CMS Software and Offline preparation for future runs

Tommaso Boccali
INFN Pisa / CERN
CMS Computing circa 2019 the status

- LHC concluded in December 2018 its second Run (“RunII”)
- CMS very successful in the data taking and analysis operations, with computing supporting unexpected requirements
  - ParkingB: additional 12 B events collected in 2018 to support CMS B Physics; a sample 5x larger than Babar’s and Belle’s!
    - Up to 6 kHz additional rate to tape
  - HF flavour physics in Heavy Ions: 4.5 B additional Minbias events collected in Nov 2018
    - Rate to offline > 7 GB/s
- On top of that, standard pp operations (64/fb collected), analysis operations in full swing
  - 859 collider papers submitted
  - Derivative increasing!
Overall...

- Complete utilization of resources, with sizeable over-pledges in CPU
- In 2018, more than **24 B full Simulation events generated** to support the physics program
- Storage areas (disks, tape systems) well under control
  - Thanks to Dynamo/DDM, operational since the start of the Run
- CMS SW in **full-real multithreading mode since 2015** (8 threads is the default): **no memory problems**, even with PhaseII simulations

![24B Full Sim events generated in 2018](image)

**Data Deletions by DDM (2015-2018)**

- DDM deletes **~40 PB/month**
- Production varies from 1-15 threads
- Analysis still mostly single thread
So where is the problem?

- **2009-2012 (RunI):** resources somehow overprovisioned, luxury mode
- **2013-2014 (LS1):** Funding Agencies imposed a “flat funding”, which means ~20% increase/y thanks to Moore’s law (and friends)
- **2015-2018 (RunII):** resources more and more constrained, Moore’s law starting to be excessively optimistic
- **2019-2020 (LS2):** virtually no increase in resources granted
- **2021-2023 (RunIII):** in principle not incredibly different from RunII, but LHC is willing to surprise us
- **2024-2025 (LS3):** no increases?
- **2026+:** the LHC Phase II, the problem!
RunII $\rightarrow$ RunIII

- Limit in instantaneous luminosity @ $2\times10^{34}$ cm$^{-2}$s$^{-1}$, close to RunII ...
- … but levelled for most of the fill time (12 h?) $\rightarrow$ much larger average inst lumi $\rightarrow$ $<\text{PileUp}> 35 \rightarrow 55$ (?)
RunIII $\rightarrow$ RunIV

- Current modelling expects up to $7.5\times10^{34}$ cm$^{-2}$s$^{-1}$ flat luminosity during fills
- PU = $\langle PU \rangle = 200$
- CMS with an upgraded detector:
  - Many more silicon tracker measurements
  - Completely revamped forward calorimetry
  - $\rightarrow$ many more channels, much larger expected algorithmic complexity
- Physics requirements currently suggesting 5-10x increased trigger rate
  $\rightarrow$ on paper, easy to get factors 50-100x more resources needed with respect to RunII!

(see An analytics driven computing model for HL-LHC here at ACAT)
What to do?

- Even considering an optimistic technology factor $\sim 4x$ from technology, factors $> 10x$ are missing.
- Miracles apart (quantum computing anyone?), we need an intense and focused R&D program in order to allow for a(n economically) viable exploitation of HL-LHC.
- In CMS, the activity is carried out in multiple groups, with an attempt of light steering from the Evolution of Computing Model 202X → ECoM2X.
- It is not a Computing only effort, but an overall CMS effort, including:
  - Physics, Trigger, PPD and Run Coordination
  - Expert analysts
  - Representative of major funding agencies
- Activity split in 7 Working Groups

Some highlights in the next slides...
Technology tracking

- Difficult to fully predict the 2026+ technology scenario, but our analyses show that we should focus on:
  - Directly supporting **GPUs** and **FPGAs** in our software stacks
  - **TPUs** are also promising, but are there use cases apart from speeding up ML training?
- At the moment there is no convergence of **HDD** and **SDD**, with a “distance” in $$/TB$$ roughly constant in time
- Tape technologies are advancing at higher pace than **disk**; still there is a problem with decreasing user base
- Up to now no real effort to streamline **network utilization** (considered infinite)
- This will probably change by RunIV, at least for transatlantic links
  - Abandon uniform full mesh?
  - Segregate continents?
  - US (ESnet) LHConE traffic +40%/y
Physics choices, data rates and analysis model

- Investigations just started, and complex due to many stakeholders
- **Naive extrapolation from** $L=1.9\times10^{34}$ → $L=7.5\times10^{34}$ explains the expected need of HLT output at ~ 7.5 kHz, mostly coming from single object triggers
  - Unless we want to reduce / descope a part of the physics program
  - Unless we can use less inclusive trigger approaches - to be studied!

- More to be gained by **smart data handling** approaches:
  - **Park** a large fraction of triggers, and recover in the winter shutdown or LSs
  - **Scouting** datasets, with small “reco like” output to offline
  - Prompt reconstruction just in order to ensure data quality; **deferred reconstruction** for the rest
  - ...

- The amount of MC to produce has a large effect on resource needs
  - Need for a **common-HEP GAN based Fast Simulation**?
  - Event generators’ increase in resources due to $N(NN)LO$ to be kept under control - **today it is not**
Heterogeneous architectures

- There is general consensus that the best performance/$$ will not be obtained with standard CPUs
- Testbeds active on GPUs, FPGAs, ... initially as standalone exercises
- In the last year, CMS has performed a general attempt to systematically **include these in the standard CMSSW Software Framework:**
  - Allow multiple versions of “equivalent” modules, and defer the decision on which to use even very late (event by event, module by module)
  - Allow the best communication between modules exposing different interfaces (for example, aut. chain GPU modules without moving data back to the host)
  - Have CUDA as an external tool in CMSSW, for native utilization
  - Next step (in collaboration with other experiments?) is to try and have automatic code translation in place (is it even possible?)
- Examples exist; see for example **Towards a heterogeneous High Level Trigger farm for CMS** at ACAT
- Status of the framework allows to run benchmarks / compare architectures / plan for infrastructure
Reduced data formats

- The most important result of the previous ECoM17 task force has been the **definition** of a (even more) reduced data format.
- CMS **already in 2014** pioneered the definition of “small” general purpose data format, **MiniAOD @ ~ 1/10 of the AOD**
- Its adoption has been **overwhelming in RunII**: its adoption was expected to be for ~70% of analysis, it now reached 95%
- **NanoAOD** go even further: we were aiming at ~ 5 kB/ev, eventually we are at 1 kB/ev!
- **Expected analysis coverage ~50 %**: we hope to be positively surprised as for Mini!

| Data Tier          | Size (kB) |
|--------------------|-----------|
| RAW                | 100       |
| GEN                | 3000      |
| SIM                | 1000      |
| DIGI               | 3000      |
| RECO(SIM)          | 400 (8x reduction) |
| AOD(SIM)           | 800 (8x reduction) |
| MINIAOD(SIM)       | 50 (8x reduction) |
| NANOAOD(SIM)       | 1 (50x reduction) |

**Full RunII:** 25 (DT) + 35 (MC) B events expected to fit in 60 TB

**Exec Summary:** “Prevalent analysis format in CMS reduced by a factor 3000x in event size since the start of RunI”
Changing SW tools

- The CMS SW stack and Computing Infrastructure were **adequate for CMS needs in RunII**, and then some.
- We have **no real hint that RunIII would pose irresolvable problems either**; but, since RunIV is a different story, CMS plans to **try and test any disruptive technology already in RunIII**.
- Among the software tools, the biggest worries in the RunIV time scale are about **software support and sustainability**. **Common solutions with other experiments are a way to mitigate the support cost**.

CMS identified 3 initial areas where we can profit from existing OSS:

- Geometry description: **testing DD4HEP from AIDA2020**; if testing is positive, transition in ~1 y
- **CRIC from CERN** as a replacement for the Information System - already in place for the first use cases
- **Rucio** (initially from ATLAS) as the **Data Management solution** - transition and then large scale test in ~ 1 y
Changes to the infrastructure

- Reducing the needs for **data replication** is of paramount importance for a cost reduction.
- This can come from aggressive policies paired with remote reads:
  - **CMS already executes prod WFs without data locality, and explicit overflows to “close sites” via the Xrootd federation.**
- The **data lake** seems the most promising solution to-date for a safe storage of our data, with limited replication.
- It also **allows to think that most of the CPU resources can be served without managed disks**. This opens to:
  - HPC systems
  - Commercial Clouds
  - Ephemeral sites
- and drives the needs for **proper caching tools, easy to deploy**. See Using DODAS as deployment manager for smart caching of CMS data management system at ACAT.
HPC systems

- An HPC race is going on, at least between major players
- Next big thing is **ExaScale** \((10^{18} \text{ Flops} - \text{ operations per second})\)
  - Should be well available by HL-LHC
- Somehow difficult to compare, technologies / benchmarks, but
  - **LHC needs today the equivalent of \(~30 \text{ PFlops}\)**
  - A single Exascale system is ok to process 30 “today” LHC
  - **Scaling: a single Exascale system could process the whole offline HL-LHC with no R&D or model change**
- Some FAs/countries are explicitly requesting HEP to use the HPC infrastructure as ~ only funding; it is generally ok IF we are allowed to be part in the planning (to make sure they are usable for us)

**US:** apparently no current way to have a say at least on big DOE systems
**EU:** ETP4HPC has at least “asked for HEP position”
**China:** no current way to have a say
Our Funding Agencies are asking CMS to be prepared to use national HPC infrastructures for a sizeable part of our needs, by RunIV.

There are many not trivial problems to solve:

- **Data access** (access, bandwidth, ...)
- **Accelerator Technology** (KNL, GPU, FPGA, TPU, ???, ...)
- **Primary architecture** (Intel, Power9, ARM, proprietary ...)
- **Submission of tasks** (MPI vs Batch systems vs proprietary systems)
- **Node configuration** (low RAM/Disk, ...)
- **Not-too-open environment** (OS, Access policies, ...)

Since many problems are more political than technical, CMS has prepared a document to perform handshaking with HPC sites, and in order to

- Explain our needs
- Propose solutions (standard, ad-hoc)
- Discuss out-of-the-box solutions for Future systems
- (shared with the other exps to find a Common ground)

CMS plans a (virtual) trip to visit all the HPC sites, and establish direct links.

---

### Report on HPC resources integration at CMS

#### Introduction

High Performance Computing (HPC) systems are highly non-standard facilities, and are custom-built having in mind use cases largely different from High Energy Physics (HEP) ones. The utilization of these system by HEP experiments is not trivial: each HPC center is different and, of course, this increases the level of complexity from the integration and operations perspectives.
Current understanding of resource needs for PhaseII

- All, in all, **starting from the factors 50-100x as previously mentioned**, the efforts already put in place are **starting to pay off**

- Last public version of our 2027 estimates cite projected needs for
  - CPU: 44 MHS06
  - Disk: 2.2 EB
  - Tape: 3 EB
  - *(with respect to 2019 pledges, these are 22x, 13x and 15x)*

- .. with a storage **decrease by 2x** due the **modellization of NanoAOD as a tool for 50% of the analyses**, and thus reducing the needs to process and store on disk larger data formats

See **An analytics driven computing model for HL-LHC** at ACAT for more information
Conclusions

- HL-LHC is a fascinating research environment, with incredible capabilities for physics discoveries
- Unfortunately, the large amount of expected data does not fit any reasonable amount of funding, if handled via standard operation models
- CMS has started a deep and intense R&D program, involving all the stakeholders from Physics groups to Trigger experts, in order to pave a way towards an affordable HL-LHC computing
- Ideas are being collected, analyzed and formalized, with the plan to have them under test in RunIII before the final deployment in production starting from 2026