CMS Software and Offline preparation for future runs

Tommaso Boccali

INFN Sezione di Pisa, Largo B. Pontecorvo 3, Pisa 56127, ITALY. Also at CERN, Geneva, Switzerland.

E-mail: tommaso.boccali@cern.ch

Abstract. The next LHC Runs, nominally Run III and Run IV, pose problems to the offline and computing systems in CMS. Run IV in particular will need completely different solutions, given the current estimates of LHC conditions and Trigger estimates. We report on the R&D process CMS has established, in order to gain insight on the needs and the possible solutions for the 2020+ CMS computing.

1. Introduction

The CMS[1] experiment has been taking data at the CERN/LHC collider since late 2009, collecting year-over-year records on events collected and energy at the collisions. To date, CMS has published almost 1000 physics papers, at 7, 8 and 13 TeV.

The computing, based on a Tiered Distributed e-Infrastructure model, has been able to support the physics program in a quite comfortable way, allowing for unplanned operational modes in the last year of the Run-II, 2018.

2. A wrap-up of 2018 data taking

During 2018, the last year of Run II, LHC delivered to CMS integrated luminosity in excess of 67 fb^{-1}, out of which more than 64 fb^{-1}were recorded for offline utilization. Apart from the planned pp data taking at $\sim$1 kHz, during 2018 CMS experimented with two novel operational modes.

For most of the data taking period, a special trigger designed to collect "bb" events was activated, with rates up to 6 kHz in the last part of the fill (where pile-up, also called PU in the following, and event size is at the minimum). This sample is going to be the basis for an intense program on b-quark physics, to be completed largely before the restart of the LHC in 2021. In this mode, 12 billion events were collected, resulting in a b sample factors larger than the full samples collected by BaBar and Belle in their lifetime.

The end of year Heavy Ion Pb-Pb collected events at a much larger rate than previous years, with the attempt to collect a large statistics of unbiased Minimum Bias events (at 6 kHz) on top of the standard central physics. Overall, a sample of 4.5 billion Minimum Bias events has been collected, and is currently being analyzed.

Overall, the CMS Computing model has shown its maturity in 2018, allowing agile data taking operations, an intensive Monte Carlo production (totalling 24 billion full simulation events) and analysis operations at a constant level on roughly 50 thousand cores. The overall resources deployed by CMS in year 2019, according to WLCG Rebus[2], are shown in Table 1.

Table 1. WLCG Rebus CMS resource deployments for 2019.

Resource Type	**Unit of Measurement**	**Year**	**Pledged Amount**
CPU	HS06	2019	2M ($\sim$ corresponding to 200k computing cores)
Disk	PB	2019	160
Tape	PB	2019	290

3. The challenges ahead

With the performance of computing system as detailed in the last section, the situation for the near future is apparently under control.

Table 2. LHC and CMS parameters per data taking period.

Data Taking Period	Years	Max instantaneous luminosity in 10^{34} cm^{-2}s^{-1}	aver-age PU	CMS selection rate	
Run I	2009-2013	0.7	20	500 Hz	
Run II	2015-2017	2	37	1 kHz	
Run III	2021-2024	2	60	1 kHz	expected
Run IV	2026-2029	7.5	200	7.5 kHz	expected

Table 2 highlights the expected future LHC data taking periods. Indeed, the next LHC Run (Run III, 2021-2024) has projected computing needs that are similar to those of the recently completed Run II. The CMS experiment expects that the current computing model will be able to support Run III.

The longer term scenario is completely different: Run IV, starting tentatively in 2026, is currently expected to deliver instantaneous luminosities up to 7.5 10^{34} cm^{-2}s^{-1}, a factor 7.5 higher than the initial LHC design, and a factor 4 greater than that which is expected to be reachable in Run III. This translates into an average number of superimposed inelastic pp collisions per beam crossing (called pile-up) of up to 200. On top of that, the CMS experiment is currently extrapolating the need for a selection rate to offline of up to 7.5 kHz, in order to not lose performance on the relatively low mass Higgs boson precision studies.

A back-of-the envelope estimate for computing resources, even assuming that all the categories scale linearly with the number of events collected and their size, can be extrapolated from 2018 conditions by applying a factor of 7.5 for trigger rate, and a factor of 200/37 for event size and complexity, thus yielding a global factor exceeding 40. On top of that, event complexities are expected to increase due to the new detectors that CMS wants to deploy by Run IV[3], introducing another factor of up to 2. Overall, a simple estimate of the computing resource needs in case the same computing, data taking and analyses models are applied to Run IV, is close to 100 times more than the 2018 (or 2019) resources.

In perspective, technology evolution has always been considered as the main ingredient to a sustainably increasing computing infratructure, allowing in the past 1-2 decades for year by year increases in perfromance for the same price of up to +50%. Unfortunately, there is strong evidence[4] that such a steep evolution has slowed down considerably in the last 5 years, with new and more realistic extrapolations limited at +15%/y; over a period of 8 years, technology is than not expected to help for more than a factor 3x.

4. CMS R&D activities for HL-LHC

Even considering a quite optimistic technology gain factor of 3x, the HL-LHC computing would on paper be expected to need a budget in excess of today's by factors of 20-30x. These are clearly unfeasible, and would result in the experiments implementing a much smaller physics program to fit within the available resources (for example, by dractically reducing the trigger selections on low energy objects).

CMS has started an intense research program on the HL-LHC computing requirements, steered via the "Evolution of Computing Model 202X" (ECoM2X) task force, which includes efforts from the physics, the trigger and the detector communities, in a global effort in order to better understand the needs for CMS at HL-LHC, and to eventually propose solutions, changes, and other directions of study. The task force is structured into 7 working groups, covering aspects from technology tracking, modelling of computing needs, evolution of the infrastructure and of the computing environment.

In the rest of the paper we will present some highlights from the ECoM2X work in progess, and more in general from the R&D efforts CMS is putting in place in order to plan for sustainable computing operations during HL-LHC.

4.1. Technology Tracking

It is difficult to predict the 2026+ technology scenario, but trends already started are likely to negatively impact CMS computing at HL-LHC.

The most performant computing architectures (e.g. when measuread as Flops/$) will not be standard multi core CPUs, but simpler chips with a larger level of parallelism (SIMD).

General Purpose Graphics Processing Units (GPGPUs) are a derivative of the graphic cards which have seen an explosion in recent years, mostly due to the video game market. They are vector processors, with thousands of available core, and very limited capabilities for serial programming. Their utilization is best suited in extremely parallel algorithms, where the same operation has to be performed on a series of input data (SIMD[5]); they also deploy an extremely rigid memory model, with access to external memory having a larger on timing than the actual computing operation. While very difficult to objectively compare performance of CPUs and GPGPUs in general, given the different programming model, some selected applications have been ported and show large speedups; see [6, 7] for HEP (High Energy Physics) specific examples of such comparisons.

Field-Programmable Gate Arrays (FPGAs) offer a way to port algorithms in-silico, either via low level languages like VHDL[8], or via synthetization from higher level languages[9]. The main interest in the technology comes from the acquisition of ALTERA (one of the biggest FPGA producers) by Intel[10]: this paves the way for a strict integration of current x86_64 and FPGA technologies, potentially on the same chip and with large communication bandwidths. FPGA based technologies have been common since many years in the online systems of experiments; an availability on offline systems paves the way to their utilization as accelerators in standard workflows. Examples in such directions are [11, 12].

Tensor Processing Units (TPUs) are chips designed for fast matrix manipulation. While not a completely new idea, they have gained renewed interest in the last years due to the emergent sector of Artificial Intelligence, where matrix algebra is a key tool in algorithms like gradient

descent[13] for the training of Machine Learning systems. Currently, the most powerful technical implementation is by Google[14], and powers its internal tools, from search systems to decision systems; unfortunately, Google's TPUs are not available on the market, but just via direct partnership.

It is difficult to imagine a different technology emerging now and relevant for the HL-LHC timeline, so CMS has decided to focus its studies on software solutions embedding these architectures as possible targets.

For what concerns storage, the price differential between rotating disks and solid state disks is not reducing with the increasing sizes. Standard disks, using MAMR or HAMR technologies[15], are probably staying with us, with solid state disks limited to specific solutions like fast analysis systems or caches.

Tape technology is still evolving at a good pace, but is suffering from a shrinking market, and a now basically unique manufacturer[16].

Global networking availability (as bandwidth, number of links, their qualities) have never been problematic for LHC up to now, with perfromance exceeding our needs; in our computing models, networking has generally been considered as an infinitely available resource. This could change by Run IV, at least on the expensive and difficult to deploy intercontinental routes, which have seen yearly increases in traffic up to +40%/y. The need for a proper networking modelling, for example in order to avoid unnecessary multiple transatlantic transfers, has clearly emerged.

4.2. Physics choices

As detailed before, it is difficult to imagine any gain coming from a reduction of the trigger rates to offline, unless a reduction of the CMS physics capabilities is accepted.

A reduction of the computing needs could still be possible by implementing smarter data handling approaches, in principle with a small effect on the experiment's physics reach:

- Park to tape, with no prompt reconstruction, a large fraction of the selected events, processing only the fraction needed to ensure good data quaility; the rest could be processed either in the winter shutdown, or at the end of the LHC Run. While feasible and with no long term effect on physics output, it has clearly an effect in that it slows down analysis activities with respect to the competition.
- Implement scouting triggers (as was already done on small scale), for which the trigger objects are direclty used for analysis; eventually the original raw data can be discarded, making these samples not reprocessable in the future.
- Switch a large(r) part of the Monte Carlo production to fast simulation, investing for example on realistic GAN based tools[17].
- Invest in colalborations with the authors of Event Generators, in order to make sure the tools scale well on modern hardware.

4.3. Towards heterogeneous architectures

There is general consensus in the HEP community that a large help in solving the HL-LHC computing scaling can come from cost effective computing architectures, as those listed in the technology tracking section. In the last year CMS has invested in a systematic effort to include non-CPU artchitectures as first citizen in its software. The strategy usss the concept of "multiple equivalent modules" being able to perform a given task, with the module selection possible at submission, site level or even event by event basis[18]. When accelerators, or in any case technologies different from standard CPu are involved, the CMS framwork is non-blocking, allowing a full utilization of the hardware in all occasions. At the same time, CUDA[19] has been included in the standard deployment of the CMS software, in order to ease the development effort. The next step will be a survey of the available tools for automatic code translation on

Table 3. CMS analysis data formats for analysis.

CMS Data Format Name	Main analysis data format for the period	Size (kB/event)
RECO	2010-2011	3000
AOD	2012-2015	400
MINIAOD	2016-2019	50
NANOAOD	2019-	1

multiple platforms, in order to allow for a faster development with fewer physics validation concerns.

4.4. Reduced data formats

A large part of the disk storage needed in the computing operations, is to host data and Monte Carlo samples for user analysis. Along the years, with the achievement of a better undertanding of the LHC environment, the CMS detectors, and the needs for typical analysis, CMS has been able to drastically shrink the data format in which the samples are presented to analysis users.

Table 3 shows how effective such a shrink has been, for a total of a 3000x reduction since the start of beam activities. This is an ingredient of primary importance in keeping disk requests low. On the other hand, having formalized a data format like NANOAOD, currently expected to be valid for $\sim 50\%$ of the studies, also reduced the computing needs since most users will be able to find preprocessed data for their analyses. The NANOAOD format is now produced for all the CMS data and Monte Carlo samples, but its utilization is at first stages and will need to monitored in the next years.

4.5. Common tools

Even if not strictly included into computing requests, CMS needs to deploy and maintain into production a large series of software tools, for aspects concerning databases, toolkits for detector simulation and description, data and workload management infrastructures. The human cost behind these is not easy to calculate, but is not a small correction to the overall needs.

In order to reduce the longterm maintenance effort, CMS is evaluating the adoption of standard tools, at least within HEP. Current targets are:

- Rucio[20] for Data Management;
- CRIC[21] as Information System;
- DD4HEP[22] as geometry description toolkit.

CMS expects to finish evaluating and porting to such tools by 2020, in order to be ready with Run III data taking.

4.6. Changes to the infrastructure

In the context of WLCG, a general effort towards a new-generation data infrastructure is strongly active, also via groups like DOMA[23] and European Projects like ESCAPE[24] and XDC[25]. The "data-lake" approach aims to create a solid, secure and curated data infrastructure, using HEP-owned data centers, linked via middleware allowing them to appear as much as possible as single logical system. Strong reliance of network links is needed both for inter-lake communications, and for data delivery to computing centers external to the lake, and eventually being:

- standard "GRID-like" centers;

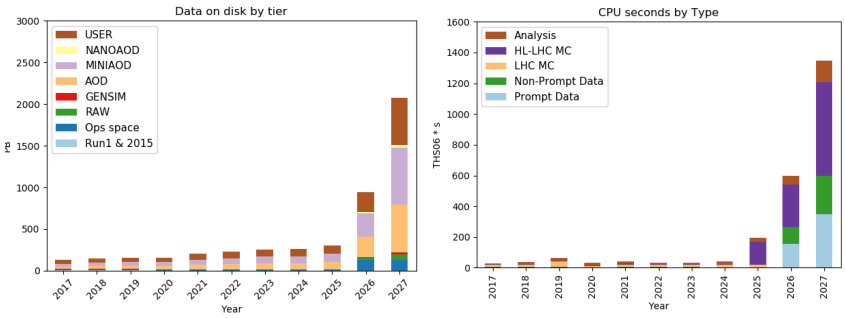

Figure 1. CMS Experiment projections for Computing resource needs in HL-LHC (from [26]).

- diskless "GRID-like" centers;
- commercial cloud providers;
- High Performance Computing centers;
- short lived centers coming from grants, collaborations, etc;

The concept behind the data-lake is to make sure our data is safe, and to profit for every computing opportunities, lowering the bar for their acceptance for what concerns lifetime, local storage and support level.

5. Current CMS extrapolated resource needs for HL-HLC

A part of the solutions described in the previous section has been already inserted in the CMS computing model simulation, used for long term planning. The last public figures are shown in Figure1. Starting from the back-of-the-envelope resource increases up to 100x with respect to 2019, current best estimates are 15x for storage and 22x for CPU (which does not include any reliance on GPGPUs or such yet); more information and details can be found in [26].

6. Conclusions

The current understanding of CMS computing needs and strategies at HL-LHC are detailed, delining a possible evolution from current infrastructure and inserting changes at the level of computing architectures, physics operations down to analyses. While CMS cannot yet demonstrate a viable solution to HL-LHC computing, a large effort is ongoing invoving all the Collaboration sub-projects, with the already evident result of year by year decreasing projected needs.

7. Acknowledgments

This paper is partially supported by the EU Project ESCAPE, G.A. 824064.

References

[1] CMS Collaboration, "The CMS experiment at the CERN LHC", JINST 3 (2008) S08004, doi:10.1088/1748-0221/3/08/S08004.
[2] `https://wlcg-rebus.cern.ch/apps/pledges/summary/`
[3] CMS Collaboration, "Technical Proposal for the Phase-II Upgrade of the CMS Detector", CERN-LHCC-2015-010 ; LHCC-P-008 ; CMS-TDR-15-0
[4] `https://twiki.cern.ch/twiki/bin/view/Main/TechMarketDocuments`, maintained by B. Panzer (CERN).
[5] `https://it.wikipedia.org/wiki/Single_instruction_multiple_data`
[6] Performance studies of GooFit on GPUs vs RooFit on CPUs while estimating the statistical significance of a new physical signal, Adriano Di Florio, Journal of Physics: Conference Series, Volume 898, Track 5: Software Development

[7] GPU-accelerated track reconstruction in the ALICE High Level Trigger, David Rohr et al, Journal of Physics: Conference Series, Volume 898, pages 032030, 2017
[8] https://en.wikipedia.org/wiki/VHDL
[9] Module-per-Object: a Human-Driven Methodology for C++-based High-Level Synthesis Design, Jeferson Santiago da Silva et al, arXiv:1903.06693
[10] https://newsroom.intel.com/news-releases/intel-completes-acquisition-of-altera/
[11] Fast inference of deep neural networks in FPGAs for particle physics, Javier Duarte et al, https://arxiv.org/abs/1804.06913
[12] Level-1 Track Finding with an all-FPGA system at CMS for the HL-LHC, Zhengcheng Tao, arXiv:1901.03745
[13] https://medium.freecodecamp.org/understanding-gradient-descent-the-most-popular-ml-algorithm-a66c0d97307f
[14] https://cloud.google.com/tpu/docs/tpus
[15] https://fstoppers.com/originals/hamr-and-mamr-technologies-will-unlock-hard-drive-capacity-year-326328
[16] https://www.theregister.co.uk/2017/02/17/oracle_streamline_tape_library_future/
[17] Fast and accurate simulation of particle detectors using generative adversarial networks, P. Musella et al, https://arxiv.org/pdf/1805.00850.pdf
[18] A. Bocci et al, "Towards a heterogeneous High Level Trigger farm for CMS", Subm. Procs to ACAT2019, Saas-Fee (CH), 10-15 Mar 2019.
[19] J. Nickolls et al, "Scalable Parallel Programming with CUDA", ACM Queue, vol. 6 no. 2, March/April 2008
[20] V. Garonne et al, "Rucio – The next generation of large scale distributed system for ATLAS Data Management", J. Phys.: Conf. Ser. 513 042021
[21] A. Anisenkov et al, "CRIC: a unified information system for WLCG and beyond", Subm. Procs to CHEP2018, Sofia (BG), 9-13 Jul 2018.
[22] M. Petric et al, "New Developments in DD4hep", Subm. Procs to CHEP2018, Sofia (BG), 9-13 Jul 2018.
[23] https://twiki.cern.ch/twiki/bin/view/LCG/DomaActivities
[24] https://lapp.in2p3.fr/spip.php?article2624&lang=en
[25] D. Cesini et al, "Advancements in data management services for distributed e-infrastructures: the eXtreme-DataCloud project", Subm. Procs to CHEP2018, Sofia (BG), 9-13 Jul 2018.
[26] https://twiki.cern.ch/twiki/bin/view/CMSPublic/CMSOfflineComputingResults