U.S. CMS Tier 3 Computing

1 Overview

Tier 3 (T3) resources play an essential role in the success of the U.S. CMS physics program by providing critical CPU and storage to meet the needs for the final stages of analysis. T3 computing is used for ntuple production, histogram creation, fitting, limit setting, and other similar activities that are necessary to complete a physics analysis. These sorts of activities rely on a flexible and interactive computing environment, and ensuring sufficient resources are available to U.S. CMS institutions is essential for guaranteeing that U.S. CMS physicists can continue to push the envelope and play a leading role in CMS physics analysis. Furthermore, U.S. CMS institutions have demonstrated the ability leverage NSF and DOE investments in T3 sites by obtaining additional contributions of funding, hardware, and services from universities.

Unlike Tier 1 (T1) and Tier 2 (T2) sites, U.S. CMS T3 sites are not centrally funded or operated by the U.S. CMS program or international CMS. Instead, individual U.S. CMS institutions are responsible for pursuing funds and managing their sites. Nonetheless, CMS does provide important support in the form of central services, software, and expertise. Specific element provided by the international CMS computing effort include the software needed to analyze CMS data and to manage CMS jobs running on T3 sites, as well as services for data location and dataset transfer between sites. The U.S. CMS S&C program is a major contributor to all of these efforts. Furthermore, the U.S. CMS S&C program provides experts specifically to advise the administrators of U.S. T3 sites and assist in debugging problems as they arise.

Planning for U.S. CMS T3 support is undertaken as a partnership between the U.S. CMS S&C program leadership and the U.S. CMS collaboration, with the U.S. CMS computing liaison providing a bridge between those two groups. As part of the planning process two surveys have been administered to members of the U.S. CMS institutions. One survey focused on analysis software usage and needs, while the second survey focused on T3 resources. This report focuses on the results from the second survey, specifically how this information would impact planning for future T3 support.

2 Usage

One element of the survey was to ask participants about the ways in which they use T3 computing to do things that are not possible (or at least not practical) using T2 resources. The survey responses included a number of consistent themes:

- One of the leading responses involved using T3 resources for interactive access to data. In nearly every analysis, there comes a point in which the analyzers require interactive access. Students and postdoctoral researches must actually login to a computer with access to the data and manipulate it in real time. By design, T2 sites do not supply this functionality, so T3 sites are the only places where interactive access can be obtained. In some cases, the needs of interactive use (especially when graphically intensive) makes a central login location like the LPC less than optimal because of the delays incurred by network lag for those logging in from a distance.

- Another consistent response was the T3 sites allow options for running more complicated, flexible, or customized workflows than are possible at T2 sites. Again, because of the scale of the computing done at T2 sites, interaction with those resources is restricted to standard OSG-based GRID tools. In contrast, using the interactive login access provided at T3 sites, analysts can create workflows that make use of the native batch schedulers at the site, incorporate specialized or alternative workflow control tools, and use non-standard software not distributed to all CMS sites, like multivariate analysis (MVA), numerical, statistical, or Monte Carlo (MC) software packages.

- T3 resources are also clearly beneficial for providing fast turnaround on critical analysis steps, especially near the end of the analysis process when computationally intensive steps must be performed in sequence with the results of the previous step providing input to the next step. One example of such a case would be a complicated sequence of computationally intensive fits. Suppose that the final fit is sensitive to the initial starting conditions, and so it's necessary to perform one or more computationally intensive preliminary fits and propagate those results to the final fit. The natural latency involved in using T2 resources would make a sequence of fits run at T2 sites impractical.

- T3 resources have the ability of being partitioned more dynamically than T2 resources. Because of the large-scale nature of T2 computing, resources tend to be allocated either in large blocks at the physics analysis group (PAG) level, or in smaller groups to each individual user. At T3 sites, it is straightforward to allocate storage or computing priority to members working on a single analysis topic, possibly just during a short but critical time period (for example, in the final push to meet a conference or paper deadline). Because T3 sites tend to be relatively small in scale, such dynamic allocations are simpler to accomplish.

- Software development, especially for more advanced algorithms or computing techniques, often requires both interactive access as well as a moderately large scale of

CPU or storage resources, possibly configured in non-standard ways. One example of such a case would be working to develop a GPU-enabled parallel processing algorithm for faster track reconstruction. Evaluating the performance of such an algorithm would require specialized resources (i.e. access to a GPU cluster) as well as a sufficient scale to process large enough chunks of data to get an accurate picture of the performance over a range of conditions. T3 sites are uniquely positioned to meet the specialized needs of these developers.

- In many cases, T3 sites are organized around the goal of providing resources to support a particular analysis activity that requires special purpose MC or data samples that are not of broad interest to the CMS community.

- T3 sites can also be used for running jobs that have non-standard resource requirements, such as very large memory or CPU time. For example, developing simulations for very high pileup scenarios often requires running on machines with larger than typical memory.

- In many cases, T3 sites provide the only really practical opportunity for inexperienced students and postdoctoral researchers to gain training in the sort of advanced computing skills that will be necessary to build future generations of HEP computing. There are two reasons for this: T3 sites provide opportunities to work on smaller, more self-contained clusters giving a similar experience to working on a detector test stand before progressing to the fully-integrated CMS detector. In addition, using local T3 resources allows a more active involvement of local experts in the student's and postdoc's educational experiences.

- Likewise, a T3 site may be the only portal to enable undergraduate students to participate in CMS data analysis without going through the process of obtaining a full set of grid credentials.

- T3 sites can serve as a nexus for fostering collaboration at institutions beyond the local CMS group. For example, T3 computing can allow CMS groups to collaborate with theorists in particular on developing new and improved MC models. Furthermore, T3 computing can be used to boost collaboration between physics and computer science (CS) faculty at an institution in working on HEP-related computing problems.

Each of the items listed above is critical for the continued success of the U.S. CMS physics program. None of these activities would be practical (if even possible) to perform using T2 resources, at least as they are presently configured. It should be emphasized that this is not a shortcoming of the T2 resources. Rather, the configuration and capabilities of T2 sites is driven by their goal of providing large-scale grid computing to the whole of the CMS collaboration. To accomplish such a goal with the given level of personnel support, it is necessary to limit the ways in which users can interact with the resources.

3 Support from International CMS

As mentioned above, although T3 sites are funded and managed by the institutions that host them, the international CMS computing effort provides critical, enabling services that make the CMS T3 strategy possible. International CMS computing contributions—including those from the U.S. CMS S&C program—can be broken down into three categories: LHC Physics Center (LPC) Central Analysis Facility (CAF), central infrastructure and services, and expertise and assistance:

LPC CAF: Although not a T3 cluster, the LPC CAF fills a similar role for institutions that lack the expertise or resources to operate their own T3 cluster or who need more resources than their local T3 can provide. During 2012, there were 676 unique users who made use of the LPC CAF, and in 2013, even after data-taking had ceased and analysis activities were ramping down, there were still 552 unique users. Although the numbers above include some users that are not part of U.S. CMS, the majority of the LPC users are part of U.S. CMS, representing a substantial fraction of the U.S. CMS physics community.

Central Infrastructure and Services: International CMS computing supports the necessary infrastructure and services that make T3 clusters at institutions across the U.S. possible. Specifically, the CMS software is deployed using the CERN Virtual Machine File System (CVMFS). Tracking of datasets and file locations is accomplished via the Data Aggregation System (DAS) and the Data Book-Keeping System (DBS). GRID job submissions to T3 sites is enabled by a centrally operated GlideIn WMS server. Data stored at T1 and T2 sites is made accessible across T3 sites via the CMS data federation created as part of the Any Data, Anytime, Anywhere (AAA) project. Job submission and workflow control is accomplished via the CMS Remote Analysis Builder (CRAB) software package. Finally, international CMS computing coordination operates a Physics Experiment Data Export (PhEDEx) server to facilitate data movement to and from T3 sites.

Expertise and Assistance: The U.S. CMS S&C program provides 1.5 FTE of personnel support for U.S. CMS T3 activities. This support is directly available to administrators and users of T3 sites to help both in setting up sites and debugging problems. These personnel run a bi-weekly meeting and respond to questions via a mailing list. They also maintain T3 documentation. In special cases, the support personnel will even travel to a site to help with a particularly challenging problem.

4 Resources

Roughly 70% of U.S. CMS institutions operate a T1, T2, or T3 site. The only T1 site in the U.S. is operated by FNAL. There are seven T2 sites operated by Cal Tech, Florida, MIT, Nebraska, Purdue, UCSD, Wisconsin plus one T2 operated by Vanderbilt dedicated to the

heavy ion program. A detailed breakdown of the fraction of CMS institutions operating a T1, T2, or T3 site is given in Figure 1. This figure also shows the fractions weighted by the number of authors at each authors at each institute. From these numbers, one can see that although 28% of CMS institutions do not operate a T3 site of any kind, only 14% of the CMS authors reside at such an institution. This suggests that the institutions without T3 sites tend to have fewer authors, and likely lack the resources and expertise to run a site of their own. The LPC CAF is particularly important for keeping this population within U.S. CMS from being disenfranchised from data analysis.

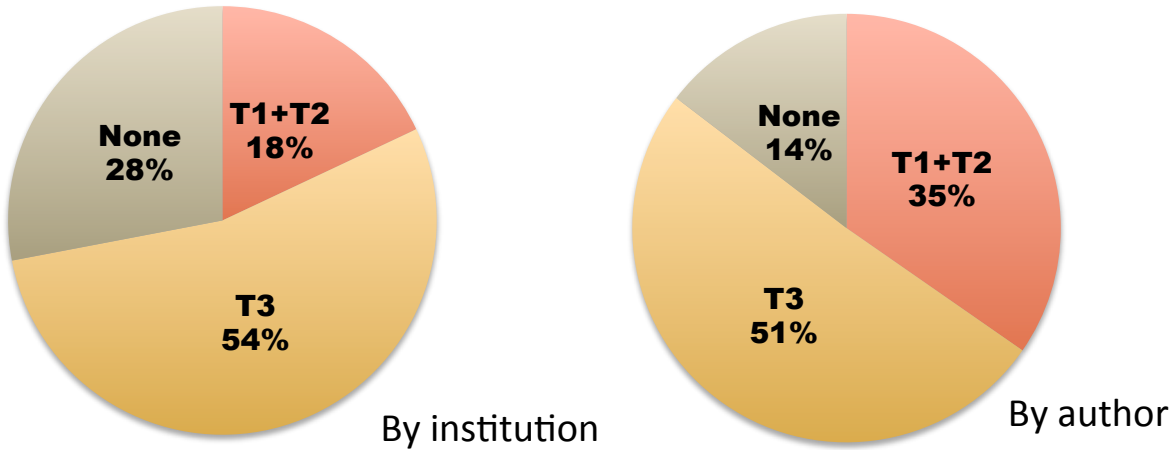
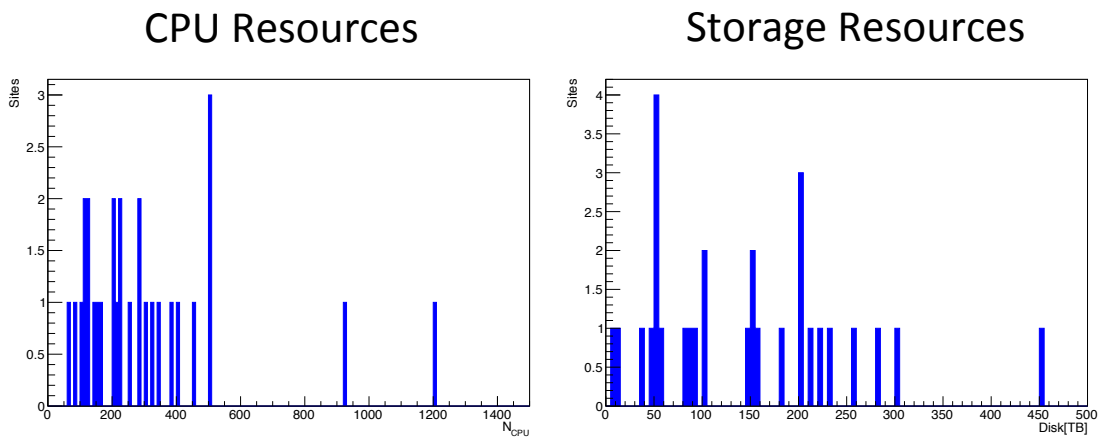


Figure 1: Fraction of U.S. CMS institutions that operate a T1, T2, or T3 site. The left plot shows the fraction counting each institution equally. The right plot shows the fractions weighting each institution by the number of authors.

Figure 2 shows the distribution of CPU cores and disk storage in terabytes (TB) for U.S. CMS T3 sites. The table below the plots also gives the number for the LPC CAF, for comparison. These plots include only dedicated T3 resources and not those shared with T2 sites or acquired opportunistically.

As can be seen from Figure 2 U.S. CMS T3 sites come in a wide range of sizes and capabilities. This range is a reflection of the variety of approaches different institutions have taken to structuring their T3 clusters. Some institutions have built systems that effectively serve as storage or interactive analysis clusters to support relatively small scale, interactive analysis for a group of users. Other sites have implemented the full CMS and OSG software stack to provide a batch system plus GRID-integrated storage, similar to the operation of a T2 but on a smaller scale. A number of sites have been built up within the larger structure of a university computing center and benefit accordingly. In Section 6 we provide several specific examples of site configurations.



Mean (per Site)	306.9
Total	8899
LPC CAF (not included)	2500

Mean (per Site)	142.9 TB
Total	4143 TB
LPC CAF (not included)	3500 TB

Figure 2: Distribution of CPU cores and disk storage for U.S. CMS T3 sites. Only dedicated T3 resources, and not shared or opportunistic resources are included. The numbers for the LPC CAF are also included for comparison.

5 Tier 3 Funding

Because the U.S. CMS T3 sites are not centrally funded, there are a variety of sources for funding those sites, including contributions both from the DOE and NSF as well as from internal university sources. Because of the variety of funding sources, this report relies on the results of the survey to capture a snapshot of how U.S. CMS T3 sites are funded. Figure 3 shows the fraction of U.S. CMS T3 sites that reported receiving some amount of funding from a given source. Note that sites can receive funding from more than one source, so the sum of the fractions for the individual categories exceeds 100%. Most sites (about 50% of those responding to the survey) receive funding from multiple sources, mainly one funding agency plus internal university funds. T3 sites receiving funding only from the DOE or NSF, but without internal university funding represent 31% of the total, while 19% of the sites are funded only by internal university sources. The disparity between DOE and NSF funding can be understood in the context of the approximately $1 million in ARRA funding distributed to U.S. CMS T3 sites via the DOE in 2010. It should be noted that hardware funded as part of ARRA is now nearing its end of life and will need to be replaced.

As can be seen from Figure 3, roughly two-thirds of U.S. CMS T3 sites receive some amount of funding from university sources. However, this number by itself does not reflect the full university contribution. In addition to providing direct funding, many universities support U.S. CMS T3 sites by providing free infrastructure or personnel. Figure 4 breaks down the categories of support provided by universities. Again, a T3 site can receive more than one type of university support so the total of the fractions for each category exceeds

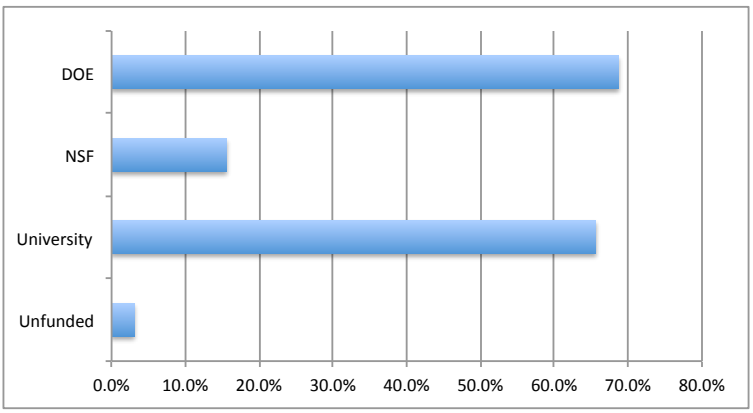


Figure 3: The fraction of U.S. CMS T3 sites funded by a given source. Note that a site can receive funding from more than one source, so the sum of the fractions shown exceeds 100%.

100%. Overall, 91% of T3 sites receive some kind of university support whether in the form of monetary support or free infrastructure or services.

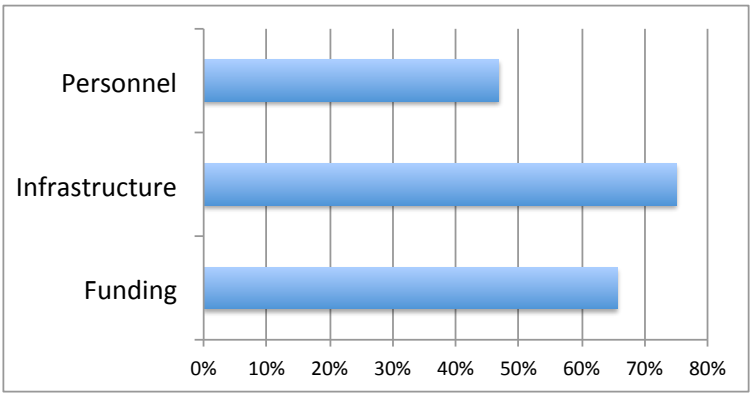


Figure 4: Types of university support provided to T3 sites. Note that a site can receive more than one type of support, so the sum of the fractions shown exceeds 100%.

6 Example T3 Sites

The best way to summarize the wide variety of U.S. CMS T3 approaches is to provide specific examples. The examples given below span the range of approaches currently employed within U.S. CMS T3. Each approach has particular strengths, and the choice of approach is dictated by the goal of the PI setting up the site and the available local resources. Included in the discussion of these example sites are ways that the choice of site implementation allows for maximizing the value of DOE or NSF investments in the site.

6.1 No T3 at Instituion

Thanks to the LPC CAF, an institution can forego operating a T3 site and still remain productive. Most of the institutions that do this are smaller institutions, but there are also a couple of larger institutions that choose to rely on LPC CAF resources alone.

6.2 Minimal T3

Relatively recent developments, primarily from the AAA project, have enable a new approach to operating a T3 cluster referred to as the "minimal configuration." Such a T3 would consist of a single server providing sufficient CPU and storage resources to support a university group doing interactive analysis. In particular, the local storage resources of this minimal T3 configuration would be used to host reduced or skimmed analysis tuples ranging in size from a few hundred gigabytes (GB) to several TB. The cluster gains access to larger CMS datasets (which can be tens of TB or more in size) via the CMS data federation mentioned above. Large scale batch computing needs are met using a combination of the LPC CAF and T2 sites as needed. As of the time of this writing, the cost for such a system (a single server with 32 CPU cores, 64 GB of memory, and 48 TB of storage) was approximately $10,000.

6.3 Co-located with a T2

Another possible model for configuring a T3 site is to arrange to have it hosted alongside a T2 site. For example, the University of Kansas has arranged to have its T3 hosted with the Nebraska T2. By working together with the Nebraska T2 to host the Kansas T3 cluster, Kansas benefits from economies of scale that would not ordinarily be available to a site the size of a T3. This includes both the increased purchasing power that comes with bulk purchasing hardware at the scale done by the Nebraska T2, as well as sharing the power and cooling infrastructure of the Nebraska T2 site. Furthermore, administration of the Kansas T3 can be handled by a relatively small fraction of the FTE of one of Nebraska's highly experienced T2 systems administrators rather than Kansas trying to hire one independently. For these reasons, the Kansas T3 is able to acquire resources at a lower cost than a stand-alone T3 would be able to.

6.4 In Partnership with a Campus Computing Center

Another option for efficient resource usage is to create a T3 site in collaboration with a broader campus computing center. In many cases, the campus center will provide infrastructure (power, cooling, etc.), support personnel, or both free of charge. Furthermore, it is often possible to purchase assistance from systems administrators, systems engineers, or programmers on an hourly or short term basis, reducing the need for employing dedicated support staff for the T3 site. Finally, in arrangements like these, it's often possible gain opportunistic access to idle cycles on other machines hosted as part of the campus center.

For example, the T3 site at the University of Notre Dame (ND) is hosted within the ND Center for Research Computing (CRC).

By locating their T3 within the ND CRC, the ND T3 benefits from economies of scale and resource sharing in much the same way that Kansas does. ND T3 hardware acquisitions can be included in bulk CRC hardware purchasing agreements to provide substantial discounts. The ND CRC provides systems administration that can be purchased in fractions of an FTE, so that an entire systems administrator does not need to be hired to do less than one FTE of work. However, the ND CMS group has also been able to obtain significant university investments both in terms of funding and resources to supplement support received from the NSF. Over the past five years, in terms of equipment, the ND T3 has received roughly $17,000 in NSF funding for servers and disk, while internal university grants have funded $120,000, primarily for storage servers, providing a total of 288 CPU cores and 350 TB usable storage. The ND CRC has also provided two racks and associated power supplies at no cost, representing roughly a $4,800 hardware investment. In terms of operational costs, the ND NSF base grant supports 0.2 FTE of an HPC engineer for T3 systems administration at a cost of roughly $20,000 per year. In return, the ND CRC provides another 0.2 FTE of an HPC engineer at no cost (equivalent to $20,000 per year), as well as power and cooling for the ND T3 hardware, which is the equivalent of approximately $30,000 per year. Also included at no cost are any number of consultations with higher-level experts on various computing and software issues. Not included in the above figures are the CRC resources that the ND T3 is able to access in an opportunistic fashion. The primary computing resources for the ND T3 comes from a general-purpose OSG cluster with 888 CPU cores run by the CRC as part of the now expired North West Indiana Computing Grid (NWICG) grant through the DOE. The compute nodes that are part of the NWICG cluster have been configured by the CRC to provide the CMS software environment and the ND CMS group is the primary user of those resources. Since the NWICG has expired, the ND CRC has taken on the expense of operating and maintaining the compute servers. Furthermore, the ND T3 can access opportunistically the pool of 20,000 CPU cores owned by other ND research groups but hosted at the CRC. On average, roughly 30% of these CPU cores are idle at any moment, and the ND T3 has demonstrated the ability to run CMS jobs successfully on up to 8000 of these CPU cores simultaneously in addition to the dedicated NWICG resources.

6.5 Full-Featured, T2-Style Site

The other end of the spectrum from the minimal T3 configuration described above, is a configuration in which the T3 site implements most or all of the functionality expected of a T2 site, including the ability to submit jobs and transfer data to and from the site via the CMS/OSG tools. Such sites tend to provide surplus resources to the broader CMS collaboration and in many ways operate like a T2 site that is not funded by the U.S. CMS S&C program. One example of such a site is the T3 from the University of Colorado, which has 1200 CPU cores and 450 TB of storage. The Colorado T3 site hosts datasets that are of broader interest to the CMS collaboration, particularly for the supersymmetry (SUSY) group, and contributes to official CMS MC production.

The Colorado T3 is also funded by a mixture of DOE and university funding. Over the past five years, the DOE has funded approximately $100,000 worth of hardware purchases through the ARRA and supports 0.4 FTE of a computer scientist who has the primary responsibility for the Colorado T3 deployment and operation, at the level of $80,000 per year. In turn, university contributions have included $500,000 for power and cooling upgrades, $100,000 for network connectivity upgrades, and $60,000 per year for hardware purchases. Thus over the past five years, a total investment of $500,000 from the DOE has leveraged a $900,000 investment from the University of Colorado. Not included in the above numbers of the university's contributions to power and cooling for the Colorado T3.

7 Utilization and Upgrade Plans

Although the plans for each individual T3 site can, in principle, evolve independently at the discretion of the PI's for the site, the U.S. CMS collaboration is making a conscious effort to coordinate plans for future utilization and upgrades of T3 resources. One of the survey questions asked, "Which option (if any) do you think would be most useful for CMS and for your research group?" The options from which respondents could select were as follows:

1. Improvements to software tools and/or support to allow better use of existing department or university computing resources as CMS T3 resources?

2. Centralized, virtual T3 resources?

3. Additional hardware distributed among the clusters of different U.S. CMS groups?

Figure 5 summarizes the responses. There was no overwhelmingly favored direction, indicating that all three directions are perceived as important for future U.S. CMS computing success. The strong support for tools development and centralized services implies the need for the U.S. CMS S&C program to continue to provide good support for T3 activities. In addition, university groups will also play an important role in driving some of these initiatives forward.

7.1 Plans for the U.S. CMS S&C Program

In light of the survey results and operational experience to date, the U.S. CMS S&C program plans include the following:

- Continue support at the 1.5 FTE level for T3 activities.

- Continue to contribute to necessary central services, such as CVMFS, DAS/DBS, GlideIn WMS, AAA, and PhEDEx.

- Upgrade the LPC CAF, expanding both the CPU and disk storage resources.

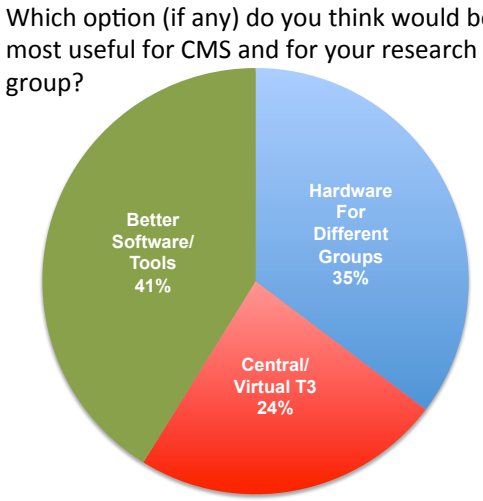


Figure 5: Summary of responses to the question "Which option (if any) do you think would be most useful for CMS and for your research group?"

- Contribute to the development of additional and improved software tools for deploying T3 functionality with as light-weight a software footprint and administrative burden as possible. In particular, OSG Connect should provide an excellent platform for developing this capability.
- Work to support expanded access to opportunistic resources, including campus clusters as well as academic and commercial clouds.

7.2 Plans for Individual U.S. CMS institutions

Although as stated above, the plans will evolve independently, common features includes the following:

- U.S. CMS institutions will continue to bear the responsibility for operating the T3 sites, including obtaining sufficient funding to maintain and expand each cluster. Emphasis will be placed on strategic investments of DOE and NSF funding to leverage university matching funds, university-provided resources, and opportunistic computing potential. From the survey, 70% of respondents reported that their university provides access to opportunistic or general purpose computing resources. Only 56% of institutions have used these resources to run CMS jobs, and only 30% have integrated such resources into their T3 operations. The potential for leveraging additional university resources can make investments in T3 resources very attractive.
- Development of software tools, especially in conjunction with local CS faculty specializing in the problems most important to HEP computing. In particular, individual university groups, or small collaborations between institutions can augment central

S&C program efforts, expanding the range of options explored. The AAA effort provides a good example of this approach.

8 Conclusions

T3 computing is essential to the success of the U.S. CMS physics program, providing capabilities that by design are not provided by T2 sites. The key uses for T3 computing include interactive data analysis, especially for activities requiring fast turn around times, non-standard resource requirements, or both. Although the U.S. CMS S&C program does not directly fund individual T3 sites, it does play a pivotal role in providing centralized services, shared infrastructure, and expert advice. Because T3 sites are typically funded by a mixture of sources, they represent an opportunity to leverage university funding through strategic investments of DOE and NSF resources. T3 sites can also serve as portals for connecting CMS computing to opportunistic campus resources. T3 management is intrinsically decentralized, but coordination between the S&C program and the collaboration as a whole is coordinated through the U.S. CMS computing liaison, allowing U.S. CMS to continue to evolve plans for T3 expansion and improvements.