<td>-</td>
<td>18</td>
<td>1</td>
<td>18</td>
<td></td>
</tr>
<tr>
<td>Inner Tracker</td>
<td>4</td>
<td>2</td>
<td>-</td>
<td>8</td>
<td>-</td>
<td>24</td>
<td>96</td>
<td></td>
</tr>
<tr>
<td>MIP Timing Det. - BTL</td>
<td>1</td>
<td>-</td>
<td>1</td>
<td>-</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>MIP Timing Det. - ETL</td>
<td>1</td>
<td>-</td>
<td>1</td>
<td>-</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>ECAL Barrel</td>
<td>12</td>
<td>1</td>
<td>-</td>
<td>12</td>
<td>-</td>
<td>9</td>
<td>108</td>
<td></td>
</tr>
<tr>
<td>HCAL Barrel</td>
<td>2</td>
<td>1</td>
<td>-</td>
<td>2</td>
<td>-</td>
<td>9</td>
<td>18</td>
<td></td>
</tr>
<tr>
<td>HCAL HO</td>
<td>1</td>
<td>-</td>
<td>1</td>
<td>-</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>HCAL HF</td>
<td>1</td>
<td>-</td>
<td>1</td>
<td>-</td>
<td>1</td>
<td>4</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>Endcap CALO</td>
<td>9</td>
<td>2</td>
<td>-</td>
<td>18</td>
<td>-</td>
<td>18</td>
<td>168</td>
<td></td>
</tr>
<tr>
<td>Endcap CALO TPG</td>
<td>12</td>
<td>-</td>
<td>1</td>
<td>12</td>
<td>2</td>
<td>24</td>
<td></td>
<td></td>
</tr>
<tr>
<td>muon DT</td>
<td>8</td>
<td>-</td>
<td>1</td>
<td>8</td>
<td>1</td>
<td>8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>muon CSC</td>
<td>2</td>
<td>-</td>
<td>1</td>
<td>2</td>
<td>6</td>
<td>12</td>
<td></td>
<td></td>
</tr>
<tr>
<td>muon GEM - GE1/1</td>
<td>1</td>
<td>-</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>muon GEM - GE2/1</td>
<td>1</td>
<td>-</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>muon GEM - ME0</td>
<td>1</td>
<td>-</td>
<td>1</td>
<td>-</td>
<td>9</td>
<td>9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>muon RPC</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Level1</td>
<td>14</td>
<td>-</td>
<td>1</td>
<td>14</td>
<td>1</td>
<td>14</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BRIL</td>
<td>1</td>
<td>-</td>
<td>1</td>
<td>-</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Contingency</td>
<td>6</td>
<td>-</td>
<td>1</td>
<td>-</td>
<td>6</td>
<td>1</td>
<td>6</td>
<td></td>
</tr>
</tbody>
</table>

Total: 113 58 68 - 560

\(^a\)Conservatively assigned the DTH-1200 to the OT. In case the throughput deviations across the BE boards from the average is small, a DTH-400 (max. 25 Gb/s bandwidth per leaf) is sufficient.

A testing platform for TCDS and DAQ functions.

- Second prototype (P2) by Q2-2020. Will feature a fully functional timing/monitoring unit and one or more fully functional DAQ modules providing complete connectivity within the crate, possibly with reduced output bandwidth. To be used for software development of a companion software stack.
- Pre-production units by Q2-2022. Could be distributed to detector groups for integration test, local setups, etc. In principle, full functionality and full performance. Update of the companion software stack.
- Full production tested and available by end 2023.

A more detailed description of the goals and constraints of the first DTH prototype is presented in Appendix C.

## 2.2 Trigger Control and Distribution

The CMS Run-2 Trigger Control and Distribution System (TCDS) evolved from three Run-1 systems [23, 24]:

- the Trigger Control System (TCS), previously part of the Global Trigger,
- the Trigger Timing and Control (TTC) system, and
2.2. Trigger Control and Distribution

- the Trigger Throttling System (TTS).

The trigger control part of the TCDS orchestrates the data-taking based on trigger conditions and subsystem readiness. The distribution part serves two main purposes: in the downstream direction it controls the distribution of Level-1 accepts and timing signals (i.e., the TTC information) to the back-end and front-end electronics of all CMS subdetectors, and in the upstream direction it collects readiness (TTS) information across all subdetectors and combines that into an overall throttling signal used to stop triggers when readout buffers are full or out of sync.

The key requirement for the Phase-1 TCDS upgrade was to continue supporting the legacy front-end and back-end electronics based on the TTCrx ASIC [25]. For Phase-2 this requirement is removed and the TCDS will only reach the backend electronics, which connects to the front-end electronics via GBT links (see Section 1.2).

2.2.1 Phase-1 TCDS

The current TCDS system has been in operation since LS1. Its design addresses certain shortcomings of the original TTC and TTS systems, and fulfills several new requirements introduced by the Phase-1 upgrades. Most notably:

- Closer integration of the trigger control functionality with the DAQ.
- Improved monitoring and correlation of the TTS and DAQ states (ongoing work).
- Integration of the Phase-1 upgrades, in particular the MicroTCA-based backend electronics.
- Addition of support for stream-based systems, e.g. BRILDAQ for the CMS luminometers (see Section 1.3.15).
- Expandability to accommodate additional and/or upgraded subsystems (number of partitions, number of timing signals, etc.).

Figure 2.3 schematically illustrates the structure of the Phase-1 TCDS system. The system is implemented in the MicroTCA standard with three different kind of boards: CPM, LPM and PI. All three boards were implemented based on common carrier boards, equipped with specialized mezzanines and custom firmware. Like the original TTC system, the TCDS is implemented as a distribution tree across partitions. Each partition corresponds to a major subsystem component capable of independent operation. Partitions can be operated in groups for tests, or all together under the Central Partition Manager (CPM) for global data-taking operation. The CMS Run Control software is responsible for configuring the CPM prior to operation. This configuration includes the triggering conditions, the list of partitions to be included or excluded, etc. Run Control starts data taking by enabling the issuing of the L1 accepts to the participating partitions. The CPM board receives physics triggers as input from the Final Decision Logic (FDL) of the L1 global trigger, mixes in calibration, random, and other special triggers, and applies constraints. It also receives beam-/timing-related information (Beam Synchronous Timing, BST [26]) from the LHC, used for event timestamping and included in the CPM event record for the DAQ.

The CPM board is connected via bidirectional links to the local partition manager (LPM) boards, which house logical firmware blocks implementing groups of subdetector interfaces (ICIs), and handle all the partition-specific logic. Each ICI is connected to a partition interface (PI) board, via bidirectional optical links. The PI board handles the specific connectivity needs of each partition. Most importantly, it provides connectivity for both legacy (i.e., non-MicroTCA-based)
Chapter 2. Phase-2 DAQ Baseline Design

Figure 2.3: Schematic view of the TCDS as currently implemented for Phase-1 CMS. CPM, LPM, and PI represent physical electronics boards. The ICI subdetector interfaces are implemented (in groups of eight) as firmware blocks inside the LPM boards.

and Phase-1 subsystems, which have a different physical link for the TTS.

The downlink from the PI to the subsystems is used for distribution of clock, L1As, and synchronisation signals using the TTC protocol [25]. Short/broadcast TTC commands are used for essential synchronisation (e.g. counter resets) and their handling on the detector side follows strict prescriptions [23]. Long TTC commands are used in specific ways by each subsystem and their handling is left to the implementer. Since Phase-1, long TTC commands are also used for the distribution of synchronization data for the BRILDAQ subsystems.

The uplink from the subsystems to the PI carries TTS state information. Each PI holds full TTS info and history for each of its connected back-end crates. The information on the TTS state of the individual leaf cards in each crate and the status of the DAQ link are not propagated to the PI. The PI just creates a TTS summary state, which is propagated to the CPM to form the global TTS state used to throttle the L1 accepts when needed.

2.2.2 Phase-2 TCDS

A schematic view of the Phase-2 TCDS system as it is currently conceived is shown in Fig. 2.4. As is the case in the current implementation, the system will be installed in the USC in a central location in dedicated racks, close to the machine interface (see Section 2.2.5).

The distribution of the timing and trigger control from the CPM to the DTH, and subsequently to the BE electronics in the various subdetector crates, will be via a high-speed serial protocol. This TCDS-2 protocol will be different from the current TTC datastream. The new datastream
2.2. Trigger Control and Distribution

will embed the low-speed LHC clock and a frame-alignment feature to allow fixed latency clock recovery with deterministic phase. The speed of this serial datastream is not yet fully defined and will depend upon the amount of payload data required by the upgraded system. The aim is to provide a line speed between 1 and 10 Gb/s, optimized for both bandwidth and clock recovery performance. This corresponds to an increase of one to two orders of magnitude in bandwidth compared to the legacy TTC stream.

The datastream from the BE electronics boards upstream to the DTH, and subsequently to the CPM will continue to be based on a high-speed serial protocol. This protocol will be used to signal the readiness of the sub-detector systems attached to the DTH for data-taking. The data payload size for this return data path remains to be determined based on the requirements on the latency for this signal channel as well as the requirements coming from the orchestration of the data-taking.

In order to fully benefit from the fact that all subsystems will be upgraded for Phase-2, backward compatibility is not considered a requirement for the TCDS timing and throttling streams. It is assumed henceforth that no TTCrx ASICs will remain in CMS after LS3. Should Phase-0 or Phase-1-based subsystems still remain in Phase-2, such subsystems should foresee implementing appropriate TTC/TTS adapters. It should be noted that maintaining the possibility to translate from the TCDS-2 to the TCDS-1 timing stream would impose constraints on the overall scheduling in the TCDS-2 timing stream.

The CMS Phase-2 distribution tree can be arranged such that any future dynamic partitioning/broadcast scheme can be accommodated. In particular, the Phase-2 TCDS hardware functionality need not be constrained by the choice of a particular technology for the optical transport layer. The choice of whether to use direct point-to-point optical links, or a Passive Optical Network (PON) [27], can therefore come late in the development process. This is facilitated

Figure 2.4: Schematic view of a preliminary TCDS system layout for Phase-2 CMS. Notable differences are the absence of the ICI and PI layers. The ICI has been absorbed in the DTH, allowing independent operation of individual crates. The PI adaptation layer is no longer necessary after the Phase-2 upgrades.
by the fact that the physical interface to the subsystems is based, in both cases, on pluggable optical modules.

### 2.2.2.1 TCDS-2 versus TCDS-1

In terms of functionality, the upgraded TCDS (TCDS-2) system will largely perform the same functions as the Phase-1 TCDS system (TCDS-1). In order to rationalize the design and increase the performance and functionality, the TCDS-2 will break with some ‘legacy traditions’ and remove most of the subsystem-specific artifacts and work-arounds.

Following the overall trend of the Phase-2 upgrades towards a more homogeneous system architecture, there is no longer need for a PI-like adaptation layer. This simplifies the TCDS-2 architecture, reducing cost as well as resources in the USC service cavern.

As far as the trigger control part is concerned, the full TCDS-1 functionality will be carried over. The support for non-event-based subsystems, as used for BRILDAQ, will remain and be expanded to be available to all CMS subsystems. The most interesting addition planned for the trigger side of the TCDS for Phase-2 is the distribution of a multi-bit trigger type for each bunch crossing. This feature has a wide variety of use cases, including:

- The CPM can globally distribute periodic ‘trace triggers’ for which all subsystems store all their data in full granularity. These events can be used for monitoring, debugging, trigger studies, etc.
- An extension of the above, ‘debug triggers’, can be used to efficiently capture data in known problematic bunch locations/trigger patterns.
- Calibration triggers can be distributed explicitly, instead of implicitly as is done currently (by the definition of a calibration trigger as ‘the first L1A after a TestEnable B-go’). Different calibration triggers in different places in the abort gap can even be envisaged for systematic studies.
- The BRIL group has shown interest in ‘lumi triggers’ for specific bunch-crossings (see Section 1.3.15). For example, for the first colliding bunch pair in the orbit, or in the first empty bunch crossing after the last bunch train in the orbit. These ‘lumi triggers’ would only be honored by luminometer subsystems.
- In anticipation of enhanced data-taking modes like the ‘40 MHz scouting’ (see Section 5.5), multiple trigger types could be defined for subsystems capable of different acquisition rates. E.g., all subsystems would honor the default trigger types, and ‘high-speed-capable’ subsystems would in addition honor a set of ‘high-speed’ trigger types.

Along the same lines, it would be advantageous to make the Level-1 candidate signal from the Global Trigger to the TCDS a multi-bit signal.

Given the current availability of high-speed serial links, it seems natural to move from a frame-based multi-channel protocol like the TTC protocol to a simple serial link transmitting N bits per bunch crossing. The above multi-bit L1A would be part of the data payload, and synchronization commands would also naturally fit such a scheme. With line rates in the range of 1–10 Gb/s and a bunch-crossing frequency of approximately 40 MHz, bandwidth would not appear to be a problem. Before committing to details about such an approach, a careful study is needed taking into account bandwidth distribution and clock recovery potential.

The most notable optimizations/simplifications on the side of the timing distribution will be on the level of customization allowed in the ‘subsystem branches’ of the timing stream. In order to
be able to provide stronger guarantees about timing performance and system stability, it will no longer be possible to modify the timing stream on a per subsystem basis as is currently the case in the ICI.\textsuperscript{4} This includes the possibility to delay signals.\textsuperscript{5} Mostly for reasons of bandwidth optimization, it is not foreseen to use the timing stream for the transmission of subsystem-specific information.\textsuperscript{6}

### 2.2.3 Trigger Rules and Deadtime

As originally designed, CMS has a need for trigger (suppression) rules of the form ‘No more than \( n \) L1As in \( N \) successive bunch crossings’ (see Section 1.2). These rules protected token rings in several subsystems from command conflicts and prevented buffer overflows at high trigger rates.

The need for trigger rules in Phase-2 is largely reduced with respect to the Phase-0/Phase-1 detector. Note that, by design, all Phase-2 subsystems should be capable of accepting triggers on two subsequent bunch-crossings. The only case considered at present is for the Inner Tracker (see Section 1.3.3). It may be necessary to introduce one or more trigger rules for proper operation of their front-end ASIC. A reasonable goal is that the total deadtime incurred due to trigger rules shall be no more than 1% of time at an L1A rate of 1 MHz. A possible trigger rule that satisfies the above criterion is at most 8 L1As in 130 successive bunch crossings. This would give a deadtime of 1% at a trigger rate of 1 MHz, or \( \approx 0.3\% \) at a trigger rate of 750 kHz.

### 2.2.4 Precision Clock Distribution

With the desire to include precise time-stamping for events in the readout data for a number of CMS subdetectors, in order to reduce the impact of pile-up, the need to define the performance of the CMS clock distribution tree has become more important. The subdetectors requiring a precise clock are the MIP Timing detector (see Section 1.3 in Ref. [9]), the ECAL barrel calorimeter (see Section 3.7.1 in Ref. [4]), and the Endcap Calorimeter (see Section 1.3.9).

A full specification of the timing system requirements will follow from studies performed by the CMS Fast Timing Working Group. This working group’s initial report [28] states that: ‘At the hardware level, it requires enhanced timing capabilities in the calorimeters, to measure photons – down to low energies – with precision of order 30 ps, and fast timing of charged hadrons’ as well as ‘clock distribution and channel response jitter, that should be kept to the level of 10 ps’ (in a dedicated fast timing layer). Further study will be necessary to qualify this adequately in terms of phase noise spectra and allowed ranges for all key quantities.

In anticipation of a full specification, the baseline solution for the TCDS-2 remains to distribute the sampling clock to be used by the detector front-ends embedded in a high-speed (i.e., multi-Gb/s) control data stream. This clock will then be re-distributed by the DTH over the ATCA backplane to the node cards. Jitter cleaners will in any case be necessary on both the DTH and all detector-specific ATCA blades.

\textsuperscript{4}Bandwidth permitting, it could be possible to include a ‘user bit field’ in the timing stream, the contents of which could be determined on a per-partition basis, possibly configured under control of subsystem software. (The approach could be similar to the one currently used for the configuration and control of the TCDS ICI.) It should be noted, however, that this is likely to add significant latency to the Level-1 trigger path; that is, unless one relaxes the requirement of adding the customizations in the same tick as the timing flags that trigger the customizations, which is a possibility.

\textsuperscript{5}It is not foreseen to re-implement the possibility to delay timing signals on a per-partition basis (as is the case in the Phase-1 ICI). Such a feature would negatively impact the ability to guarantee a well-functioning timing system without scheduling conflicts. If such behavior is deemed critical for subsystem functioning, it can always be implemented on the receiving/subsystem side.

\textsuperscript{6}This is currently done for the strip tracker front-end pipeline addresses.
Given that the development of all the ASICs in the clock distribution chain, including those to be used on-detector, will still take a number of years, it is worth investigating alternative schemes to the one outlined in the preceding paragraph. To this end, the timing performance of various FPGA-based front-end control boards is being evaluated in order to be able to guide the requirements on any new timing system that may be developed. In parallel, work is being carried out in simulation to understand the physics goals that set the performance benchmarks for these systems. Furthermore, system-level experiments and simulations will be needed to guide the requirement-setting, in particular with respect to how external factors (e.g., temperature, power supply voltage) affect the overall timing performance of the various upgrading sub-detector systems.

### 2.2.5 TTC Machine Interface

CMS is synchronized to the LHC based on several signals received from LHC Point-4 on optical fibers. These signals are decoded, converted to electrical signals, and switched in the TTC machine interface (TTCmi) crate. The TCDS-1 inherited the unmodified TTCmi (VME) hardware from LHC Run-1. The software was rewritten to integrate with the rest of the TCDS online software.

At least part of the TTCmi electronics will not be able to provide a clock of sufficient quality to serve as precision clock and, given the age of the electronics, it should be foreseen to fully replace the machine interface crate for Phase-2\(^7\). Just as was the case for the original TTC machine interface, CMS will, where possible, follow the LHC-wide RF/timing developments for this TTCmi upgrade.

### 2.2.6 Roadmap for the TCDS

A preliminary design of the TCDS for Phase-2 and the specification of the TCDS-2 protocol is planned by the end of 2017. This includes a discussion of the multi-bit trigger type assuming a certain bandwidth for the serial datastream.

A first generation of TCDS demonstrator (mid-2018 to mid-2019) will be based on the first prototype DTH (P1) board. An integration test of the full chain will include:

- the CPM, using a DTH-P1 board;
- the DTH itself, in an ATCA crate emulating a BE electronics crate;
- an ATCA board emulating a sub-detector BE board;
- an evaluation board emulating the front-end electronics and connected to the BE emulator board via a GBT link.

The primary goal of this setup is the investigation of the timing performance. Both the DTH-P1 and the GBT link are expected to become available mid-2018. The outcome of these tests will be used to choose the bandwidth of the TCDS-2 stream and to assess the clock precision achievable.

A second aim is to produce a DTH-kit encompassing the firmware and software to support single crate basic DAQ operation. Target date is mid-2019.

Further generations of the TCDS demonstrator will

- use the next versions of the DTH board.

---

\(^7\)This is considered a part of the TCDS upgrade for Phase-2, and included in the central DAQ upgrade plan.
• provide a multi-crate setup, where a CPM distributes to two ATCA crates with their DTH board,
• provide additional functionality, such as BRILdaq and support for partitions.

A further goal is to assess the implementation of the CPM hardware with the required connectivity, and in particular whether an existing board can be re-purposed, or if a dedicated board needs to be developed. Target date for this decision is the TDR (Q2-2021). In case a dedicated board should be required, sufficient time for development and manufacturing will be available before the installation in early 2025.

2.3 Event Builder

The current CMS DAQ Event Builder [29] uses InfiniBand FDR switches to build events of up to 2 MB 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 200 Gb/s interconnect technology (e.g. InfiniBand 200G HDR) would require about 500 I/O servers receiving data from the D2S system and 500 switch ports at 200 Gb/s to interconnect them. This assumes the same effective interconnect efficiency of approximately 50% as in the current system. In order to reduce the number of switch ports required, it is assumed that it will be possible to use the same I/O servers as both input and output for the event builder, such that each processor builds events receiving fragments from all the others, and provides input fragments for events being in built in other nodes. This geometry, henceforth indicated as “folded event builder”, helps exploiting at best the bidirectional interconnects and is also planned e.g. for the LHCb upgrade [30]). The adoption of the “folded event builder” is the main difference between the baseline for Phase-2 and the current event builder. Events will then be distributed to the HLT over 100 Gb/s links through small TOR switches similar to those used for input.

In the current CMS DAQ, full events are built and buffered in the Builder Unit (BU) on a large RAMdisk (256 GB). This relatively large local buffer guarantees sufficient time decoupling to absorb fluctuations in the HLT service time, particularly when loading conditions and other parameters from the database. The current system provides up to 60 seconds of local storage. It is desirable to maintain this capability, which will require, for each I/O server, about 0.67 TB of fast access local storage. In the current system, event files in the RAMdisk are accessed by individual HLT processing nodes, connected to the BU over an inexpensive 10/40 Gb/s Ethernet switch, using the NFSv4 distributed file system. The performance of NFSv4 is at the moment the bandwidth-limiting factor in distributing data to the HLT nodes. An appropriate replacement will depend on how exactly the local fast buffer is implemented.

2.3.1 Roadmap for the Event Builder

Test equipment for the most relevant HPC (High Performance Computing) interconnect technologies (Infiniband HDR and Intel Omnipath) will be acquired in the course of 2017. In principle, the current Event Builder software stack can make use of faster interconnects in both technologies with minimal modifications. Verifying the functionality and performance in a small setup is a first necessary step in view of the possible adoption of these new interconnects already for Run-3, in the context of the regular replacement of obsolete or end-of-life network and computing equipment. Time slices on larger setups for scaling tests can be subsequently negotiated with the vendors. Installing a new interconnect in the course of LS2 will allow exhaustive testing of the folded option at a larger scale. It is not excluded that a folded event
builder be adopted already for Run-3. In general, as new solutions become mature, the Run-3 DAQ system can represent an invaluable real-life testing ground.

At the same time, alternative solutions to the current low-level transport libraries used in the Phase-1 DAQ will be studied. In particular, with an eye to adopting widely supported frameworks for HPC process interconnects, such as the Message Passing Interface (MPI, [31]). The value of adopting externally supported software, with the goal of optimizing person-power and keeping into account the need to support the running system while developing the next generation DAQ, cannot be overestimated.

Alternatives to RAM for the fast access local buffer, including Non-Volatile Memory express (NVMe) modules, as well as other forms of addressable non-volatile memory, must be studied. This is a necessary preliminary step in view of identifying an access pattern with better performance than NFSv4 for the access to input data by the HLT processes. One concern with solid state storage devices and NV memory is endurance, since the read/write pattern of the event builder local buffer is radically different, and more challenging, than a typical storage device. In the period between the completion of the present document and the DAQ TDR, a small number of interesting technologies are expected to become available for testing, in particular second-generation addressable NV memory with low latency.

Finally, it is expected that servers with PCIe Gen4 become available in the course of 2018. Testing PCIe-Gen4-based motherboards in a small event builder setup is a first important step to achieve better-than-marginal performance with 200 Gb/s interconnects, and results are expected in the timescale of the DAQ TDR.

### 2.4 High Level Trigger

The HLT algorithms are implemented in the offline CMSSW framework. Since the beginning of CMS the HLT was running on dual-CPU (x86) general purpose servers. This has also been the platform for general purpose and offline computing at the CERN-IT data center. An extrapolation of the expected performance of similar dual-CPU (x86) general purpose servers to 2027 can be used to translate the CPU requirements discussed in Section 1.5.1 into baseline parameters of a Phase-2 CMS data center.

#### 2.4.1 Evolution of HLT Processing Nodes

The evolution of the benchmark performance of dual-CPU servers and the Thermal Design Power (TDP) of the processors used in the CMS HLT, compared to the ones used in the CERN IT data center, are illustrated in Fig. 2.5. The gains in processing power have come mainly from an increased number of cores, a trend that appears to continue. From 2012 onwards, Intel introduced a large variety of Xeon(2S) processors with different numbers of cores and frequencies within the same family. The particular processor choices made for HLT and the CERN-IT data center are based on considerations including performance per node, power consumption, memory per processor core, and follow a different trend. The trend for the HLT nodes is roughly described by an increase of 72 HS06 per year and well parametrized by a second order polynomial. This trend predicts an increase in performance by a factor 2.6 in 2026 compared to the last generation of HLT server nodes (based on dual E5-2680v4 launched in Q1-2016, rated at 659 HS06).

The number of HLT servers and the total cooling power required for the ultimate HLT farm of Phase-2 are illustrated in Table 2.3. This assumes the power consumption of 340 W observed for the last generation HLT nodes stays constant. It is interesting to note that following the
2.4. High Level Trigger

Figure 2.5: Evolution of the CPU HS06 benchmark (left y-axis) and TDP (right y-axis) of CMS HLT servers (red), compared to servers of the CERN-IT data center (blue). The black solid curve shows the HS06 trendline for CMS HLT servers.

trend in the data of the nodes for the CERN-IT data center would instead require about two times more nodes with approximately half the power consumption each, resulting in about the same requirements for total power.

Table 2.3: Phase-2 HLT datacenter size and total power projections for the ultimate 9.2 MHS06 requirement in 2026 and, respectively, 4.5 and 9.2 MHS06 at the end of LS3 and LS4 in the staged scenario.

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Ultimate</th>
<th>Staged</th>
</tr>
</thead>
<tbody>
<tr>
<td>CPU Required (MHS06)</td>
<td>9.2</td>
<td>4.5</td>
</tr>
<tr>
<td>benchmark ratio (w.r.t. 2016)</td>
<td>2.6</td>
<td>2.6</td>
</tr>
<tr>
<td>number of servers</td>
<td>5570</td>
<td>2718</td>
</tr>
<tr>
<td>total power (MW)</td>
<td>1.9</td>
<td>0.9</td>
</tr>
</tbody>
</table>

The current CMS data center has a total cooling and power capacity of 1 MW. Deploying the large number of servers indicated in Table 2.3 with a much larger total power dissipation would require a new data center or a major upgrade of the existing infrastructure. This will be discussed in Chapter 4. It is therefore paramount to investigate ways to reduce the power consumption of the final system without compromising the physics program.

2.4.2 Roadmap for HLT

Full simulation of the Phase-2 detector has recently become available in the context of the studies for the TDRs. Furthermore, design studies have been done for the L1 trigger with algorithms and a strawman menu for $\langle PU \rangle = 200$ [5]. Work to be done in the coming years towards the DAQ/HLT TDR around 2020 include:
Chapter 2. Phase-2 DAQ Baseline Design

- Reconstruction for HLT leveraging the forthcoming reconstruction code for the Phase-2 detector
- Studies of an HLT menu with a rejection factor of 1 in 100
- Studies of the timing of such algorithms
- Investigation of alternative computing platforms to reduce cost and total power

2.5 Storage Manager and Transfer System

The Storage Manager and Transfer System (SMTS) is responsible for collecting events accepted by the HLT processes and for aggregating monitoring information produced by the HLT reconstruction algorithms. It buffers the events on a parallel filesystem from where they are transferred to Tier 0 for offline processing or to local consumers for data-quality monitoring or fast calibration.

The current system [19] is composed of a Lustre parallel file system, served by three storage appliances organized in active/passive high availability server pairs, and one metadata server with a secondary passive server for high-availability. Each storage appliance contains a NAS system with six RAID-6 groups of HDDs (Hard Disk Drives). The total amount of formatted Lustre storage is approximately 350 TB. The applications to handle the data flow are implemented in python and run as data-driven services. The application level throughput is about 4.5 GB/s, i.e. data can be aggregated and written, and transferred concurrently, at 4.5 GB/s.

The required performance of the SMTS is determined by the overall HLT output throughput. Currently, the event size is reduced by the compression at the HLT output by a factor of about 0.8. On the other hand, the volume of the HLT output is increased by about a factor 1.5 due to the addition of the HLT reconstruction products, as well as data from triggers used for calibrations, monitoring and diagnostic. This results in an overall factor of about 1.25 compared to the HLT throughput.

The current SMTS, based on a parallel file system, can scale in throughput and storage capacity by adding storage appliances. We assume that a similar technique will be used for Phase-2. The Phase-2 SMTS will see an increase in the HLT output bandwidth from 4.5 GB/s to 31–61 GB/s and a substantial increase in the number of HLT processes and event builder nodes. This implies that the storage system has to be scaled up by a factor 7–14 compared to the current system.

2.5.1 Roadmap for SMTS

LS2 will offer, in the context of the regular replacement of obsolete or end-of-life equipment, the opportunity to perform a market survey for storage equipment to replace the existing SMTS hardware. The evolution of parallel file systems will also be examined with the aim of identifying competitive solutions in view of the possible replacement of Lustre for Phase-2 (see Section 3.5.2).

Concerning the medium term time scale, an evaluation during 2018–2019 will select the technology for Run-3. For the timescale of Phase-2, it is planned to assemble a TDAQ-slice demonstrator to develop the full DAQ chain. This will include a small-scale SMTS and a test of up-to-date hardware is foreseen for 2023, in order to decide on the purchase of the production system.
2.6 Software Infrastructure, Control and Monitoring

The CMS online software follows a Service Oriented Architecture [32] where several intercommunicating application processes are executed over a network to achieve the data acquisition task. The software relies on three cornerstones:

1. XDAQ is a C++ platform designed specifically for the development of distributed data acquisition systems [33]. It follows a layered middleware approach, providing support for communication, hardware access, monitoring, error reporting, logging, to facilitate the interoperation of different DAQ components (Trigger Control and Distribution, Data to Surface, Event Building). XDAQ builds upon industry standards, open protocols and libraries (e.g. TCP, HTTP, XML, Apache Xerces, etc.), and is used by both CMS Central DAQ and CMS sub-detectors/sub-systems.

2. The Run Control System (RCMS) [34] is a Java framework based on web applications running inside container instances (apache tomcat), which provides the building blocks to compose a distributed hierarchy of nodes to control and monitor the state of XDAQ applications. Control nodes are based on state machines, with business logic implemented in Java. They are steered and monitored through web-based GUIs. RCMS includes database schemas to hold the configuration of all software components, to define hardware configurations, and to manage the complex interconnects required for the current two-stage event builder. Extensive tools for configuration management are available.

3. The file-based filter farm uses a lightweight infrastructure [18] to control and monitor the execution of the HLT processes, to aggregate the output of the selection process, and to transfer the data for offline processing or to online consumers. It consists solely of daemon processes implemented in python, is entirely data-driven, and integrates the collection of monitor information into a modern analytics engine: Elasticsearch [35].

The above components are in use and expected to be maintained and developed throughout the operation of CMS and up until the beginning of the HL-LHC program. They constitute a solid baseline around which to build the software for the Phase-2 DAQ. The system also relies on commercial Off-The-Shelf and other external software, including, among others, Apache HTTP Server, Elasticsearch, Oracle, WinCC OA.

2.6.1 Online Software Services

The distributed processing infrastructure is made scalable by the ability to partition DAQ related applications into smaller functional units that can be distributed over multiple processes. The CMS online software services operate in two different modes: on demand services, where applications are executed for a given period of time like cronjobs, RCMS jobs, etc. (the lifetime of these can vary from seconds to days); and permanent services, running at all times.

The lifetime of on demand services is in most cases defined by the RCMS, which decides when they are started and stopped. Examples of on demand services are the event builder, device configuration and readout. Configurations for these services are provided by the RCMS dynamically. Alternatively, on demand services are started and stopped by Linux crontab or manually.

Examples of permanent services are monitoring, error reporting, database access, DCS access, etc. They are statically configured for their whole lifetime, are autonomous and do not need
external control. Nevertheless they are under the supervision of the systemd init system. They can be automatically restarted on failure according to configuration.

XDAQ based services are configured using a homogeneous approach via XML configuration files. This approach is used for both on demand and permanent services. Configurations of on demand services are generated, kept in the database and deployed by RCMS. For permanent services, namely XaaS (XDAQ as a Service), XML configuration files are kept under version control and managed by centrally. These services are deployed via the RPM package management system.

2.6.2 Possible development directions for overall CMS online software

Despite the availability of the above common frameworks and tools, the current online software implementations show a certain degree of heterogeneity across CMS sub-systems. In particular, implementations differ in whether XDAQ is only used at the leaf nodes of the control hierarchy or also at an intermediate layer, in whether applications are dynamically started through RCMS or are always running as a system service, and in the usage of the database for configuration management. In part, these differences can be explained by person-power available to central and to sub-system specific development, by online software development starting prior to the availability of common solutions, and by the large number of specialised hardware boards developed by different groups. A higher level of homogeneity would be highly desirable in the online software of the Phase-2 upgrade.

The Phase-2 upgrade of CMS will likely be based on relatively few types of multi-purpose processing boards, mostly in the ATCA standard, featuring large FPGAs and interconnected by high-speed optical links. These boards will be specialised to perform their task through customised algorithm firmware. Common functionalities such as link management, the TCDS interface and the DAQ interface, may be implemented in firmware blocks that can be common to each type of processing board or even to multiple types of processing boards. A new software stack will have to be developed to operate these new processing boards. This will give CMS the opportunity to develop a common online software framework that provides out-of-the-box support for such common firmware functionality and a framework for the implementation of algorithm-specific control, monitoring and configuration management.

While requiring a larger centralized effort, such a solution would significantly reduce the overall development cost and the long-term maintenance of the online-software. The firmware and online software of the Phase-1 trigger upgrade [36] are a successful example of such an approach.

2.6.3 Software Roadmap

To enable R&D activities in the area of D2S, TCDS, and the event builder, the core software components will have to be extended to provide support for new network protocols, new custom hardware and new functionality of the underlying application components, such as for example the extended TCDS functions and the folded event builder architecture.

The Phase-2 DAQ software development will be focused on the creation a uniform stack providing DAQ and TCDS functionality for test-bench setups, standalone single-crate systems, small scale DAQ systems (aka miniDAQs), and global DAQ. This software stack should provide enough flexibility to accommodate subsystem specific modules for test and debugging setups as well as for calibration purposes.

In view of the projected lifetime of the Phase-2 CMS, it is clearly necessary to plan for general
2.7 Detector Control System

The Detector Control System (DCS) ensures the reliable, uninterrupted operation of the experiment, while providing monitoring data which is vital for detector performance and safety, as well as physics analysis.

The detector controls are distributed over widely different hardware, from power supplies to temperature sensors to modular electronics crates. The DCS data are processed in the WinCC OpenArchitecture (WinCC OA) SCADA system. All devices are continuously monitored by WINCC OA and acquired values are compared to predefined thresholds. In case of significant deviations from nominal settings, the SCADA system can take automatic remedial action, or alert the operator. The operational limits, device settings and configuration parameters are stored in an Oracle database. Different applications, called drivers, map the values in WinCC OA to the parameters of the controlled hardware. The drivers leverage the reusable components provided by the Joint Controls Project (JCOP) toolkit.

The different systems are modeled using a state machine toolkit, and state machines are arranged in a hierarchical structure, representing the detector structure. A qualitative summary is passed to the parent at each level, leading to a global status of the detector components, and the experiment as a whole.

The computing infrastructure currently consists of about 70 windows servers and a number of database schemas to maintain data history, and is managed centrally by the DAQ DCS team.

2.7.1 DCS Roadmap

The DCS system has helped operating the CMS experimental apparatus with efficiencies close to 100% and is a good example of collaboration among the different experiments. As such, the DCS common components are expected to continue being developed and maintained through LS2 and LS3, and the system to naturally scale up to the final Phase-2 needs. Supporting the assembling and the commissioning of the upgraded detectors will require the DCS system to be available throughout the entire upgrade and commissioning process.

2.8 Baseline Bill of Material

The CMS baseline DAQ architecture for the Phase-2 upgrade is feasible with readily available technology. An overall cost estimate for the baseline architecture is discussed in Chapter 6. In the following, we detail the different components that enter into that estimate.

The number of different DTH units required is obtained from Table 2.2. As discussed in Section 2.1, the DTH will come in two configurations with moderate and maximum bandwidth: the DTH400 accommodates the needs of most of the low-throughput detectors, while

---

8Supervisory Controls And Data Acquisition.
the DTH1200 is needed for the high-throughput detectors such as Tracker, ECAL and HCAL barrel, and Endcap Calorimeter.

The TCDS comprises the custom electronics boards, CPM and the TCDS module inside the DTH boards, interconnected in a tree structure with optical links (see Section 2.2.2); and the LHC machine interface (see Section 2.2.5). Note that a parallel precision timing distribution system, should it be shown to be required, still needs to be studied (see Section 2.2.4).

The number of D2S links (long-range transceiver pair and fibre between USC and SCX) must match the required number of DTH 100 Gb/s output links\(^9\).

A total of 500 I/O servers is needed for event building. The number of I/O servers can be slightly lower than the number of D2S links, as some level of data concentration is possible in the intermediate TOR D2S switches. These intermediate D2S 100 Gb/s Ethernet TOR switches will serve a double purpose: allow some level of data concentration and rerouting of D2S links to I/O servers for redundancy. Additional intermediate HLT TOR switches will be used to connect the I/O servers with the HLT nodes. An Ethernet NIC with dual 100 Gb/s ports will connect the I/O server with the two intermediate TOR switches.

An event builder HPC interconnect switch with 500 external ports is required. The exact structure of the interconnect network and the number of switch layers required will depend on the specific interconnect that is adopted.

The baseline storage system is inspired by the current Lustre-based system, scaled up to match the Phase-2 requirements.

The baseline DAQ bill of material is summarized in Table 2.4.

\(^9\)Total D2S links in Table 2.2.
### Table 2.4: Phase-2 baseline DAQ bill of material

<table>
<thead>
<tr>
<th>item</th>
<th>component</th>
<th>type</th>
<th>Qty</th>
<th>notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>ATCA DAQ-TCDS Hub (DTH)</td>
<td>DTH1200</td>
<td>58</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>DTH400</td>
<td>68</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>TCDS</td>
<td>CPM ATCA boards</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>optical links CPM-DTH</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>LHC machine interface</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>D2S links</td>
<td>LR optics (Transceiver pair + fibre)</td>
<td>560</td>
<td>((a))</td>
</tr>
<tr>
<td>4</td>
<td>I/O servers</td>
<td>PC server + 1 TB buffer</td>
<td>500</td>
<td>((b))</td>
</tr>
<tr>
<td>5</td>
<td>Intermediate switch for D2S</td>
<td>100 GbE TOR switch 32 ports</td>
<td>36</td>
<td>((c))</td>
</tr>
<tr>
<td>6</td>
<td>Intermediate switch for HLT</td>
<td>100 GbE TOR switch 32 ports</td>
<td>64</td>
<td>((d))</td>
</tr>
<tr>
<td>7</td>
<td>Ethernet NIC to TOR switches</td>
<td>NIC Dual port 100 Gb/s + cable</td>
<td>500</td>
<td>((e))</td>
</tr>
<tr>
<td>8</td>
<td>NIC for EVB HPC interconnect</td>
<td>NIC 200 Gb/s + cable</td>
<td>500</td>
<td>((f))</td>
</tr>
<tr>
<td>9</td>
<td>EVB HPC interconnect switch</td>
<td>Number of ports</td>
<td>500</td>
<td>((g))</td>
</tr>
<tr>
<td>10</td>
<td>Storage Element</td>
<td>Disk appliance + metadata server</td>
<td>20</td>
<td>((h))</td>
</tr>
<tr>
<td>11</td>
<td>HLT</td>
<td>Server</td>
<td>5570</td>
<td>((i))</td>
</tr>
</tbody>
</table>

*\(a\)Transceiver pair + fibre 100 Gb/s for USC-SCX (200 m), e.g. 100 GbE with CWDM4 and single mode fiber.
*\(b\)PC server with 100 Gb/s Ethernet in/out and 200 Gb/s port to EVB interconnect.
*\(c\)Ethernet TOR switch with 100 Gb/s ports for D2S: 14 ports to I/O server, 18 ports from D2S.
*\(d\)Ethernet TOR switch with 100 Gb/s ports for HLT: 8 ports to I/O server, 24 ports (break-out) to HLT nodes.
*\(e\)Ethernet NIC with dual 100 Gb/s ports.
*\(f\)NIC with 200 Gb/s HPC port + 50 m cable.
*\(g\)Unit is 200 Gb/s bidirectional port, used 50%.
*\(h\)One storage element can do 25 Gb/s effective throughput.
*\(i\)MHS06 from Table 2.3.
Chapter 3

Design Optimization and Technology Survey

3.1 Data to Surface

The current Phase-1 D2S is based on an FPGA hardware implementation of a simplified \(^1\) (uni-directional) TCP/IP protocol at 10 Gb/s [21]. The TCP/IP protocol features a flow control\(^2\) and congestion control\(^3\) mechanism. The congestion control allows transmitting multiple TCP streams on the same link and allows to merge several links, which are not fully utilized, together using a switch. This in turn allows to minimize the number of switch ports and to equalize data traffic by merging streams of unequal size. The Phase-1 hardware implements a simplified congestion control. One of the key aspects of this implementation is the optimization of the buffer memory required for the flow control and congestion control to efficiently run the TCP streams at maximum speed while avoiding backpressure. A congestion window with a maximum size of 180 kB was found to give optimal efficiency and minimize switch saturation in a wide range of working conditions, from a single TCP stream configuration, with a throughput of 9.7 Gb/s, up to configurations merging 16 TCP streams. The TCP flow control takes advantage of the full amount of memory installed on the D2S board, enabling to avoid backpressure up to 7 ms at 10 Gb/s. A subsequent version of the D2S board, used for the Pixel Phase-1 upgrade readout, has larger buffer memory and bundles four 10 Gb/s links to feed a single 40 Gb/s port.

The current R&D studies for Phase-2 are aimed at testing and validating the Phase-1 implementation with 100 Gb/s streams and produce a similar analysis of the buffer memory needs for 100 Gb/s links. It is anticipated that the amount of memory will largely exceed the standard on-chip memory of the FPGA, and external components will be necessary. The recently introduced High Bandwidth Memory (HBM), which stacks DRAM together with the FPGA, may not provide a cost-effective solution. A promising technology is serial memory (see Appendix C). A test stand using an FPGA evaluation kit and a 100 Gb/s NIC is currently being used to assess the above.

The exact network protocol and speed used for D2S will not need to be fixed until the DTH prototyping work reaches an advanced stage. Should enough logics become available in upcoming FPGA families to implement an HPC protocol, D2S could plug directly into the switched network used for event building, thus eliminating the need for intermediate D2S TOR switches. It

\(^1\)The TCP/IP protocol simplifications are compatible with RFC 793, and therefore a PC with a standard Linux TCP/IP stack can be used as a receiver without any modifications.
\(^2\)The TCP/IP flow control issues a back-pressure notification to the sender when the buffer occupancy of the receiver is high. It will automatically decrease the sending rate or completely stop the transfer allowing the receiver to process received data during occasional busy periods.
\(^3\)The TCP/IP congestion control limits the rate of data sent to the network below a level that would create a network congestion.
is important to maintain flexibility in the overall design of the D2S and the EVB systems. The choice of protocol and link speed will be made on the base of cost/benefit, taking into account also other options which could emerge from ideas discussed in Chapter 5.

3.2 Trigger Control and Distribution

A number of hardware-related technology investigations have to be carried out in order to optimise the Phase-2 system. The most important involves the investigation of the timing performance of the proposed baseline Phase-2 TCDS, i.e. what level of timing stability can be provided to the ATCA leaf cards downstream of the DTH. Related to this is the choice of optical transport layer between the TCDS and DTH. In order to account most appropriately for the needs of the Phase-2 TCDS, the relative merits of the two options: direct point-to-point, and Passive Optical Network (PON) [27] must be examined.

3.3 Event Builder

Optimisation of the baseline Event Builder design will be considered in view of the foreseeable evolution of hardware over the timescale of the Phase-2 upgrades.

Regarding cluster interconnects, multi-100 Gb/s is starting to appear on the market and can be expected to become affordable over the next five years, allowing the reduction of the number of switch ports and, possibly, EVB nodes. A corresponding evolution of I/O on-board interconnects can be expected: PCIe 4.0, with a speed of about 16 Gb/s per lane, will become a standard on the server market over the same timescale. On the other hand, server memory bandwidth increase is driven by the needs of many-core CPUs and could well exceed that required to fully exploit top-of-the-line single bidirectional links on a relevant timescale for the CMS Phase-2 upgrade. An hypothetical 2025 I/O server, like the one sketched in Fig. 3.1, could handle multiple input D2S links and still perform event building in a folded architecture, which could decrease the number of I/O processors by up to an order of magnitude.

Figure 3.1: Possible features of a 2025 I/O processor.
A potential reduction of the number of interconnect switch ports and/or I/O servers does not necessarily imply a reduction in the cost of the system. Such an optimization could however have broader scope when considered in the context of the overall architecture, on one hand, and the total cost of ownership on the other. Overall, it will be useful to identify additional tasks that could be delegated to more powerful I/O processors. These could include elementary pre-processing of input data, to lift, for example, the load of raw data reformatting from the HLT, but also more complex tasks involving preemptive reconstruction and creation of feature indices, which can be used to speed up the HLT. At the other end of the spectrum, a reduction in system complexity can represent an advantage in terms of maintenance and operation load and possibly optimize its impact on the technical infrastructure (power distribution and cooling), and on the IT infrastructure (installation, system administration, configuration management).

3.4 Heterogeneous Architectures for the HLT

One of the main strategic early choices of the CMS DAQ design was, for the HLT, to use the exact same framework and algorithms used offline [6]. 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 CPUs and has had limited success in optimisation campaigns meant to exploit the most modern features like vectorisation and large-scale parallelism.

A fair amount of R&D work has been already directed at exploiting coprocessors and GPUs as offload engines for specific reconstruction algorithms (see Refs. [37–39]). The deployment of coprocessors in the HLT farm could proceed in three different ways.

1. Coprocessor-equipped HLT nodes. The first most conceptually straightforward option is to equip every processing node with relevant coprocessors (Fig. 3.2 bottom left) and integrate offloading of computation in the HLT reconstruction framework, maintaining the normal behavior for standard execution nodes. This will require committing to a single accelerator architecture for a medium period of time, despite the fact that different algorithms may benefit from different accelerator types; it would also require care in balancing the amount of work done by the CPUs and accelerators, or risk over-dimensioning one of the resources. On top of that, it has proven particularly important, in the context of the CMS HLT, to be able to rapidly deploy additional CPU power when conditions warrant. Installing new nodes with potentially different coprocessors, due to e.g. obsolescence of the coprocessors themselves, or of the expansion bus used to connect them, would require complex porting and revalidation of algorithms on new hardware. Offloading will also need particular care in the development of an accelerator-aware scheduling algorithm, to avoid a livelock situation where all computing threads end up waiting for the offload computations to finish, without the capability to take on more work, thereby defeating the purpose of the offload.

2. Coprocessor network offload service. A second option would be offloading suitable algorithms to a farm of specific machines carrying the appropriate coprocessors over a net-

\[4\] I.e. as dedicated processors to which a certain portion of computation within a larger program, which can be optimally executed on that specific hardware, is delegated for speed optimization.
Figure 3.2: Different options for offloading HLT algorithms to specialized coprocessors. Bottom left: coprocessor-equipped HLT nodes; top: I/O nodes with coprocessors as part of the event builder; bottom right: coprocessor network offload service.

work service (Fig. 3.2 bottom right). The advantage of such a solution is that the special hardware can be confined to a dedicated set of nodes, simplifying their management and upgrade. This requires the corresponding offload farm to be interconnected to every HLT node via a high-bandwidth, low-latency interconnect, such as the one envisaged for the event builder network. The additional number of switch ports required may render this option economically unattractive. While the considerations above concerning the need for full and transparent framework integration of such a networked offload, and the need for dynamic thread management, remain, it must also be taken into account that the additional network layer and its latency, however small, may compound additional pitfalls and make the livelock problems more dire.

3. I/O nodes with coprocessors integrated in the event builder. An alternative option, explored for one particular case in Ref. [40], is to run algorithms preemptively on a dedicated farm receiving input directly from the event builder network (Fig. 3.2 top), making the results available as part of the input data for the HLT. One advantage of this approach is that the algorithm latency could be absorbed by the Event Building buffer, whereby HLT processing would only start on an event when the output of the offloaded algorithms becomes available. It must be noted, however, that such an approach might require maintaining special versions of the algorithms capable of running in lightweight processes, such as those typically used to operate GPGPUs, working on raw-data without the use of offline-type framework infrastructures. Such algorithms would have to be testable and reproducible offline on a general purpose CPU, and monitored separately to maintain complete control on the final selection efficiency.

The latter approach could be particularly suitable when combined with detector data federation and indexing, as discussed in Section 5.4, making best use of large amounts of low-latency, addressable non-volatile memory.

The three architectures discussed above have each their own advantages and drawbacks. In particular, the different I/O overheads, the work required for software integration, and the complexity of the configuration management will need to be carefully evaluated in the three cases and will constitute important criteria to choose one or the other.
3.4.1 Case Study of using GPUs

As an example of the kind of CPU savings that are possible with the use of coprocessors, a recent case study [41] uses a setup where pixel raw-data is fed to a battery of eight GPUs (GTX 1080 Ti), installed in a standard dual-socket server with two 12-core Intel CPUs. The same algorithm, based on a cellular automaton approach, as used for the pixel detector pattern recognition in the HLT, is run in parallel in the CPU and the GPU from the same input data. The resulting processing rate of the 8 GPUs is 10.6 times that of the 24-core server. This includes data transfer from the host memory to one of the GPUs, track reconstruction execution, and copying the result back into host memory. In this case study, several operations overlap for the most efficient use of the hardware.\(^5\)

Using today’s list prices for the two elements, the authors conclude that the GPU can process 4.6 times the rate of the CPU per unit cost. In terms of power consumption, preliminary results indicate a 30% power saving by using GPUs instead of CPUs. In concrete terms, using the hardware discussed in this work, 14 GPU-equipped servers would be sufficient to process the current CMS pixel data at the current maximum L1 accept rate, whereas 128 standard servers would be required to perform the same task without GPUs.

3.5 Storage Manager And Transfer System

3.5.1 Data Format

In the Run-2 DAQ system, all FED data are assembled into continuous memory before being handed to the HLT processing. This data format is called FED raw data (FRD) and is only used between the event builder and the HLT processors. Accepted events are written in a custom, streamed data format storing serialized and compressed event ROOT objects. These data are written into a filesystem as individual files by each HLT process. The custom data format is very efficient for online use as multiple output files can be concatenated into contiguous objects (files) without involving potentially costly merging by a ROOT-based executable. The concatenated events in streamer format are buffered at the experimental site before being transferred to tier-0. The first step of the offline processing is the repacking of the streamer format into CMSSW ROOT format at tier-0. This format is then published, stored to tape, and used for any offline reconstruction or for detector studies.

Depending on the architecture of the Phase-2 HLT farm, the use of these different data formats may not be the best solution. The storage manager might need to assemble and convert event fragments originating not only from the HLT processes running CMSSW, but also from coprocessors and/or from pristine FED binary payloads on non-volatile memory. In this case, it might be beneficial to directly produce final format files, which could be fed directly into the standard tier-0 reconstruction. This would require a careful evaluation of the resource allocation between tier-0 and the storage manager system in terms of required person-power and hardware.

3.5.2 Storage Technology

Optimisation of the SMTS design will consider parallel file systems and storage technologies based on HDD, SSD, or a combination of them. Properties to be evaluated include throughput, storage capacity, endurance, ease of management, and cost.

\(^5\)Specifically, while input event A is being copied from the host memory to the device, results for another event B are being copied back from the device to host memory, while multiple instances of the track reconstruction kernel function are executed on the device’s cores.
Various technologies underlying the Lustre file system are being evaluated. A proof-of-concept study was realized in order to understand if a system based on ScaleIO [42] could complement or replace the current hardware solution. The performance and the resilience of ScaleIO have been confirmed, but the implications of supporting such a highly distributed system have to be carefully considered. The ScaleIO high availability also relies on a significant amount of reserved spare space (basically the double of the space effectively used). Possible drawbacks in terms of maintenance and availability in case of power cuts, cooling issues and similar events need to be examined. Finally, ScaleIO will introduce an additional layer to the debugging of possible problems or performance issues.

Other, more extreme options are also under consideration. For example, Intel has advertised new extremely dense devices with enormous potential bandwidth. The devices are still in the design phase, but it is claimed that storage densities up to 1 PB per rack unit with up to 48 GB/s raw total bandwidth for a single device could be achieved. Such systems would also provide a DAOS (Distributed Application Object Storage) access layer. One of the characteristics to be evaluated is the overhead of the access layer and the optional parallel file system on top.

### 3.6 Conclusions

Reduction of complexity and cost of the Phase-2 DAQ with respect to the baseline may result from technology tracking and late adoption of maturing technologies, in particular multi-100 Gb/s links and high bandwidth I/O servers with large amounts of addressable non-volatile memory.

Prototyping and exploratory work will aim at developing efficient D2S solutions and studying different alternatives for the Event Network. 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 years, as well as studying the application of heterogeneous architectures to offload suitable algorithms, may help reducing the Total Cost of Ownership (TCO). The adoption of new programming styles more suitable for truly distributed systems with large amounts of non-volatile memory, such as e.g. distributed container-query models, in-memory databases, and position independent data structures should also be considered and studied in more detail. Some of these techniques could be tested already in Run-3.

---

6On non-volatile memory, sequential and random access essentially converge, while the read-write asymmetry inherent to these devices, combined with the possibility of reusing in-memory data across the lifetime of a process, call for a fresh approach to the implementation of large data structures.
Chapter 4

Online Infrastructure at Point 5

4.1 Introduction

The upgrade of the DAQ system has implications on the technical and IT infrastructure necessary to operate it. This chapter presents the requirements of the new system and the different options being considered to fulfil them.

4.2 DAQ infrastructure in USC

As discussed in Section 2.1, the baseline DAQ design envisages the installation of the DTH boards in the subsystem back-end crates. A complete recabling of the DAQ links from the subdetector crates to the surface counting room will be required. Considering the occupation of the cable trays in USC-S1 and S2, it will not be possible to install the new DAQ system without fully removing existing cables. This de-commissioning will include cables installed by other subsystems. This is a delicate task that will have to be carefully planned in the context of the LS3 schedule.

Provided that all sub-detectors upgrade their back-end modular electronics to the Phase-2 standard (one or two DTH boards in the hub slots of each ATCA crate), it will be possible to reattribute the racks currently used for the D2S boards (FRL/FEROL), the legacy TTS system, and their controller PC, for other needs. These are detailed in Appendix D.1. Should there be cases where the installation of the DTH boards in the hub slot of sub-detector crates is not possible, corresponding rack space for dedicated DAQ ATCA crates, in an appropriate location, will be necessary.

Other rack space used for general DAQ services in the USC (notably, TCDS) will continue to be needed.

4.3 Online Data Center

The surface counting room in SCX has the characteristics of a small data center. Power distribution and cooling will have to be adapted to fit the needs of the Phase-2 system as presented in Chapter 2.

The current CMS surface counting room has a surface of about 300 m$^2$ and houses 145 racks with active cooling doors. The total cooling capacity, guaranteed by a mixed water system, is approximately 1.2 MW, with a total input water flow of 170 m$^3$/h. This system provides an inlet water temperature to the racks of 14°C. The data center achieves a Power Usage Effective-
ness (PUE)\(^1\) of approximately 1.4. Electrical distribution is provided by a Canalis 160 A/250 A system running on the top of the rack groups. More technical details can be found in Appendix D.2.

### 4.3.1 Technical Infrastructure Roadmap up to LS3

For Run-1, 145 racks were installed in two successive stages to match the evolving needs of the DAQ system. The first phase used Gen-1 racks, whereas the second phase used Gen-2 racks. In particular, the cooling capacity of the Gen-2 racks has been increased (to 16 kW) to match evolving hardware, and the depth has been increased for mechanical compatibility and to accommodate an increasing cabling volume inside the cabinet. The parameters of all the generations of racks are detailed in Table D.1 of Appendix D.2.

Even the 16 kW cooling capacity of the Gen-2 racks limits their efficient use, especially for the HLT farm, which consists of quad-server chassis with a height of 2U (rack units\(^2\), yielding a density of one motherboard per half rack unit) with a total power of 1.5 kW. Only a maximum of 9 chassis can be installed in a rack, while the rack height would allow 18–20.

In the course of 2017, 29 Gen-1 racks were removed and twelve new Gen-3 racks with a cooling capacity of 45 kW have been installed.

Eleven more Gen-3 racks will be added during LS2 in 2019, bringing the total to 1104 rack units specifically for the HLT servers, with an allowed power density of 1 kW per rack unit. The necessary additional upgrade of the piping will be ready for implementation in 2018. A modified powering scheme is envisaged to replace the current Top Of the Rack (TOR) distribution boxes and 1U Power Distribution Units (PDUs), such that all the rack space will be available for servers\(^3\). Such racks will be able to accommodate the expected power dissipation of the servers projected for Phase-2 (see Section 2.4.1). The current counting room with these 23 high power racks will be sufficient throughout Run-3 and up to LS3.

### 4.3.2 Phase-2 Online Data Center Options

In what follows, we consider the different options for the evolution of the surface data center at P5 in view of the needs of the Phase-2 CMS DAQ, in particular for the HLT CPU as outlined in Table 2.3. The projected power and cooling required will be approximately 1–2 MW for the HLT and 0.5 MW for the rest of the DAQ and IT infrastructure.

#### 4.3.2.1 Upgrade of the Existing Data Center

An upgrade of the existing data center, in its current location, implies:

1. Installation of an additional transformer on site;
2. Replacement of the final distribution cabinet, and of the 160 A and 250 A canalis feeds, with 400 A feeds;

\(^1\)The Power Usage Effectiveness (PUE) metric is a popular method of calculating energy efficiency. PUE is the ratio of total amount of energy used by a computer data center facility to the energy delivered to computing equipment.

\(^2\)Rack Unit (U): unit of measurement defined as 44.50 mm (1.752 in), used as measurement of the overall height of electronics and computing racks, and of the equipment that fit inside them, expressed in multiples of rack units.

\(^3\)Initially, only up to 20 quad-server 2U chassis will be installed per rack, reserving the remaining 8U for control and data switches.
3. Increase of the total cooling capacity of the room to 1.5–2.5 MW. A new water distribution system with capability of approximately 350 m$^3$/h and a maximum flow of 4–5 m$^3$/h per rack will be required.

The existing cooling towers on the CMS site do not provide sufficient cooling capacity for the required power increase. Hence, additional capacity must be created. Free cooling with hybrid cooling towers is discussed in detail in Appendix D.3. This technology can achieve a PUE of 1.1, and, hence, has a significantly lower power consumption compared to the existing system with a PUE of 1.4. An additional advantage is the decoupling of the data center operation from the yearly maintenance of the main cooling towers.

An important drawback of this option is the need to schedule an extended outage of the network starpoint of the CMS site, currently located in the surface counting room. This starpoint is necessary in order to maintain a minimal DAQ functionality and to provide network connectivity for the whole site during LS3.

Removal of old piping and electric distribution, removal of false floor support to match the new rack footprint, re-installation of the new infrastructure, and finally installation and commissioning of the new DAQ system will require careful planning to fit in the global LS3 schedule.

**4.3.2.2 Container-based Data Center for HLT**

The computing for the HLT could be provided by industrial container-based data centers. The existing data center would host the computing and networking infrastructure, the DAQ I/O nodes and the storage, whereas the HLT nodes would be located in containers at the P5 site. A typical industrial container can house 15 racks, 500 kW power and is based on free air cooling (see Appendix D.3). The required HLT capacity could be hosted in 2–4 such containers.

This option implies:

1. Installation of an additional transformer on site;
2. Provision of infrastructure to place the computing and power containers and electrical connections to the transformer;
3. Provision of network connections to the data center in SCX.

A drawback of this solution is the physical separation of the HLT from the rest of the DAQ system and the longer distance required for the network interconnects. Depending on the location of the containers, this might require more costly transceivers and/or limit the connectivity for possible HPC interconnects.

**4.3.2.3 New Data Center at P5**

Feasibility studies for a new Data Center on the CMS site have been carried out. The initial features of the new data center would be similar to those of the current counting room after refurbishment, with several added benefits:

1. Evolutionary design of the infrastructure to allow novel solutions for cooling and power distribution as they become available;
2. Maintained DAQ and network services functionality from the existing counting room during LS3. The value of this functionality for installation and commissioning in the very tight LS3 schedule, for DAQ and other projects, cannot be overestimated;
3. Availability of the current counting room for other uses after LS3.

4.3.2.4 Remote Data Center in Prevessin

Discussions have been held concerning the possible construction of a new Data Center on the CERN Prevessin site, which would potentially host the online farm of the experiments. Since at the moment there are no concrete proposals made to the collaboration as to how this option would be implemented and financed, we postpone its discussion until it becomes a realistic possibility.

4.4 IT Infrastructure and Systems

For Phase-2, the number of nodes for DAQ and HLT will increase by a factor of 4–5 compared to the present system. This has implications on the ability of the infrastructure to scale by a similar factor. Furthermore, the greater use of embedded devices (in particular, well over 1000 for the ATCA hardware), implies an increase by a factor of 20–40 in the number of sub-detector DAQ control devices with respect to the current system.

This shift towards having PC-like control hardware (with an OS, reasonable CPU, and memory), closer to the traditional electronics hardware, means extending centralized management to these devices. A clear set of requirements on how and by whom the devices will be managed must be elaborated and agreed upon before the TDR. The level of integration at the operating system, computer infrastructure, and software layer will also have to be addressed. Moving the sub-detector DAQ applications to these platforms without any changes is only possible if one can use existing builds for the software, which implies an Intel based architecture with sufficient CPU and memory. Alternate platforms/architectures would require support in the DAQ online software framework and an increased load on the team providing the corresponding builds.

Reliable operation of the DAQ is crucial to take efficient, high quality data to the full potential of CMS. To achieve this, one should maintain a robust configuration management system and powerful monitoring system, which extends and encompasses also those embedded devices.

In the next sections, the various aspects of the IT infrastructure will be discussed and plans for future support examined.

4.4.1 Control Network

The CMS control network is a high performance distributed network, isolated from the CERN General Purpose Network. It is configured and managed by CERN IT, according to specifications set by the CMS online system administration, to provide the required functionality for the different online areas. For Phase-2, a complete replacement of the star points and their routers on the surface (SCX5 and some outlying buildings) and underground (USC55) will be required. The replacement of the distributed TOR switches according to need must be anticipated. An upgrade of the inter-router links, to follow the evolution of the overall traffic bandwidth, will be necessary, and will entail the upgrade of the fiber infrastructure (from 10 Gb/s to 40 or 100 Gb/s). A general upgrade and expansion of the fiber infrastructure in the data center should be foreseen to fulfil the needs of the core IT infrastructure services. In addition, the

---

4Reasons for replacing the switches include: end of warranty, increased network speed, increase in number of ports, change of location, etc.
integration of ATCA hardware into the CMS control network needs to be investigated and addressed\textsuperscript{5}. These boundary and usage conditions will need to be known in advance (probably at least one year before the start of LS3) in order to dimension the core network hardware and have the work done by IT or support services.

### 4.4.2 Operating Systems

The CMS online computing system runs the CERN supported version of Linux, currently CERN CentOS 7 (CC7), and it is expected that it will continue following the evolution of the CERN supported version of Linux, and use the latest version available before LS3. However, it must be noted that, for reasons of reproducibility, maintainability and security, not all software which can run on this version will be supported\textsuperscript{6}.

### 4.4.3 Sub-detector Hardware Support Infrastructure

Today, sub-detector hardware support is centrally provided for VME crate controller PCs and MicroTCA client hosts, as well as for MicroTCA network connectivity. Furthermore, all network connectivity needs for the sub-detector hardware (power supplies, PLCs, etc.) are provided for in the control network.

Since more and more hardware is coming with embedded OSs (usually Linux-based but often proprietary distributions, and on different architectures), these also will require regular maintenance and management (e.g. security patches, etc.). In the context of efficient and reliable data-taking, configuration management is a necessity. Historically, the OS management has been taken care of by the DAQ system administration (mainly Linux) and the DCS team (mainly Windows), while sub-detector groups have managed the firmware versioning and related hardware. Centralized management is desirable, but it will not be possible to fully support and centrally manage a large number of different embedded operating systems.

The number of ATCA boards, each with an embedded OS, will be of order 2000, which is about the scale of the present centrally managed PC nodes. It will be necessary to establish rules and requirements for embedded systems to operate in the P5 network. Clear procedures should be put in place for configuration and patching of these embedded systems (at least for central configuration aspects such as network, computing services, network disks, etc.). Central mechanisms for version control and configuration management will be proposed. Clearly, to limit the risks and the costs to fulfil the requirements, it is desirable (if not required) to minimise the diversity of embedded systems, across DAQ and sub-systems.

#### 4.4.3.1 ATCA-based hardware

As sub-detectors are moving to ATCA based hardware, an increase in required network connectivity for crates should be foreseen. As a baseline, it is assumed that Ethernet connectivity, in a homogeneous, centrally managed and monitored system, will be provided as part of the CMS control network. Maintenance and upgrade of any embedded network switches in ATCA crates will have to be closely coordinated between users, CMS sysadmins, and CERN IT.

\textsuperscript{5}Considerations include integration of ATCA crate embedded switches, number of required connections, bandwidth requirements, etc.

\textsuperscript{6}One such example is PIP installed Python modules, which are not supported in the online computing cluster after discussion with control system experts, CERN Linux and security teams. Users wanting such features will be requested to package the corresponding module in the standard supported way (currently RPM), and from then on take full responsibility for its maintenance and updates. This is only an example, and the guidelines and restrictions will evolve with subsequent versions of the operating system and development tools. Some investigations in this area could be undertaken for common tool sets.
The availability and pricing of small embedded processor systems should allow the Phase-2 electronics designs to embed board management and monitoring functionality onto each ATCA blade, thereby removing the need for a corresponding network of control PCs. It should be anticipated to equip each ATCA blade in CMS with a management and monitoring mezzanine that also provides the network connectivity. It is desirable that such mezzanines use an industry standard form factor, with standard connectivity. Evaluation of some products is ongoing. This includes an assessment of the use of the configuration management system deployed currently for all PCs in P5, but also the integration with user management, network file systems, etc. One interesting option is commercial mezzanines based on the COMExpress [43] (Computer on Module) PICMG [44] standard, with PCIe for high speed communication with the ATCA onboard electronics.

4.4.4 Core Computing Services

Core services required for the computing infrastructure encompass DNS, DHCP, LDAP, Kerberos, proxies, user gateway nodes, general user login nodes, user management, home directory management, etc. These will continue to be provided inside the CMS computing domain by the central DAQ system administration team, in collaboration with CERN IT (for certain services such as Windows Domain controllers, etc.).

The core services listed above are provided as redundant services, as detailed below.

4.4.4.1 Service Redundancy

Currently service redundancy is achieved in multiple ways. Multiple servers with load balancing and failover are widely used\(^7\). High Availability mechanisms for starting/moving services across machines on failover scenarios\(^8\) are a second option, and finally virtual machines (VMs) with failover, which will restart the VM and, if necessary, migrate it, a third. Currently most core services are redundant to rack failure (failover to neighboring rack) or power failure (see next section). These or any other standard OS mechanisms that might appear in the future should be leveraged to obtain service redundancy for core services that require high availability.

4.4.4.2 Power Redundancy

Core services (as listed above) are also designed to survive power outages, and are therefore dual powered from normal power and UPS power\(^9\).

4.4.4.3 NAS storage

Today, CMS uses a fully redundant NAS storage system for all the core storage needs, such as users’ home directories, sub-system project areas and general central storage needs. This storage is exposed to all Linux and Windows PCs. Similar or increased requirements for Phase-2 are expected to be fulfilled with hardware existing or to appear, with preference for solutions used elsewhere at CERN (as now). The actual total size and I/O requirements for the lifetime of the equipment should be assessed as close to purchase time as possible.

---

\(^7\)For example using HA proxy [45].
\(^8\)For example using the ClusterLabs stack [46] using corosync and pacemaker.
\(^9\)General services, which are not critical to safety systems of CMS, are on a UPS branch covering 10 minutes. Services (DNS, DHCP, core networking, etc.) which are critical for the safety systems, are on a UPS branch which will continue to operate after the UPS batteries ran out of power, due to a diesel generator which takes the load within one minute.
4.4.4.4 Configuration Management System

CMS has been using a configuration management system since its inception. Currently this is based on Puppet [47]. The benefits of using such a system are indisputable. It is foreseen to continue using such a system in the future, selecting among the open source tools available at the time and taking into consideration their adoption across CERN.

4.4.4.5 Monitoring System

Currently the monitoring system is used to verify the correct operation of the computing infrastructure. It is based on industry standard open source tools\textsuperscript{10}, which is also expected to be the case in the future. The evolution of the tools and the monitoring cluster size will be done to match the evolving needs of the system.

4.5 Conclusions

The IT infrastructure required to operate the Phase-2 DAQ and HLT will have to scale up to deal with an increase by a factor five in the number of computing nodes and an even larger factor in that of networked embedded devices.

Detailed cost estimates of the three options for the upgrade of the data center technical infrastructure will be available for the final DAQ/HLT TDR, when a decision will be taken.

The upgrades of the IT and technical infrastructure need careful planning within the general technical coordination schedule of CMS to ensure continued support for the commissioning and installation activities during LS3.

\textsuperscript{10}Icinga [48], graphite [49] and grafana [50].
Chapter 5

Development Directions

In this chapter we discuss development directions that can either simplify the DAQ/HLT design, or provide new interesting features.

5.1 Event Building over D2S

As discussed in Chapter 2, the baseline D2S system envisions an Ethernet network with limited switching capability between the DTH boards and the I/O processor nodes performing event building, and the use of a reliable protocol such as TCP/IP. In this scheme, which is essentially identical to the one of the current CMS DAQ system, the data sources are configured to send to a fixed destination and event building is performed at a subsequent stage on the I/O nodes communicating over a powerful HPC interconnect.

An interesting alternative design would be to incorporate the event-building logic into the D2S system. In this design, an external controller instructs the DTHs for each (group of) triggers to which destination I/O node the fragments have to be sent. This requires additional FPGA logic and memory in the DTH, not only for the more complex protocol and its associated latency, but also for sophisticated error reporting and diagnostic. The controller is naturally the TCDS, where the CPM receives the trigger information, adds the destination assignment and broadcasts this information to the TCDS module in all the DTH boards. The DTH provides a direct path to throttle triggers if the event building process cannot keep up. Methods to provide dynamic load balancing across the destination I/O nodes must be developed, to avoid the slow down of the overall event building process due to one slower I/O node.

This alternative design requires full non-blocking interconnectivity between the DTH and the I/O nodes, but removes the need for an additional Event Network connecting the I/O nodes among each other.

5.2 TCDS Support for Prioritized Throttling

One of the main new features of the Phase-2 TCDS is the capability to distribute a trigger type for each bunch crossing (see Section 2.2.2.1). This feature provides the possibility to group subdetectors according to their acquisition rate capability. Each rate group could have a corresponding set of trigger types. To first order, the trigger throttling system (TTS) would not need any modifications for this approach: any subsystem asserting not-ready would stop the trigger flow for all rate groups. A more efficient extension of the TTS could be to ‘prioritize’ TTS states across the rate groups, allowing targeted throttling of single rate groups based on only their own TTS state.
5.3 TCDS Support for Latency Measurement

In both Run-1 and Run-2 significant time has been spent on measuring, consolidating, and verifying the latency in various paths of the Level-1 systems.

The overall latency is composed of three contributions (see Fig. 2.4):

1. The Level-1 latency, measured between the time of the (physics) collision and the corresponding Level-1 decision reaching the input to the TCDS (“L1A” in Fig. 2.4).

2. The TCDS latency incurred due to trigger control and scheduling. This is measured between the input and the output of the TCDS (“CPM” in Fig. 2.4).

3. The ‘return’ latency, measured between the output of the TCDS and the input to the subsystem DAQ backend (“CPM” and “BE blade” in Fig. 2.4).

The TCDS latency (2) could be measured by instrumenting the firmware of the CPM.

The ‘return’ latency (3) from the CPM down to any end-point in the distribution tree could be determined from the round-trip time between the CPM and the end-point. Such a system would imply the need for a return path up the distribution tree to the CPM.

The most complex measurement to perform is probably the Level-1 latency (1). This is composed of multiple steps across different boards and links in the Level-1 systems. An unambiguous measurement would require the distribution of an independent time reference for time-stamping events along the Level-1 signal path.

It would be highly advantageous to have a built-in way to measure and verify the latency. The feasibility of implementing support for measurement of one or more of the above latencies (1–3) has to be studied. The conclusion of these studies should become part of the Phase-2 DAQ TDR.

5.4 Federated Event Data Indexing

Leveraging the computing power potentially available in the I/O processors, as well as ancillary engines, such as GPUs and FPGAs, installed as coprocessors in the I/O nodes or in separate nodes connected to the event network, to perform pre-processing and organise data, it is possible to envision a further alternative approach to the event building, where event fragments, both raw from the sub-detectors and resulting from processing in dedicated “algorithm farms”, 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 [51].

This scheme is innovative with respect to the more “classic” ones described so far, and does away with full event building altogether. It assumes that processing happens both on the I/O nodes, with co-processors taking most of the computational load (for example to convert raw data to convenient formats, apply calibrations, do low-level standalone reconstruction), and on dedicated nodes that combine information from different sources (for example tracks and calorimeter clusters, to perform particle-flow reconstruction), but without ever explicitly moving all data for one event into the memory of a single computer. Instead of relying on the cluster interconnect for full event building, this design would use the bandwidth and low latency to “bring the algorithm to the data”. The net result would be that the entire online farm, made of the I/O processors and additional servers used for high-level object reconstruction, would provide a distributed index of physics objects and an endpoint to query this index. The HLT would
then consist of a set of queries that are submitted to this endpoint using a “search-engine”-type API, which would return only the (list of) events that satisfy the query, for which raw data and all the intermediate products could then be moved directly to storage (independently, by the nodes that hold them, using the same HPC interconnect).

### 5.5 Partial Acquisition at Bunch Crossing Rate

As seen in Section 2.1, in the upgraded CMS several detectors will be read out in streaming mode, i.e. for every bunch 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-$p_T$ 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 produce fast calibration or monitoring for streaming detectors, or to enable the study of physics channels lacking a well defined signature for Level-1 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 (see Fig. 5.1). We call this system “Level-1 scouting” in analogy to Ref. [52]. 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 [51]. 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. This could be used to assist “standard” scouting in the HLT, which is usually limited by the inferior quality of the calibrations and alignment typically available online.

![Figure 5.1](image_url)

**Figure 5.1:** A cartoon of the Phase-2 CMS DAQ including the 40 MHz scouting system.

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. This idea presents interesting challenges and the feasibility and actual physics potential of such a system should be studied in more detail.

One important aspect to note is that such a system can be implemented progressively, and it does not impose any constraint on the upgrade schedule. Indeed, a 40 MHz scouting system making use of data from a single detector or trigger subsystem could prove already extremely useful. On the other hand, it is important to take this possibility into account when designing the upgraded detector backend electronics and the Level-1 trigger system, to avoid design decisions that would make the extraction of intermediate data at full bunch crossing rate impossible.
Chapter 6

Organization, Schedule and Costs

6.1 Organization

The CMS DAQ project is responsible for development, commissioning, operation, maintenance and upgrade of:

- the “core” DAQ, to transfer event data from the sub-detector back-end electronics to the HLT facility and save the HLT accepted events to storage;
- the Trigger and Control Distribution System (TCDS);
- the IT infrastructure (computing and network services) at point-5;
- the Detector Control System (DCS).

This includes the hardware, firmware and software needed for operating the DAQ system and for monitoring the performance of the data taking. The DAQ and TCDS Hub (DTH) board implements a common interface to the subdetector specific backend electronics (see Section 2.1.1) and is a deliverable of the DAQ project. Note that the development of the HLT algorithms is not in the scope of the DAQ project, but the responsibility of the Trigger Coordination inside the CMS organization structure.

Due to its central role in DAQ, TCDS, DCS and IT infrastructure, the DAQ project is the “owner” of the interfaces between the central system and all subdetector systems. The definition of the interfaces, the use of common solutions across subsystems and the development of the R&D program are coordinated by the Electronics and Online System Working Group (EOSWG), reporting to CMS upgrade coordination.

The institutes involved in the DAQ project of the initial CMS detector and the Phase-1 upgrade are CERN and a group of US institutes (Fermilab, MIT, UCSD, Rice University, and until recently UCLA). A few other institutes have contributed over the years in well defined areas. It is expected that the current institutes will continue their involvement for Phase-2.

6.2 Schedule

The schedule for the DAQ activities until the completion of the Phase-2 upgrade is shown in Fig. 6.1. The DAQ system follows the schedule of the LHC running periods, and is upgraded during the LS periods. The DAQ-n system is the system deployed during LHC run-n. Note that running conditions, in terms of instantaneous luminosity and pile-up, are not expected to change significantly between Run-2 and Run-3, prior to the HL-LHC upgrade. According to the current LHC schedule, the HL-LHC will commence physics operation in Q4/2026 (“HL-LHC run 4” in Fig. 6.1). The aim is to have the DAQ system installed as soon as possible during LS3 in order to allow sufficient time for commissioning. The date for submission of the DAQ
TDR is Q2/2021. The milestones planned in preparation for the TDR are listed in Table 6.1.

<table>
<thead>
<tr>
<th>Milestone</th>
<th>Date</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>L1.1</td>
<td>Q3/2017</td>
<td>Phase-2 Upgrade DAQ Interim Technical Design Report</td>
</tr>
<tr>
<td>DTH.0</td>
<td>Q4/2017</td>
<td>Initial specification of the DTH (DAQ and TCDS Hub)</td>
</tr>
<tr>
<td>DTH.1.1</td>
<td>Q2/2018</td>
<td>DTH-P1 prototype-1 board</td>
</tr>
<tr>
<td>DTH.1.2</td>
<td>Q2/2019</td>
<td>Evaluation of functions and performance of the DTH-P1</td>
</tr>
<tr>
<td>DTH.2.1</td>
<td>Q2/2020</td>
<td>DTH-P2 prototype board</td>
</tr>
<tr>
<td>DTH.2.2</td>
<td>Q2/2021</td>
<td>Demonstration of full functional and performance requirements of DTH-P2</td>
</tr>
<tr>
<td>TCDS.1</td>
<td>Q2/2019</td>
<td>Test of timing performance with DTH-P1</td>
</tr>
<tr>
<td>TCDS.2</td>
<td>Q1/2021</td>
<td>Demonstration of TCDS multi-crate setup</td>
</tr>
<tr>
<td></td>
<td></td>
<td>with CPM and few DTHs</td>
</tr>
<tr>
<td>TDAQ.1</td>
<td>Q1/2021</td>
<td>Demonstration of TDAQ slice with full chain</td>
</tr>
<tr>
<td></td>
<td></td>
<td>TCDS-D2S-EVB-HLT-5SM</td>
</tr>
<tr>
<td>L1.2</td>
<td>Q2/2021</td>
<td>Phase-2 Upgrade DAQ Technical Design Report</td>
</tr>
</tbody>
</table>

Table 6.1: Milestones till the TDR for the Phase-2 DAQ.

The DAQ system deals with two types of components: custom electronics, encompassing the TCDS and the DTH boards; and commercial computing, networking and storage equipment for the data transport, event building, HLT and storage. The systems based on these two types of components follow an appropriate process and schedule.

During the years 2018–2019, demonstrators for the DAQ-3 system will be built ("DAQ-3 Demo"). The D2S system, with custom electronics, will not be changed, but replacement of the (by then mostly obsolete) D2S Ethernet switches, event builder nodes and network, HLT farm, and storage is planned. Elements of the Phase-2 DAQ baseline architecture will likely be included, albeit not at the full Phase-2 performance and scale. Results from the DAQ-3 demonstrator will be used for the Phase-2 TDR.

The roadmap for the DTH is outlined in Section 2.1.3. A DTH specification is planned by the end of 2017. The aim is to have an initial definition of the interfaces so that the subdetectors can proceed with the design of their prototype backend boards. The DTH-P1 prototype will be manufactured during Q1-Q2/2018 ("P1 manufacturing"), followed by firmware and software development and testing till mid 2019. ("P1 testing"). A second prototype cycle is scheduled two years later ("P2"). Results will be incorporated in the TDR. After an Electronic System Review (ESR) in Q4/2022, manufacturing of the production DTH is planned, during Q1-Q2/2023 ("Production manufacturing"), followed by testing in Q3-Q4/2023 ("Production testing"). Pre-production of a small series is foreseen one year earlier.

The roadmap for the TCDS is outlined in Section 2.2.6. The R&D on various aspects of the Phase-2 TCDS can be done with the P1 prototype of the DTH ("Test with DTH-P1"), followed by R&D on a multi-crate setup ("TCDS Demo-1") in time for the TDR. A decision for the implementation for the CPM hardware with an existing-board or a new dedicated board is foreseen at the time of the TDR. A demonstrator system based on an existing board or on a prototype of a dedicated board ("TCDS Demo-2 & CPM") will be used for development till the ESR in Q4/2023, reviewing the full TCDS system. In case a dedicated board should be required, sufficient time for development and manufacturing will be available before the installation in early 2025.

Three DAQ-4 demonstrators are planned to test commercial equipment, develop software, and
Figure 6.1: Schedule for the DAQ activities and the Phase-2 upgrade.
perform integration. The DAQ-4 EVB will concentrate on the event builder (see Section 2.3.1), and possibly the integration of co-processors (see Section 3.4.1), using commercial computing and networking equipment. The complementary EVB+D2S demonstrator will focus on the integration of the DTH output with the EVB. The TDAQ Slice aims for integration of the full chain encompassing TCDS, D2S, EVB, HLT and Storage Manager. These demonstrators will evolve (“Demo 1–3”) following various generations of custom electronics, commercial equipment, and software functionality.

There will be an update of the DAQ-4 demonstrators in 2023 (“Demo-3”), with state of the art commercial equipment, in order to prepare the final equipment specification. Selection and purchase of the commercial computing equipment (“DAQ-4 EVB-SM”) for data transport, event builder, and storage is foreseen for 2024, followed by installation in early 2025. The HLT farm will be installed gradually. It is assumed that commissioning activities and first operation with beams in Q3-Q4/2026 can be done with the HLT equipment from Run-3.

There is no detailed LS3 schedule of the technical coordination for DAQ related items yet. In the above schedule for the DAQ, it is assumed that the full year 2024 will be needed to de-commission the old system in USC and the datacenter in SCX, and install new basic infrastructure (“USC+SCX infrastructure de-commissioning”). The first half of 2025 is foreseen for installation and central-DAQ internal commissioning (“cDAQ commissioning”). From the second half of 2025, commissioning with subdetectors can commence (“global commissioning”). This schedule will be updated in due course according to the overall CMS schedule.

### 6.3 Cost Estimate

The estimated CORE\(^1\) cost for the DAQ/HLT is shown in Table 6.2. The estimate is based on the bill of material for the baseline design discussed in Section 2.8.

<table>
<thead>
<tr>
<th>Area</th>
<th>Accounted in project</th>
<th>Cost (MCHF) DAQ + HLT 750 kHz (LS3)</th>
<th>Cost (MCHF) DAQ + staged HLT 500 kHz (LS3) + 250 kHz (LS4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>custom DTH boards</td>
<td>sub-detector</td>
<td>(1.0)</td>
<td>(1.0)</td>
</tr>
<tr>
<td>TCDS and DAQ</td>
<td>DAQ</td>
<td>5.5</td>
<td>5.5</td>
</tr>
<tr>
<td>HLT</td>
<td>DAQ</td>
<td>13.4</td>
<td>7.1</td>
</tr>
<tr>
<td>Total (DAQ/HLT)</td>
<td></td>
<td>18.9</td>
<td>12.6</td>
</tr>
</tbody>
</table>

The DTH (see Table 2.4, item 1) is a deliverable of the DAQ project, but accounted for in the sub-detector project cost. The cost is based on an initial estimate of the number of units and the cost of the DTH-400 and DTH-1200 board\(^2\).

The cost of the TCDS, DAQ data transport to surface, event builder and storage (see Table 2.4, items 2–10) is estimated based on the number of units, the estimated cost of the custom electronics boards and the cost of commercial processors, network and storage items and their price/performance evolution based on past purchases.

As discussed in Section 1.5.1, the estimated HLT computing needs for running at 750 kHz and PU = 200 amounts to 9.2 MHS06. Taking into account that HL-LHC instantaneous luminosity is

\(^1\)The concept of CORE cost has been developed for the CMS Construction Project MoU and LHCC Cost Review Committee. CORE costs are defined as M&S (Materials and Services) costs for the production phase of the project. Costs to support long term Maintenance and Operation (M&O) are not included.

\(^2\)Includes cost of DTH board and optics for connection to back-end boards.
expected to increase through the initial HL-LHC operation and that the design Level-1 trigger rate includes a 50% margin, a staged deployment of the HLT is proposed as a baseline plan, with 4.5 MHS06 installed in LS3 (in 2026) to run at 500 kHz and PU = 140, and an additional 4.7 MHS06 staged to LS4 (in 2030) for running at the full 750 kHz and PU = 200. For the HLT cost estimate, presented in Table 6.2, it is also assumed that a constant M&O-A\(^3\) budget will continue to be available, and that the HS06 unit cost will follow an average of the 1.1 to 1.2 performance/cost improvement per year projection, tracked by the CERN-IT department in the context of the Worldwide LHC Computing Grid (WLCG) [53]. This leads to a HS06 unit cost of approximately 2 CHF and 1.5 CHF respectively in 2026 and 2030. Note that the bandwidth of the core DAQ is not staged. Full connectivity to all back-end electronics is required from the start and experience has shown that early running is done with larger than nominal event sizes until the new detector is well understood. Furthermore, it cannot be excluded that alternative computing technologies will provide a cost effective solution to operate the HLT at 750 kHz, with a broader physics reach. Section 3.4.1 describes the use of coprocessors in a heterogeneous HLT architecture and the performance/cost for one particular use case. Further studies are needed to assess the effort required to re-engineer the HLT software and the potential overall cost savings in computing equipment for an HLT based on such architectures.

Not included in these CORE costs are the replacement of the IT infrastructure (computing and network services), database, and DCS system. These are assumed to be covered by the regular M&O-A replacement. Also not included is the cost associated with the upgrade of the cooling and power infrastructure of the data center (see Section 4.3). The cost and the source of funding for the latter two items have not yet been established. They will be evaluated for the TDR with a better estimate of the CPU needs and considering the four options outlined in Section 4.3.2.

It should be noted that the uncertainties on these cost estimates are large, as the majority of the cost results from pricing of commercial computing and networking equipment and uncertain estimates of their price/performance evolution over a long timescale. Furthermore, there are also large uncertainties on the HLT computing needs. Improved estimates are expected to become available at the time of the TDR. The DAQ/HLT system has a scalable design, which enables adaption to evolving needs.

---

\(^3\)M&O-A is Maintenance and Operation, Category A (CMS wide, i.e. not sub-detector specific).
Appendix A

Optical Link Technologies for LHC Upgrades

Due to the presence of radiation and a strong magnetic field inside HEP detectors, as well as the large physical separation between the detector front- and back-ends, almost all communication between the detector and the back-end electronics is based on custom developed optical links. As was the case for the initial construction of the LHC experiments, common projects are established to develop optical transceivers as well as transceiver ASICs. This appendix provides an overview of some of these projects relevant to the CMS upgrades.

A.1 Phase-1 and Phase-2 Upgrades

The life span of the LHC and most of its experiments contains two major upgrade phases after the initial construction: Phase-1 (i.e., installation during LS2), and Phase-2 (i.e., installation during LS3). Though the actual installation planning may vary for the different experiments and/or projects, the following text refers to the terms Phase-1 and Phase-2 in their original meaning.

A.2 Optical Link Systems

The links between the on-detector front-end electronics and the back-end electronics in the counting room typically carry three types of data:

- Clock, trigger, and timing information from the back-ends (down) to the front-ends. Typically low- and fixed-latency.
- Slow-control data (e.g., configuration and monitoring commands) from the back-ends (down) to the front-ends.
- Slow-control/monitoring information from the on-detector electronics (up) to the back-ends.
- High-volume, low-latency DAQ and trigger data from the on-detector electronics (up) to the back-ends.

The initial design of the LHC experiments used common developments, like the RD12 TTC system for the distribution of clock, trigger, and timing information [25, 54]. The slow-control and DAQ/trigger read-out solutions used were more diverse.

For the Phase-1 upgrades, the GigaBit Transceiver (GBT) [55] chipset was developed, together with the Versatile Link (VL) [56] series of optical transceivers. This transceiver-and-chipset pair was designed to be tolerant to radiation levels encountered in calorimeters of LHC experiments. The GBT chipset combines the clock and timing, slow-control, and DAQ data into a single link, removing the need for separate solutions. These combined Phase-1 developments
Appendix A. Optical Link Technologies for LHC Upgrades

<table>
<thead>
<tr>
<th>Era</th>
<th>Component</th>
<th>Line speed</th>
<th>Capacity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phase-1</td>
<td>GBT</td>
<td>4.8 Gb/s</td>
<td>DAQ: 2.56 Gb/s</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>TTC: 640 Mb/s</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Slow-control: 160 Mb/s</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Down: 2.56 Gb/s</td>
<td>User: 1.28 Gb/s</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>ASIC control: 80 Mb/s</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>External control: 80 Mb/s</td>
</tr>
<tr>
<td>Phase-2</td>
<td>IpGBT</td>
<td>Up: 5.12/10.24 Gb/s</td>
<td>DAQ: 3.84(4.48) Gb/s with FEC12 (FEC5)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>ASIC control: 80 Mb/s</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>External control: 80 Mb/s</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Down: 2.56 Gb/s</td>
<td>User: 1.28 Gb/s</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>ASIC control: 80 Mb/s</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>External control: 80 Mb/s</td>
</tr>
</tbody>
</table>

Table A.1: Current understanding of front-end link chipsets developed as common projects for the Phase-1 and Phase-2 upgrades of the LHC experiments. For the corresponding transceivers, see Table A.2. The IpGBT implements two different Forward-Error-Correction schemes: FEC5 aimed at maximum throughput, and FEC12 aimed at maximum resilience against single-event upsets. For details, please refer to the IpGBT documentation.

<table>
<thead>
<tr>
<th>Era</th>
<th>Component</th>
<th>Max. link speed</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phase-1</td>
<td>Versatile Link (VL)</td>
<td>5 Gb/s</td>
<td>Single-mode or multi-mode</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Package: Tx+Rx or 2Tx</td>
</tr>
<tr>
<td>Phase-2</td>
<td>Versatile Link Plus (VL+)</td>
<td>Rx: 2.5 Gb/s, Tx: 10 Gb/s</td>
<td>Multi-mode (850 nm)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Package: 4Tx+1Rx</td>
</tr>
</tbody>
</table>

Table A.2: Current understanding of optical transceivers developed as common projects for the Phase-1 and Phase-2 upgrades of the LHC experiments, corresponding to the front-end chipsets in Table A.1.

also clearly demonstrate the approach for link solutions for the HL-LHC upgrades: custom, radiation tolerant transceivers paired with custom, radiation tolerant transceiver chipsets in the front-ends, and standard commercial transceivers combined with standard FPGAs in the back-end electronics.

For the Phase-2 upgrades, successors of the GBT and Versatile Link are under development. The Phase-2 transceiver ASIC is the low-power GBT (IpGBT) [57], providing higher bandwidth and adequate radiation tolerance for use in the LHC Phase-2 trackers. The Versatile Link Plus (VL+) [58] replaces the original Versatile Link, matching the design goals of the IpGBT for Phase-2 upgrades.

Tables A.1 and A.2 show a brief summary of the characteristics of the transceivers and chipsets discussed above. For a more comprehensive discussion, please refer to the relevant references.
Appendix B

Modular Electronics

The original CMS systems used a mixture of different modular electronics standards, ranging from fully custom systems (mostly for on-detector electronics), through VME (for sub-detector back-ends), to compact-PCI (for the data to surface FRL/FEROL and TTS systems). It is worth noticing that in no case were the crate data busses used for actual data transportation, which was, in general, realized with dedicated front panel connections or, in some cases, with the addition of custom backplanes.

With progressive phasing out of parallel buses, and driven by the need for ever higher bandwidths, the subsequent upgrades have focused on the use of standards supporting high-speed telecom-grade serial backplanes.

B.1 MicroTCA

The Advanced Mezzanine Card (AMC) and MicroTCA standards [59, 60] were selected for the Phase-1 upgrades, in particular for the detector back-end electronics systems.

A full size, double height MicroTCA crate, also known as shelf, as adopted by CMS, is schematically depicted in Fig. B.1.

Figure B.1: Schematic view of a MicroTCA crate and its typical use in a CMS back-end system.
Elements of a MicroTCA shelf implement the requirements of an AMC carrier board as defined in the AMC standard. The term “carrier” refers to functions needed to provide an infrastructure that supports multiple AMCs. The crate features a redundant backplane consisting of point-to-point “fabric” interconnects. A fabric interconnect provides the main connectivity among the AMCs. This interconnect consists of a central switch and a number of high speed serial lanes to each AMC position. Lanes on MicroTCA are differential high speed serializer/deserializer interconnects (SerDes), with bidirectional bandwidth capability of up to 10 Gb/s per lane. The MicroTCA Carrier Hub (MCH) combines the control and management infrastructure and the interconnect fabric resources to support up to twelve AMCs in a single module, with the possibility to include two MCHs for high-availability applications. This latter feature is not used in CMS.

The fabric component of an MCH is the hub of a star network. The interconnect protocol used depends upon the specific format expected by the AMCs and is implemented on the MCH.

The AMC is unique in its dual specification as the mezzanine card on ATCA carriers as described in the ATCA standard [8, 61] and as the blade, or main electronics module, in MicroTCA systems.

### B.1.1 Phase-1 CMS MicroTCA systems

In a typical CMS MicroTCA back-end crate, a commercial MCH module is used, in combination with a Module Management Controller (MMC) core implemented in the different AMC modules. For the MMC, multiple different designs were retained for the different cards, resulting in two different approaches for the management of the individual mezzanines.

The second MCH slot houses a standardized DAQ interface board (AMC13 [20]), which is responsible for the aggregation of data from the individual AMC slots over the backplane, the distribution of clock and TTC signals, and the collection of the TTS status from each AMC. This is illustrated in Figure B.1. This is the case for the back-end crates of the Phase-1 HCAL (Ref. [10], Section 6), Pixel, and GEM upgrades. A similar configuration is used for the Phase-1 Level-1 Trigger. The TCDS crates use an AMC13 hardware module with a modified firmware set.

### B.2 Phase-2 Modular Electronics

With the Phase-2 upgrade of CMS, the number of links from the front-end systems, as well as the overall bandwidth requirements for DAQ and Trigger data, will increase radically. This poses both electrical and mechanical problems for the use of MicroTCA. Firstly, the aggregation of many serial input links requires large FPGAs, which need both more power and more real estate on the individual boards; secondly, the large number of optical connectors requires front panel real estate, which is not feasible, even on the double height AMC boards. These shortcomings, compounded with the limited commercial success of MicroTCA, and its uncertain future, motivated the decision to adopt the ATCA standard.

### B.3 ATCA

The ATCA standard crate specifies two types of boards: “hub” and “node”. The node blades in an ATCA crate are connected to the hub(s) over the shelf backplane by multiple point to point high speed serial links under various topologies. A common shelf specification, which will most likely be the one adopted by CMS, features two hub slots and 12 node slots and is
Figure B.2: Top: Proposed ATCA crate layout for use in a typical CMS back-end system. Bottom: Provisional CMS-defined ATCA backplane connections.

illustrated in Figure B.2. The backplane data transport interface consists of a Base Interface (BI, two duplex serial links, usually for common connectivity such as control and monitoring) and a Fabric Interface (FI, four duplex serial links used for high speed data transfer). From each hub slot these two interfaces are routed to each node slot. The BI is wired as a Dual-Star with redundant fabric hub slots at the core. The FI supports many different Fabrics (protocol-agnostic) and can be wired as a dual-star, full-mesh or other architecture. The maximum specified link speed of the FI is 10 Gb/s with a recent upgrade to 25 Gb/s. At the time of writing there is no true 25 Gb/s-link backplane available on the market, although this may change shortly.

Each board is networked via one or both hub boards, which must provide a Fabric switch. The hub board must also implement the control functionality of a Shelf Manager as described in the ATCA specifications.

An ATCA board can act as an AMC carrier board, and therefore implements the same interface specifications as a MicroTCA crate.
Appendix C

Prototype P1 for the DAQ and TCDS Hub

The requirements and design considerations for the DAQ and TCDS Hub (DTH) board are described in Section 2.1.1. The first DTH prototype (P1) is due by mid 2018. P1 will assess the following aspects:

- ATCA common services needed on a hub board:
  - Power converters for both common services and application components;
  - IPMC\(^1\) integration with the shelf manager;
  - COM Express PC integration (see Section 4.4.3) for remote control/operation of the application;

- TCDS unit:
  - shall provide basic TTC and TTS functionality;
  - will be used to evaluate timing performance with the TCDS datastream distributed over the ATCA backplane.

- DAQ module:
  - feasibility of an FPGA TCP/IP engine operating at 100 Gb/s;
  - usage of fast serial external memories;
  - usage of high-speed mid-board transceivers (e.g. FireFly, Amphenol-Leap);
  - PCB board layout with 25 Gb/s transmission lanes.

To design and construct this first prototype P1, readily available components, satisfying the technical requirements, have to be chosen. The subsequent prototypes and preseries will not necessarily use the same components.

For the ATCA common services, P1 will rely on the ongoing common CMS effort to implement and demonstrate such services on a generic board. The TCDS unit will leverage the firmware development done for the existing Phase-1 TCDS (Fig. C.1), while the DAQ unit will re-use the developments for the MicroTCA-based FEROL40 modules in the existing DAQ2 system.

The DAQ module will be based on a Xilinx Kintex UltraScale+ KU15P FPGA with footprint A1760. This FPGA features\(^2\) 32 transceivers @28 Gb/s and 44 transceivers @16 Gb/s. The usage of these transceivers is described in Table C.1. The selected fast serial memories for P1 are Hybrid Memory Cube (HMC) gen2 from Micron [62], with 2 GB capacity and a maximum DRAM bandwidth of 240 GB/s. This design should allow full speed operation of the links, i.e. concurrent 400 Gb/s input/output. The optical mid-board interconnect components will

---

1Intelligent Platform Management Controller, an ATCA standard.
2For speed grade -2.
Appendix C. Prototype P1 for the DAQ and TCDS Hub

Figure C.1: Block diagram of the prototype DTH TCDS module. SFP-based inputs are provided for both legacy and Phase-2 TCDS (right-hand side), as well as an SMA input for injection of a precision clock. The clock routing and jitter cleaning circuitry is expected to evolve before final layout.

Table C.1: Usage of FPGA transceivers on a single DAQ module.

<table>
<thead>
<tr>
<th>KU15P</th>
<th>Available</th>
<th>PCI Gen3</th>
<th>TCDS</th>
<th>Serial memory</th>
<th>4 x 100 GbE output</th>
<th>Serial inputs</th>
<th>Remaining</th>
</tr>
</thead>
<tbody>
<tr>
<td>Serdes 28 Gb</td>
<td>32</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>16</td>
<td>16</td>
<td>0</td>
</tr>
<tr>
<td>Serdes 16 Gb</td>
<td>44</td>
<td>1</td>
<td>1</td>
<td>32</td>
<td>0</td>
<td>8</td>
<td>2</td>
</tr>
</tbody>
</table>

be FireFly 28 Gb/s in a four-fold (×4) optical duplex package. Six FireFly connectors for this package will be mounted for each DAQ module input, to address two different use cases. One use case is to equip only four connectors with FireFly 28 Gb/s. Another use case is to equip all six connectors with FireFly 16 Gb/s. Hence, the DAQ unit will be capable of 400 Gb/s (Ethernet) output and up to 24 inputs, with a total input throughput not to exceed 400 Gb/s. Figure C.2 depicts a principle block diagram of the P1 DTH DAQ module prototype.

Dedicated lines are foreseen for on-board communication between DAQ and TCDS modules. Overall on-board communication for configuration and monitoring will be based on PCI-express.
Figure C.2: Block diagram of a prototype DTH DAQ module. Each DAQ module is built around a KU15P FPGA, receiving data on multiple high-speed serial lanes from the detector back-end electronics (bottom left), and driving four 100 Gb/s lines to the central DAQ (top left). High-bandwidth HMC serial memory is used to buffer the output data.
Appendix D

Technical Infrastructure and Cooling

D.1 Current Infrastructure in the Service Cavern (USC)

The current DAQ infrastructure in USC consists of:

- **Lower floor (S1):**
  - 13 electronic racks\(^1\), including 11 racks for FRL/FEROL crates and 2 racks for the TCDS system;
  - 2 PC racks for FRL control.

- **Upper floor (S2):**
  - 4 electronic racks for FRL/FEROL crates
  - 17 PC racks, including 16 racks for FEROL40/VME/DCS/services, and 1 rack for FRL control;
  - 4 passive racks for optical patch panel.

D.2 Parameters of the Current Surface Counting Room

The basic parameters of the CMS surface counting room (B3562-0001) are as follows:

- **Room dimensions:** 30m x 10m, or about 300 m\(^2\);
- **False floor:** 1m depth, 3 layers of cable trays;
- **Floor tile resistance:** 300 kg/m\(^2\);
- **Cooling system total input water flow:** 170 m\(^3\)/h;
- **Active cooling:** thermic machine allowing a \(\Delta T\) of 6 \(^\circ\)C (hence a total cooling capacity of approximately 1.2 MW);
- **Average water flow per rack:** 1 m\(^3\)/h (limited by distribution piping);
- **Electric distribution:** via the Canalis 160 A or 250 A system running on the top of a group of racks, usually eight racks per Canalis;

During Run-1, the rack room was equipped with 145 racks with active cooling doors, of two different generations (Gen-1, Gen-2). The main difference between the two is that Gen-2 racks allowed a higher cooling capacity and more space for cabling (see Table D.1). During Run-2, 29 Gen-1 racks were removed to recover space for 12 newer Gen-3 racks with higher cooling capacity and yet more space for cables and power distribution. The current piping grid has been locally modified to allow a water flow of up to 4 m\(^3\)/h specifically for the Gen-3 racks. The total available flow remains, however, 170 m\(^3\)/h. Regulation valves are tuned according to

---

\(^1\) Standard LHC experiment electronic 52U rack with vertical airflow.
individual rack consumption. At the start of LS2, in 2019, 11 more Gen-3 racks will be installed in anticipation of the requirements for DAQ3 in Run-3. The number of racks and the detailed rack parameters for each generation of racks can be found in Table D.1.

<table>
<thead>
<tr>
<th>Year</th>
<th>Gen-1</th>
<th>Gen-2</th>
<th>Gen-3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dimensions W (cm) × D (cm) × H (U)</td>
<td>60 × 90 × 47</td>
<td>60 × 100 × 47</td>
<td>80 × 120 × 48</td>
</tr>
<tr>
<td>Water flow (m$^3$/h)</td>
<td>1</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>Max cooling capacity (kW)</td>
<td>10</td>
<td>16</td>
<td>30</td>
</tr>
</tbody>
</table>

### D.3 Free Cooling for a Data Center in the Geneva Area

Free cooling is an approach to lowering the air temperature in a data center by using air (or water) at ambient temperature instead of mechanical refrigeration. Full free cooling using air is achievable in the Geneva area. The extreme climatic condition observed over the last 35 years for Geneva are summarized in Table D.2.

<table>
<thead>
<tr>
<th>Extreme temperature and the corresponding RH</th>
<th>Extreme Relative Humidity (RH) and the corresponding RH</th>
</tr>
</thead>
<tbody>
<tr>
<td>Winter -18 °C / 76% RH</td>
<td>-6 °C / 100% RH</td>
</tr>
<tr>
<td>Summer +38 °C / 21% RH</td>
<td>+18 °C / 100% RH</td>
</tr>
</tbody>
</table>

During summer, the maximum temperature observed is 38 °C / 21% RH. This translates into a wet bulb temperature of 21 °C. Under these conditions, modern hybrid cooling towers can produce water at a temperature 2–3 degrees above the wet bulb temperature. This implies that, with free cooling only, the maximum inlet water temperature to the racks is 23–24 °C on extremely hot days. Current servers are specified to operate permanently at up to 40 °C ambient maximum, and transiently, for few days, at 45 °C, without change in the reliability figures. It is to be noted that the trend in HPC operation is to increase the permanent working temperature of the servers in order to allow the use of free cooling at any location on the planet.

Table D.3 shows that the Gen-3 racks described in Table D.1 are well dimensioned to dissipate more than 40 kW under free cooling conditions with an acceptable water flow of 3–4 m$^3$/hour.

---

2A hybrid cooling tower is a combination of a classical dry closed water cooler (outdoor air used as a coolant) and a water sprinkler system activated only during the warm days, when the air is not cold enough. The water evaporation (adiabatic cooling) allows to reach temperatures close to the wet bulb temperature of the air.
Table D.3: Performance summary of Gen-3 racks

<table>
<thead>
<tr>
<th>Inlet water temperature (°C)</th>
<th>Outlet water temperature (°C)</th>
<th>∆T (°C)</th>
<th>Water flow (m³/hour)</th>
<th>Duty (kW)</th>
<th>Pressure drop (kPa)</th>
</tr>
</thead>
<tbody>
<tr>
<td>15</td>
<td>31</td>
<td>16</td>
<td>3.5</td>
<td>65</td>
<td>37</td>
</tr>
<tr>
<td>15</td>
<td>28</td>
<td>13</td>
<td>3.0</td>
<td>45</td>
<td>28</td>
</tr>
<tr>
<td>15</td>
<td>22</td>
<td>7</td>
<td>6.0</td>
<td>49</td>
<td>101</td>
</tr>
<tr>
<td>15</td>
<td>21</td>
<td>6</td>
<td>4.0</td>
<td>28&lt;sup&gt;a&lt;/sup&gt;</td>
<td>42&lt;sup&gt;b&lt;/sup&gt;</td>
</tr>
<tr>
<td>25</td>
<td>35</td>
<td>10</td>
<td>4.0</td>
<td>46</td>
<td>42</td>
</tr>
<tr>
<td>25</td>
<td>37.5</td>
<td>12.5</td>
<td>3.0</td>
<td>43</td>
<td>25</td>
</tr>
</tbody>
</table>

<sup>a</sup>Calculated value.
<sup>b</sup>Calculated value.
# Glossary of Special Terms and Acronyms

<table>
<thead>
<tr>
<th>Term</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>AMC</td>
<td>Advanced Mezzanine Card.</td>
</tr>
<tr>
<td>AMC13</td>
<td>A MicroTCA data concentration and clock distribution card specifically designed for the CMS experiment.</td>
</tr>
<tr>
<td>APD</td>
<td>Avalanche photo diode.</td>
</tr>
<tr>
<td>ASIC</td>
<td>Application-specific integrated circuit.</td>
</tr>
<tr>
<td>ATCA</td>
<td>Advanced Telecommunications Computing Architecture.</td>
</tr>
<tr>
<td>B-go</td>
<td>A timing signal in the TCDS system.</td>
</tr>
<tr>
<td>B-command</td>
<td>A timing signal in the TCDS TTC stream.</td>
</tr>
<tr>
<td>BCM1F</td>
<td>Fast beam conditions monitor.</td>
</tr>
<tr>
<td>BE</td>
<td>Back-end.</td>
</tr>
<tr>
<td>BI</td>
<td>Base Interface.</td>
</tr>
<tr>
<td>BRIL</td>
<td>Beam Radiation Instrumentation and Luminosity, a CMS project.</td>
</tr>
<tr>
<td>BRILDAQ</td>
<td>BRIL DAQ Readout.</td>
</tr>
<tr>
<td>BST</td>
<td>Beam Synchronous Timing.</td>
</tr>
<tr>
<td>BU</td>
<td>DAQ Builder Unit.</td>
</tr>
<tr>
<td>BX</td>
<td>Bunch crossing.</td>
</tr>
<tr>
<td>CMSSW</td>
<td>Compact Muon Solenoid software, the CMS experiment’s software package.</td>
</tr>
<tr>
<td>COTS</td>
<td>Commercial (or commodity) Off-The-Shelf, about hardware and software.</td>
</tr>
<tr>
<td>CPM</td>
<td>Central Partition Manager, part of TCDS.</td>
</tr>
<tr>
<td>CPU</td>
<td>Central Processing Unit.</td>
</tr>
<tr>
<td>CSC</td>
<td>Cathode Strip Chamber.</td>
</tr>
<tr>
<td>CTP7</td>
<td>Calorimeter Trigger Processor 7 card, featuring the Xilinx FPGA Virtex-7 chip.</td>
</tr>
<tr>
<td>CT-PPS</td>
<td>CMS-TOTEM Precision Proton Spectrometer.</td>
</tr>
<tr>
<td>D2S</td>
<td>Data to Surface.</td>
</tr>
<tr>
<td>DAOS</td>
<td>Distributed Application Object Storage.</td>
</tr>
<tr>
<td>DAQ</td>
<td>Data acquisition.</td>
</tr>
<tr>
<td>DB</td>
<td>Database.</td>
</tr>
<tr>
<td>DCS</td>
<td>Detector control system.</td>
</tr>
<tr>
<td>DRAM</td>
<td>Dynamic Random-Access Memory.</td>
</tr>
<tr>
<td>DSS</td>
<td>Detector safety system.</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Description</td>
</tr>
<tr>
<td>--------------</td>
<td>-------------</td>
</tr>
<tr>
<td>DT</td>
<td>Drift tubes.</td>
</tr>
<tr>
<td>DTC</td>
<td>Data, Trigger and Control board.</td>
</tr>
<tr>
<td>DTH</td>
<td>DAQ and TCDS Hub.</td>
</tr>
<tr>
<td>DTH-P1</td>
<td>DTH Prototype 1, by mid-2018, ATCA based with limited performance.</td>
</tr>
<tr>
<td>DTH-P2</td>
<td>DTH Prototype 2, by Q2-2020, full functionality with possibly reduced output bandwidth.</td>
</tr>
<tr>
<td>ECAL</td>
<td>Electromagnetic Calorimeter (of CMS).</td>
</tr>
<tr>
<td>EE</td>
<td>Endcap Electromagnetic Calorimeter.</td>
</tr>
<tr>
<td>EVB</td>
<td>Event Builder, or Event Building.</td>
</tr>
<tr>
<td>EVM</td>
<td>Event Manager, arbitrator of the Event Building.</td>
</tr>
<tr>
<td>FDL</td>
<td>Final Decision Logic.</td>
</tr>
<tr>
<td>FE</td>
<td>Front-end.</td>
</tr>
<tr>
<td>FEC</td>
<td>Forward error correction (as used in, e.g., optical link protocols).</td>
</tr>
<tr>
<td>FED</td>
<td>Front-end driver.</td>
</tr>
<tr>
<td>FeROL</td>
<td>FED Readout Optical Link.</td>
</tr>
<tr>
<td>FI</td>
<td>Fabric Interface.</td>
</tr>
<tr>
<td>FPGA</td>
<td>Field-Programmable Gate Array.</td>
</tr>
<tr>
<td>FRD</td>
<td>FED Raw Data.</td>
</tr>
<tr>
<td>FSM</td>
<td>Finite State Machine.</td>
</tr>
<tr>
<td>GbE</td>
<td>Gigabit Ethernet, a term describing various technologies for transmitting Ethernet frames at a rate of a gigabit per second.</td>
</tr>
<tr>
<td>GBT</td>
<td>Gigabit Transceiver. A project developed at CERN, source of the GBTx and LpGBT chips.</td>
</tr>
<tr>
<td>GBTx</td>
<td>Common term for the family of Gigabit Transceiver ASICs developed at CERN.</td>
</tr>
<tr>
<td>GEM</td>
<td>Gas Electron Multiplier. A type of gaseous ionization detector used in nuclear and particle physics.</td>
</tr>
<tr>
<td>GLIB</td>
<td>Gigabit Link Interface Board. A MicroTCA card developed by the CERN microelectronics group for testing and operating with the GBT link.</td>
</tr>
<tr>
<td>GPU</td>
<td>Graphics Processing Units.</td>
</tr>
<tr>
<td>GPGPU</td>
<td>General-purpose computing on graphics processing units.</td>
</tr>
<tr>
<td>HCAL</td>
<td>Hadronic calorimeter (of CMS).</td>
</tr>
<tr>
<td>HDD</td>
<td>Hard Disk Drive, traditional computer spinning drive.</td>
</tr>
<tr>
<td>HDI</td>
<td>High density interconnect, a very thin flexible PCB.</td>
</tr>
<tr>
<td>HEP</td>
<td>High energy physics.</td>
</tr>
<tr>
<td>HepSpec06</td>
<td>Computing benchmark for generic HEP applications.</td>
</tr>
<tr>
<td>HF</td>
<td>Hadron forward, very forward portion of HCAL covering pseudorapidity between 3 and 5.</td>
</tr>
<tr>
<td>HF-lumi</td>
<td>Sub-project in HF concerned with the luminosity measurement.</td>
</tr>
<tr>
<td>HGCAL</td>
<td>High Granularity Calorimeter, part of the CMS detector.</td>
</tr>
<tr>
<td>Term</td>
<td>Definition</td>
</tr>
<tr>
<td>------</td>
<td>------------</td>
</tr>
<tr>
<td>HL-LHC</td>
<td>High Luminosity Large Hadron Collider.</td>
</tr>
<tr>
<td>HLT</td>
<td>high level trigger, a collection of software trigger algorithms.</td>
</tr>
<tr>
<td>HPC</td>
<td>High Performance Computing.</td>
</tr>
<tr>
<td>HPD</td>
<td>Hybrid photodiode.</td>
</tr>
<tr>
<td>HS06</td>
<td>See HepSpec06.</td>
</tr>
<tr>
<td>I²C</td>
<td>Inter-Integrated Circuit, a chip-to-chip communication protocol.</td>
</tr>
<tr>
<td>ICI</td>
<td>A translation unit in the TCDS (of which there is one per partition) with as main task the translation of (CMS-wide) B-gos into detector specific B-commands.</td>
</tr>
<tr>
<td>I/O</td>
<td>Input/output.</td>
</tr>
<tr>
<td>IP</td>
<td>Interaction point, or internet protocol, or intellectual property.</td>
</tr>
<tr>
<td>IP</td>
<td>block Electronics building block that can be included in a chip.</td>
</tr>
<tr>
<td>IPBus</td>
<td>A protocol to control and communicate with Ethernet-attached xTCA hardware.</td>
</tr>
<tr>
<td>IPMC</td>
<td>Intelligent Platform Management Controller, an ATCA standard.</td>
</tr>
<tr>
<td>IT</td>
<td>Inner Tracker.</td>
</tr>
<tr>
<td>JCOP</td>
<td>Joint Controls Project, a CERN based project to provide common controls software.</td>
</tr>
<tr>
<td>L1</td>
<td>Level-1 trigger.</td>
</tr>
<tr>
<td>L1A</td>
<td>Level-1 accept.</td>
</tr>
<tr>
<td>L1</td>
<td>tracks Trigger primitives of the tracker.</td>
</tr>
<tr>
<td>LHC</td>
<td>Large Hadron Collider.</td>
</tr>
<tr>
<td>LHCC</td>
<td>Large Hadron Collider Experiments Committee.</td>
</tr>
<tr>
<td>LpGBT</td>
<td>Low-power Gigabit Transceiver, low power version of the GBTx.</td>
</tr>
<tr>
<td>LPM</td>
<td>Local Partition Manager, part of TCDS.</td>
</tr>
<tr>
<td>LS1</td>
<td>Long Shutdown 1, first LHC long shutdown from beginning 2013 to end of 2014.</td>
</tr>
<tr>
<td>LS3</td>
<td>Long Shutdown 3, third LHC long shutdown scheduled for 2024–2026.</td>
</tr>
<tr>
<td>LVDS</td>
<td>Low voltage differential signaling, a standard for differential, serial communication.</td>
</tr>
<tr>
<td>MCH</td>
<td>MicroTCA Control Hub, a shelf controller.</td>
</tr>
<tr>
<td>MIP</td>
<td>Minimum ionizing particle.</td>
</tr>
<tr>
<td>miniDAQ</td>
<td>Small scale DAQ systems used in CMS by sub-detectors.</td>
</tr>
<tr>
<td>MPI</td>
<td>Message Passing Interface, a standardized and portable message-passing system.</td>
</tr>
<tr>
<td>NFS</td>
<td>Network File System, a distributed file system protocol.</td>
</tr>
<tr>
<td>NV</td>
<td>Non-Volatile.</td>
</tr>
<tr>
<td>Term</td>
<td>Definition</td>
</tr>
<tr>
<td>------------</td>
<td>--------------------------------------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>NVMe</td>
<td>NVM Express, or Non-Volatile Memory Host Controller Interface Specification (NVMHCI) is a logical device interface specification for accessing non-volatile storage media attached via a PCI Express (PCIe) bus.</td>
</tr>
<tr>
<td>NVMemory</td>
<td>Non-Volatile Memory, see NVRAM.</td>
</tr>
<tr>
<td>NVRAM</td>
<td>Non-Volatile Random-Access Memory, is RAM that retains its information when power is turned off.</td>
</tr>
<tr>
<td>OT</td>
<td>Outer Tracker.</td>
</tr>
<tr>
<td>P5</td>
<td>LHC Point 5, the location where the CMS detector is installed.</td>
</tr>
<tr>
<td>PCB</td>
<td>Printed circuit board, is used to mechanically support and electrically connect electronic components.</td>
</tr>
<tr>
<td>PCI</td>
<td>Peripheral Component Interconnect, a computer bus.</td>
</tr>
<tr>
<td>PCIe</td>
<td>PCI Express, update in bandwidth and speed to the PCI standard, with multiple generations.</td>
</tr>
<tr>
<td>Phase-0</td>
<td>Time before Long Shutdown 1, also the respective tracker detector configuration.</td>
</tr>
<tr>
<td>Phase-1</td>
<td>Time between Long Shutdown 1 and Long Shutdown 3, also the respective tracker detector configuration.</td>
</tr>
<tr>
<td>Phase-2</td>
<td>Time after Long Shutdown 3, also the respective tracker detector configuration.</td>
</tr>
<tr>
<td>PI</td>
<td>Partition Interface, part of TCDS.</td>
</tr>
<tr>
<td>pileup</td>
<td>Overlapping of multiple soft interactions during a single LHC beam crossing.</td>
</tr>
<tr>
<td>PLC</td>
<td>Programmable Logical Controller.</td>
</tr>
<tr>
<td>PLL</td>
<td>Phase-Locked Loop.</td>
</tr>
<tr>
<td>PLT</td>
<td>Pixel Luminosity Telescope.</td>
</tr>
<tr>
<td>PMC</td>
<td>PCI Mezzanine Card.</td>
</tr>
<tr>
<td>PON</td>
<td>Passive Optical Network.</td>
</tr>
<tr>
<td>$p_T$</td>
<td>Transverse momentum of a physics object.</td>
</tr>
<tr>
<td>$\langle PU\rangle$</td>
<td>Pileup, average number of hard particle collisions per bunch crossing.</td>
</tr>
<tr>
<td>PV</td>
<td>Primary vertex.</td>
</tr>
<tr>
<td>RAID</td>
<td>Redundant Array of Independent Disks: is a data storage virtualization technology that combines multiple physical disk drive components into a single logical unit. for the purposes of data redundancy, performance improvement, or both.</td>
</tr>
<tr>
<td>RAM</td>
<td>Random Access Memory.</td>
</tr>
<tr>
<td>RAMdisk</td>
<td>An in-memory disk available in the Linux OS.</td>
</tr>
<tr>
<td>RCMS</td>
<td>Run Control and Monitoring System, part of the XDAQ software suite. R&amp;D research and development.</td>
</tr>
<tr>
<td>RF</td>
<td>Radio Frequency.</td>
</tr>
<tr>
<td>ROOT</td>
<td>An Object-Oriented modular scientific software framework developed by CERN.</td>
</tr>
<tr>
<td>RPC</td>
<td>Resistive Plate Chamber.</td>
</tr>
</tbody>
</table>
RPM  Red Hat Package Manager. A software package management system.
RTM  Rear Transition Module.
RU   Readout Unit
Rx   Optical receiver.
SCADA Supervisory Control And Data Acquisition.
SERDES Serialisation and deserialisation.
SFP  Small form-factor pluggable, a standard for optical and other transceivers which connect to a standard-defined socket.
SFP+ Enhanced small form-factor pluggable, supports up to 10 Gb/s data rates.
SiPM Silicon PhotoMultiplier.
SLINK A CERN specification for an easy-to-use FIFO-like data-link.
SMTS Storage Manager and Transfer System.
SSD  Solid State Drive, a solid-state computer storage device (disk).
stub Pair of hits in the two sensors of a $p_T$ module compatible with particles above the chosen threshold.
TCDS Trigger Control and Distribution System.
TCO  Total Cost of Ownership.
TCS  Trigger Control System.
TDP  Thermal Design Power, the power envelope of a CPU.
TEDD Tracker Endcap Double-Discs.
TEPX Tracker Endcap Pixel Detector.
TFPX Tracker Forward Pixel Detector.
TMUX Time-multiplexing.
TOB  Tracker Outer Barrel, a part of the present CMS tracker.
TOR  Top Of the Rack.
TTC  Trigger, Timing and Control, a system for distribution of clock and control.
TTcci TTC CMS interface.
TTCmi TTC machine interface.
TTCrxi TTC receiver ASIC.
TTS  Trigger Throttling System.
Tx   Optical transmitter.
UI   User interface.
UPS  Uninterruptible Power Supply.
USC  Underground Service Cavern, where the CMS counting room is located.
UXC  Underground Experimental Cavern, where the CMS detector is located.
VL, VL+ Versatile Link (plus), CERN projects.
<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>VM</td>
<td>Virtual Machine.</td>
</tr>
<tr>
<td>VME</td>
<td>Short for VMEbus (Versa Module Europa bus), is an ANSI/IEEE computer hardware bus standard.</td>
</tr>
<tr>
<td>VTRx</td>
<td>Versatile Transceiver plus, optical transceiver developed by the CERN Versatile Link plus project.</td>
</tr>
<tr>
<td>Xaas</td>
<td>XDAQ as a Service.</td>
</tr>
<tr>
<td>XDAQ</td>
<td>Cross-DAQ, a data acquisition software framework.</td>
</tr>
<tr>
<td>YETS</td>
<td>Year-end technical stop, a relatively brief stop of the LHC, typically less than three months in length, during the winter holidays.</td>
</tr>
</tbody>
</table>
References


[58] C. Soos, “Versatile Link+”. Presented at the 2016 ACES workshop at CERN.


CMS Collaboration

Yerevan Physics Institute, Yerevan, Armenia
A.M. Sirunyan, A. Tumasyan

Institut für Hochenergiephysik, Wien, Austria

Institute for Nuclear Problems, Minsk, Belarus
V. Chekhovsky, V. Mossolov, J. Suarez Gonzalez

Universiteit Antwerpen, Antwerpen, Belgium

Vrije Universiteit Brussel, Brussel, Belgium

Université Libre de Bruxelles, Bruxelles, Belgium

Ghent University, Ghent, Belgium

Université Catholique de Louvain, Louvain-la-Neuve, Belgium
H. Bakhshiansohi, O. Bondu, S. Brochet, G. Bruno, C. Caputo, A. Caudron, P. David, S. De Visscher, C. Delaere, M. Delcourt, B. Francois, A. Giammanco, M. Komm, G. Krintiras, V. Lemaitre,
A. Magitteri, A. Mertens, M. Musich, K. Piotrzkowski, L. Quertenmont, A. Saggio, M. Vidal Marono, S. Wertz, J. Zobec

Centro Brasileiro de Pesquisas Fisicas, Rio de Janeiro, Brazil

Universidade do Estado do Rio de Janeiro, Rio de Janeiro, Brazil

Universidade Estadual Paulista a, Universidade Federal do ABC b, S˜ ao Paulo, Brazil
S. Ahuja a, C.A. Bernardes a, T.R. Fernandez Perez Tomei a, E.M. Gregores b, P.G. Mercadante b, S.F. Novaes a, Sandra S. Padula a, D. Romero Abad b, J.C. Ruiz Vargas a

Institute for Nuclear Research and Nuclear Energy, Bulgarian Academy of Sciences, Sofia, Bulgaria
A. Aleksandrov, R. Hadjiiska, P. Iaydjiev, M. Misheva, M. Rodozov, M. Shopova, G. Sultanov

University of Sofia, Sofia, Bulgaria
A. Dimitrov, L. Litov, B. Pavlov, P. Petkov

Beihang University, Beijing, China
W. Fang5, X. Gao5, L. Yuan

Institute of High Energy Physics, Beijing, China

State Key Laboratory of Nuclear Physics and Technology, Peking University, Beijing, China
Y. Ban, G. Chen, J. Li, Q. Li, S. Liu, Y. Mao, S.J. Qian, D. Wang, Z. Xu, F. Zhang5

Tsinghua University, Beijing, China
Y. Wang

Universidad de Los Andes, Bogota, Colombia
C. Avila, A. Cabrera, C.A. Carrillo Montoya, L.F. Chaparro Sierra, C. Florez, C.F. González Hernández, J.D. Ruiz Alvarez, M.A. Segura Delgado
References

University of Split, Faculty of Electrical Engineering, Mechanical Engineering and Naval Architecture, Split, Croatia
B. Courbon, N. Godinovic, D. Lelas, I. Puljak, P.M. Ribeiro Cipriano, T. Sculac

University of Split, Faculty of Science, Split, Croatia
Z. Antunovic, M. Kovac

Institute Rudjer Boskovic, Zagreb, Croatia
V. Brigljevic, D. Ferencek, K. Kadija, B. Mesic, A. Starodumov, T. Susa

University of Cyprus, Nicosia, Cyprus

Charles University, Prague, Czech Republic
M. Finger, M. Finger Jr.

Universidad San Francisco de Quito, Quito, Ecuador
E. Carrera Jarrin

Academy of Scientific Research and Technology of the Arab Republic of Egypt, Egyptian Network of High Energy Physics, Cairo, Egypt
E. El-khateeb, S. Elgammal, A. Ellithi Kamel

National Institute of Chemical Physics and Biophysics, Tallinn, Estonia
R.K. Dewanjee, M. Kadastik, L. Perrini, M. Raidal, A. Tiko, C. Veelken

Department of Physics, University of Helsinki, Helsinki, Finland
P. Eerola, H. Kirschenmann, J. Pekkanen, M. Voutilainen

Helsinki Institute of Physics, Helsinki, Finland

Lappeenranta University of Technology, Lappeenranta, Finland
T. Tuuva

IRFU, CEA, Université Paris-Saclay, Gif-sur-Yvette, France
Laboratoire Leprince-Ringuet, Ecole polytechnique, CNRS/IN2P3, Université Paris-Saclay, Palaiseau, France

Université de Strasbourg, CNRS, IPHC UMR 7178, F-67000 Strasbourg, France

Centre de Calcul de l’Institut National de Physique Nucléaire et de Physique des Particules, CNRS/IN2P3, Villeurbanne, France
S. Gadrat

Université de Lyon, Université Claude Bernard Lyon 1, CNRS-IN2P3, Institut de Physique Nucléaire de Lyon, Villeurbanne, France

Georgian Technical University, Tbilisi, Georgia
T. Toriashvili¹³

Tbilisi State University, Tbilisi, Georgia
Z. Tsamalaidze⁷

RWTH Aachen University, I. Physikalisches Institut, Aachen, Germany

RWTH Aachen University, III. Physikalisches Institut A, Aachen, Germany

RWTH Aachen University, III. Physikalisches Institut B, Aachen, Germany
G. Flügge, B. Kargoll, T. Kress, A. Künsken, T. Müller, A. Nehrkorn, A. Nowack, C. Pistone, O. Pooth, A. Stahl¹⁴

Deutsches Elektronen-Synchrotron, Hamburg, Germany
M. Aldaya Martin, T. Arndt, C. Asawatangtrakuldee, K. Beernaert, O. Behnke, U. Behrens,

University of Hamburg, Hamburg, Germany

Institut für Experimentelle Teilchenphysik, Karlsruhe, Germany

Institute of Nuclear and Particle Physics (INPP), NCSR Demokritos, Aghia Paraskevi, Greece
G. Anagnostou, G. Daskalakis, T. Geralis, A. Kyriakis, D. Loukas, I. Topsis-Giotis

National and Kapodistrian University of Athens, Athens, Greece
G. Karathanasis, S. Kesisoglou, A. Panagiotou, N. Saoulidou

National Technical University of Athens, Athens, Greece
K. Kousouris, I. Papakrivopoulos

University of Ioánnina, Ioánnina, Greece

MTA-ELTE Lendület CMS Particle and Nuclear Physics Group, Eötvös Loránd University, Budapest, Hungary
M. Csanad, N. Filipovic, G. Pasztor, O. Surányi, G.I. Veres

Wigner Research Centre for Physics, Budapest, Hungary
G. Bencze, C. Hajdu, D. Horvath, Á. Hunyadi, F. Sikler, V. Veszpremi
Institute of Nuclear Research ATOMKI, Debrecen, Hungary  
N. Beni, S. Czellar, J. Karancsi, A. Makovec, J. Molnar, Z. Szillasi

Institute of Physics, University of Debrecen, Debrecen, Hungary  
M. Bartók, P. Raics, Z.L. Trocsanyi, B. Ujvari

Indian Institute of Science (IISc), Bangalore, India  
S. Choudhury, J.R. Komaragiri

National Institute of Science Education and Research, Bhubaneswar, India  

Panjab University, Chandigarh, India  

University of Delhi, Delhi, India  
Ashok Kumar, Aashaq Shah, A. Bhardwaj, S. Chauhan, B.C. Choudhary, R.B. Garg, S. Keshri, A. Kumar, S. Malhotra, M. Naimuddin, K. Ranjan, R. Sharma

Saha Institute of Nuclear Physics, HBNI, Kolkata, India  

Indian Institute of Technology Madras, Madras, India  
P.K. Behera

Bhabha Atomic Research Centre, Mumbai, India  
R. Chudasama, D. Dutta, V. Jha, V. Kumar, A.K. Mohanty, P.K. Netrakanti, L.M. Pant, P. Shukla, A. Topkar

Tata Institute of Fundamental Research-A, Mumbai, India  

Tata Institute of Fundamental Research-B, Mumbai, India  

Indian Institute of Science Education and Research (IISER), Pune, India  
S. Chauhan, S. Dube, V. Hegde, A. Kapoor, K. Kothekar, S. Pandey, A. Rane, S. Sharma
References

Institute for Research in Fundamental Sciences (IPM), Tehran, Iran
S. Chenarani\textsuperscript{25}, E. Eskandari Tadavani, S.M. Etesami\textsuperscript{23}, M. Khakzad, M. Mohammadi Najafabadi, M. Naseri, S. Paktinat Mehdia\textsuperscript{26}, F. Rezaei Hosseinabadi, B. Safarzadeh\textsuperscript{27}, M. Zeinali

University College Dublin, Dublin, Ireland
M. Felcini, M. Grunewald

INFN Sezione di Bari \textsuperscript{a}, Università di Bari \textsuperscript{b}, Politecnico di Bari \textsuperscript{c}, Bari, Italy
M. Abbrescia\textsuperscript{a, b}, C. Calabria\textsuperscript{a, b}, A. Colaleo\textsuperscript{a}, D. Creanza\textsuperscript{a, c}, L. Cristella\textsuperscript{a, b}, N. De Filippis\textsuperscript{a, c}, M. De Palma\textsuperscript{a, b}, F. Errico\textsuperscript{a, b}, L. Fiore\textsuperscript{a}, G. Iaselli\textsuperscript{a, c}, S. Lezki\textsuperscript{a, b}, G. Maggi\textsuperscript{a, c}, M. Maggi\textsuperscript{a}, G. Miniello\textsuperscript{a, b}, S. My\textsuperscript{a, b}, S. Nuzzo\textsuperscript{a, b}, A. Pompili\textsuperscript{a, b}, G. Pugliese\textsuperscript{a, c}, R. Radogna\textsuperscript{a}, A. Ranieri\textsuperscript{a}, G. Selvaggi\textsuperscript{a, b}, A. Sharma\textsuperscript{a}, L. Silvestris\textsuperscript{a, 14}, R. Venditti\textsuperscript{a}, P. Verwilligen\textsuperscript{a}

INFN Sezione di Bologna \textsuperscript{a}, Università di Bologna \textsuperscript{b}, Bologna, Italy
G. Abbiendi\textsuperscript{a}, C. Battilana\textsuperscript{a, b}, D. Bonacorsi\textsuperscript{a, b}, L. Borgonovi\textsuperscript{a, b}, S. Braïant-Giacomelli\textsuperscript{a, b}, R. Campanini\textsuperscript{a, b}, P. Capiluppi\textsuperscript{a, b}, A. Castro\textsuperscript{a, b}, F.R. Cavallo\textsuperscript{a}, S.S. Chhibra, G. Codispoti\textsuperscript{a, b}, M. Cuffiani\textsuperscript{a, b}, G.M. Dallavalle\textsuperscript{a}, F. Fabbrini\textsuperscript{a}, A. Fanfani\textsuperscript{a, b}, D. Fasanella\textsuperscript{a, b}, P. Giacomelli\textsuperscript{a}, C. Grandi\textsuperscript{a}, L. Guiducci\textsuperscript{a, b}, S. Marcellini\textsuperscript{a}, G. Masetti\textsuperscript{a}, A. Montanari\textsuperscript{a}, F.L. Navarria\textsuperscript{a, b}, A. Perrotta\textsuperscript{a}, A.M. Rossi\textsuperscript{a, b}, T. Rovelli\textsuperscript{a, b}, G.P. Siroli\textsuperscript{a, b}, N. Tosi\textsuperscript{a}

INFN Sezione di Catania \textsuperscript{a}, Università di Catania \textsuperscript{b}, Catania, Italy
S. Albergo\textsuperscript{a, b}, S. Costa\textsuperscript{a, b}, A. Di Mattia\textsuperscript{a}, F. Giordano\textsuperscript{a, b}, R. Potenza\textsuperscript{a, b}, A. Tricomi\textsuperscript{a, b}, C. Tuve\textsuperscript{a, b}

INFN Sezione di Firenze \textsuperscript{a}, Università di Firenze \textsuperscript{b}, Firenze, Italy
G. Barbagni\textsuperscript{a}, K. Chatterjee\textsuperscript{a, b}, V. Ciulli\textsuperscript{a, b}, C. Civinini\textsuperscript{a}, R. D’Alessandro\textsuperscript{a, b}, E. Focardi\textsuperscript{a, b}, P. Lenzi\textsuperscript{a, b}, M. Meschini\textsuperscript{a}, S. Paoletti\textsuperscript{a}, L. Russo\textsuperscript{a, 28}, G. Sguazzoni\textsuperscript{a}, D. Strom\textsuperscript{a}, L. Viliani\textsuperscript{a}

INFN Laboratori Nazionali di Frascati, Frascati, Italy
L. Benussi, S. Bianco, F. Fabbrini, D. Piccolo, F. Primavera\textsuperscript{14}

INFN Sezione di Genova \textsuperscript{a}, Università di Genova \textsuperscript{b}, Genova, Italy
V. Calvelli\textsuperscript{a, b}, F. Ferro\textsuperscript{a}, F. Ravera\textsuperscript{a, b}, E. Robutti\textsuperscript{a}, S. Tosi\textsuperscript{a, b}

INFN Sezione di Milano-Bicocca \textsuperscript{a}, Università di Milano-Bicocca \textsuperscript{b}, Milano, Italy
A. Benaglia\textsuperscript{a}, A. Beschi\textsuperscript{a}, L. Brianza\textsuperscript{a, b}, F. Brivio\textsuperscript{a, b}, V. Ciriolo\textsuperscript{a, b, 14}, M.E. Dinardo\textsuperscript{a, b}, S. Fiorendi\textsuperscript{a, b}, S. Gennai\textsuperscript{a}, A. Ghezzi\textsuperscript{a, b}, P. Govoni\textsuperscript{a, b}, M. Malberti\textsuperscript{a, b}, S. Malvezzi\textsuperscript{a}, R.A. Manzoni\textsuperscript{a, b}, D. Menasce\textsuperscript{a}, L. Moroni\textsuperscript{a}, M. Paganoni\textsuperscript{a, b}, K. Pauwels\textsuperscript{a, b}, D. Pedrini\textsuperscript{a}, S. Pigazzini\textsuperscript{a, b, 29}, S. Ragazzi\textsuperscript{a, b}, T. Tabarelli de Fatis\textsuperscript{a, b}

INFN Sezione di Napoli \textsuperscript{a}, Università di Napoli ‘Federico II’ \textsuperscript{b}, Napoli, Italy, Università della Basilicata \textsuperscript{c}, Potenza, Italy, Università G. Marconi \textsuperscript{d}, Roma, Italy
S. Buontempo\textsuperscript{a}, N. Cavallo\textsuperscript{a, c}, S. Di Guida\textsuperscript{a, d, 14}, F. Fabozzi\textsuperscript{a, c}, F. Fienga\textsuperscript{a, b}, A.O.M. Iorio\textsuperscript{a, b}, W.A. Khan\textsuperscript{a}, L. Lista\textsuperscript{a}, S. Meola\textsuperscript{a, d, 14}, P. Paolucci\textsuperscript{a, 14}, C. Sciacca\textsuperscript{a, b}, F. Thyssen\textsuperscript{a}

()}
INFN Sezione di Padova, Università di Padova, Padova, Italy, Università di Trento, Trento, Italy

INFN Sezione di Pavia, Università di Pavia, Pavia, Italy

INFN Sezione di Perugia, Università di Perugia, Perugia, Italy

INFN Sezione di Pisa, Università di Pisa, Scuola Normale Superiore di Pisa, Pisa, Italy

INFN Sezione di Roma, Sapienza Università di Roma, Rome, Italy

INFN Sezione di Torino, Università di Torino, Torino, Italy, Università del Piemonte Orientale, Novara, Italy

INFN Sezione di Trieste, Università di Trieste, Trieste, Italy
S. Belforte, M. Casarsa, F. Cossutti, G. Della Ricca, A. Zanetti.

Kyungpook National University

Chonbuk National University, Jeonju, Korea
A. Lee.
Chonnam National University, Institute for Universe and Elementary Particles, Kwangju, Korea
H. Kim, D.H. Moon, G. Oh

Hanyang University, Seoul, Korea
J.A. Brochero Cifuentes, J. Goh, T.J. Kim

Korea University, Seoul, Korea

Seoul National University, Seoul, Korea

University of Seoul, Seoul, Korea
H. Kim, J.H. Kim, J.S.H. Lee, I.C. Park

Sungkyunkwan University, Suwon, Korea
Y. Choi, C. Hwang, J. Lee, I. Yu

Vilnius University, Vilnius, Lithuania
V. Dudenas, M. Janulis, A. Juodagalvis, D. Simelevicius, J. Vaitkus

National Centre for Particle Physics, Universiti Malaya, Kuala Lumpur, Malaysia

Centro de Investigacion y de Estudios Avanzados del IPN, Mexico City, Mexico

Universidad Iberoamericana, Mexico City, Mexico
S. Carrillo Moreno, C. Oropeza Barrera, F. Vazquez Valencia

Benemerita Universidad Autonoma de Puebla, Puebla, Mexico
J. Eysermans, I. Pedraza, H.A. Salazar Ibarguen, C. Uribe Estrada

Universidad Autónoma de San Luis Potosí, San Luis Potosí, Mexico
A. Morelos Pineda
University of Auckland, Auckland, New Zealand
D. Krofcheck

University of Canterbury, Christchurch, New Zealand
P.H. Butler

National Centre for Physics, Quaid-I-Azam University, Islamabad, Pakistan
A. Ahmad, M. Ahmad, Q. Hassan, H.R. Hoorani, A. Saddique, M.A. Shah, M. Shoaib, M. Waqas

National Centre for Nuclear Research, Swierk, Poland
H. Bialkowska, M. Bluj, B. Boimska, T. Frueboes, M. Górski, M. Kazana, K. Nawrocki, M. Szleper, P. Zalewski

Institute of Experimental Physics, Faculty of Physics, University of Warsaw, Warsaw, Poland
K. Bunkowski, A. Byszuk, K. Doroba, A. Kalinowski, M. Konecki, J. Krolikowski, M. Mistura, M. Olszewski, A. Pyskir, M. Walczak

Laboratório de Instrumentação e Física Experimental de Partículas, Lisboa, Portugal

Joint Institute for Nuclear Research, Dubna, Russia

Petersburg Nuclear Physics Institute, Gatchina (St. Petersburg), Russia
Y. Ivanov, V. Kim, E. Kuznetsova, P. Levchenko, V. Murzin, V. Oreshkin, I. Smirnov, D. Sosnov, V. Sulimov, L. Uvarov, S. Vavilov, A. Vorobyev

Institute for Nuclear Research, Moscow, Russia

Institute for Theoretical and Experimental Physics, Moscow, Russia
V. Epshteyn, V. Gavrilov, N. Lychkovskaya, V. Popov, I. Pozdnyakov, G. Safronov, A. Spiridonov, A. Stepenov, M. Toms, E. Vlasov, A. Zhokin

Moscow Institute of Physics and Technology, Moscow, Russia
T. Aushev, A. Bylinkin
P.N. Lebedev Physical Institute, Moscow, Russia
V. Andreev, M. Azarkin\textsuperscript{37}, I. Dremin\textsuperscript{37}, M. Kirakosyan\textsuperscript{37}, A. Terkulov

Skobeltsyn Institute of Nuclear Physics, Lomonosov Moscow State University, Moscow, Russia
A. Baskakov, A. Belyaev, G. Bogdanova, E. Boos, A. Ershov, A. Gribushin, V. Klyukhin, O. Kodolova, I. Lokhtin, O. Lukina, I. Miagkov, S. Obraztsov, V. Savrin, P. Volkov, V. Volkov

Novosibirsk State University (NSU), Novosibirsk, Russia
V. Blinov\textsuperscript{40}, D. Shtol\textsuperscript{40}, Y. Skovpen\textsuperscript{40}

State Research Center of Russian Federation, Institute for High Energy Physics of NRC &quot;Kurchatov Institute&quot;, Protvino, Russia

University of Belgrade, Faculty of Physics and Vinca Institute of Nuclear Sciences, Belgrade, Serbia
P. Adzic\textsuperscript{41}, P. Cirkovic, D. Devetak, M. Dordevic, J. Milosevic, V. Rekovic

Centro de Investigaciones Energéticas Medioambientales y Tecnológicas (CIEMAT), Madrid, Spain

Universidad Autónoma de Madrid, Madrid, Spain
C. Albajar, J.F. de Trocóniz, M. Missiroli

Universidad de Oviedo, Oviedo, Spain

Instituto de Física de Cantabria (IFCA), CSIC-Universidad de Cantabria, Santander, Spain

CERN, European Organization for Nuclear Research, Geneva, Switzerland

Paul Scherrer Institut, Villigen, Switzerland

ETH Zurich - Institute for Particle Physics and Astrophysics (IPA), Zurich, Switzerland

Universität Zürich, Zurich, Switzerland

National Central University, Chung-Li, Taiwan

National Taiwan University (NTU), Taipei, Taiwan

Chulalongkorn University, Faculty of Science, Department of Physics, Bangkok, Thailand
B. Asavapibhop, K. Kovitanggoon, G. Singh, N. Srimanobhas

Çukurova University, Physics Department, Science and Art Faculty, Adana, Turkey
References


Middle East Technical University, Physics Department, Ankara, Turkey
G. Karapinar, K. Ocalan, M. Yalvac, M. Zeyrek

Bogazici University, Istanbul, Turkey
E. Gülmez, M. Kaya, O. Kaya, S. Tekten, E.A. Yetkin

Istanbul Technical University, Istanbul, Turkey
M.N. Agaras, S. Atay, A. Cakir, K. Cankocak, I. Köseoglu

Institute for Scintillation Materials of National Academy of Science of Ukraine, Kharkov, Ukraine
B. Grynyov

National Scientific Center, Kharkov Institute of Physics and Technology, Kharkov, Ukraine
L. Levchuk

University of Bristol, Bristol, United Kingdom

Rutherford Appleton Laboratory, Didcot, United Kingdom

Imperial College, London, United Kingdom

Brunel University, Uxbridge, United Kingdom
J.E. Cole, P.R. Hobson, A. Khan, P. Kyberd, I.D. Reid, L. Teodorescu, S. Zahid

Baylor University, Waco, USA
A. Borzou, K. Call, J. Dittmann, K. Hatakeyama, H. Liu, N. Pastika, C. Smith
Catholic University of America, Washington DC, USA
R. Bartek, A. Dominguez

The University of Alabama, Tuscaloosa, USA
A. Buccilli, S.I. Cooper, C. Henderson, P. Rumerio, C. West

Boston University, Boston, USA
D. Arcaro, A. Avetisyan, T. Bose, D. Gastler, D. Rankin, C. Richardson, J. Rohlf, L. Sulak, D. Zou

Brown University, Providence, USA

University of California, Davis, Davis, USA

University of California, Los Angeles, USA

University of California, Riverside, Riverside, USA

University of California, San Diego, La Jolla, USA

University of California, Santa Barbara - Department of Physics, Santa Barbara, USA

California Institute of Technology, Pasadena, USA

Carnegie Mellon University, Pittsburgh, USA
References


University of Colorado Boulder, Boulder, USA
J.P. Cumalat, W.T. Ford, F. Jensen, A. Johnson, M. Krohn, S. Leontsinis, T. Mulholland, K. Stenson, S.R. Wagner

Cornell University, Ithaca, USA

Fermi National Accelerator Laboratory, Batavia, USA

University of Florida, Gainesville, USA

Florida International University, Miami, USA
Y.R. Joshi, S. Linn, P. Markowitz, J.L. Rodriguez

Florida State University, Tallahassee, USA

Florida Institute of Technology, Melbourne, USA

University of Illinois at Chicago (UIC), Chicago, USA
M.R. Adams, L. Apanasevich, D. Berry, R.R. Betts, R. Cavanaugh, X. Chen, O. Evdokimov,

**The University of Iowa, Iowa City, USA**

**Johns Hopkins University, Baltimore, USA**

**The University of Kansas, Lawrence, USA**

**Kansas State University, Manhattan, USA**
A. Ivanov, K. Kaadze, Y. Maravin, A. Mohammadi, L.K. Saini, N. Skhirtladze

**Lawrence Livermore National Laboratory, Livermore, USA**
F. Rebassoo, D. Wright

**University of Maryland, College Park, USA**

**Massachusetts Institute of Technology, Cambridge, USA**

**University of Minnesota, Minneapolis, USA**

**University of Mississippi, Oxford, USA**
J.G. Acosta, S. Oliveros

**University of Nebraska-Lincoln, Lincoln, USA**

**State University of New York at Buffalo, Buffalo, USA**
J. Dolen, A. Godshalk, C. Harrington, I. Iashvili, D. Nguyen, A. Parker, S. Rappoccio, B. Roozbahani

**Northeastern University, Boston, USA**

**Northwestern University, Evanston, USA**

**University of Notre Dame, Notre Dame, USA**

**The Ohio State University, Columbus, USA**

**Princeton University, Princeton, USA**

**University of Puerto Rico, Mayaguez, USA**
S. Malik, S. Norberg

**Purdue University, West Lafayette, USA**

**Purdue University Northwest, Hammond, USA**
T. Cheng, N. Parashar, J. Stupak

**Rice University, Houston, USA**

University of Rochester, Rochester, USA
A. Bodek, P. de Barbaro, R. Demina, Yt. Duh, T. Ferbel, M. Galanti, A. Garcia-Bellido, J. Han, O. Hindrichs, A. Khukhunaishvili, K.H. Lo, P. Tan, M. Verzetti

The Rockefeller University, New York, USA
R. Ciesielski, K. Goulianos, C. Mesropian

Rutgers, The State University of New Jersey, Piscataway, USA

University of Tennessee, Knoxville, USA
A.G. Delannoy, J. Heideman, G. Riley, K. Rose, S. Spanier, K. Thapa

Texas A&M University, College Station, USA

Texas Tech University, Lubbock, USA

Vanderbilt University, Nashville, USA

University of Virginia, Charlottesville, USA
M.W. Arenton, P. Barria, B. Cox, R. Hirofsky, M. Joyce, A. Ledovskoy, H. Li, C. Neu, T. Sinthuprasith, Y. Wang, E. Wolfe, F. Xia

Wayne State University, Detroit, USA

University of Wisconsin - Madison, Madison, WI, USA
M. Brodski, J. Buchanan, C. Caillol, S. Dasu, L. Dodd, S. Duric, B. Gomber, M. Grothe, M. Hern-

†: Deceased
1: Also at Vienna University of Technology, Vienna, Austria
2: Also at IRFU, CEA, Université Paris-Saclay, Gif-sur-Yvette, France
3: Also at Universidade Estadual de Campinas, Campinas, Brazil
4: Also at Federal University of Rio Grande do Sul, Porto Alegre, Brazil
5: Also at Université Libre de Bruxelles, Bruxelles, Belgium
6: Also at Institute for Theoretical and Experimental Physics, Moscow, Russia
7: Also at Joint Institute for Nuclear Research, Dubna, Russia
8: Now at Ain Shams University, Cairo, Egypt
9: Now at British University in Egypt, Cairo, Egypt
10: Now at Cairo University, Cairo, Egypt
11: Also at Université de Haute Alsace, Mulhouse, France
12: Also at Skobeltsyn Institute of Nuclear Physics, Lomonosov Moscow State University, Moscow, Russia
13: Also at Tbilisi State University, Tbilisi, Georgia
14: Also at CERN, European Organization for Nuclear Research, Geneva, Switzerland
15: Also at RWTH Aachen University, III. Physikalisches Institut A, Aachen, Germany
16: Also at University of Hamburg, Hamburg, Germany
17: Also at Brandenburg University of Technology, Cottbus, Germany
18: Also at MTA-ELTE Lendület CMS Particle and Nuclear Physics Group, Eötvös Loránd University, Budapest, Hungary
19: Also at Institute of Nuclear Research ATOMKI, Debrecen, Hungary
20: Also at Institute of Physics, University of Debrecen, Debrecen, Hungary
21: Also at Indian Institute of Technology Bhubaneswar, Bhubaneswar, India
22: Also at Institute of Physics, Bhubaneswar, India
23: Also at University of Visva-Bharati, Santiniketan, India
24: Also at University of Ruhuna, Matara, Sri Lanka
25: Also at Isfahan University of Technology, Isfahan, Iran
26: Also at Yazd University, Yazd, Iran
27: Also at Plasma Physics Research Center, Science and Research Branch, Islamic Azad University, Tehran, Iran
28: Also at Università degli Studi di Siena, Siena, Italy
29: Also at INFN Sezione di Milano-Bicocca; Università di Milano-Bicocca, Milano, Italy
30: Also at Laboratori Nazionali di Legnaro dell’INFN, Legnaro, Italy
31: Also at Purdue University, West Lafayette, USA
32: Also at International Islamic University of Malaysia, Kuala Lumpur, Malaysia
33: Also at Malaysian Nuclear Agency, MOSTI, Kajang, Malaysia
34: Also at Consejo Nacional de Ciencia y Tecnologia, Mexico city, Mexico
35: Also at Warsaw University of Technology, Institute of Electronic Systems, Warsaw, Poland
36: Also at Institute for Nuclear Research, Moscow, Russia
37: Now at National Research Nuclear University ‘Moscow Engineering Physics Institute’ (MEPhI), Moscow, Russia
38: Also at St. Petersburg State Polytechnical University, St. Petersburg, Russia
39: Also at University of Florida, Gainesville, USA
40: Also at Budker Institute of Nuclear Physics, Novosibirsk, Russia
41: Also at Faculty of Physics, University of Belgrade, Belgrade, Serbia
Also at University of Belgrade, Faculty of Physics and Vinca Institute of Nuclear Sciences, Belgrade, Serbia
Also at Scuola Normale e Sezione dell’INFN, Pisa, Italy
Also at National and Kapodistrian University of Athens, Athens, Greece
Also at Riga Technical University, Riga, Latvia
Also at Universität Zürich, Zurich, Switzerland
Also at Stefan Meyer Institute for Subatomic Physics (SMI), Vienna, Austria
Also at Adiyaman University, Adiyaman, Turkey
Also at Istanbul Aydin University, Istanbul, Turkey
Also at Mersin University, Mersin, Turkey
Also at Cag University, Mersin, Turkey
Also at Piri Reis University, Istanbul, Turkey
Also at Izmir Institute of Technology, Izmir, Turkey
Also at Necmettin Erbakan University, Konya, Turkey
Also at Marmara University, Istanbul, Turkey
Also at Kafkas University, Kars, Turkey
Also at Istanbul Bilgi University, Istanbul, Turkey
Also at Rutherford Appleton Laboratory, Didcot, United Kingdom
Also at School of Physics and Astronomy, University of Southampton, Southampton, United Kingdom
Also at Instituto de Astrofísica de Canarias, La Laguna, Spain
Also at Utah Valley University, Orem, USA
Also at Beykent University, Istanbul, Turkey
Also at Bingol University, Bingol, Turkey
Also at Erzincan University, Erzincan, Turkey
Also at Sinop University, Sinop, Turkey
Also at Mimar Sinan University, Istanbul, Istanbul, Turkey
Also at Texas A&M University at Qatar, Doha, Qatar
Also at Kyungpook National University, Daegu, Korea