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

December 7, 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 the 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 the 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.

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. 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 the 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. As the beam energy has reached more or less its maximum expected, it is highly likely that from here on out analyses will tend to use all the data accumulated over time, from the 3 fb^{-1} at 13 TeV accumulated in 2015 to the 300 fb^{-1} expected by the end of Run 3 in 2025. While the analysis and Monte Carlo production computing needs scale roughly linearly with integrated luminosity, the reconstruction time per event for both data and simulation grows roughly exponentially with instantaneous luminosity. As a result, the overall computing needs outpace expected technological advances and reasonable funding scenarios by a very large factor, requiring significant innovations in order to guarantee that the physics potential of the data taken is fully realized.

Computing environment and facilities for the CMS experiment are continually evolving to meet the requirements of the collaboration and to take advantage of the evolution of technology within and beyond the high-energy physics community. The Moore's law scaling of computing capabilities and evolution of storage have slowed down from x2 gains every 15-18 months 10 years ago to a modest x2 gain every 4-7 years expected in the future. Moore's law alone is thus quite unlikely to meet the CMS computing challenge under constant budgetary levels. Innovations in resource utilization, adaptation to modern computing architectures, and improved workflows, will need to make 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.

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.

Primary goal of our program is to empower physicists at all 48 US-CMS member institutions to conveniently analyze CMS data. For the next 5 years, we propose to capitalize on NSF investment in networking at US universities as well as developments from other NSF projects [AAA, gWMS, OSG, PRP] to integrate much more tightly resources at US-CMS member institutions beyond Tier-1 and Tier-2 centers into the centrally operated services infrastructure of the CMS experiment. In this context effort funded via the Tier-2 program will become responsible to maintain infrastructure in the Science DMZs of US-CMS member institutions jointly with IT professionals at those institutions. The effort funded via this proposal will provide consultation to campus IT organizations and ultimately maintain services on hardware inside the various Science DMZs, in order to support the desired integration of campus IT with CMS IT.

This proposal focusses on the NSF supported University based computing, especially for the most diverse physicist-driven scientific data analysis activities. 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 discussed.

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**. Resources available at these centers funded through prior NSF support are summarized in Tables 1 and 2.

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, OSG and 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, and provide prompt feedback for operations.
  - Faculty collaborations at the University level can bring in additional campus or cloud resources

  - Opportunistic computing resources at the Universities amount to 37%.

- Cost of infrastructure is subsidized at the Universities.
- Cost of personnel is also lower.
- Friendly competition amongst the sites results in increased productivity.

CMS computing workflows fall under few broad categories, namely, prompt calibration and reconstruction, which is primarily a Tier-0 functionality, centrally scheduled reconstruction of LHC data and Monte Carlo, which can be distributed world-wide at all tiers, centrally scheduled production of simulated data, and chaotic user analysis, which is primarily done at Tier-2s and any opportunistically available resources.

CMS data is organized in several tiers ranging from RAW data acquired from the detector or simulated, to RECO format for reconstructed data, FEVT combining the two, full set of analysis objects (AOD) and compressed AOD, i.e., miniAOD. Ubiquitous access to AOD and miniAOD for the analysts is the key enabler for prompt production of physics results.

1.2.1 Current Status

The seven US Tier-2s rank amongst the top ten providers of the 50 such CMS centers world-wide. Together they provide about 35% of CMS Tier-2 resources, outlined in Tables 1 and 2. The compute resources at Tier-2s serve both production and physicist analysis cases. The resource utilization at the US Tier-2s in the past month is summarized in the Figure 1.2.1. The top and bottom panels show the counts of the successfully processed production and analysis jobs respectively, which add up to about 30,000 jobs in steady state. **fkw: Do you really want these plots in here? How do they help the proposal get funded?** These centers together are hosting 10 PB of CMS centrally and user produced data on their storage systems as shown in Figure 1.2.1.

The US CMS Tier-2s not only maintain resources, but also provide many additional services. **fkw changed this - need to decide on XX:** XX FTE across more than one person at each center are necessary to provide high-quality service that results in very high availability, upwards of 95%. The personnel are responsible for all aspects of provisioning these resources, from specifications through deployment to operations, taking advantage of local considerations.

In addition to the facilities maintenance and operation, the two people funded at each Tier-2 also take on other roles within the larger US-CMS software and computing project. CMS benefits because of their innovations and pioneering deployments, such as the most recent work in testing and commissioning of the world-wide CMS data federation using AAA technologies.

1.2.2 Future Plans

Computing Resources

To define plans for the future, we start with an extrapolation of needs from the present.

For the sake of simplicity, CPU requirements are estimated in units of number of batch slots needed based on the following assumptions:

- 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.

Tier 2 Center	Number of Job Slots		
	Purchased	**Opportunistic**	**Total**
Caltech	5,780	384	6,164
Florida	4,126	6,068	10,194
MIT	5,200	2,056	7,256
Nebraska	5,840	3,717	9,557
Purdue	6,636	9,581	16,217
UCSD	5,256	SDSC	5,256
Wisconsin	7,860	2,713	10,573
Total	**40,698**	**24,502 +**	**65,200**

Table 1: **fkw: should this table be updated to the end of FY2015 purchased infrastructure?** Useable batch slots currently deployed at US Tier2 centers. The San Diego Supercomputer Center has in the past provided access to resources via the NSF XRAC allocation process, and is committed to in addition provide spare capacity on an opportunistic basis in the future.

Tier 2 Center	Storage (TB)
Caltech	TBD
Florida	TBD
MIT	TBD
Nebraska	TBD
Purdue	TBD
UCSD	TBD
Wisconsin	2300
Total	**TBD**

Table 2: Useable storage space currently deployed at US Tier2 centers.

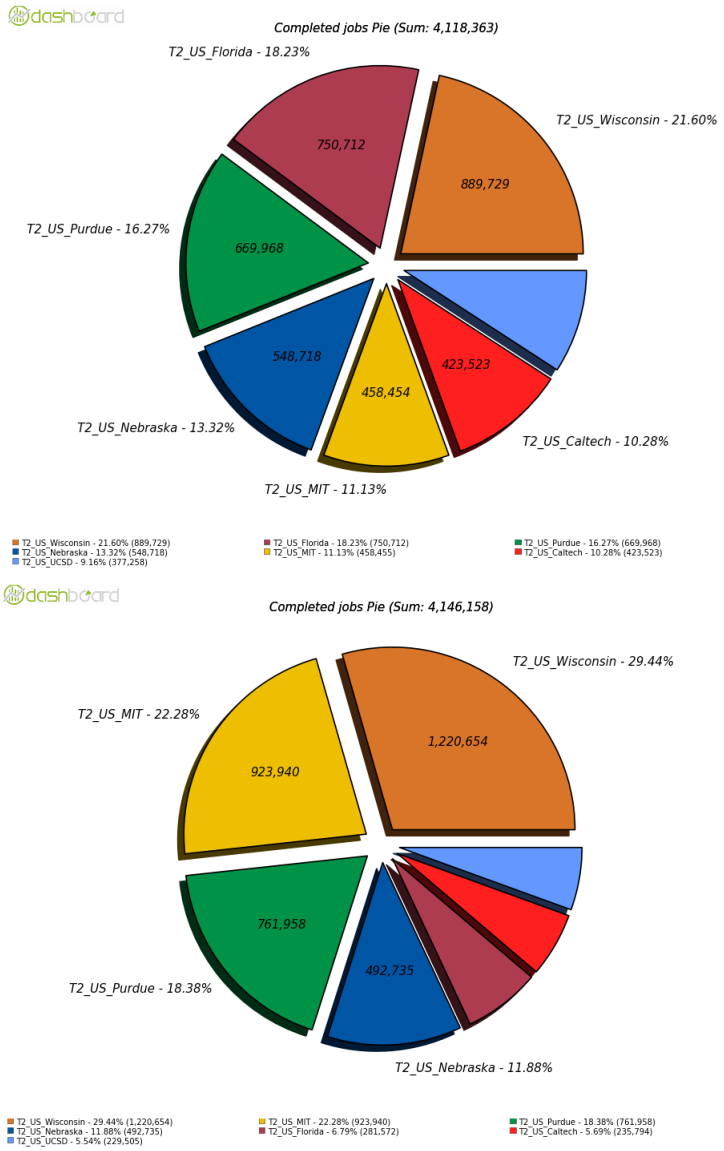

Figure 1: Counts of completed production (top) and analysis (bottom) jobs at the US Tier-2s in the past month (November 2015).

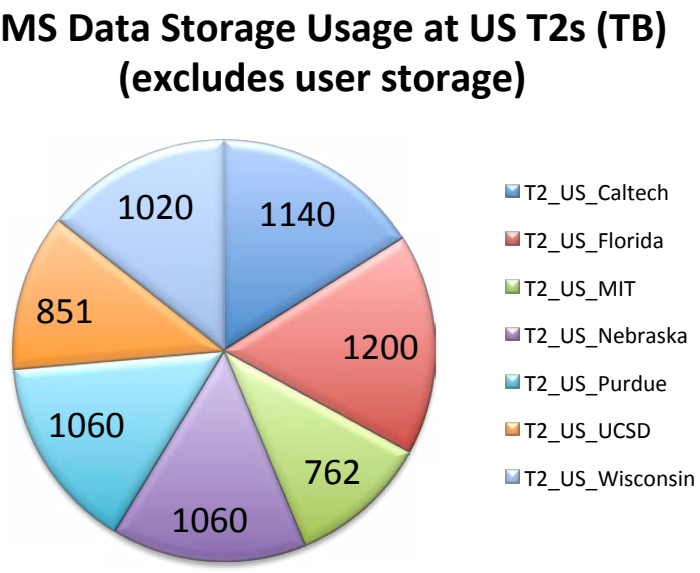

Figure 2: Current usage of storage resources for centrally managed data at the US Tier-2s in the past month (November 2015) in TBs. Physicist data storage at each site at the level of about 500 TB is additional. **fkw: What does this add to the previous table? How does this help the proposal get funded?**

- 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. While we unfortunately collected only 2 fb^{-1}, we nevertheless consider 10 fb^{-1} the appropriate integrated luminosity baseline for the needs scaling into the future.
- 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.

Storage Resources

During LS1, CMS introduced a refined analysis format called "MiniAOD". This format is 1/10th the size of the AOD, and typically requires somewhere between 1/2 to 1/5th the CPU power to analyze. These reductions are accomplished by storing only more refined information, requiring remaking of the MiniAOD in response to improved calibrations, or significant improvements of physics objects. During Fall of 2015 CMS went through two iterations of MiniAOD, with at least one more to follow before Moriond 2016 conference results. MiniAOD is expected to satisfy the needs of at least 90 % of all physics analyses. I.e. it is envisioned that a few analyses will require more detailed information not present in the MiniAOD, and will thus need to access the AOD or maybe even the RAW format.

The operational model today is that Analysis groups process these "primary data" to produce custom Ntuples for their analyses at a event processing rate of 1-10Hz or so. The custom Ntuples

are then typically analyzed at x100 larger event rates. The transformation from primary data useful to the entire collaboration to custom data useful only to a handful analyses is thus dramatically accelerating science and reducing computing costs with the drawback that non-negligible amounts of disk space at Tier-2s need to be provisioned to host that custom data. US-CMS organizes this by assigning a fixed set of University groups to each Tier-2 as their host Tier-2 for these custom data samples.

The estimated miniAOD size for 300-fb^{-1} of integrated luminosity is: (50kB)(1kHz) $\times$ (10^7s/year) $\times$ (6y) =3×10^9 MB=3PB. Typically to analyze this data requires an associated MC sample which is three times the size of the detector data, leading to a $\sim$ 12PB disk storage requirement for the CMS data federation for hosting one replica of one version of all MiniAOD datasets. Since the MiniAOD is highly processed it is expected that improvements to the reconstruction, calibration and other changes will necessitate remaking of MiniAODs. As 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 an expected disk space requirement of 20-30 PB for MiniAOD for 300-fb^{-1} of integrated luminosity.

In addition to the MiniAOD, we expect to require disk space to accommodate $\sim$ 10% of the data in AOD format on disk to allow the $\leq$ 10% of analyses that require it. For an AOD event size of $\sim$ 500 kB this amounts to $\sim$ 10PB of storage for 300-fb^{-1} of integrated luminosity, or 50 PB total for MiniAOD and AOD combined, allowing some modest level of replication across sites within the US.

For the custom data we project a need of 30PB across the US-CMS Tier-2s based on our Run1 experience. This data is not replicated, nor is it assumed to be backed up to tape. Implicit in this operational model is the assumption that the data can be reproduced easily enough if lost due to disk failures at a Tier-2.

In addition to data analysis activities of the Run2 and Run3 data, we also must budget for disk space for the CMS upgrade activities. We expect event sizes to be larger, and a more heavy use of AOD as upgrade activities will be focused around custom simulations and detailed reconstruction development. We budget 20 PB for the purpose of the present needs assessment.

Finally, there is a need for some RAW data on disk, plus staging space to in front of the Tier-1 tape archive, as well as at the Tier-2s for staging data as it is being reconstructed or simulations as they are being produced. Based on our present experience, we assume 10PB to be sufficient to accommodate this.

Adding it all up, we arrive at a total disk space need of 110PB at US-CMS facilities for 300 fb^{-1} of integrated luminosity, and we are proposing to locate 12PB of this at each of the seven US Tier-2s. There is no tape storage at Tier-2 centers today, and we expect to keep it that way. Tape needs are thus out of scope for the present proposal and thus not discussed here.

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 expect multi-100 Gbps connections in the coming years. Up to now, networking at Tier-2 centers has always been funded via sources outside the US-CMS software and computing project. We expect this to stay that way, and are thus not budgeting any costs for networking as part of this proposal.

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

In the past, resources beyond the traditional Tier-1 and Tier-2 sites were generically lumped into the category of Tier-3. The prevailing model of these resources was that they were structured more or less as small Tier-2 sites operated independently of the Tier-2 program by dedicated local administrators. With the rise of significant computing capabilities across U.S. university campuses, and in particular, driving by substantial NSF-ACI investment into networking infrastructure across more than 100 campuses nation wide, this model is changing. The Tier-3 site now functions more as a portal. It provides a local portal for university researchers to access larger-scale U.S. CMS computing resources. It also provides a portal for the CMS central computing infrastructure to access campus computing resources. We propose to strengthen and expand this new model for non-traditional computing resources by seamlessly integrating CMS and campus IT infrastructures in a way that minimizes administrative effort while maximizing flexibility. This approach is based on the NSF-funded "Pacific Research Platform" and the CMS Connect effort which utilizes the OSG's CI-Connect platform.

The central hub of our proposal is a single node that will be deployed at each participating institution. This node should be viewed as a "Tier-3 in a box," a single, self-contained appliance that when deployed into a campus Science DMZ will bridge the CMS and campus infrastructures, enabling local CMS researches to access both sets of resources via a single portal. This node will provide interactive data analysis, batch submission, CVMFS software cache, XRootD data cache, and XRootD an server to export local data. The HTCondor batch systems implemented on these nodes are all connected to the global CMS HTCondor pool via glideinWMS. Similarily any University computing resources are integrated requiring nothing more than ssh access to a US-CMS account on the local University cluster. Local CMS university groups will thus be 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 by the node as needed. Private data produced by the local university group is served out to the Tier-1 and Tier-2 system via the XRootD server integrated into the node. Each Tier-2 will also have an XRootD cache in order to transparently cache the private data of any of the local university groups to avoid IO latencies.

Deployment and maintenance of these nodes will be undertaken as a partnership between local campus IT and U.S. CMS Tier-2 personnel, following the model of the PRP. The PRP deploys single nodes into the Science DMZs 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. In this model, local campus IT is responsible for the maintaining the hardware and local user accounts, while OS and software service (including necessary OSG and CMSSW element) maintenance is undertaken by a collaboration of U.S. CMS Tier-2 personnel and local campus IT effort. 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.

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 $\sim$\$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 collaborating institutions within US-CMS. We thus augment it with an additional hosted service—CMS Connect—built on the OSG-connect/CI-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. University groups will generally be better served by the more flexible and customized Tier-3 in a box approach, but with the combination of the two approaches ensures that all groups will be served.

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)

In addition to operating the Tier-2 facilities, personnel supported by this project contribute to the operations of the distributed computing system of the CMS experiment. The tasks performed by these staff members support the efficient processing of data and successful execution of both production and analysis computing jobs.

1.3.1 Current Status

U.S. CMS personnel fill a variety of roles in CMS computing operations. Staff MIT support Tier-0 operations for the experiment, overseeing the day-to-day operation of the facility, which is of critical importance for the operation of the experiment. Other MIT personnel play leading roles in operating the experiment's data transfer system and providing support for the distributed grid infrastructure. UCSD maintains the CMS job submission infrastructure. Nebraska provides support for AAA operations and for network performance reliability. Johns Hopkins supports the operation of the Frontier system that provides run conditions and other configuration information for reconstruction and analysis jobs running on the distributed infrastructure. Florida takes responsibility for software distribution throughout the grid sites of the experiment.

1.3.2 Future Plans

All of these activities are expected to continue in the coming years, as they will always be necessary to the operation of the experiment. They will become even more critical to the success of CMS as the number of sites (including opportunistic sites) grows and highly distributed storage access over the WAN using AAA increases. Additional operations support for smooth operation of U.S. university portals (Tier-3-in-a-box) and efficient harnessing of opportunistic resources is also anticipated.

1.4 Computing Infrastructure and Services (WBS 4)

CMS as a global experiment depends on a variety of computing infrastructure services (CIS), several of which have been long term US-CMS commitments. In particular, US-CMS is traditionally leading in the areas of data management and centralized production.

To meet two of the three goals outlined in the introduction, increased efficiencies across Tier-1 and Tier-2 and integration of additional resources outside those tiers, requires CIS to become substantially more agile. During LS1, substantial improvements towards a more agile data management infrastructure were made by creating a global XRootd Data Federation, and development & deployment of a Dynamic Data Placement and Management (DDM). The NSF funded effort led both of these. Data management and workflow management development is done at Cornell, xrootd development at UCSD, and dynamic data placement development at MIT. Additional support was provided by NSF through the "Any Data, Anytime, Anywhere" (AAA) project to Nebraska, UCSD and Wisconsin.

In the following, we briefly summarize those recent achievements, and then describe necessary future work to be done within the present proposal.

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.

DDM uses PhEDEx to manage the actual transfers. It thus replaces decisions made by human operators during Run1 with algorithms based on dataset "popularity", i.e. use.

CMS Data Federation (AAA)

CMS Data Federation is built to provide seamless international-scale data access under the auspices of Any Data, Anytime, Anywhere (AAA) project. 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.

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 clusters controlled by non-CMS entities are easily integrated in the CMS computing environment. Temporary access to dedicated large resources can be purchased on commercial clouds or obtained from national or campus research facilities.

The main advantages of AAA vs DDM are that AAA fully supports partial file reads, a typical analysis job accesses less than 10% of the data in each file, and ease of integration of non-traditional resources. The drawback of AAA as deployed today is that data still needs to be placed first, and replication of datasets need to be controlled according to needs. The two systems AAA and DDM thus augment each other perfectly.

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.

During Run1 each Tier was used for only a small subset of all workflows. This led to inefficiencies and delays in processing due to inflexibility. During LS1, we expanded dramatically what workflows can be run where, making the overall system much more flexible. In addition, all resource usage,

distributed analysis and centralized production, is now scheduled via a single global HTCondor pool, allowing for relative prioritization of different activities.

Resources beyond the Tiered System

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. Campus clusters not purchased via US-CMS funds are accessible both through their OSG connection if existing generically The CMS High Level Trigger farm (HLT) has been integrated into offline computing operations via its OpenStack Cloud interface. HLT resources are now available for general processing anytime the DAQ is not running.

1.4.2 Future Plans

The PhEDEx data management system has served CMS extremely well for more than 10 years. However, it is ill equipped for the more agile needs of the future. E.g. it's internal mechanisms for source selection for a given transfer is much less agile than the bit torrent-like multi source client used in Xrootd. It's internal back-off mechanism for handling transfer failures leads to long transfer tails, and it is generally very difficult to impossible to fill modern large bandwidth network pipes using PhEDEx.

A significant redesign of PhEDEx is necessary. Such a redesign also provides an opportunity to consider discontinuing poorly supported protocols like SRM, as well as duplication of protocols like maybe replacing gridftp with Xrootd, as the latter is required anyway for agile operations.

Among the AAA toolset, a proxy-cache was developed that is not yet used in CMS. Initial tests indicate that the cache system performs exceptionally well. As we gain more experience with this technology, we may want to transition a sizable fraction of the US-CMS disk space at Tier-2s and Tier-3s into XRootd proxy caches to gain additional efficiencies.

The entire end-to-end centralized data production process still has far too many human effort intensive aspects. Significantly more automation is needed to make the overall system both more agile and more efficient. Today it is still not uncommon that some resources, especially in US-CMS, are oversubscribed while others, especially elsewhere in the world, remain unused. Enforcing dynamically changing processing priorities is also still very difficult due to multiple layers of queuing and workflow restrictions. Significant efficiencies are yet to be gained here.

Finally, commissioning of resources beyond the tiered system is still at the very beginning, and while having large potential for additional resources, will require significant effort still.

1.5 Software and Support (WBS 5)

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. Keeping up with technology evolution in the market requires continues investigation and CMS framework and utilities software development. Cornell, Princeton and UCSD groups are engaged with central CMS in this essential software development and support.

1.5.1 Current Status

A systematic effort to make the core CMSSW thread-safe and has successfully deployed it in the past year. The event display for CMS has been reworked to work on a variety of platforms conveniently.

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 400 kB size per event, which when scaled to 300 fb^{-1} results in unaffordable data volume. Further, rate of event processing matters in time to production of physics results, so size of event being small is also beneficial for computational loads.

The US CMS personnel supported by NSF played key role in development of the MiniAOD. 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. 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 replica counts will become small.

1.5.2 Future Plans

FIXME: We need to list what is required to be done by the people supported by this WBS.

1.6 Technologies and Upgrade R&D (WBS 6)

The main thrust of the R&D effort of the project is to control the rate of growth of computing required, and thus cost incured, and to retain flexibility with regard to possible future changes in computer architecture. There are fundamentally two largely independent effects that drive cost, those that scale with event complexity, i.e. average pile-up (PU) per event, and those that scale with integrated luminosity, or total data volume.

The cost of event reconstruction is driven by occupancy of the tracker. Higher instantaneous luminosity leads to higher pile-up thus higher occupancy and with it a near exponential growth in CPU time per event in the pattern recognition step of the track reconstruction. For example, an increase in the average number of PU events from 20-30 was measured to result in an increase of x3 of event reconstruction time in the current CMS software release. This range of PU matches the expected running conditions during 2015/16. To set the scale, the time to reconstruct a 13TeV top pair event exceeds the time to simulate the same event at an average PU $\sim$ 25 which we expect to reach early on in 2016. Any speed-ups of the reconstruction software, especially the tracking pattern recognition, thus directly translate into computing cost savings.

Analysis, MC production, and data reprocessing all scale roughly linearly with total integrated luminosity, or total data volume, leading to the x30 increase from the beginning of Run 2 in 2015 to the end of Run 3 in 2023 mention previously. The software and computing R&D program for the next 5 years is geared towards two timelines. It is meant to engage in fundamental R&D towards solving the challenges of scaling out computing for the High Luminosity LHC era (2025-2035) but also to provide near term improvements that can be put into production during LS2 (2019/20) in order to address the challenges of Run 3 (2021-23). Given limited resources in personnel, our

strategy is to focus on the long term with an eye towards adopting lessons learned in this process to address the Run 3 challenges.

fkw: the prose below is probably meant to go into the introduction in some form? High-Luminosity is the least pleasant way to go exploring. Unlike high(er) energy, one has to cope with increased event size (due to pile-up), pile-up complexity increases due to many overlapping events, data set size increases due to long running period required, impacting CPU, storage and network resources. For example, an increase in the average number of PU events from 20-30 was measured to result in increase of a x3 of event reconstruction time. Such size increases (or larger) are unaffordable and must be prevented.

1.6.1 Current Status

During LS1, US-CMS drove multiple developments all focused on overall cost reductions in computing. We led the algorithmic improvements and code optimizations in the pattern recognition software that reduced the reconstruction time per event by xY for an average PU $\sim$ 30. We instigated the introduction of MiniAOD, reducing the event size by x10 and the average analysis processing time per event by xYY **fkw: figure out what a defensible number of this is given the data we have from the Data Popularity service**. We prepared the core framework to be multithreaded blablabla — Sridhara to fill in text from David Lange here ... And we transitioned the computing infrastructure and services that CMS depends towards a suite of services that are much more agile.

US-CMS was the primary driver in all of these, and thus has a solid track record of innovation to be able to do more science given fixed hardware investments. We propose to continue being an innovation leader in global CMS.

1.6.2 Future Plans

Reconstruction Software: Cornell, Princeton, and UCSD are collaborating on an ambitious R&D program towards redesigning the core Kalman filter tracking algorithms of CMS for parallel architectures While the bulk of this R&D is funded to the tune of 3 FTE for 3 years via an independent NSF grant [cite the PIF], the present proposal includes a modest effort of XX focused on deriving short term benefits from the independently funded long term focused R&D agenda.

Deriving short term benefits is particularly interesting in light of the planned roll-out of large Supercomputers at both DOE and NSF based on the next two generations of Intel MIC processors. E.g. Cori Phase 2 at NERSC is expected to include 9,300 Intel Knights Landing processors by 2017. Aurora at ANL is expected to deploy 50,000 Intel Knights Hill processors in 2018. Similar plans exists in the NSF for the Stampede Supercomputer at TACC. The degree to which CMS can benefit from these large scale resources for its core processing needs in Run2 and Run3 will depend crucially on successfully transitioning lessons learned from the externally funded long term R& D program into production. This transition is within the scope of the present proposal.

R&D towards a new data analysis model: For the HL-LHC era, CMS might need to contemplate a fundamental shift in the boundary between "primary data" and "custom data". Already in Run 1, the custom data Ntuples typically were analyzed at event rates ranging from 100 Hz to 10 kHz. Ntuple analysis is even today in many cases IO rather than CPU limited. In contrast, the production of these custom Ntuples is almost always CPU limited. Even for the MiniAOD of Run 2, typical event processing rates reach little more than a few Hz. The trade-off

at work here is flexibility versus speed. A data format for the entire CMS collaboration must be flexible in content for two reasons. First it needs to satisfy many types of data analyses, and second it must be "forward compatible", i.e. a MiniAOD produced today must still be useful a few months from now when the state of the art in physics object definition, jet energy calibrations, etc. etc. have changed to incorporate a variety of improvements. The R&D questions here include: Can we speed up MiniAOD to the kinds of event processing rates typical for custom Ntuples? If we can, what does it mean for the Tier-2 infrastructure to support IO limited jobs at large scale? E.g. do we need specialized batch system entry points for IO limited jobs for which disks need to be co-scheduled? Can we reuse some of the industry standard products for IO limited jobs, or is this impossible because we would loose the benefits of partial file reads in ROOT IO?

1.7 Coordination with CMS (WBS 7)

US CMS S&C personnel are well inetgrated in the CMS-wide coordination efforts and hold management positions.

1.7.1 Current Status

Current support under this category includes S&C coordination at Princeton, reconstruction coordination at Wisconsin and UCSD, and user support at UCSD.

1.7.2 Future Plans

It is anticipated that approximately a third of the management positions in CMS are held by US personnel, of which NSF computing supported personnel needs will have to be covered by this project. The need is likely to remain approximately constant.