



Dec 30, 2003

User Analysis Facility Requirements 2004 Version 1

Introduction:

US-CMS deployed a User Analysis Facility (UAF) in 2003. The first UAF was a relatively small computing facility, as compared to computing centers used for production, and it had a modest number of users. Between now and the start of the experiment US-CMS intends to build the scale, capability, and usage of the UAF to support a world class physics analysis center. In 2004 there are several analysis activities scheduled that require US-CMS to increase services and capabilities offered by the UAF. This document describes the requirements of the facilities and services for the UAF in 2004. Given the dynamic and evolving requirements of the user community, a reevaluation of the UAF requirements will be performed on a yearly basis.

US-CMS has established the role of user services coordinator. He/she will serve as the liaison to the PRS groups. We will use direct contact with PRS leaders, active analysis users in CMS, and analysis experience from running experiments to establish the requirements of the UAF. These should be updated regularly and feedback should be provided from the user community to the user services coordinator. This document will have a version number and should be updated throughout the year.

The four primary activities this year are the analysis components of the 2004 data challenge (DC04), the analysis required to complete the Physics Technical Design Report (TDR), the support for testbeam activities and analysis, and the general support for detector and physics groups for analysis.

The analysis components of DC04, which is currently scheduled for March of 2004, are fairly straightforward. DC04 was primarily specified as a test of the CMS core software and the initial components of the CMS data flow between the various computing tiers. For analysis, the Tier1 center will have to demonstrate the ability to analyze quantities derived from the incoming data stream. These quantities can be fed into toy calibration programs or simply displayed on web pages with regular updates.

The Physics TDR, which will be submitted in 2005, will form the basis of the analysis performed in the first year of the experiment. The document will include basic analyses that indicate the reach of the detector in interesting channels, as well as complete detailed analysis of particular channels. The physics groups have requested approximately 70 million events so far, divided into datasets of a few thousand to a few million events

each. The requests for official event production are expected to increase throughout 2004.

The detector groups have testbeam activities planned for the summer of 2004. The total amount of testbeam data collected is relatively small. However, the short timescale of the testbeams themselves and precious nature of the data collected requires the support services to be in place well in advance of the activities. There are a number of important services the UAF can provide to testbeam groups, which overlap with other facility activities and provide a motivated user base.

In general for Fermilab to build up a reasonable analysis center and offer reasonable support to detector and physics groups, the UAF must improve its capability to support a growing community of users performing analysis. CERN is struggling to meet demands of CMS physics needs, including US users at CERN. We expect that some of those needs will be met by the UAF at Fermilab, by enabling remote use within the US and from CERN. This will help build a credible analysis center at Fermilab and allow the evaluation and development of larger scale analysis services.

This document is broken into five sections. The first describes the common services and technical capacity required by all the analysis activities. The following four sections describe services that are specific to one or more analysis activities.

Common Requirements:

The common requirements for the predicted 2004 analysis activities are mainly related to the technical capacities of the UAF computing facility itself. The storage capacity, the data access performance, and the CPU available. There are a few facility services, but they are closely related the computing resources. Most of the higher-level analysis services are identified with a specific analysis activity, and described later in the document, though they may be useful to several.

Storage

The expected number of active analysis users for 2004 across all activities is between 20 and 40 individuals. This is a substantial increase from the number of active users this year. US-CMS expects approximately 500 physicists at the start of the experiment, so US-CMS should prepare for doubling of the number of registered users each year until data taken. There is a wide range of activity level for the analysis users. The storage requirements and access capability is given in Table 1.

Service Type	Value	Service Quality
Storage Quantity		
User Updateable Storage	100GB per user	8 x 5 support Automated Failover Catastrophic Recovery
User Writeable Mass Storage	1TB per user	8 x 5 support Automated Failover Scalable to 10 users

Official Simulation Staging Areas	15TB	8 x 5 support Automated Failover
Official Simulation Mass Storage Allocation	75TB	8 x 5 support Automated Failover
Storage Access		
Read/Write Access to User Updateable Storage	5 MB/s read per process 3 MB/s write per process	8 x 5 support Automated Failover, with service degradation Scalable to 15 processes simultaneously
Read/Write Access to User Writeable Storage	8MB/s read per process 5 MB/s write per process	8 x 5 support Automated Failover, with service degradation Scalable to 10 processes simultaneously

Table 1: Common User Storage Requirements

The storage quantities for updateable user space in Table 1 were estimated from an estimate of a reasonable analysis object collection from a CMS simulated dataset and requests for space that US-CMS has already received from analysis users. The mass storage allocation was estimated by assuming that a small percentage of users would be creating custom datasets that are typically a few hundred gigabits to a few terabytes in size. The allocation sizes for official CMS simulation caching and archiving were estimated from the total size of the pre-challenge production for DC04, which is the most recent group of samples created for general analysis, and a fraction of the Objectivity based data US-CMS expects to serve from Fermilab. In addition mass storage was allocated for continuous official production expected throughout the year. The disk caching space assigned to official datasets was chosen based on 20% of the total stored data.

The data access requirements were estimated using the expected size of the CMS DST data and summarized analysis object data. The CMS DST is estimated at 0.5MB per event and the current analysis object data, stored frequently as Root trees, are estimated at 50-100kB per event. CMS datasets are currently between one hundred thousand to several million events. In order to loop over a dataset in a reasonable length of time, like an afternoon, 5MB/s is required. The write speed is estimated to allow a reasonable rate for deep copying the data from the database and reasonable archiving speeds for custom user data sets.

Processing

The estimated common CPU requirements are shown in Table 2. The number of interactive systems was estimated by assuming 20 – 40 active users with a 25 – 50% duty cycle. Each user is assumed to fully utilize one CPU when active. The number of batch nodes is less well motivated, but the cluster must be large enough to be an attractive place to work with enough resources for the active user community. The allocation between

interactive and batch nodes should be reassessed throughout the year as usage patterns are better understood.

Service Type	Value	Service Quality
Processing Services		
Interactive Systems	5 dual Xeon CPU nodes	8x5 support Failover with reduction of quantity of service
Batch Systems for User Analysis Work	30 dual Xeon CPU nodes	8x5 support Failover with reduction of quantity of service
Computing elements with Redhat 6.2 OS	10 dual CPU P3 nodes	8x5 support

Table2: Common CPU Resource Requirements

The final group of systems in Table 2 is assigned for a specific task. The data simulated before the summer of 2003 was persistently stored using an Objectivity database. The CMS license agreement terms, negotiated through CERN, specifies that only Redhat Linux 6.2 systems can be used. It is considerable effort with little return to install older versions of the operating system on new hardware, so we specify that P3 based systems will be used for this purpose. The UAF does not expect to get many requests for the Objectivity data services, but in case the physics groups need to perform consistency checks with the new data, the service needs to be offered. We expect to shut the service down after the proper validation of the GEANT4 based simulation, which should be completed by the beginning on the summer.

The interactive nodes in the UAF need to have load balanced access to the systems. The technique used in 2003 of utilizing the batch system was quite successful. Architecturally, this technique introduces a single log in point and subsequently a single point of failure. It also required development to ensure that all the features of an interactive log in are available through the batch system interface. At the same time, the batch system allows detailed accounting of user processes and utilization. For 2004 in the interest of scaling to very large numbers of systems in the future, the UAF will attempt to deploy the interactive load balancing techniques based on DNS servers that have been successfully deployed at CERN. These have been successfully scaled to very large numbers of systems and can introduce the same failover protection used for the normal DNS servers.

Networking

In order to meet the data access requirements to storage outlined above in Table 1, it will be necessary for all CPU nodes assigned to analysis to be connected over gigabit links. It is possible for a single process to meet the requirements outlined above, but the current dual Xeon systems are enabled to run up to four processes simultaneously requiring improved bandwidth. In general the systems at the Tier1 center will be connected over gigabit links. 100Base-T network interfaces were commonly available in 1996 when the

typical CPU speeds were 300-400MHz and the system memory bandwidth was less than a tenth of a modern system. With CPU speeds of 3GHz it is only logical to remove this bottleneck to the system where economically feasible. The networking requirements for the UAF are given in Table 3.

Service Type	Value	Service Quality
Networking		
UAF to dCache Pool Storage Resources	1 Gigabit	Each UAF cell should have 1-2 pools connected at true gigabit speed, remaining Pools can be oversubscribed up to 300% 8x5 support
UAF to afs servers	300Mbit	Failover protection 8x5 support
UAF to Enstore	2 Gigabit	Total bandwidth from Enstore to UAF switch Failover protection with service degradation 8x5 support
UAF to offsite	100Mbit	8x5 support

Table 3: Networking Requirements for the UAF in 2004

The values in table 3 were chosen based on the requirements and the performance of current UAF hardware components. The speed from UAF cells to dCache Pool resources was chosen to enable the current generation of pool systems to reach peak performance and to allow the worker node to storage access rates shown in Table 1. The UAF to afs server rate is lower, because CMS does not expect large amounts of user data to be stored and retrieved from afs space. Primarily afs is used for the software environment and login space and needs to support multiple users with compilation and code development space. Given the current caching size compared to dataset size of 20% and the number of 9940B drives accessible to CMS, we estimate 2Gb to Enstore is sufficient to prevent bottlenecks in the mass storage system. The CMS Tier1 facility has a variety of activities planned for 2004, which will exercise wide area networking, but the UAF offsite network requirements are primarily associated with interactive login and 100Mbit is sufficient.

We expect the same excellent level of networking support on the UAF that US-CMS currently enjoys on the production systems. The choices of network technology have been made with performance and support in mind.

Software Environment

In 2003 afs and features of the CMS software configuration tool, SCRAM, were exploited to duplicate the CERN software environment on Fermilab systems. This has been very successful in allowing all releases of the software to be immediately available

to Fermilab users. This technique will continue to be employed. During 2004 US-CMS should develop a test harness to verify the local software environment. A chain of CMS software from generation, to GEANT4 simulation, to ORCA reconstruction, and finally to IGUANA visualization should be executed against production releases in a, hopefully, automated way and the results published on the UAF web pages.

The software environment relies on afs services to provide the software and SCRAM to provide the local configuration. The afs support is handled centrally at Fermilab and CERN. The support level of this activity is defined by the quality of afs support, which is currently excellent. The SCRAM components of the environment and supported locally and will be dealt with on an 8 x 5 basis.

Documentation

There is a common need for detailed, web-based documentation for users. The documentation currently available has been very good, with getting started instructions and basic physics examples for new users. The documentation should be expanded and updated in 2004. This is especially important as more members of the physics community begin to work at Fermilab. In addition to UAF effort, representatives of the physics groups should prepare examples and tutorials for the UAF documentation.

The web-based documentation is supported on an 8 x 5 basis, but the servers themselves should be enabled with failover protection to better provide support for remote participants in different time zones.

Most of the UAF services will be supported on an 8 x 5 basis. There are some components that may need high availability and we may specify longer support periods, especially during scheduled activities. We will try to use fail-over protection to maintain services even at degraded performance levels to ensure that services are available to users in many time zones.

DC04 Analysis Services

There are a small number of analysis services that need to be developed for the UAF to effectively participate in the analysis aspects of DC04. During DC04 data is reconstructed at CERN at a rate of 20Hz and made available for transfer to the Tier1 centers. The Tier1 centers will update a central catalog when the data has been successfully received and archived. The analysis components are designed to show that the Tier1 centers can access and process the incoming data in real time. During the challenge detailed physics analysis is not required; data access is sufficient.

There are several services required for DC04 that will be supported on a 24 x 7 basis. These deal with transferring data from CERN and archiving it at Fermilab. The output buffers at CERN are deep enough to handle an overnight failure, but it is difficult to catch up with the transfers if a service has been lost for too long. The analysis components can be supported on an 8 x 7 basis. The data processing and updating should not be idle for the weekend due to a service failure, but the analysis activities can recover from the loss of services during the night.

Data Discovery Services

During the challenge the arrival of new data needs to be cataloged and published. Data will ship in two formats: a portion of the raw data and a complete set of the reconstructed data (DST). Data will be sent in multiple data streams broken into physics processes at CERN. The catalog and publishing tools used at the Tier1 for analysis should be as similar as possible to catalogs used in other aspects of the challenge. The plan is to use a combination of Replica Location Service Catalogues¹ ²(RLS) and Storage Resource Broker³ (SRB) catalogs. The catalogs used at the Tier1 need to be queried by processes and experimenters and to be compatible with the CMS software.

The requirements of cataloging and discovery are not very taxing. The current estimates are for 1 2GB file of raw data every 2 to 3 minutes, with smaller reconstructed data files 2 –3 times per minute. This rate can be handled by a number of technologies.

Automated Processing Services

Once data arrives at the Tier1 center and is published in the catalog, a process must be run to access the data. This process has the potential for being very simple; it may just pick a single reconstructed quantity and make a histogram. If there is the desire from the physics groups to perform more advanced calculations, it should be possible, but not necessary for the success of the challenge. The important aspect is the real time nature of the analysis. The data should be analyzed within 12 hours of arrival. This could be handled manually, but it would be preferable to have the analysis handled automatically.

The automated processing challenge is well within the technical capabilities of the UAF. The analysis should expect to loop over up to 1TB of data per day, most of which can be analyzed from disk before archiving to tape. The length of time the analysis step requires defines the CPU capacity needed.

Automated Updating

In order to demonstrate that the data has been analyzed in a reasonable length of time, the updated results must be published. This could be updating a web site with a new histogram or e-mailing a revised set of calibration constants. There are frameworks like Clarens⁴, which would help automate this process.

Physics TDR

The analysis activities planned for the physics TDR represent a large increase for CMS in terms of number of events and data analyzed, the number of experimenters participating, and the complexity and depth of the analysis activity. In addition to the core services for storage, CPU, and user support, the analysis activities for the Physics TDR are in need of Data Cataloging and Publishing Services.

Data Cataloging and Publishing Services

In order for physics groups to work effectively at the UAF they need to know what datasets from the official production are available. A data catalog must be provided that has web, human, and process interfaces to allow data discovery. There are currently

approximately 70 million events produced or in production forming 700 datasets from the four physics groups. The total sample is about 50 thousand files. Though FNAL will not be expected to host the entire sample, we should be prepared to host a large fraction of it.

50 thousand files in a catalog is not a technically difficult challenge, but it does eliminate some simple solutions. With the number of datasets expected, querying and selection tools will be required. The UAF should examine using the cataloging and replication tools planned for the data challenge and grid activities, including RLS. The data cataloging and publishing services should be supported on an 8 x 5 basis with automated failover to better support remote collaborators.

Testbeam Services

The total amount of testbeam data is small, but the analysis application is similar to experiment running and provides a realistic use case for developing computing services. The data is collected at CERN and there is a community of users in the US and elsewhere anxious to retrieve, analyze, and archive the data.

Bulk Data Transfer Services

Data is collected at CERN and needs to be transferred to the UAF for analysis. The total amount of data is typically only a few hundred gigabits, but this is sufficiently large to merit using high performance transfer tools. The Tier1 facility operates a gridFTP door directly into mass storage at Fermilab, which could be used for archiving and transfer. Tests have been performed between CERN and FNAL using SRM interfaces. Both of these tools have successfully demonstrated more than 1TB per day transfer capability. For the testbeam community it is important to establish stable and semi-permanent services sufficiently far in advance of testbeam operations to train and support participants.

Data cataloging and discovery

A similar set of tools as those being deployed for DC04 should be deployed for the testbeam activities to allow discovery and cataloging of available data. These are services that would be deployed at the testbeam site, but supported through the UAF to enable remote participants to discover new datasets and automatically trigger the bulk data transfer services to replicate data to Fermilab.

The data transfer and data cataloging services can be supported on an 8x5 basis because the data rate is sufficiently low. The output buffer depth can ensure no data loss and the speed of the transfer tools allows for quick recovery.

Services for Detector and Physics Group Analysis Support

The final section is support for services for detector and physics groups. Most of these are covered in the general services section, but there is one important additional service for Monte Carlo simulation that the UAF can provide to this community. Currently the physics groups make prioritized requests for large production samples that are processed centrally by the production team. The time between requesting a sample, completing all

the steps, and receiving the data for analysis can be three to six months. The production team runs the events in the most efficient way to complete the entire sample. To enable rapid development it would be useful to the physics groups to enable individual users to process small simulation requests themselves.

Monte Carlo Generation Services

The tools developed for centralized productions can be deployed for individuals as well. There is currently effort in the distributed applications group to enable the MC_RunJob production environment to run as a web service through the Clarens framework. Individual users could submit requests for samples of events, which would be run on batch analysis resources or centralized production resources if available. This would empower individual physicists to make rapid progress by providing simulation samples in a matter of days. It would also allow individuals to make more personal decisions about the priority of production requests.

The Monte Carlo generation services should be supported on an 8x5 basis. The current batch analysis farm at the UAF could generate approximately 10,000 events per day, depending on the channel, for custom user simulation.

¹ European Data Grid RLS Web Page. (2003) http://eu-datagrid.web.cern.ch/eu-datagrid/Intranet_Home.htm

² Globus Project RLS Web Page (2002). <http://www.isi.edu/~annc/RLS.html>

³ Storage Resource Broker (2003). <http://www.npaci.edu/DICE/SRB/>

⁴ Steenberg, Conrad (2003) Clarens Web Page. <http://clarens.sourceforge.net>