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 18, 2015

1 Software and Computing

1.1 Introduction

The NSF support for software and computing at U.S. 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 the possible discovery of new physics. In this section we briefly describe the current status and future plans of the U.S. CMS software and computing project, focusing on its Tier-2 program, on which U.S. and international CMS physicists rely 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 technical design report. The discovery of Higgs at low mass and continued investigation of electroweak scale physics requires low trigger thresholds. During the recently started new phase of data acquisition, 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 arises from the increase in integrated luminosity, a factor of three comes from the increased trigger output rate to facilitate continued access to electroweak scale physics, and a final factor of three or so increase is due to greater event-complexity from increased energy and instantaneous luminosity leading to event pileup. As the beam energy has reached more or less the 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. Unfortunately, Moore's law scaling of computing capabilities and evolution of storage have slowed down from $\times 2$ gains every 15-18 months ten years ago to a modest $\times 2$ gain every 4-7 years expected in the future. As a result, the overall computing needs outpace expected technological advances and reasonable funding scenarios significantly by the end of Run 3, with even larger shortfalls projected for the HL-LHC era.

Our vision for meeting this challenge is threefold. First, we will gain efficiencies by being more agile in the way we use the traditional Fermilab-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 U.S. CMS universities, DOE and NSF supercomputing centers, and commercial cloud providers as much as possible. In this context, effort funded via the Tier-2 program will become responsible to maintain infrastructure in the Science DMZs of U.S. 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. Finally, 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 contain the growth in computing needs.

The primary goal of our program is to empower physicists at all 48 U.S. CMS member institutions to conveniently analyze CMS data. This proposal thus focuses on the NSF supported university-based computing, especially for the most diverse physicist-driven scientific data analysis

activities, as they will require the bulk of the S&C resources requested in this proposal. Other supported activities are also discussed, including a brief look at the computing R&D necessary for the HL-LHC phase (2025+), during which another order of magnitude in data volume (3000 fb^{0-1}) is expected.

1.1.1 CMS computing model details

The tiered computing model of the LHC experiments, based on a distributed infrastructure of regional centers outlined by the MONARC project **ref**, includes a Tier-0 center at CERN, seven Tier-1 centers (including one at Fermilab) and about fifty Tier-2 centers (including eight in the U.S., seven of which will be supported by this award). [1] The original model envisioned that only a very specific set of tasks would be performed at each tier of the infrastructure, but in recent years, much more flexibility has been introduced into the system. The technical changes leading to this have largely been pioneered within the U.S., in many cases by university-based personnel supported by the previous NSF award for U.S. CMS operations, especially those at the Tier-2 centers. Some of the innovations have included commissioning the global data transfers matrix across all Tier-1 and Tier-2 sites, initiating the centralized Monte Carlo production, leading the commissioning of the grid worldwide for CMS, introducing the concepts of late-binding WMS, CMS data federation, dynamic data placement, MiniAOD, and global queue across analysis and production processing.

CMS uses a variety of data formats tailored for different purposes. Ongoing Run 1 analyses are based on the Run 1 AOD format, comprising 220 kB per event on average. Average AOD event sizes in Run 2 are roughly 500 kB. The growth is partly due to increased pileup, and partly by design, to allow re-running particle-flow reconstruction in its entirety from the AOD. In addition, the high level trigger output rate was increased by a factor of three between Run 1 and Run 2. To contain costs and speed up physics analysis, U.S. CMS initiated the introduction of a refined analysis format called "MiniAOD". Implemented in collaboration with CERN, this format is 1/10 the size of the AOD, and typically requires somewhere between 1/2 to 1/5 the CPU power to analyze. 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 transformation from "MiniAOD" to custom data for further user analysis is necessary but has the drawback that non-negligible amounts of disk space at Tier-2s also needs to be provisioned to host that custom data.

1.2 University Facilities (WBS 2)

1.2.1 Current Status

The primary university-based facilities are the Tier-2 centers at Caltech, Florida, MIT, Nebraska, Purdue, UC San Diego and Wisconsin. Resources available at these centers funded through prior NSF support are summarized in Table 1. Faculty collaborations at the university level have brought access to significant opportunistic resources on their campuses and on the cloud (∼37% of the total available to CMS). Often cost of infrastructure at the university Tier-2s is subsidized and the cost of personnel is lower than at DOE labs. Further, friendly competition amongst the sites results in increased productivity of U.S. Tier-2s overall. Finally, the US CMS Tier-2 institutions also provide

[1] The Fermilab facilities are covered by S&C WBS 1; they receive no NSF support and are not discussed further in the proposal. The eighth U.S. Tier-2 center, at Vanderbilt, is supported by the DOE Nuclear Physics program.

Number of Job Slots		Storage (PB)	
Available as of 11/2015			
Purchased	40,698	Purchased	14
Opportunistic	24,502	Opportunistic	-
Total	65,200	Total	14
Average Usage (11/2015)			
Production	15,000	Hosted	10
Analysis	15,000	Local	4
Total	30,000	Total	14

Table 1: Currently available and used number of job slots and storage totals for all the US Tier-2 sites.

connections to strong physics groups at those universities as well, resulting in prompt feedback for operations with their intense analysis activities.

The seven U.S. Tier-2s rank amongst the top ten providers of the 50 such CMS centers worldwide. Together they provide about 35% of CMS Tier-2 resources. 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 adds up to about 30,000 jobs in steady state split equally between production and analysis workflows. These centers together are hosting 10 PB of centrally and 4 PB of user produced data on their storage systems.

Not only does the U.S. CMS Tier-2 program provide the premier Tier-2 centers in all of CMS, the staff at these centers, working in partnerships with their universities, also provide the intellectual leadership to make radical changes such as those mentioned in Section 1.1.1. With the current proposal we continue this transition, focusing on maintaining leadership in operations of a production infrastructure, and leadership in engineering changes to afford doing ever more demanding physics with fixed hardware budgets. Thus, the investment in personnel at the universities sites brings at least as much value as the investment in the computing equipment.

1.2.2 Future Plans

Tier-2 Personnel: We continue to request in this proposal 2 FTE at each of the seven Tier-2 institutions. On average, 1.6 FTE of these are necessary at each center 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. The remaining 7×0.4 FTE of effort take on other roles within the larger U.S. CMS software and computing project as described in the later sections of the proposal. CMS benefits from this close connection because of innovations and pioneering deployments initiated at U.S. Tier-2s, such as the most recent work in testing and commissioning of the world-wide CMS data federation using AAA technologies.

Tier-2 Resources: We use a simple model that scales the present usage of job slot count and storage to future years based the expected LHC luminosity plan (see Table 2). The CPU requirements are estimated in units of number of batch slots needed and the storage is determined by the amount of data anticipated to be accumulated and equal amount of MC simulated. In addition to

the MiniAOD, we expect to require disk space to accommodate ∼10% of the data in AOD format on disk to allow the ≤ 10% of analyses that require such detail.

In Table 2, we start by showing the actual numbers for 2015 scale of operations. We have accumulated about 3 fb^{-1} of data from LHC runs and produced 10 fb^{-1} worth simulated data. Including the data simulated and accumulated in Run 1, we are running 30,000 production jobs on 1 pb of data hosted at U.S. Tier-2s currently, equally split between production and analysis jobs. We show the current average availability of job slots and storage at the U.S. Tier-2s to start the scaling. We then extrapolate to out years using the LHC run plan, which sets the scale for integrated luminosity.

We assume that the production job slot usage scales as a function of incremental luminosity expected that year. The analysis job slot usage scales as a function of the accumulated by that time, because the analysts will be combining all of the 13 TeV data in Run 2 and Run 3. The MiniAOD format data is expected to be used by most analysts reducing the needed storage volume. We multiply this expected MiniAOD volume by 1.5 to account for the need to support CMS upgrade activities, custom physics Ntuples, staging space for processing and reprocessing, and to allow for some replication of popular datasets to improve access and redundancy. To arrive at the total disk space across the 7 Tier-2's in the US, we further assume that 10% of the AOD is also on disk.

In this simplistic projection, we have not accounted for increased reconstruction times per event as a result of increased complexity of the events due to increased pile-up at higher instantaneous luminosity, nor the corresponding increase in event sizes. Nor have we made any assumptions about software performance improvements, or other increased efficiencies, are taken into account decommissioning of old hardware. The model is thus admittedly subject to some uncertainties, though probably no more so than the expected luminosity profile of the LHC, which we take as an input.

To meet the needs as calculated with this model requires an average annual hardware investment per Tier-2 between $300k and $500k depending on whether we assume a 20% or 10% cost reduction per year due to Moore's law. The hardware budget in the present proposal matches these needs for the lower end of this range.

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 U.S. 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.

In the remainder of the proposal, we will, among other items, describe R&D investments towards cost savings that are essential for the HL-LHC. If we successfully transition some of this R&D into production already for Run 3, then this will allow us to tolerate the slower Moore's law curve of only 10% per year cost reductions that is presently deemed to be the more likely in projections at CERN.

Non-traditional resources: 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. In most places, these Tier-3's were either poorly, or not at all integrated within the larger context of campus IT infrastructures due to the idiosyncrasies and lack of agility of the technologies available to CMS.

Year	Luminosity (fb^{-1})		Total for US Tier-2s				Per Tier-2	
			Job Slots		Storage (pb)			
	Incr.	Cumul.	Prod.	Ana.	AOD	MiniAOD	Job Slots	Storage (pb)
2015	3	3	15000	15000	3.8	13	5814	2.0
2016	37	40	18000	20000	9	31	5429	5.7
2017	40	80	20000	40000	18	62	8571	11
2018	40	120	20000	60000	27	92	11429	17
2019	0	120	10000	60000	27	92	10000	17
2020	0	120	10000	60000	27	92	10000	17
2021	60	180	30000	90000	40	138	17143	25
2022	60	240	30000	120000	54	184	21429	34
2023	60	300	30000	150000	67	230	25714	42

Table 2: Projection of resources by the year from actual usage in 2015 for out years through 2023, based on LHC luminosity expectation is shown.

As these technologies have advanced, and NSF-ACI has made substantial investment into networking infrastructure across more than 100 campuses nationwide, we are proposing here a significantly more integrated Tier-3 program.

The Tier-3 of the future functions more as a portal that integrates campus IT infrastructure with global CMS infrastructure seamlessly. It provides a local portal for university researchers to access their campus IT resources as well as larger-scale U.S. CMS computing resources. It also provides a portal for the CMS central computing infrastructure to access campus computing resources to the extend allowed by local policies. This is accomplished, following approaches being pioneered by the NSF-funded "Pacific Research Platform" and the CMS Connect effort which utilizes the OSG's CI-Connect platform, to minimize administrative effort while maximizing flexibility.

The central hub of our proposal is a "Tier-3 in a box" that will be deployed at each participating institution. This node is a single, self-contained appliance that when deployed into a campus Science DMZ will bridge the CMS and campus infrastructures. This node will provide interactive data analysis, batch submission, CVMFS software cache, XRootD data cache, and XRootD server to export local data. The HTCondor batch systems implemented on these nodes are all connected to the global CMS HTCondor pool via glideinWMS. Similarly any University computing resources are integrated requiring nothing more than ssh access to a U.S. 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 to increase processing efficiency. 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. Any data, privately or centrally produced, will thus be available anytime anywhere.

We are proposing to scale this "Tier-3 in a box" model to serve as many U.S. CMS institutions as possible but focussing initially on 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 University Facilities portion of this proposal. The entire system of Tier-3 services across all institutions will thus be centrally managed by Tier-2 personnel such that

local IT effort at the 25+ institutions is restricted to managing local accounts, initial hardware deployment, and hardware replacement upon failures. 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 five years of this proposal. We fully understand that the above model will not be appropriate for all collaborating institutions within U.S. 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. Development work on CMS Connect is led by Notre Dame.

Finally, in the latter years of this proposal, we will fully integrate cloud services access into this infrastructure in such a 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. 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. We expect to be collaborating on this functionality with the HEPCloud project at Fermilab as well as the Open Science Grid (OSG).

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. MIT staff support Tier-0 operations for the experiment, overseeing the day-to-day operation of the facility, which is of critical importance. 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 via the CVMFS caching system.

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 and services (CIS), several of which have been long-term U.S. CMS commitments. In particular, U.S. CMS has traditionally been a leader in the areas of data management and centralized production workflow management.

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 and deployment of a dynamic data placement and management system (DDM). Both efforts were led by NSF-funded personnel. Data management and workflow management development is done at Cornell, XRootD development at UCSD, and DDM development at MIT. Additional support was provided by NSF through the "Any Data, Anytime, Anywhere" (AAA) project to Nebraska, UCSD and Wisconsin.

1.4.1 Current Status

Dynamic Data Placement and Management (DDM): The PhEDEx data distribution system, which underpins the CMS computing infrastructure, was largely manually operated during Run 1. Operators moved large chunks of data on command across the grid as necessary to enable subsequent processing. This labor intensive and lead to inefficient use of disk resources. The DDM software was developed and commissioned during LS1 to address these shortcomings. DDM is now deployed at all tiers to automatically place data where it is meant to be processed at the time it needs to be there, and then prune the unused but archived data using well-defined policies. DDM uses PhEDEx to manage the actual transfers. It thus replaces decisions made by human operators during Run 1 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 the AAA project. AAA removes the requirement of co-location of storage and processing resources through an infrastructure that is transparent to users and highly reliable. It enables greater access to the data, in fact, **A**ny data can be accessed **A**nytime from **A**nywhere (AAA) with an internet connection. 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 CMS data federation is now fully deployed across all tiers of the global computing infrastructure.

Improvements to CMS Workflows: The main objectives of the workflow management middleware is to process data as quickly as possible, maintain uniform load across all resources and enable fast recovery in case of service interruptions. During Run 1 each tier was used for only a small subset of all workflows. During LS1, we expanded dramatically what workflows can be run where, making the overall system much more flexible, thus decreasing processing delays due to inefficiencies. 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. At this point, CMS is in an exploratory phase for smoothly integrating resources beyond the tiered system for production and routine use. Such resources include the CMS High Level Trigger Farm (HLT) whenever the DAQ is not running, DOE and NSF HPC facilities like NERSC and SDSC, commercial cloud resources such as AWS, and campus IT resources across the nation.

1.4.2 Future Plans

The PhEDEx data management system has served CMS extremely well for more than ten years. However, it is ill-equipped for the more agile needs of the future. Its internal mechanisms for source selection for a given transfer is much less agile than the bit-torrent-like multi-source client used in XRootD. Its internal back-off mechanism for handling transfer failures leads to long transfer tails, and it is generally very difficult if not impossible to fill modern large-bandwidth network pipes using PhEDEx. A re-design of PhEDEx to address these issues also provides an opportunity to consider discontinuing poorly supported protocols like SRM, as well as duplicative protocols such as gridftp when XRootD is required for agile operations. The MIT group will take on this task.

The AAA toolset contains a proxy-cache 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 U.S. CMS disk space at Tier-2s and Tier-3s into XRootD proxy caches to gain additional efficiencies. Development work to this end will be pursued at UCSD.

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. Efforts on workflow management systems will continue at Cornell and Purdue.

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. Wisconsin will work on these issues.

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, require adaptation of our software to suitable multithreaded frameworks. Keeping up with technology evolution in the market requires continuous investigation and CMS framework and utilities software development. The Cornell, Princeton, Florida and UCSD groups work with international CMS on this essential software development and support. have focused on a number of well contained projects.

1.5.1 Current Status

Development of Multi-threaded CMSSW Applications: A systematic effort to make the core CMSSW, and the reconstruction application thread-safe has been successfully completed and deployed in the past year. Both the HLT and Tier0 were able to use this work to great advantage. In addition to the development of the framework itself, deployment of applications requires making the algorithmic physics code "thread friendly", meaning adjusting it to follow the rules of the multi-threaded framework, such that the entire application can safely run with multiple concurrent threads. This past year developers at Cornell have been involved in this work. **Support and Maintenance of Fireworks:** The event display for CMS has been reworked to adapt to the DAQ system of Run 2 and to work on a variety of platforms conveniently. **Support and Maintenance Build Systems:** In order to take advantage of developments on alternative architectures as outlined in the R&D section there must be support in the build tools for cross compilation and/or ports to these new architectures. This past year Princeton personnel have had major success in

porting both CMSSW and the OSG stack to ARM64 and Power8. **Development of MiniAOD:** The development of the MiniAOD format described in Section 1.1.1. The MiniAOD is now the main data format that has been used for early results on the 13TeV data from CMS.

1.5.2 Future Plans

Development of Multi-threaded CMSSW Applications: Future steps in this area are the extension of the number of other CMSSW applications that can be run multi-threaded. In 2016 we are still on target to make the digitization thread friendly. This work is being carried out at Cornell. As CMSSW evolves, constant attention must be given to keep the code thread friendly. Personnel at Florida will be helping with this task. **Support and Maintenance of Fireworks:** As Fireworks is the main event display for CMS, new feature requests, and updates for new operating systems are needed. Personnel at UCSD will continue to provide this support. **Support and Maintenance Build Systems:** As innovations in new architectures and techniques from the R&D area become beneficial to the operating program, it is important to have personnel that can make that transition happen. Princeton personnel have demonstrated success in this area, and will continue to provide this support.

1.6 Technologies and Upgrade R&D (WBS 6)

The R&D effort within the project aims to control the rate of growth of computing required, and thus cost incurred, and to retain flexibility for future changes in computer architectures and software technologies. This relatively small WBS area aims to leverage for CMS developments from elsewhere in HEP and beyond.

Two largely independent effects drive cost: those that scale with event complexity, or average pileup (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 pileup and 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 to 30 was measured to result in a three-fold increase of event reconstruction time in the current CMS software release. This range of PU matches the expected running conditions during Run 2. To set the scale, the time to reconstruct a 13 TeV $t\bar{t}$ event exceeds the time to simulate the same event at an average PU $\sim$ 25, which we expect to reach in 2016/17. 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 $\times$30 increase from the beginning of Run 2 in 2015 to the end of Run 3 in 2023 mentioned previously. The software and computing R&D program for the next five 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 HL-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.

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 x2-5. 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 on towards a suite of services that are much more agile as described in Section 1.4.

1.6.2 Future Plans

Reconstruction Software: Cornell, Princeton, and UCSD are collaborating on an ambitious R&D program to redesign the core Kalman filter tracking algorithms of CMS for parallel architectures. While the bulk of the long term R&D is funded via an independent NSF collaborative research award (PHY-1521042,1520942,1520969), the present proposal includes effort to derive short-term benefits from the independently-funded long term focused R&D agenda. This 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. For example, Cori Phase 2 at NERSC plans for 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. Whether CMS can benefit from these large scale resources for its core processing needs in Run 2 and Run 3 depends crucially on successfully transitioning lessons learned from the 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 must 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 I/O 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. There is a trade-off between flexibility and 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 have changed to incorporate 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 I/O limited jobs at large scale? Do we need to schedule disks for I/O limited jobs? Can we reuse some of the industry standard "Big Data" products, or is this impossible because we would loose the benefits of partial file reads in ROOT I/O? CMS will also benefit from the separately funded NSF SI2 DIANA/HEP (Data Intensive ANAlysis for High Energy Physics) project (ACI-1450310, 1450319, 1450323, 1450377) involving PIs working on CMS, Atlas and LHCb. DIANA/HEP will develop state-of-the-art software tools for analysis and improve interoperability of HEP tools with the larger scientific software ecosystem. A modest effort has been included in this proposal to integrate those advances into CMS and its analysis model.

1.7 Coordination with CMS (WBS 7)

U.S. CMS S&C personnel are well integrated 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 software coordination at Wisconsin and UCSD, U.S. CMS Tier-2 program coordination at UCSD, and submission infrastructure coordination at UCSD.

1.7.2 Future Plans

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