NSF Cooperative Agreement - Computing Section

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 21, 2015

1 Software and Computing

1.1 Introduction

The NSF support for software and computing at the US Universities played a crucial role in the success of CMS program, having contributed to almost all the published work thus far, including the discovery of the Higgs boson that completed the Standard Model of particle physics. Continued NSF support for software and computing is mandatory for future successes, including perhaps discovery of new physics. In this section we briefly describe the current status and future plans of US CMS software and computing project, focussing on its Tier-2 program, on which US and international CMS physicists rely upon for extracting physics from the expected large CMS datasets.

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.

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.

Our vision for meeting the challenge of growth of computing needs beyond what is affordable via a simple Moore's law extrapolation is threefold. First, we will gain efficiencies by being overall more agile in the way we use the traditional FNAL based Tier-1 and seven university based Tier-2s center resources. Second, we will grow the resource pool by more tightly integrating resources at all other US-CMS universities, DOE and NSF supercomputing centers, and commercial cloud providers as much as possible. And third, we will pursue an aggressive R&D program towards improvements in software algorithms, data formats, and procedural changes for how we analyze the data we collect and simulate, in order to significantly reduce the computing needs.

1.2 University Facilities (WBS 2)

The tiered computing model of the LHC experiments, based on a distributed infrastructure of regional centers outlined by the MONARC project **ref**, includes Tier-0 center at CERN, one US based Tier-1 center at FNAL (WBS 1) and seven US university based Tier-2 centers (WBS 2) at **Caltech, Florida, MIT, Nebraska, Purdue, UC San Diego and Wisconsin**.

The original MONARC 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. Nevertheless, dedicated facilities at the existing Tier-2s to address the core analysis computing needs must be met.

The advantage of strengthening the existing 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.

1.2.1 Current Status

1.2.2 Future Plans

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.
- 10 PB for AOD (for a fraction of datasets)
  - 10 PB fraction should satisfy users requiring full AOD access.
- 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: 110 PB
- Proposed Approximate Disk Storage Distribution: 30 PB disk and 110 PB tape at T1 and 12 PB at each at seven US T2s

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 30 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.
- Bulk of the 15,000 analysis jobs running at Tier-2s recently, are identified as 13-TeV MC jobs. The MC for 2015 was generated with the anticipation that we collect 10 fb^{-1} this year. Unfortunately, we only collected 2 fb^{-1}. Nevertheless, the analysis effort presently is equivalent to 10 fb^{-1}.
- Therefore, scaling by luminosity, 300 fb^{-1} vs 10 fb^{-1} collected to date, we should expect to support about 900,000 jobs at T2s at steady state and 300,000 jobs at T1.
- Proposed job slot availability: 300,000 for production at T1 and 100,000 through each of the seven US T2s.

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.

Non-traditional resources beyond Tier-1 and Tier-2

In the last few years, NSF-ACI made some very substantial investment into networking infrastructure across more than 100 campuses nationwide. Among them are 25 of the 40 collaborating universities in US-CMS. We propose to build on this NSF investment by working with all of them, as well as any of the remaining 15 universities interested, to fully integrate their campus IT operated hardware infrastructures and ScienceDMZs into the US-CMS Tier-2 infrastructure. This will be done following the model of the NSF funded "Pacific Research Platform" (PRP) using Open Science Grid (OSG) tools and processes. The PRP deploys single nodes into the ScienceDMZs of 20 institutions across the West Coast, including the US-CMS institutions UC Davis, UC Santa Barbara, UC Riverside, Caltech, and UC San Diego. These pieces of hardware are collaboratively maintained between the campus IT organizations and the PRP and SDSC teams at UCSD such that local IT is responsible for hardware and user account maintenance, and UCSD is responsible for all OS and software service maintenance. The functionality implemented includes interactive data analysis, batch submission, CVMFS software cache, XRootd data cache, and XRootd server to export local data. The hardware is effectively a Tier-3 in a box without any of the human maintenance needs from the local CMS community. The deployment model includes careful custom integration into any existing University clusters accessible to the local group. This is made manageable with minimal effort beyond the initial deployment by management of OS, US-CMS and OSG services, and local configurations via a central Puppet infrastructure at UCSD.

The local CMS community is thus empowered to transparently use any and all local resources the University allows them to share in combination with the entire Tier-1 and Tier-2 system. Official CMS data is cached locally as needed. Private data by the local community is served out to the Tier-1 and Tier-2 system via XRootd servers. Each Tier-2 will also have an XRootd cache in order to transparently cache the private data of any of the local communities to avoid IO latencies due to WAN reads given the finite speed of light. The HTCondor batch systems implemented on these pieces of hardware are all connected to the global CMS HTCondor pool via glideinWMS. Similarily any University clusters are integrated requiring nothing more than ssh access to a US-CMS account on the local University cluster. Sharing policies are controlled locally following local rules at each University. We expect that some Universities will enable access to all of US-CMS to share their spare capacity, while others will be more restrictive. All of this is presently already deployed and operated by PRP and SDSC for the US-ATLAS group at UC Irvine, and is being deployed at the CMS institutions listed above. Operations for the UC groups is funded via a mix of NSF and state funds. We are proposing to scale out deployment and operations of this model across the US to as many US-CMS institutions as possible, focusing on the 25 institutions that have received ScienceDMZ funding from NSF-ACI since 2012. The hardware costs as well as the human effort to deploy and operate this system will be borne out of the Tier-2 portion of this proposal. At a cost of ~$10,000 per Tier-3 in a box, this is a modest fraction of the total Tier-2 hardware budget across the seven Tier-2s and the 5 years of this proposal.

We fully understand that the above model will not be appropriate for all 40 collaborating institutions within US-CMS. We thus augment it with an additional hosted service build on the OSG-connect model pioneered by the University of Chicago OSG/ATLAS group. This service will provide identical functionality to the Tier-3 in a box for institutions that are either lacking appropriate network connectivity or a local IT organization that would be capable and/or willing

to collaborate on the hardware and user account maintenance. There will be only a single instance of this "CMS-Connect" infrastructure for all these remaining groups. Groups within US-CMS are thus generally better off with a Tier-3 in a box, especially when they have sizable private data collections and large groups of students and post-docs.

Finally, we will fully integrate cloud services access into this infrastructure in such way that local University groups can use local funds to purchase cloud resources to augment their personal access to computing resources, and thus accelerate their science. We expect to be collaborating on this functionality with the HEPCloud project at FNAL as well as the Open Science Grid.

In addition to all of the above functionality geared towards data analysis, we propose to also integrate Supercomputing resources at DOE and NSF funded national facilities mostly for the purpose of simulation and reconstruction, i.e. the production of the official CMS datasets. Again, we expect to collaborate heavily with HEPCloud and OSG on the detailed access mechanisms and policies. At this point, December 2015, HEPCloud is focused on AWS, while OSG is working with Comet (NSF) and Cori (DOE) to understand the technical, operational, and security processes for use of these supercomputers via OSG interfaces.

Facilities Support Personnel

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.

1.3 Operations (WBS 3)

Operations that is not the operation of the facilities themselves. We are paying for Frontier ops/support at Johns Hopkins, software distribution at Florida, glide-in and submission infrastructure operations at UCSD, AAA operations at Nebraska, T0 operations at MIT, and a ton of COLA for MIT students who are working on various aspects of CompOps. Also part of Ajit is in here but if you want him to work on non-owned computing we will probably move him somewhere else.

1.3.1 Current Status

1.3.2 Future Plans

1.4 Computing Infrastructure and Services (WBS 4)

The US CMS S&C institutions continue to develop 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. For instance, data management and workflow management development is done at Cornell, xrootd development at UCSD, and dynamic data placement development at MIT. Additional support is provided by NSF through AAA project to Nebraska, UCSD and Wisconsin.

1.4.1 Current Status

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.

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.

Data Management and Workflow Management

CMS data is organized in several tiers ranging from RAW data acquired from the detector, to RECO format for reconstructed data, FEVT combining the two, full set of analysis objects (AOD) and compressed AOD, i.e., 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 400 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 original AOD format will be supported as needed. In rare cases FEVT access may also be needed. As the miniAOD improves we anticipate AOD replicas will be small.

CMS would like to ensure that the CMS Data Federation hosts the compact MiniAOD with appropriate level of replication for the popular data and a fraction of the AOD data. The estimated miniAOD size of the data federation for 300-fb^{-1} luminosity data sample is: (50kB)(1kHz)*(10^7s/year)*(6y)=3*10^9 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 AOD data tier is currently used by many analysis groups. It is anticipated there will remain significant usage for certain analysis which requires detailed understanding of the detector. Complicated final states with missing transverse energy, multiple low P_T jets, τ final states, etc. The AOD is 410 kB per event, and using 30% as the fraction of data that needs to be hosted on disk in this format we get 10PB storage.

The tape archived raw data size for 300 fb^{-1} is 60PB. It is assumed that the reconstructed FEVT objects are only available in the intermediate stages of processing and not archived. However, it is anticipated that a fraction of RAW data, at about 10 PB is expected to be stored on disk for special reconstruction techniques, e.g., those with heavy slow particles, and for cache.

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.

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.

1.4.2 Future Plans

1.5 Software and Support (WBS 5)

All things software. There is analysis tools development at Princeton, multi-threaded framework work at Cornell, and a lot of event display stuff at UCSD. I am trying to get some of the latter reclassified as operations of the event display, so it might move.

1.5.1 Current Status

1.5.2 Future Plans

1.6 Technologies and Upgrade R&D (WBS 6)

This is a huge grab-bag of R&D, some of which is poorly defined, but we can make up a good story. There is a pile of stuff at UCSD that Frank and Avi want to classify as cost containment work new computing architectures, re-engineering of reconstruction for use on HPC machines, impact of reduced data formats like miniAOD and beyond. At Princeton Pete has efforts in reconstruction on new architectures and the like. Brian has some Brian things. Cornell also has small efforts on new architectures and data mining. And there are some things at Caltech I am either trying to get better defined or moved elsewhere.

1.6.1 Current Status

1.6.2 Future Plans

1.7 Coordination with CMS (WBS 7)

This is just where we file the effort that is working on CMS-wide coordination rather than US CMS tasks. On the NSF side this is S&C coordination at Princeton, reconstruction coordination at Wisconsin and UCSD, user support at UCSD, and a small amount of money that goes to Harvey (let us not talk about it).

1.7.1 Current Status

1.7.2 Future Plans