Response to Request for Information on U.S. CMS Software and Computing Resource Planning

U.S. CMS Operations Program

Abstract

This note provides a response to the request for information from DOE to the U.S. CMS Operations Program to give an estimate of the yearly computing resource needs in CPU, disk and tape up to 2027, with a possible extension to 2030; and to discuss which of these needs can be fulfilled by HPC centers and/or commercial clouds.

The document gives an introduction to the CMS workflow types, followed by a brief status report of our capabilities to use Intel's Knights Landing (KNL) CPU architecture and progress in using HPC centers and commercial clouds. We introduce the input parameters to the model used to estimate resource needs for Run 3 and the HL-LHC era, followed by a discussion of the resulting overall CMS yearly resource needs. We conclude by discussing the year-by-year computing resources to be provisioned by U.S. CMS, and the possible utilization of HPC and commercial cloud resources.

Workflows

CMS executes a variety of tasks on its distributed computing infrastructure, for reconstructing collision data recorded by the detector, simulating collision data, and analyzing both. In the following, we call all information of a collision, both for data from the detector and for simulations, an event. CMS uses an object-oriented event data model to persist information of an event in a file based on ROOT. Standardized groups of objects are persisted in different files, for example all objects that are related to the raw detector information. Files with the same event content are described by a common "data tier" identifier. CMS distinguishes six main types of computing tasks.

The **gridpack integration** workflow is a preparation step for simulating, or generating, events from theoretical principles. It performs the phase space integration for specific physics processes before the actual events are generated to save processing time later. It is executed by CMS physicists and not by central production teams as it needs close supervision and normally requires several iterations to guarantee the correctness of the physics processes. In the following, we do not show the resource needs of the gridpack integration separately and assume that it is included in the analysis CPU resources. The output is small, of the order of tens of gigabyte per gridpack, and is provided to generation applications through CVMFS. The CPU needs for this step are negligible compared to the other steps.

The **generation** task uses software packages provided by the theoretical particle physics community to generate events from basic theoretical principles. The events are stored in text files in a standardized format called the LesHouches Event accord (LHE) and consist only of a few numbers per event. If possible, these generators use the aforementioned gridpacks to avoid repetition of the phase space integration. In most cases, the generation task is executed by central production teams as part of the full simulation workflow (generation + simulation + digitization + reconstruction, see later). In some special cases, the generation is not run centrally. In this case, CMS physicists execute the generation step like in the gridpack integration step to guarantee correctness of the generated physics processes. These special cases are also included in the analysis resources. CMS is currently not using generators that benefit from massive MPI execution (Sherpa, Alpgen). CMS's primary generator is aMC+Madgraph and CMS' use of generators today is a negligible part of its total CPU needs given the generators it uses. We note that the CPU time per event during generation differs by more than an order of magnitude between different generators simulating identical physics. There is a fundamental problem with next-to-leading order (NLO) generators in that their use of negative event weights leads to anywhere between $\times 3$ to $\times 25$ the number of events required for leading order simulations to achieve the same statistical precision. In all our projections for CPU needs we assume that the problem of negative weights in NLO generators will be solved by the phenomenology community, which develops these generators. If it is not solved, the CPU needs for simulation from generation through reconstruction will increase by an order of magnitude.

The **simulation** task is part of the full simulation workflow and uses Geant4 to simulate the energy deposition of the generated particles in the CMS detector components. It is centrally executed. The time to simulate an event for LHC Run 2 beam conditions is roughly the same as for the HL-LHC — the additional complexity comes through the Pile-up simulation, which is done in a later step. The output of the simulation task is called the GEN-SIM data tier.

The **digitization** task simulates detector component signals from the energy depositions determined during the simulation task . The digitization task also simulates the LHC beam conditions, characterized by the number of additional collisions happening when the two proton bunches collide in the detector, called pileup. CMS distinguishes two pileup simulation modes. The standard pileup mixing combines a number of individually simulated minimum bias events as a secondary input to the digitization task along with the primary simulated event. In LHC Run 2, about a hundred minimum bias events are needed to digitize one generated event. This increases to the order of a thousand for HL-LHC. Reading hundreds or thousands of simulated minimum bias events to digitize a primary event causes significant I/O load for the applications, making streaming impossible, especially in the HL-LHC case. To lower the I/O load, CMS developed pre-mixing of pileup, where the combination of minimum bias events is done in a separate step. These pre-mixed events can be re-used in the simulation of different generated samples. The digitization step in pre-mixing mode combines a simulated primary event with a pre-mixed secondary event. The ability to pre-mix pileup requires an excellent understanding of the detector component behaviors under LHC running conditions. The pre-mixing mode of pileup in LHC Run 2 consumes $\frac{1}{2}$ of the CPU time of standard pileup per event. For the HL-LHC, digitization dominates the CPU budget per event for the sequence of generation, simulation, and digitization. At present, CMS is unable to use pre-mixing for HL-LHC simulations. The projections thus assume standard pileup mixing.

The **reconstruction** task combines all algorithms to reconstruct detector component signals as well as global event quantities. It is the same task both for simulated and recorded events. The output of the reconstruction task is optimized for the analysis task that follows. Apart from the RECO data tier that is used only in rare cases for detector component commissioning, the main output data tiers are AOD (Analysis Object Data) and MINIAOD (a smaller subset of the AOD). The reconstruction time per event increases non-linearly with number of pileup events per collision. This explains the large difference in CPU time per event between LHC Run 2 and HL-LHC. To reduce this increase significantly would require changes in the detector geometry and/or giving up on low pT tracking. The latter is expected to compromise the physics performance provided by particle flow algorithms.

The **analysis** task reduces the output of the reconstruction through slimming and skimming. Analysts also calculate specific event properties that are then combined with the centrally provided information to produce plots and tables. Analysts can also write out ntuples at the end or during the reduction process; the corresponding data tier is called USER. The total disk budget of CMS worldwide is dominated by the needs of analyzers to easily and reliably access the reconstructed data and to be able to process it quickly and efficiently. The present model of needs for HL-LHC assumes analysis can be done using MINIAOD. If instead, a significant fraction of the analyses require AOD, then this estimate of needs is too optimistic. If, on the other hand, a format can be developed that is significantly smaller than the MINIAOD then the needs would shrink. CMS is starting R&D in 2017 to define such a smaller format, currently referred to as MICROAOD. The CPU budget for analysis is driven by the average number of events per second analyzed. At present, analysis on MINIAOD achieves an average of 5 Hz. If a more refined data format could be developed that requires less detailed computations from the end user then the analysis CPU needs might decrease significantly.

These tasks are combined into three primary workflow types, although all tasks could be executed individually or in different combinations. The full simulation workflow consists of the generation + simulation + digitization + reconstruction tasks and produces AOD and MINIAOD output. The reconstruction workflow uses the RAW data tier as input and produces AOD and MINIAOD output. The analysis workflow is the most diverse workflow and is not handled centrally, using primarily MINIAOD as input, and can produce USER ntuples apart from plots and tables. CMS assumes that all intermediate outputs of the workflows remain on the compute resource, and has the ability to choose which outputs of a given workflow are archived to storage.

Status of Using HPC and Commercial Clouds

Executing CMS Applications on Intel Knights Landing Architectures

We have successfully demonstrated the multi-threaded execution of CMS' physics software on Intel Knights Landing (KNL) CPU architectures, by running the reconstruction algorithms on input data and producing analysis output data. Comparing event throughput per thread, the KNL-based host has 5x less throughput than a 2010-era Xeon host. A cost extrapolation suggests that, dollar-for-dollar, the KNL platform is far less efficient than a traditional Intel Xeon host.

CMS' physics application demonstrates near-linear scaling up to 64 threads; we fully utilize a current KNL host by running 4 multi-threaded physics applications in parallel. The factor limiting the number of threads is the need to serialize writing of results to storage, using the third-party ROOT libraries. Moving to a parallel-friendly I/O system would allow CMS to fully utilize the system with just one job instance.

To improve overall throughput on KNL hosts requires major refactoring of CMS software, to make better use of vectorization. This mean a significant overhaul of millions of lines of code. Currently the "hot spots" account for less than 5% of the total processing time, suggesting a broad end-to-end redesign of reconstruction algorithms would be needed to better leverage vectorization and increase memory locality. As such a code overhaul would also benefit the traditional Xeon platforms, we believe the performance gap would decrease but not close.

HPC Centers

U.S. CMS has demonstrated production-use of HPC centers with an emphasis on the main central CMS workflows, starting with the re-reconstruction of LHC Run 1 data at SDSC, to the the recent production use of allocations at NERSC for full Monte Carlo production workflows. Access to HPC resources is provided through HEPCloud at Fermilab for work submitted to the CMS global glideinWMS pool.

In the last months we made significant progress in our ability to use NERSC HPC resources for full simulation workflows, at the relatively low level of running about 5,000 jobs in parallel. CPU efficiency is reasonably high on standard Xeon architectures and the workflows behave similar to running on regular Grid sites. We are using pre-mixed pileup workflows that stream the pileup events from Fermilab. We have been able to run at much higher scales, achieving this for short amounts of time, but are starting to run into overload situations of the shared NERSC infrastructure.

This situation would worsen if we were forced to use standard pileup mixing to minimize network data access to Fermilab, and a solution would have to be found. We have a very good relationship to NERSC staff and are working with them to remove these limitations. A bottleneck we are frequently encountering is ramp up time, limiting the "elasticity" of resource provisioning for overflowing to HPC resources. It takes significant time to have glideinWMS pilots start at NERSC and provide significant resources compared to standard grid sites. This challenge will have to be addressed to enable substantial use of these resources.

Commercial Clouds

U.S. CMS has demonstrated that we can execute large-scale simulation workflows [1] generation + simulation + digitization + reconstruction) on the Amazon Elastic Compute Cloud (EC2) and Google Cloud Platform (GCP), via the Fermilab HEPCloud pilot project. Workloads were sustained at 60,000 concurrent cores on EC2 (increasing CMS global resource use by 25%) and 160,000 concurrent cores on GCP (increase of ~100%). An important factor in the success of these studies was that ESnet established high-speed (100 Gb/s) peering points with the commercial cloud providers.

The cost per utilized core-hour was estimated to currently be 60-80% higher than the cost of executing these workflows at the Fermilab Tier-1 facility; we anticipate cloud cost to decrease while industry continues to invest $30B/quarter into their infrastructure and service offerings. As in the HPC case, HEPCloud at Fermilab is used to provide access to commercial cloud resources for work submitted to the CMS global glideinWMS pool. The HEPCloud program is scheduled to transition to production/operations by the end of 2018.

Estimate of LHC and CMS performance

The overall scale of CMS' required computing resources is driven by the total number of events, both simulated and recorded. For our estimates, we follow the current long-range LHC schedule, which indicates shutdowns in 2019-20 and 2024-25, with otherwise full operational years, except in 2026 when HL-LHC will be commissioned. We include a Hübner factor of 0.6 at a typical 150 data taking days per year to determine the number of running seconds each year (7.8 million).

The CMS trigger rate for physics is currently 1 kHz, which we assume to remain constant until the HL-LHC era, when it is expected to increase to 10 kHz.A significant amount of computing resources go into Monte Carlo simulations. We assume that twice as many events will be simulated as recorded in any given running year.

We expect to reconstruct a given LHC year's data and to generate, simulate and reconstruct the same year's Monte Carlo samples within the same year. The CPU times required per event are given in Table 1, in units of HS06 seconds, for both the LHC (pre-2026) and HL-LHC (2026 onwards) eras. We note that the fact that HS06 seconds per event for the simulation workflow is the same as the reconstruction workflow is pure coincidence. Moreover, each simulated event must also be reconstructed. The CPU cost per event for simulation is thus twice that for RAW data. As we expect to simulate twice as many events per year than we take in data, this implies that the total CPU cost of data processing is ¼ that of simulation processing per year.

CMS has historically made continual improvements in the execution time of simulation and reconstruction through improved algorithms and software engineering. These are included in the model, assuming steady but moderate year-by-year improvements for the LHC software in the near term, and then more aggressive improvements in the HL-LHC software starting in 2024, as the collaboration becomes more focused on HL-LHC preparations.

[1] Given that these platforms charge an egress fee per GB that leaves their network, these workflows were chosen specifically to optimize cost, maximizing CPU use while minimizing the size of the output data.

Workflow	LHC events	HL-LHC events
Generation/simulation/ digitization [HS06*s]	600	4000
Reconstruction [HS06*s]	250	4000

Table 1: Current processing times per event for different workflows in HS06 seconds.

The estimated sizes of each data tier for LHC and HL-LHC events are given in Table 2.

Data Tier	LHC events	HL-LHC events	Use Case
RAW	1 MB	5 MB	LHC raw data
GENSIM	1 MB	5 MB	Simulated events
AOD	400 kB	2 MB	Input for 10% of analyses
MINIAOD	40 kB	200 kB	Input for 90% of analyses
USER	4 KB	10 kB	Unique per analysis

Table 2: Sizes of each data tier for LHC and HL-LHC events.

The annual CPU requirements are primarily driven by the product of the number of events that must be processed and the processing time per event. Several additional factors are included to make the model more realistic. Some additional resources are needed for ancillary activities, such as alignment and calibration and are included with the prompt reconstruction of detector data in the category of "Prompt Data" below. Resources are allocated for both re-reconstructing detector data and re-simulating Monte Carlo events in response to improved understanding of detector calibrations and improved algorithms. These include "legacy" reconstruction a year's samples at the end of each year, and samples for the entire run in the first year of a shutdown.

Resources for data analysis are estimated at a level that is consistent with current experience in the experiment, with increases in future years as LHC data accumulates. In the pre-HL-LHC era we expect to simulate and reconstruct some number of HL-LHC events for detector and physics studies. For the first few years this is projected as being just a few percent of the number of events that will be required during HL-LHC running, with increases as we get closer to the HL-LHC.

The total disk and tape space needs are driven by the number of events recorded and simulated. We take into account the fact that multiple disk replicas are needed of all the data in use for analysis in the distributed computing system. In general, we assume that more disk replicas are kept of data recorded in recent years than of data from earlier years. The detailed assumptions chosen in the model are based on our Run 2 experience. The baseline assumption is that the MINIAOD tier will continue to be used for the vast majority of analyses.

Estimation of CMS Resource Needs

We estimate the yearly CPU time needs of CMS according to the above model, with the results shown in Figure 1. The increases in needs seen in 2019 and 2024 are due to the planned legacy reprocessing of Run 2 and Run 3 data and simulation samples. There is a significant increase in needs at the start of the HL-LHC era, driven by the complexity of reconstruction of collisions of

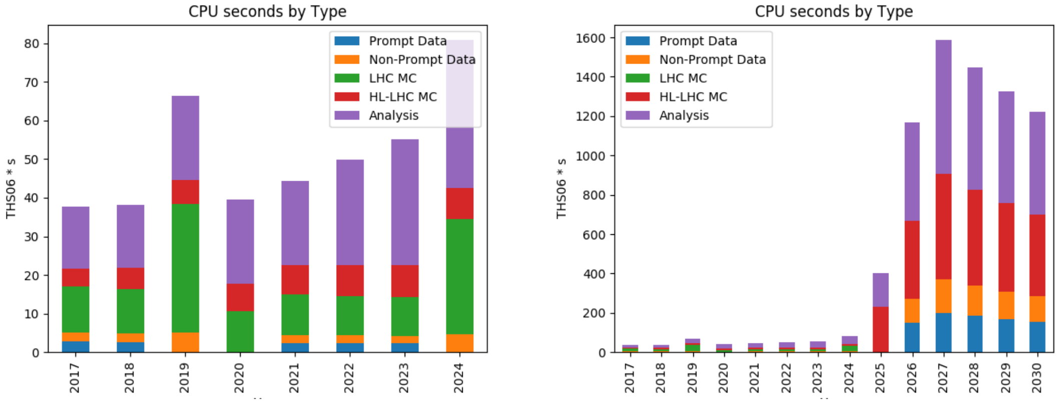

Figure 1: Estimated CPU time needs for CMS in THS06 s for the years 2017-2024 (left) and 2017-2030 (right) for the different major workflow types.

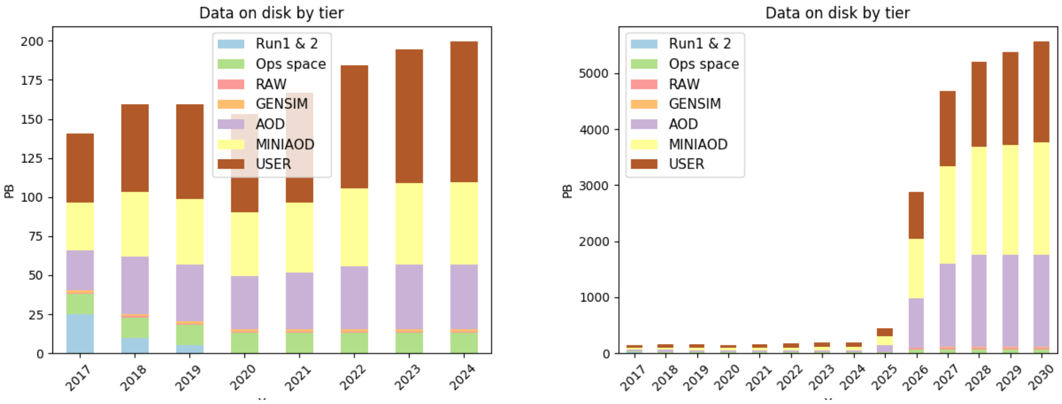

Figure 2: Estimated disk needs for CMS for the years 2017-2024 (left) and 2017-2030 (right) in Petabyte.

high beam intensities in both recorded and simulated events, and the ten-times increase in HLT output rate.

Figure 2 shows the estimate the disk needs of CMS after calculating the recorded and simulated number of events per year and applying the assumed event sizes per data tier. While disk needs only grow modestly in the LHC era, they are predicted to become unsustainable at the HL-LHC in the current model, even with a fairly aggressive policy of reducing versions and replicas.

The CMS tape needs are estimated with the same model and shown in Figure 3. We use the tape space occupied by LHC Run 1 and LHC Run 2 years 2015,2016 as Run 1 and Run 2 legacy data and assume that it cannot be reduced any further. We store two RAW copies on tape at different sites. The tape estimates include using tape for cold storage of AOD and MINIAOD samples.

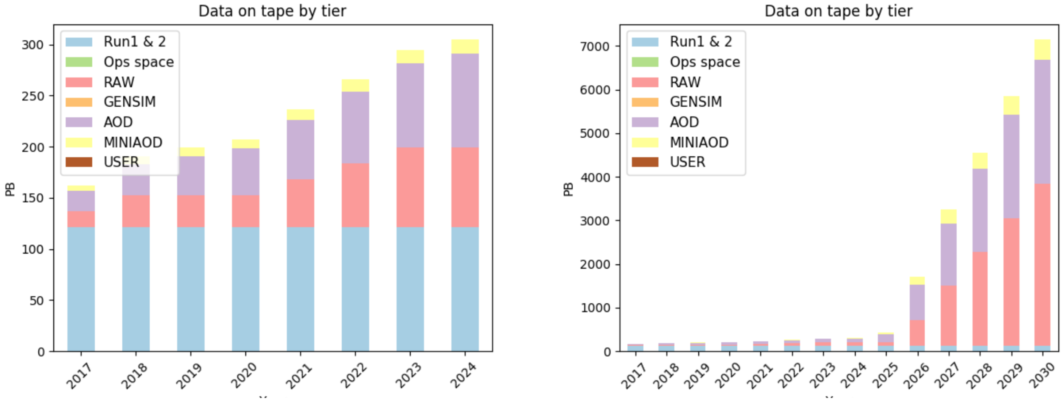

Figure 3: Estimated tape needs for CMS for the years 2017-2024 (left) and 2017-2030 (right).

Resource Needs in the U.S.

The previous sections gave detailed estimates for resource need profiles for CMS, covering the years leading up to the HL-LHC and the initial running years of high-luminosity operations. To estimate the computing resource cost for the U.S. we assume, based on the U.S. commitment for Run 1 and Run 2 to provide 40% of CMS Tier-1 resources, that for HL-LHC the U.S. will continue to provide 40% of the CMS production computing resource needs.

In addition, the U.S. CMS program has to provide computing resources to support data analysis for U.S. scientists. Ten years of historic data show that data analysis resource needs closely track production computing resource needs. U.S. scientists have consistently used about 50% of the total available computational and data storage resources in U.S. CMS, across the Tier-1, Tier-2, and LPC-CAF analysis facilities. Our resource plans for U.S. CMS thus include computational and disk storage resources for analysis at about equal size to the corresponding production resources the U.S. is providing. This estimate is consistent with the modeled analysis needs as shown above (considering that >30% of CMS analysts are U.S. scientists, and recognizing the leadership role of the U.S. in physics analysis and use of data analysis resources).

U.S. CMS CPU needs

Table 3 gives the estimate of U.S. CMS CPU needs through 2030. In addition, we estimate what fraction of these resources can be provided by HPC centers. We assume that all workflows except for the prompt reconstruction of detector data and user analysis can be performed at the centers, but not necessarily at the same event throughput per dollar as described above. The challenges of using HPC centers for these workflows is described in detail below.

As alternative scenarios, we consider possible improvements to the CPU needs due to HL-LHC software improvements. At the moment, HL-LHC digitization is done through standard pileup mixing. If pre-mixing can be implemented for the HL-LHC simulation, the digitization time could be reduced by 50%, leading to an approximately 10% reduction in overall CPU needs. Alternatively we can consider more aggressive improvements in the reconstruction algorithms that could lead to 20% decreases in CPU time. This would lead to an approximately 15% reduction in overall CPU needs. These two steps together could thus reduce CPU needs by 25%, as they have independent impacts. Such aggressive improvements would require significant investments in software engineering. To this end, the U.S. CMS Operations Program intends to add more effort in this area as we approach the HL-LHC era.

Year	CPU time (THS06 * s)	Fraction that could be run on HPC or commercial clouds
2017	15.1	50%
2018	15.3	50%
2019	26.6	67%
2020	15.8	45%
2021	17.7	45%
2022	19.9	40%
2023	22.1	36%
2024	32.3	53%
2025	161	57%
2026	467	44%
2027	634	44%
2028	579	44%
2029	531	44%
2030	488	44%

Table 3: U.S. CMS CPU requirements estimate. The integrated CPU time needed each year is given in Tera-HEPspec06 seconds (1 THS06s roughly corresponds to 100 Billion "Service Units" or CPU seconds on a Intel Xeon-type CPU), together with the fraction of CPU time that could be provided through commercial cloud services or through allocations at HPC centers, (with the provisos given in the text).

Many steps can be considered to reduce the requirements for analysis computing and to make HPC centers more functional for production computing, as we describe in the following sections.

U.S. Analysis CPU Needs

During LHC Run 1 and Run 2 about half of the experiment's computing resources have gone into data analysis and processing performed by individuals and small working groups, as opposed to the *production workflows* run by the central computing operation teams.

For HL-LHC we expect the analysis resource need to continue to scale in step with the production resource needs, and possibly more steeply. CMS does not yet have a bottom-up model to assess future analysis needs. Given the increased scales involved it is clear that data analysis at HL-LHC appears to become a huge problem without an obvious solution.

To continue to support hundreds of U.S. scientists in data analysis, optimizations of the current approach are needed to reign in the vast multiplication of data transformations and subsequent intermediate storage of a large number of differing data products for individual further analysis. Analysis requires on-demand execution, i.e. reasonably fast turnaround. Programs are individually written, adjusted and modified, less debugged than production workflows, and not tuned for efficient execution. Analysis activity tends to spike before major conferences.

New analysis paradigms and strategies will be required to address the analysis needs of the HL-LHC. We propose to invest R&D effort in dedicated analysis facilities that could streamline some of the individual work- and data flows. Optimizing data analysis would also bring about an evolution towards interactive "data mining" access to centralized data repositories hosted at the large data centers, with massive computational capabilities for latency-optimized fast-turnaround data transformations.

Improving turnaround times and latencies in performing data analysis, even as the data sample sizes increase massively, would significantly improve scientific productivity minimizing the "time to insight" for CMS analyses. For data analysis, capacity of storage space, performance of data access, and in particular I/O performance and latency, become the key concern. This is different from the production processing needs for elastic capacity and optimizing for overall throughput, where latency of individual processes are less of a concern.

Data analysis typically involves plowing through ever-increasing event samples to map and reduce data down to histograms and to perform statistical analysis. Orders of magnitude decreases in latency and event-throughput are needed to address this issue. This will require a more centralized end-to-end engineered approach to data analysis facilities to improve on the current capabilities and to enable analysis of much larger datasets than today.

R&D will be needed into data formats, compression algorithms, and new ways of storing and accessing data for analysis, to investigate optimizing the storage systems and data representation on disk together, and also to facilitate the utilization of new additional storage layers like SSD storage and NVRAM-like storage that exhibit different characteristics than the currently dominant spinning disk installations. These developments should include a fresh look at the concept of "virtual data", considering re-computing quantities instead of storing pre-computed values. Issues to be studied are relative cost and performance in optimizing resource use and performance parameters like throughput and latency of response. We propose prototyping such analysis facilities to develop the capabilities and the required ecosystem to support the analysis use cases.

It is of yet unclear what role HPC installations can play in this area, if any. In the current model these HPC facilities are available to our community mainly through providing allocation-based resources, instead of highly "elastic" on-demand, low latency response resource pools. Data analysis seems to be most challenging for the current HPC facilities model, and may be best addressed by dedicated analysis facilities with tight CPU to storage connection. Towards HL-LHC we envision dedicated data analysis facilities for experimenters that provide an extendable environment that provide fully functional analysis capabilities. In such an environment HPC facilities could play a central role as "data transformation engines," if a host of technology, data access and security issues can be resolved.

U.S. CMS Production CPU needs

Table 3 shows the fraction of production workflows that could be executed on HPC machines and commercial cloud resources. The main workflows to consider are data re-reconstruction and full simulation. Our model assumes that on the HL-LHC time scale the relevant HPC machine architecture will be based on Intel KNL CPUs or their successors. We note that it is quite likely that not all national scale HPC resources will be deployed using processor architectures that our applications can be ported to with reasonable amounts of effort. As a result, it is quite likely that we will be able to use only a subset of the available national scale HPC resources. Executing our production workflows on HPC machines and commercial clouds providers poses specific challenges that we would need to overcome to effectively make use of these resources.

The first challenge in using HPC resources is the distributed nature of the CMS collaboration and how the collaboration's software is developed. Hundreds of developers from over 40 countries contribute to the development of the CMS software that is used to execute the main production workflows. This is in conflict with the user model that HPC installations are used to, in which only a few collaborators are responsible for the software that is executed. We need to make sure that the requirements of the HPC centers continue to be compatible with the software development process of a large HEP collaboration.

The allocation-based resource assignment on HPC centers poses its own challenges to CMS. CMS is accustomed to receiving constant yearly allocations of computing resources and has the ability to use temporarily unused resources opportunistically. In contrast, HPC centers assign allocations to users which can be used in a specific time frame. The absence of a guaranteed resource assignment is a challenge for CMS to accommodate. On the other hand, HPC centers with significantly more resources than the CMS allocation offer the possibility of elastic use of allocations in shorter time frames to handle peaks in resource needs. We would like to work with the HPC centers to handle allocation-based resources for HL-LHC computing needs that best meet the operational needs of CMS.

The CMS software framework is currently used in multi-threaded mode in production with 4-8 threads. At 8 threads, we start to observe inefficiencies due to the current output I/O handling. These inefficiencies become more severe at higher thread counts, as observed when running on KNL architectures. We have begun R&D this year to address this bottleneck and will need continued support to effectively use KNL architectures in the HL-LHC era.

The product of event throughput rate and memory required is the appropriate benchmark to measure the effective use of a CPU resource. A multi-threaded application will not have the same event throughput as the corresponding number of single-threaded applications but will use significantly less total memory. On modern high-core CPU architectures, it is of paramount importance to run in multi-threaded mode because of the limited memory per core. In addition, the individual cores in KNL architectures are optimized for vectorized code. As stated above, the CMS algorithms are currently not optimized for vectorization, thus the event throughput on KNL is 5 times smaller than on Xeon architectures at comparable thread counts. R&D will be needed to vectorize reconstruction algorithms and engineering effort will be needed to implement the algorithms within the optimized framework.

HPC centers are designed for large applications spanning many individual nodes. CMS' application is most efficient single threaded and only to optimize memory consumption and to exploit new architectures is it using multi-threading and vectorization. The CMS application will be able to reach modestly high thread counts but will never be efficient at extremely high number of cores. Therefore the CMS application will not reasonably be able to use more cores than on a single node on an HPC machine, and might even need to run multiple applications depending on the characteristics of the nodes. To efficiently make use of allocations at HPC centers, CMS will need to be able to start sufficient numbers of applications and sustain these numbers to achieve good throughput. Current experience shows that this is not always easy and R&D will be needed to optimize this.

The ramp-up time is currently a limiting factor as well. Improvements would allow CMS to exploit elasticity if resources are available. Because experiment software and alignment and calibration constants need to be finalized and certified before production starts, shortening the production time significantly would mean that improvements in software and conditions could be integrated into physics analyses much quicker. Therefore being able to exploit elasticity would increase the physics output.

This would better match CMS's operations mode. Final software versions and calibration and alignment constants need to be available well in advance to allow for the production of the samples. If the production time can be shortened, this would increase physics output.

Data handling at the scales required for LHC and HL-LHC science will be challenging for HPC centers. In all cases, sufficient network bandwidth in and out of the HPC center will be needed to either use local disk for caching or streaming access to data. All of CMS's workflows need to write significant amounts of output either locally or directly off-site. Further R&D and integration work will be needed to identify and remove bottlenecks.

The full simulation workflow does not need arge amounts of input data like the data reconstruction workflow does, but does need access to minimum bias samples for pileup simulation. For pre-mixed pileup access, streaming could be feasible, but at very high core counts this would need significant network bandwidth to serve pileup to the applications. A single pre-mixed pileup minimum bias sample for the simulation of one LHC Run 2 running period requires about 500 TB of disk space and could be stored locally if disk space is provided. However, standard mixing will be used in HL-LHC simulations for some time to come, requiring the input of the order of 1000 minimum bias events in addition to the primary interaction. Streaming in this scenario becomes close to impossible. The minimum bias sample for standard mixing in Run 2 has an average size of 50-100 TB. Providing low latency and high throughput access to this data at the HPC center would be needed to run the full simulation workflow efficiently. Reading many small events requires lower latency than reading one larger pre-mixed event to keep the efficiency high, even if it is significantly larger in size.

Commercial clouds are currently 1.5 times more expensive to provision and operate than HEP owned resources, based on using the most cost effective cloud resources available, including the spot price market of AWS. But commercial cloud providers are ideal for elastic scale out as demonstrated in the CMS pilot runs on EC2 and GCP.

Data access from the applications running in commercial clouds can be provided through streaming as well as local access. Providers charge for local disk space, and local data needs to be replicated to different geographical regions to reach sufficient scale. While the transfer of data into the clouds is free, sufficient network peering with the scientific networks needs to be provided. Local space would be needed for workflows where streaming input is not an option, such as the full simulation workflow in standard mixing mode.

We do not see technical differences between running production workflows on HPC machines and commercial clouds. Although the business model of commercial clouds is different, our workflows with characteristics of small output volume to CPU fraction are not expected to accrue significant additional cost due to using commercial cloud providers. We also expect that the current business model of charging for data egress, although we do qualify for waivers in certain cases, will evolve in the future and that new commercial cloud providers will be available that are even more suited for our workflows.

U.S. CMS Disk Needs

Table 4 shows the U.S. CMS disk requirements, which reach into the exabyte range in the HL-LHC era. The *High* scenario represents our baseline model. The *Low* scenario assumes that the AOD and MINIAOD event sizes can be reduced by 20%, which leads to an approximately 15% reduction in the overall disk needs. It is hard to see how such disk storage needs for the HL-LHC could be accommodated. Transformative changes will be required.

Year	Low [PB]	High [PB]
2017	56	56
2018	63	64
2019	63	64
2020	60	61
2021	65	67
2022	72	74
2023	76	78
2024	78	80
2025	160	178
2026	995	1,150
2027	1,614	1,870
2028	1,795	2,079
2029	1,864	2,152
2030	1,933	2,225

Table 4: U.S. CMS disk requirements estimates, in units of petabytes.

The U.S. CMS disk needs are dominated by the need to provide access to data for analysis, which is traditionally very hard to model and predict.

The current analysis access model has worked well for LHC Run 1 and 2. The model is based on efficient and performant local access at the Grid sites along with the ability to stream data from the same resources to all other Grid sites. The latter requires strong network capabilities. The disk needs are determined by how many replicas are needed for efficient global access as well as how many different versions (from different software versions and constant sets to different file contents) as well as the main analysis file content and its size. The extrapolation of the current model based on MINIAOD as the main event format will require significant disk resources for HL-LHC. As outlined above, the HL-LHC analysis model needs to undergo significant changes to be sustainable and performant for the masses of data and simulation events. One aspect could be the development of a new, much smaller primary analysis event format, called MICROAOD. It would not only reduce the amount of disk needed to provide access to data and simulations for analysis, it would also allow changes to the analysis flow and reduce the USER space needed for ntuples. Preliminary estimates show a total reduction of needed disk space for HL-LHC by a factor of approximately 5 are possible. To achieve this goal, effort would be needed to improve the production infrastructure to output this new format much more quickly and efficiently than before, as well as effort to facilitate the usage of this new format using new analysis facility concepts as outlined above.

U.S. CMS tape needs

Table 5 presents the U.S. CMS tape requirements. Tape archive space is mostly for disaster recovery of RAW data. The amount of capacity needed is determined by the accelerator performance and the trigger rate and is irreducible; it reaches exabyte scale in the HL-LHC era. We assume that tape will continue to be the technology of choice for disaster recovery into the late 2020s. We plan to investigate cold storage technologies, including tape technology, to further optimize disk needs for data that is used rarely. The idea is to remove rarely used data from disk and store it on cold storage, and retrieve it when it is needed while retiring other rarely used data, therefore reducing the overall disk needs. The presented tape estimate includes a model to use tape as cold storage for AOD and MINIAOD samples.

Year	Nominal [PB]
2017	65
2018	76
2019	80
2020	83
2021	95
2022	106
2023	118
2024	122
2025	172
2026	687
2027	1,299
2028	1,823
2029	2,341
2030	2,858

Table 5: U.S. CMS tape requirements estimates, in units of petabytes.

Conclusions

This note presents an initial estimate of overall CMS computing resource needs through 2030, and the expected U.S. CMS share. We emphasize that these estimates have not been vetted by CMS management and have significant uncertainties, especially for the HL-LHC era, which is sufficiently far off that there could be disruptive technology changes. While the needs appear to be manageable for Run 2 and Run 3, given the general stability of LHC conditions and our experience with them, Run 4 is expected to be much more challenging, and transformative changes will be required to provide sufficient data analysis resources to fully exploit the physics potential of the HL-LHC era. We described a number of possible changes, including the development of new data formats and specialized computing facilities for analysis. Our estimates of resource needs also include the fraction of the processing that could be completed at HPC centers or commercial clouds. U.S. CMS has demonstrated the use of these facilities for the main CMS workflows, but has also identified a number of issues that need to be solved through R&D efforts to use HPC centers and commercial clouds efficiently.