White Paper on Future US CMS Computing

Lothar Bauerdick[1], Ken Bloom[2], Sridhara Dasu[3], Peter Elmer[4], David Lange[5], Kevin Lannon[6], Salvatore Rappoccio[7], Liz Sexton-Kennedy[1], Frank Wuerthwein[8] and Avi Yagil[8]

[1]Fermi National Accelerator Laboratory
[2]University of Nebraska – Lincoln
[3]University of Wisconsin – Madison
[4]Princeton University
[5]Lawrence Livermore National Laboratory
[6]University of Notre Dame
[7]State University of New York – Buffalo
[8]University of California – San Diego

November 12, 2015

1 Introduction

Completion of the Standard Model of particle physics with the discovery of the Higgs boson at M=125 GeV at the LHC behoves us to continue its detailed study at low energy scales, while continuing to explore higher energy new physics. The elusive TeV-scale super-symmetric or other exotic states that we hope to discover also seem to require significantly more collision data. The CMS experiment at thee LHC has resumed data-taking after a two-year shutdown this Fall.

Computing environment and facilities for the CMS experiment are continually evolving to meet the requirements of the collaboration and to to take advantage of the evolution of technology within and beyond the high-energy physics community. During the recently started new phase of data acquisition, i.e., Run-2 (2015-18) and Run-3 (2021-23) of LHC about 300 fb^{-1} will be accumulated. This 300-fb^{-1}-dataset presents two orders of magnitude increase in data volume compared to Run-1 (2009-12) dataset: An order of magnitude increase in integrated luminosity, a factor of three increase in trigger output rate to facilitate continued access to electro-weak scale physics, and a factor of three or so increase in event-complexity due to increased energy and instantaneous luminosity leading to event pileup. We note that this new beginning of the LHC data taking period as an opportune time to take stock of this evolution.

The Moore's law scaling of computing capabilities and evolution of storage are quite unlikely to meet this challenge under constant budgetary levels. Fortunately, innovations in resource utilization, adaptation to modern computing architectures, and improved workflows, are making up for the limitations in raw scaling of resources. We briefly describe these evolutionary changes in the offing and project how agile computing, utilizing owned, opportunistic and commercial cloud resources, with dynamic data management and just-in-time data movement over wide-area networks, will work to meet our challenge. In particular, this document focusses on the chaotic physicist-driven scientific data analysis activities rather than the dedicated prompt data production facilities. A brief look at the computing R&D necessary for the HL-LHC phase (2025+), during which another two orders of magnitude in data volume is expected, is also provided.

2 Motivation and Goals

(Barberis, Bauerdick, Bloom, Dasu, Wuerthwein, Yagil)

```
Notes from past meetings:
   Brief note on physics goals noting computing challenges, e.g.,
      Distributed seamless data access to all centers of computing
      High-throughput access to mini-AOD data for sparse event selection
       Ability to manage small selected data sets conveniently for analysis teams
       Ability to cope with custom data seamlessly, if they are necessary for analysis
   Brief note on changes to trigger and other running conditions (PU) affecting the input
      Scale of Run-2 data sample and processing needs
      Outlook for beyond Run-2
      Upgrade simulations need
      Computing upgrade R\&D itself
      Brief note on computing landscape now and immediate outlook
      Tier-1 scale and near term planning
```

```
       (primarily defined by CMS DAQ output and policies, LHC performance
        and especially re-reconstruction)
      Tier-2 landscape and where we are possibly headed
  (primary purpose of this document and defined by analysis plans/needs)
Cloud resources:
Handling of peak needs using non-owned, short-term and rented resources.
Projections of campus grids, OSG and other national grids, commercial clouds
  High level themes
      Agile computing
      Location independence
      Automatic optimization of CPU vs Storage for various data processing stages
```

3 Description of Run-1 Computing

(Dasu, Bloom)

```
Brief history of Monarch model
Describe entirety of computing in a page or two if they embed.
```

3.1 Evolution of computing during Run-1

3.2 Lessons learned

4 Developments during LS1

(Bloom, Dasu, Elmer, Lange)

During the LS1, the US CMS institutions have developed new solutions for CMS computing problems that result in more efficient use of both storage resources, using dynamic data management and seamless wide-area network access to our storage (AAA and XROOTD), and compute resources, using improved workflows, multi-threaded framework, GPUs and ability to exploit cloud technologies for CMS workflows.

The original "Monarch" model of organization of CMS computing resources in a tiered structure is now dated. While we retain Tier-0 at CERN for prompt processing, both calibration and reconstruction, the functionality at higher tiers is changing. Especially at the Tier-2s, we are evolving to a set of institutions providing portions of resources, focusing on local expertise, in a continuum infrastructure of services. Here we will develop an evolving new model for optimal use of resources to provide for CMS computing needs.

4.1 Dynamic Data Placement and Management

Underpinning the computing infrastructure in Run-1 was data distribution mechanism implemented using PhEDEx software. The workflow involved operators moving large chunks of data on command down the hierarchical grid so that after initial calibration and reconstruction at Tier-0, the raw data were moved to be archived at Tier-1 and reprocessed as necessary. The (re)reconstructed data from Tier-1s were further processed to obtain lower volume Analysis Objects Data (AOD) versions whose copies were transferred to Tier-2s and placed on disk storage for random access for chaotic analysis workflows. Similarly the Monte Carlo (MC) simulation data was produced at Tier-2s, aggregated, reconstructed and archived at Tier-1s. The MC AODs were then placed on command by the operators at various Tier-2s. The net result was multiple copies of data placed statically at various facilities around the globe. It was observed that much of the disk volume was occupied by often rarely used data.

Dynamic data placement and management (DDM) software was developed and commissioned during the LS1 to address these shortcomings. DDM is now deployed at all tiers to automatically prune the unused but archived data using well defined policies. For example, the archived full event, i.e., raw plus reconstructed quantities (FEVT) is pruned from disk to keep sufficient space for the Tier-0 and Tier-1 reconstruction workflows to execute smoothly. Most importantly we are able to keep at least one copy of all AOD on disk somewhere in the CMS data federation, and duplicate multiple times as needed for popular data.

4.2 CMS Data Federation (AAA)

CMS Data Federation is built to provide international-scale data access infrastructure under the name Any Data, Anytime, Anywhere (AAA). AAA removes the requirement of co-location of storage and processing resources. The infrastructure is transparent, in that users have the same experience whether the data they analyze is halfway around the world or in the room next door. It is reliable, in that end users never see a failure of data access when they run their application. It enables greater access to the data, in that users no longer have the burden of purchasing and operating complex disk systems. In fact, any data can be accessed anytime from anywhere with an internet connection. The key to success of AAA is the improved wide-area network access due to enhancements made to our dedicated LHC network.

AAA is made possible by XRootd software, which allows the creation of data federations. A data federation serves a global namespace via a tree of XRootd servers. The leaves of this tree are referred to as data sources, as they serve data from the local storage systems. Each storage system is independent of the others, allowing for a broad range of implementations and groups to participate in the federation as long as they expose an agreed-upon namespace through the XRootd software. The non-leaf nodes have no storage, but may redirect client applications to a subscribed data source that has the requested file. Each host is subscribed to at most one redirector, called a manager; loops are disallowed. If the requested file is not present on a server subscribed to the redirector, then the client will be redirected to the current hosts manager. The manager continues the process until either a source is found or the client is at the root of the tree. An application may thus be redirected to any host in the federation, irrespective of the branch point it initially accesses.

CMS Data Federation is now fully deployed across all tiers of its computing infrastructure. Easy access to this data federation across the wide-area network is democratizing the computing abilities of University groups across the world. Local campus grids controlled by non-CMS entities are easily integrated in the CMS computing environment. Temporary access to dedicated large resources can be purchased on commercial grids or obtained from national or campus research facilities.

4.3 Development of MiniAOD

Physics analysis often involves a much smaller portion of reconstructed data than is available. While the raw data acquired from CMS is about 1 MB per event, the reconstructed objects more than double that size typically. The AOD defined for Run-1 was successful but was designed in a rather lax way resulting in a 200 kB size per event. Careful pruning of unused collections of objects, packing them in appropriately sized containers resulted in redefined miniAOD which is less than 50 kB per event. Rate of event processing matters so size of event being small is also beneficial for computational loads. The miniAOD is now visualized as the main data format that will be used by bulk of CMS analysts, while niche usecases involving the FEVT format will be supported as needed.

CMS would like to ensure that the CMS Data Federation hosts the compact MiniAOD with appropriate level of replication for the popular data. The estimated size of the data federation for 300-fb^{-1} luminosity data sample is: (50kB)(1kHz)*(10^{7}s/year)*(6y)=3*10^{9}MB=3PB. Typically to analyze this data requires an associated MC sample which is three times its size, leading to a 10PB disk storage requirement for the CMS data federation for an instance of MiniAOD. Since the MiniAOD is highly processed it is expected that improvements to the reconstruction, calibration and other changes will necessitate remaking of MiniAODs. Unfortunately, support of multiple versions of MiniAOD for several versions, say three, are needed in order to accommodate the analyst needs. Not all analysts can migrate from one version to another immediately. We estimate that two to three MiniAOD versions are needed at any time, with the older data deprecated dynamically as they become stale and unused. This results in expected data volume of 20-30 PB of storage for MiniAOD. The corresponding (tape) archived raw data size is 60PB. It is assumed that the reconstructed FEVT objects are only available in the intermediate stages of processing and not archived.

4.4 Improvements to CMS Workflows

The main objectives of the workflows management middleware is to process data as quickly as possible, maintain uniform load across all resource sites and enable fast recovery in case of a site service interruption, e.g., by relocating jobs on an alternate site, while keep track of the integrity of the combined dataset. Recent developments in workflow management enable CMS to utilize impermanent opportunistic resources for data production. These workflow changes are enabling CMS to take better advantage of owned and opportunistic resources. For instance the ability to use the high-level trigger farm for MC production and reconstruction during the down periods of time. Recent tests indicate ability to switch to offline processing workflows within several minutes. Improved data transfer technology and remote Xrootd access to CMS data federation are enabling technologies.

4.5 Opportunistic Resources

CMS is in exploratory phase for smoothly integrating opportunistic resources for production and routine use. National research computing sites such as NERSC and SDSC have large resources, but often requiring additional work in adaptation of our software suite to smoothly operate there. Some access restrictions are worked around with user level code, e.g., CVMFS through Parrot and Docker/Shifter containers on Cray supercomputers. Commercial clouds such as AWS have also been used, but have cost-implications placing constraints on workflows. For example, the stage-out of data over the network is expensive. We have adapted by chaining various stages of workflow so that the smallest useable unit, say miniAOD is the only output that is staged out to the CMS data federation. Non-owned campus clusters are accessible both through their OSG connection if existing generically, or by placing suitable head-nodes at the participating university CMS group facilities. This latter use is of particularly important for analysis groups at access their home resources seamlessly processing data from the central CMS data federation using centrally supported code and conditions repositories using technologies such as CVMFS and caching SQUIDs. The CMS HLT cloud using virtual machines technology is able to quickly bring in very large resource during data taking down periods for offline workflow processing. Some of the innovations made there are useable elsewhere.

Final stages of physics analysis often involve workflows that are not centrally managed CMSSW framework jobs. Technologies such as CMS Connect are able to use campus grid and department level computer clusters to bring additional opportunistic uses for these cases.

4.6 Improvements to CMSSW Framework (multi-threading)

Multicore computing systems have become ubiquitous in the past decade. However, efficient use of available resources, especially memory volume and access, required adaptation of our software to suitable multithreaded frameworks. CMS engaged in a systematic effort to make the core CMSSW thread-safe and has successfully deployed it in the past year. Further tuning of software to increase the resource usage efficiency is underway.

4.7 Evolution of the computing hardware: CPU, Storage and Network

(Elmer)

5 Current S&C R&D

(Elmer)

```
Talk about future directions too ...
```

6 Requirements of Run-2 Computing Systems

The scale of computing resources necessary is directly coupled to the foreseen output from the detector. The trigger rates have been increased by an order of magnitude compared to the original goals at the time of CMS computing TDR. The discovery of Higgs at low mass and continued investigation of EWK scale physics requires low thresholds. Understanding the trigger plans and organizing the data in an appropriate way, for example in high and low priority processing-streams, may be a new direction to explore to flatten the computing resource needs.

6.1 Physics Analysis Considerations

(Barberis, Yagil, Sal)

```
Location independence
Rate of event processing
```

6.2 Data Reconstruction Considerations

(Sal)

```
Line between official and private data
More flexibility and faster turn around in official production.
Private data is more IO limited than today.
```

6.3 Monte Carlo Production Considerations

(Sal)

7 Proposed Solutions

Vision for US CMS computing environment for 2017-21 and beyond, comprises strengthening of its current FNAL based Tier-1 and seven university based Tier-2s, all functioning as portals to the nation-wide research and commercial clouds. These service providers will also enable all US universities to connect to the seamless CMS cloud, by providing them suitable "headnodes", democratizing access CMS computing. The resource provisioning targets steady-state operations level at the owned facilities in FNAL T1 and University Tier-2s, while peak fluctuations are handled by overflowing to the clouds. Non-owned opportunistic resources at all campuses are integrated in the CMS cloud.

The advantage of strengthening the university sites is multi-fold:

- Each university group brings unique experience and expertise to bear
  - MIT: Dynamic data management and production operations expertise
  - Nebraska: Dr. Bockleman et al, brought in numerous innovations to CMS middleware
  - San Diego: Connections to SDSC, Connections to core CMS software developers
  - Wisconsin: Connections to HT-Condor and OSG core-developers
- Connection to strong physics groups at the universities
  - Student and postdoc physics analysts exercise the system providing appropriate usecases for tuning.
  - Faculty collaborations at the University level can bring in additional campus or cloud resources
- Cost of infrastructure is subsidized at the Universities.
- Cost of personnel is also lower.
- Friendly competition amongst the sites results in increased productivity.

7.1 Storage Resources

The storage resource requirements are estimated by scaling current data set of 30 fb^{-1} acquired in 2010-2012 (Run-1) and 2015 (Early Run-2) to the full dataset of 300 fb^{-1}. MiniAOD usage reduces the needed resources significantly but poses, as indicated earlier. However, it poses data versioning issues resulting in multipliers. The user data storage space, is somewhat ill defined, but it does scales with the luminosity. We use the current usage at US T2s with the luminosity ratio to estimate this.

- 20-30 PB for MiniAOD (disk for one copy)
  - 40 PB for MiniAOD for analysts including replication.
- Currently user space at T2s is typically 0.5 PB with 10% of data acquired, i.e., 30 fb^{-1}. Projecting to full dataset one gets 5 PB per T2.
  - 30 PB of storage for user data (non-archived)
- CMS upgrade design tuning requires custom simulations, which are much more storage resource intensive than Run-2/3 data. The pileup level will be an order of magnitude higher.
  - 20 PB of storage for upgrade data
- Analysis workflows and improvements to MiniAOD require access to RAW data
  - 60 PB for RAW (tape archive) for full 300 fb^{-1}
  - 10 PB for RAW for special analyses cache
- Total Disk Storage: 100 PB
- Proposed Approximate Disk Storage Distribution: 30 PB disk and 100 PB tape at T1 and 10 PB at each at seven US T2s

7.2 Computing Resources

Unlike for Run-1, the Run-2/3 Tier-2 compute resource provisioning details are not concretely specified to allow local optimization based on cloud and opportunistic resource availability. However, the non-owned storage resource use is not yet well defined. Therefore, it is visualized that much of the storage, which is pared down to the minimum with improvements made in LS1, is owned and operated at the Tier-1s and Tier-2s.

The CPU requirements are estimated, in units of number of slots needed, by scaling current usage up by a factor of 10 accounting for increase in expected integrated luminosity. The increase in complexity of analysis is assumed to be compensated by the improved framework job efficiency and advances in computing power of individual machines. Use of CMS data federation across the wide-area network at owned resource sites and the clouds in general, opens up the possibility of provisioning needed compute resources in a flexible way depending on cost/benefit. However, it is visualized that a fraction of resources are housed at existing T2s.

- Currently 30,000 jobs, averaged over the past month, run at the seven US T2s equally split between production and analysis and 10,000 production jobs at FNAL T1.
- Scaling by luminosity, 300 fb^{-1} vs 30 fb^{-1} collected to date, we should expect to support 300,000 jobs at T2s at steady state and 100,000 jobs at T1.
- Proposed job slot availability: 100,000 for production at T1 and 40000 through each of the seven US T2s.

7.3 Network Resources

The network bandwidth requirement will also scale with increased data size and wide-area distributed computing. Typically sites are connected through 100 Gbps network presently, and we visualize multi-100 Gbps connections in the coming years and that they are funded through separate initiatives.

7.4 Non-owned Resources Usage Plan

Need to explain why we need to provide infrastructure. Why not amazon computing cards for all.

7.4.1 Opportunistic

7.4.2 Commercial

7.5 Middleware / Software

7.6 Support personnel roles

Two persons at each facility are necessary to provide full coverage. However, recent experience indicates that about 30-50% of those person's effort can be freed up for other work. Most of the effective people involved in CMS computing are former HEP physicists, who have now become experts in computing. They are able to provide wide-ranging expertise in physics software development. The additional services we expect Tier-2 personnel to provide are in the areas:

- Support for non-Tier-2 university portals to CMS cloud
  - We expect each Tier-2 to support about 7 universities in their neighborhood.
- Computing services for CMS upgrades and research to address future needs
  - Development of simulation program for upgrade detectors
  - Production of simulation data for upgrade detectors
  - Participation in computing research
  - Participation in DIANA/HEP and other community wide computing projects for future.

7.7 Physics Driven Datapaths

```
Datasets of trigger paths.
On-demand processing for some paths, etc.
```