CDF/DOC/CDF/PUBLIC/10922
Version 1.0
September 17, 2012

Report of CDF Data Preservation task force

S. Amerio[1], S. Behari[3], J. Boyd[3],Y.C. Chen[2,3], R. Culbertson[3], C. Group[4], S.Lammel[3], N. Moggi[5], R. Snider[3], J. Strologas[3], T. Parsons[3], D. Torretta[3], S. Wolbers[3]

[1] *INFN Padova (Italy),* [2] *Institute of Physics, Academia Sinica (Taiwan),* [3] *Fermilab (USA),* [4] *University of Virginia (USA),* [5] *University of Bologna (Italy)*

Abstract

This report summarizes the first outcomes of the studies carried on by the CDF long term data preservation task force between June and August 2012. The task force analyzed the current CDF computing model and identified the minimum requirements to preserve complete analysis capability in the long term future (> 10 years from now). Five main areas were investigated: data access, code, job submission, monte carlo production and documentation. For each area the current status is described, followed by the requirements for the long term future. Possible technical solutions are discussed and a tentative resource estimate and timing schedule are presented. The possibility of using CDF data for outreach purposes is also discussed.

Contents

1 Introduction

Data collected in High Energy Physics (HEP) experiments are the result of a significant human and financial effort. The preservation of HEP data beyond the lifetime of the experiment is of crucial importance for several reasons:

- long term completion and extension of scientific programs: it is demonstrated (LEP experiments, BaBar, H1 and ZEUS) that about 5 to 10% of total scientific production of a collaboration is produced after the end of data taking; often these analyses are more sophisticated and can exploit the physics potential of the entire collected dataset;

- continue to perform cross collaboration analysis, analyzing data from several experiments at once; in this way statistical and/or systematic uncertainties of single experiments can be reduced, and new analyses performed. Such an effort to combine analyses is already ongoing, for example between the H1 and ZEUS collaborations, and an evaluation of such an approach is underway between the Belle and BaBar collaborations.

- Perform new analysis on data from past experiments with new theoretical models or new analysis techniques; this can lead a significant increase in precision for the determination of physical observables. As an example data from the JADE experiment at the PETRA e^+e^- collider have been recently re-analyzed leading to several precise measurements proving the running of the strong coupling constant in a unique energy range [1]. ALEPH data have recently been re-analyzed to search for a low mass Higgs super-symmetric partner that may be produced in pairs and would be able to decay to four tau leptons. This configuration and the corresponding decay channel were not explored during the collaboration lifetime and are now shown to cover a new domain in the parameter space [2].

- Education, training and outreach: past HEP experiment data can be analyzed by graduate and undergraduate students from institutions beyond those which collaborated on the experiment; this is a unique opportunity to reach a wide audience, and an invaluable tool for particle physics classes.

Since 2008 a study group on Data Preservation in High Energy Physics (DPHEP) has been constituted to investigate the technical and organizational aspects of HEP data preservation. DPHEP identifies different models of data preservation, of increasing complexity [3], from the simpler preservation of the capability of providing additional documentation on published analysis (extra-data tables, internal notes, etc.), to the more complex preservation of the reconstruction and simulation software and basic

level data. The latter is recommended, as the only way to provide the full physics analysis chain and retain full flexibility for future use.

The experiments BaBar, Belle, BES-III, CLAS, CLEO, CDF, D0, H1 and ZEUS are represented in DPHEP, joined also by the LHC experiments ALICE, ATLAS, CMS and LHCb. The associated computing centers at CERN (Switzerland/France), DESY (Germany), Fermilab (USA), IHEP (China), JLAB (USA), KEK (Japan) and SLAC (USA) are all also represented in DPHEP.

Since Spring 2012 a dedicated task force has been created at CDF to study the current CDF computing model and develop a proposal for long term preservation of data and analysis capabilities. The task force works in collaboration with the DPHEP group and aims at identifying all the necessary requirements for CDF data analysis in the long term future, propose possible solutions and evaluate their technical feasibility. In this document the outcome of the task force studies are summarized.

2 Tevatron data: motivations for long term preservation

CDF and D0 experiments at the Tevatron collected about 10 fb^{-1} of data during RunII. These data are still yielding valuable information on the nature of physics phenomena and will continue in the future, especially in light of new discoveries by LHC or other experiments and advances in theoretical models. Additionally, as no plans are foreseen for a proton-antiproton collider in the future, Tevatron data will be unique for a very long time in terms of initial state particles and measurements of effects enhanced by $q\bar{q}$ interactions. In these areas, the Tevatron will remain competitive with the LHC.

As examples within the realm of top physics, $t\bar{t}$ production asymmetry measurements have shown a discrepancy with the Standard Model which could hint at new physics [4]. At the LHC such effects are more difficult to observe, as symmetric gg-initiated events dominate top pair production. Moreover, differing production mechanisms in the two environments test distinct aspects of the Standard Model and require different analysis strategies, as for example in $t\bar{t}$ spin correlation measurements [5]. Tevatron data will also remain of importance for single-top searches [6], particularly in the s-channel which is more challenging at the LHC than at Tevatron, because the relative cross section is much smaller. The mass of the top quark is now known with a relative precision of 0.54%, limited by the systematic uncertainties, which are dominated by the jet energy scale uncertainty [7]. This is the result of the combination of several measurements made by CDF and D0 in different $t\bar{t}$ decay channels on data samples with integrated luminosity up to 5.8 fb^{-1}. Uncertainty on jet energy scale is expected to improve as analysis are performed on the full data samples, since analysis techniques constrain the jet energy scale using kinematical information from $W \to qq'$ decays. For the first time the total uncertainty of the combination is below 1 GeV; such a level of precision requires additional theoretical study of the exact renormalization scheme definition corresponding to the current top mass measurements. We

have entered the era of precision measurements in the top sector, and the Tevatron will provide a substantial contribution to the top mass world average for many years to come.

In the electroweak sector, one of the most important measurements made at the Tevatron is the precise determination of the W mass. In conjunction with the top mass, the W boson mass constrains the mass of the Higgs boson, and possibly new particles beyond the standard model. The measurement is very challenging due to presence of an undetected neutrino from the W decay, which makes it impossible to fully reconstruct the final state unambiguously. Recently CDF and D0 have measured the most precise values of the W mass to date [8], achieving a total uncertainty of 19 and 23 MeV/c^2 respectively, dominated by the uncertainty on parton distributions functions (PDFs). It will be difficult for the LHC to overtake this precision for at least several years. Moreover, the current measurements are based on exposures of 2.2 and 4.3 fb^{-1} in integrated luminosity at CDF and D0, respectively. With the full data sample, these measurements could constrain systematic uncertainties and, in principle, reach a precision of 10 MeV/c^2. Attaining this precision will require considerable effort, especially in reducing the uncertainty on PDFs.

Heavy flavor physics has several potential analyses that can be carried out in the coming years. Some of the ideas that have emerged include measuring $A_{\rm SL}$ in B^0 and B_s^0 decays, studying the forward-backward asymmetry in charm and bottom production, measurements of the interference between scalar and vector resonances in B decays, and measuring production cross sections and polarizations (where possible) for as many heavy flavor states as possible. There are numerous decay modes that can be extracted from the data, some which will likely be overlooked by other experiments.

More generally, Tevatron data might be useful to cross check results from other experiments. This may be particularly important in the light of new discoveries at the LHC, which may require CDF data to be revisited, possibly with new, more advanced analysis techniques. This was recently demonstrated with the evidence for a CP asymmetry difference between $D^0 \to K^+K^-$ and $D^0 \to \pi^+\pi^-$ decays from the LHCb experiment that was soon after confirmed by CDF [9].

It as also to be stressed that Tevatron measurements are made in a unique energy domain, which will be no longer available; therefore QCD measurements performed on Tevatron data will continue to be as valuable in understanding QCD as on LHC data. Examples are measurements on diphoton cross sections, $Z/W+$ jets, underlying and minimum bias events, diffractive W and Z production. For the same reason, before the definitive shutdown of the accelerator, data were collected at two different energy points, 900 and 300 GeV. The data samples, collected by minimum-bias and selective triggers, will provide some valuable legacy measurements in non-perturbative QCD, soft and strong interactions.

These are only some highlights of the enormous potential of Tevatron data. The Tevatron will keep producing high quality scientific results, though at a lower rate in the coming years. But Tevatron will also serve as a fundamental point of comparison for LHC.

3 CDF task force charge

The CDF Collaboration aims at preserving the capability to perform data analysis on the full Run II data set in the future. Our data sample has some unique features and it has the potential for providing useful scientific information even many years from now. Any CDF member should be able to access the data in the future with the same level of effort as it is required at present, and on a longer timescale, we want to allow our data to enter the public domain.

Our goal is to achieve the level of data preservation defined as *Level 4*[1] by the DPHEP standard [3] for as long as the CDF Collaboration will continue to exist. We also want to prepare for the following phase, when the CDF data set will enter the public domain. In this last phase, access to the data should be preserved at *Level 2*[2] at least with special consideration to its potential value for outreach and education. The charge to this Committee is to investigate the exact terms in which this can be accomplished, formulate an appropriate plan, oversee its implementation and ensure a smooth transition to maintenance. The Committee is expected to report periodically to the Spokespersons, and submit for approval by the Collaboration a written report describing the plan.

Specific points the report should address are:

- Requirements
  Spell out in detail what functions, resources, validation procedures and documentation need to be preserved in order to maintain the capability of analyzing CDF data. Production of physics results should not be more difficult in the future than it is at present despite the expected diminishing support and reduced availability of expert advice. Ways to streamline the analysis process should be investigated. The possibility of implementing value-added procedures such as RECAST should also be considered. The Committee will perform this task in consultation with data analysis experts in the Collaboration.

- Feasibility
  Analyze the technical implications of the requirements and determine if all of them or which subset can actually be reasonably implemented. Provide a high level of confidence though detailed plans, analysis of existing systems, or a new prototype system, as necessary. The Committee is expected to iterate with the Collaboration through the Spokespersons and seek their approval if some compromise on the requirements must be made in order to ensure feasibility.

- Synergy
  Suggest how the CDF data preservation effort would profit from the synergy with similar efforts by other Collaborations facing similar problems. In particular it

[1] Preserve the reconstruction and simulation software as well as the basic level data
[2] Preserve the data in a simplified format

would be useful to understand which subset of our requirements may be common to multiple experiments and which instead are expected to be specific to CDF.

- Detailed action plan
  This should be technically as detailed as possible, including a description of the resources needed and an approximate schedule for implementation.

After the delivery of the report and its approval by the Collaboration, the Committee will remain in existence to oversee the implementation of the plan until completion.

4 Organization of the work

Fermilab is committed to support CDF data analysis in the coming years. The exact resource allocation will depend on the analysis load. Likely there will be little analysis activity beyond 2015. We assume the CDF collaboration will still exist at that time, but access to data and analysis will be reduced to a small number of users. In the very long term future (more than 10 years from now) we expect the request to analyse CDF data will be intermittent.

The preservation of the full analysis chain capability means that we need to preserve the data and the technology to access them, the analysis code, computing resources to run analysis jobs and all the necessary knowledge to work on the data and perform an analysis. The CDF task force has thus identified four different main areas of investigation: data access, code maintenance, job submission framework, and preservation of documentation. A fifth area, which relies on all the previous four, but deserves a specific section, is Monte Carlo production, which will be necessary in the future in case new theories need to be tested on CDF data.

In the following, for each area of work, we will describe the current system used at CDF, identify the necessary requirements for the long term future and propose possible solutions. A preliminary time schedule and resources evaluation is also presented.

This document will be discussed within the CDF collaboration and with the Fermilab Computing Sector (CS) and the D0 experiment; it will serve as a starting point to develop a concrete plan together with Fermilab CS and other computing centers (CNAF, KISTI) which currently support CDF computing and may want to be involved in CDF data preservation project. The final plan has to be designed, approved and start to be implemented by the end on 2012.

5 Data Access

The term *data* comprises both the events recorded by the detector and all the related data and metadata, stored in databases and in other file formats.

Data group	Volume (TB)
MC (raw data)	1125
MC (ntuples)	608
Data (raw)	2193
Data (production)	3821
Data (ntuples)	1395
TOTAL	9142

Table 1: CDF data volume.

5.1 Current system

5.1.1 Data events

Three different data formats exist at CDF: *raw data*, *production data*, which have undergone a first reconstruction of physics objects and assigned to a specific dataset depending on the triggers satisfied, and *ntuple-level data*, in three main flavours, developed for different analysis groups (top physics, B physics and generic). CDF data volume (collected data and Monte Carlo simulated data) is summarized in Table 1 and amounts to about 9 PB. All data are stored on tape (LTO3, LTO4 and T10K) within a dedicated library at Fermilab controlled by the Enstore mass storage system[10]. The data handling system of CDF is based on the Sequential Access Model, SAM [11]. Metadata describing the contents of each data file is stored in the SAM data catalogue. The data handling system can be used to define datasets based upon metadata queries within the catalogue. The files within such a datasets can then be delivered upon demand from tape to worker nodes for processing via a 800 TB dCache-based disk cache. dCache [12] fetches files requested by the users and stores them on a distributed pool of disk servers for the user to access over the network. The disk servers are managed as simple cache with the least access file deleted in case space to fetch new files is needed. dCache uses the namespace, PNFS/Chimera, of Enstore and supports multiple access protocols (dcap, kerberized FTP, SRM and Gridftp).

5.1.2 Metadata

Database Management Systems are used throughout the CDF experiment to manage a large variety of metadata. Oracle databases are used for detector and data metadata and MySQL databases for document and user information. Two sets of Oracle databases exist, the online databases which are filled with run condition, configuration, trigger, luminosity, alignment and calibration information and the offline databases which replicate some of the online information required for data processing and analysis and store additional luminosity, dataset, and file information collected offline. The online production database is hosted on `fcdfora1`. As far as the offline DB is concerned, `fcdfora3` and `fcdfora7` host the production and replica versions

respectively. Both machines date to 2009. An offline development replica also exists, hosted on fcdfora8, installed in 2009 as well.

CDF analysis programs access the database information through a multi-tier web-based system, DB FroNtier. The client translates the query to a DB FroNtier URL and sends it to local Squid servers. The Squid servers provide caching for frequent queries of static database information. If the information is not available, the request is sent to dedicated DB FroNtier servers which translates the URL into an SQL query and pass it to the Oracle database. The result of the query is encoded and sent back on the http protocol through the squid server. This multi-tier system is used to prevent the database being overloaded with identical queries coming from thousands of batch worker nodes. CDF has three FroNtier servers (fcdfdbfrontier1, fcdfdbfrontier2 and fcdfdbfrontier3) and a FroNtier monitor (fcdfdbfrontier4).

Dataset, fileset, file, and run-section information is managed and accessed via SAM, the sequential access model used for data handling. The data handling I/O modules of the analysis framework use wrappers around the batch system to schedule CDF jobs and are setup to communicate with SAM for dataset translation and to coordinate file delivery. SAM services are hosted on two dedicated machines, fcdfsam1 and fcdfsam2. A third machine (fcdfsam3) acts as monitor.

The MySQL databases live on the web server (fcdfweb) and are accessed (for query and update) through web interfaces.

5.2 Future

For the long term future we can envisage data access mechanism remaining the same, except possibly where we have identified four vulnerabilities: physical tape, SAM code maintenance, dCache and DB support.

5.2.1 Data events

As far as hardware is concerned, the preservation of CDF data will require the regular migration to new tape technology when current technology will become obsolete and new will be available. The current plan is that all CDF data will be copied to T10KC tapes in the next two years. With the increasing density of storage systems, the preservation of CDF data will require a smaller fraction of the total capacity of the storage systems, at a smaller cost and fewer cartridges and tapedrives.

Currently CDF data are only stored at Fermilab in two copies, in FCC and GCC computing centers. Small subsets of data (a few hundred TB) are available at the CNAF and KISTI computing centers. Any project of data preservation needs to ensure some redundancy of data: it would be useful if a copy of some subsets of the CDF data was also stored offsite, in one or more computing centers. This will ensure maximum protection against accidental data loss or corruption. Within the CDF data preservation task force we are investigating the possibility of copying some subsets of CDF data to the CNAF and KISTI computing centers. Fermilab and CNAF comput-

ing experts are preparing a testbed to test the maximum data transfer rate achievable between Fermilab and CNAF and the minimum amount of needed resources. Test results are expected by the end of October.

An issue that has to be carefully addressed is access to data. CDF data access is based on SAM and dCache; SAM, developed at Fermilab and also used by the D0 experiment, has been running for more than ten years and has been very successful. Fermilab CS is planning to use the SAM data handling system for the Intensity Frontier (IF) experiments. This SAM - IF SAM - will not be used in its current version. Some of its features will be heavily improved [13]. As an example, in IF SAM a C++ client is provided with an API for various actions, like defining a dataset and requesting a file. In the current version of SAM, this component has to be integrated with the experiment's framework, and this requires a SAM developer to have expert knowledge of the framework as well as an experiment expert having deep knowledge of the SAM C++ API. In the future SAM version, a HTTP based interface will be introduced to alleviate the difficulties in framework integration and SAM client deployment. Instead of integrating a SAM API piece into the experiment framework for communicating with SAM, communication is performed via HTTP protocol, thus effectively decoupling SAM from the experiment code, with advantages in terms of development and maintenance.

Given this situation, two main options for the long term future can be envisaged:

- maintenance of the current SAM data handling system as it is; in the short term this solution will surely require a minimum effort. But in the long term the support load will significantly increase, as it will be necessary to keep old services running, current SAM experts may not be around and knowledge transmission may not be comprehensive. Moreover, this solution depends on the long term availability of external software on which SAM relies upon (like omniORB).

- migration to the IF SAM data handling system; this solution will require immediate effort to migrate CDF code to the new system. But in the long term future it will require low support level on the CDF side, as it will be using the same software as IF experiments.

As far as fetching files from tape is concerned, for the short term future when a significant data access is foreseen, we can still rely on dCache. In the long term future, when the access to data will become intermittent, we should consider the possibility of abandoning dCache and use SAM cache instead. SAM cache is currently being used by D0 and by offsite CDF SAM stations at CNAF, KISTI and Karslrhue, for example. This option will require to modify CDF code and support systems such as SAM stations and dCache file servers.

5.2.2 Metadata

Metadata stored in the offline and SAM Oracle databases have to be preserved in case data need to be reprocessed or new MC samples generated. If Oracle will be supported

in the long term future by Fermilab CS, we may continue to use it for our DBs. If Fermilab plans to abandon Oracle and move to other products, CDF can have two options:

1. port all DBs to the new system: this solution will require a big effort from both CS and CDF collaboration; CS can help with the conversion, but a CDF expert is needed to validate the new DB. Moreover, the CDF software itself will need to be modified for the new backend;

2. freeze our DBs in their current status and keep them in read-only mode. This may be a drawback for SAM database, as it will not be possible to update it in case new MC samples are generated or data is reprocessed.

Other options may be possible; we will need support from the CS DB experts to investigate them in detail.

Another issue regards the DB machines: offline DB may be migrated to other machines in the future, but hardware and software support from the CS will be needed in any case. The same applies to the SAM database machines.

5.3 Resources and timing

Data migration to new tape technologies will be taken care by Fermilab CS. If data have to be copied offsite, support from the CS storage and network departments will be needed; as we are planning to exploit current CDF resources for the copy, the load on Fermilab CS will be minimum. As far as data handling is concerned, both options (keeping the current CDF SAM or migrating to new IF SAM) will require immediate action by CDF and CS: to adapt CDF code to the future SAM, the collaboration of a SAM developer and CDF code expert is needed. This is also true in case we plan to maintain the current system: a much more detailed documentation of the current system has to be started immediately, providing also tests to be regularly performed on the system and recovery procedures.

On the hardware side, CDF will need disk space to host the SAM cache. The exact amount will depend on the actual data access, and may be regularly re-negotiated with the CS.

Machines to host the online, offline and SAM databases will also need to be guaranteed.

6 Experiment Software

The term “experiment software” is intended to include all offline software that is needed to conduct an analysis and that is supported by CDF, or that is supported externally but is required to build or run that software. A significant fraction of this software is in the familiar reconstruction and simulation packages, and the analysis software. Equally important are the infrastructure packages, such as the analysis framework, the

CDF variant of Frontier, and the interface code to the data handling system. Each of these present different types of support challenges under a preservation regime. All, however, rely on a common set of underpinning services that define the preservation problem: a software repository, a distribution system, a build system and environment, and a run-time environment, all running on one or more platforms. Note that the build and run-time environments in many instances need not be identical.

6.1 Current system

CDF-supported offline software is archived as a set of packages in a `cvs` repository (see [14]). The reconstruction, simulation, data reduction, and analysis code within the repository belong to versioned releases of the offline code. The content of each release is subject to a controlled process of selective updates and validation[3] prior to release and use in primary reconstruction and MC production processing. This procedure is managed by the offline operations group in coordination with the physics groups. Procedures for building the executables, including any patches required, are well documented on the data production and MC production web pages([16] and [17]).

Secondary reconstruction and data reduction executables are typically based upon frozen releases with a controlled and validated set of updates managed by the physics groups. In all cases, the content of the software, building and most validation procedures are documented.

All subsequent tertiary processing is carried out by analysis groups or individuals. The code management practices of these groups and individuals varies widely, although it is likely that the majority of the code is in the repository. Some groups tag the versions of the code used to perform the final analysis chain. It is unknown how widespread this practice is among the groups. Where documentation and tagging practices tends to fail, even for groups that otherwise follow good code management practices, is near the terminal points of an analysis, where late changes needed to satisfy internal review issues, for instance, may not get tagged, or worse, may not get checked into the repository at all.

6.1.1 Externally supported software

In addition to the software directly supported by the experiment, there are a number of products and tool-kits written and supported by third parties that either have been integrated into the CDF offline release and analysis structure, provide essential functionality, or have been requested by one or more collaborators to support analysis. A few familiar examples include the physics event generators (see the MC production web page above), root [18], which provides not only the most broadly used analysis tool-kit, but also the data persistency format, and the Neurobayes neural network product [19].

[3] Many details of the validation procedure and other code management processes and procedures can be found on the code management web pages [15]. Note that the validation procedures may be incomplete.

Some less familiar ones include various infrastructure products supported by Fermilab (SAM, `fcp`, `encp`, etc.) or CDF (e.g., `frontier`), CDF-supported interface packages to those products, special versions of debuggers, and special versions of `Perl` and `Python` needed to use some of the other products, to name a few.

Each of these externally supported products is packaged and distributed via the Fermilab-supported `ups/upd` products [20] from the central `upd` server, `fnkits.fnal.gov` . The packaging is performed by a combination of CDF and Fermilab personnel, depending upon the product in question.

Updates to most external products and offline releases are managed asynchronously, with the notable exception of the version of root used to define the data format. (Note that `ups` allows multiple versions of a package to be installed on a machine, so the version of root needed for offline releases need not be the same as the version analyzers use as an analysis toolkit.) Most products are actively supported.

6.1.2 Software distribution and build infrastructure

The build configuration tool-kit is based upon the `SoftRelTools` and `ups` products supported by Fermilab. Scripts in the `Distribution` and `Release` offline packages (which are supported by CDF) automate the process of creating releases from the tagged contents of the `cvs` repository, building the code, and creating tarballs for distribution. The packages also include bootstrap scripts for `ups/upd` and the other installation scripts, which can then be used to install and configure the software in an arbitrary location. All files are served from either `fnkits.fnal.gov` or `cdfkits.fnal.gov` via anonymous `ftp`.

The installation, build, and distribution procedures are well documented within the `Release` and `Distribution` packages, and on the code management web pages.

The only build platform currently supported is Scientific Linux Fermi (SLF) v5 [21], plus a set of additional packages required by CDF, run on x86_64 compatible architectures. The additional packages are documented under the code management web page [22]. The only compiler in use is GNU gcc v4.1.2, the native compiler of SLF v5.

At present, the run-time environment is also SLF v5.x with the CDF-specific addons. In grid applications, non-standard run-time libraries are shipped with the executable. The code can run under SLF v6.x in a compatibility mode.

6.2 Future

The primary drivers of changes in the software environment are the march of advances in hardware technology; the development of more complex, more capable operating systems needed to utilize the latest hardware; and changes to the functionality, interfaces, or support for external products. The offline group routinely managed and coordinated the effort required to keep up with these changes while the experiment was is in operation. In a preservation regime, the effort and expertise needed to follow this evolution is severely limited or completely missing.

The challenges for preservation will be discussed across the following areas: preserving the body of software, preserving external package compatibility, and ensuring that there are compatible build and run-time platforms and environments which preserve the exact functionality of the software.

6.2.1 Software preservation

Preserving the software within the `cvs` repository is the most straight-forward problem to solve. Fermilab is committed to preserving the contents of the repository indefinitely. While no further releases of `cvs` are anticipated, the more capable, open-source replacement system, the `Subversion` (`svn`) product (http://subversion.apache.org), provides a simple migration path and supports the same commands as `cvs`. While effort is available, CDF should investigate the risk of preemptive migration to `svn` versus waiting to a later time to determine whether migration will be necessary within the context of the full preservation regime.

Since the above applies only to software within the repository, CDF should make an effort now to ensure that all analysis software that needs to be preserved is archived in the repository.

6.2.2 External products

Externally supported products present a diverse range of risks to preservation. The packages supported by Fermilab, for instance, will continue to have some level of support for an indefinite period of time. CDF should discuss the status of each of these products with Fermilab, and wherever possible, obtain an explicit support policy for any that does not already have one.

The `root` product will continue to evolve throughout the LHC era at a minimum, although the support horizon for legacy versions is not long. The version of `root` used for data persistency (only recently updated to the v5 series) is already well off the most recent production release. Due to significant interface and functionality changes, Fermilab has advised against upgrading beyond the version of `root` currently used for data persistency due in part to a lack of the requisite effort and expertise on the CDF side. There are no impediments to further upgrades of `root` as an analysis toolkit.

The products supported by CDF face the same issues as the offline release and analysis software, and will be discussed together later in this section.

Of the remaining products, some will continue to evolve and have support into the future. A number of the generator packages, for instance, fall into this category. As is the case for `root`, support for any particular version is unlikely to continue for more than a few years. One strategy for dealing with this problem is to move to standalone generators or use generated event libraries so that we are not tied to legacy operating systems for event generation. It may be necessary in this case to support software to convert the generated output from some future data structure into one that CDF code can read, or to provide a direct interface between the new structures and the CDF simulation code. A few alternative strategies are to continue updating

the CDF software to more recent versions, which carries the risk that it may require more effort or expertise than is available at some future time; to migrate the build of the current product versions to newer operating systems and compilers, which is possible only in those cases where none of the build issues encountered requires effort or expertise beyond our means; or to preserve the existing binaries as-is. This last strategy was used for some products during the migration to SLF5, and can work in principle as long as future operating systems and compilers remain compatible with current x86_64 compliant instruction sets. CDF should perform a systematic analysis of each product to assess the risk associated with pursuing each of the above strategies. A few products have known risks to preservation. The `cafclient` product is not currently supported and cannot be built against or upgraded to newer versions of `Python`. The `dcap` package, which is part of the interface between the data handling systems and CDF software, is unsupported. Support for both of these products is the responsibility of CDF. Other products supported by CDF that have limited expertise still on the experiment include `discache_i` and `frontier_client`[4]. The legacy CERN libraries have been unsupported for a number of years already, but have continued to work in an as-is state.

6.2.3 Preserving build and run-time environments

As of the time of writing, SLFv5 expected end-of-life is 2017, while 2020 is the current end-of-life date for SLFv6 [23]. Migrating to SLF v6 will require a change in the compiler from `gcc` 4.1 to 4.4. By far the most challenging problem is preservation of the functionality of the offline release and analysis code through such changes in the underlying operating system and compilers. Unlike code supported by Fermilab or other institutions, the offline and analysis code does not have an organization that can provide for continuity of expertise or effort as the resources on the experiment migrate away.

The first strategy for preserving the build and run-time environments is to port the code to new operating system and compilers. Difficulties in this process typically arise from the greater standards compliance of newer compilers, or changes in implementation-dependent compiler choices, coupled with code that may be poorly written or that was in many cases originally written before the C++ standard was finalized. A minor additional complication is that some code is still written in Fortran.

Work on porting the most recent offline release from SLF5 to SLF6 is already well under way. Looking forward, it is instructive to examine the work that has been required to port code through the various operating systems to the present. The first point revealed is that finding sufficient expertise within the collaboration to fix compilation and validation issues has been difficult for the past several years, even

[4] The reader may notice that CMS uses a product called `frontier` that performs the same function as the CDF `frontier`. It is important to understand, however, that the architecture of the two systems differ in significant ways that would require significant code modifications in order to migrate from one to the other. It might be advisable to investigate the effort required to perform this migration.

before the attrition that is now occurring since the end of data-taking. Most porting fixes do not require much effort for someone familiar with the software. In cases where no expert could be found, however, code managers could not always verify that the changes preserved the intended behavior of the code. By some estimates, the port to SLF v6 is the last that is likely to be completed given the effort and expertise currently available within the Collaboration. We recommend that CDF adopt the porting strategy as far into the future as possible. In preparation for a time when this is not possible, however, we must consider alternative strategies. The most promising of these is to preserve the existing build and run-time platforms using virtualization. This strategy eliminates the need to maintain expertise in the CDF software in exchange for introducing a new requirement to maintain the virtual machine needed to execute the builds and run the applications.

Regardless of the strategy, we need a well documented and tested validation procedure that will validate the entire reconstruction, simulation, and analysis chain from raw data to final plots. Some number of unit tests is advised so that the source of a problem with the global test can be isolated.

There are a number of problems associated with virtualization for the purpose of preservation. Running legacy operating systems, for instance, exposes a site to security risks. As a result, most sites will not allow legacy operating systems to be run on machines that are accessible from the network. While fire-walling a private network can be used to isolate machines running legacy systems, such dedicated facilities are not an economically viable solution over long periods of time. If virtualization is to be considered a serious part of long-term preservation, it is important to develop the capability to isolate virtual machines running legacy operating systems within a shared grid or otherwise distributed computing infrastructure. This capability must be available by the time that support ends for the operating system of the last migration.

While we have been discussing the software within offline releases, it important to note that these same issues pertain to analysis software that is not maintained as part of offline releases, and to external products that are supported by CDF. Since support for most of this code is not centrally managed, the main challenges are to identify the critical code and the required unit and end-to-end validation procedures that can be used to ensure that the software operates in the intended way.

The final challenge facing preservation of the functionality of the software across changes in the operating system is the documentation of the maintenance and validation procedures within the CDF web pages. CDF should mount a systematic effort to consolidate the documentation required to carry out the installation, building, distribution, and end-to-end and unit validation of the software, and to eliminate extraneous, out-dated, redundant, or irrelevant information. This task must be completed as soon as possible before the expertise and effort drops significantly from current levels.

6.3 Resources and timing

As explained in detail in the previous section, the long term preservation of CDF code will require immediate effort from the CDF collaboration in the following tasks: investigate the usage of `svn` as software repository, investigate different strategies for external and internal software preservation, build and test appropriate validation procedures for the code and prepare extensive documentation for maintenance and validation.

For the first two items we will need significant support from the CS. On the CDF side, we estimate that at least 0.5 FTE x 1 year will be required, plus a similar amount of effort directed at constructing validation suites and ensuring complete and coherent documentation.

7 Job submission

7.1 Current system

The CDF job submission system is a product of the evolution of many systems and software products over the past 15 years, beginning with the initial planning of the Computing Division (CD)-CDF-D0 joint Run 2 computing projects in the late 1990's. The current system represents a good match to the computing systems in place at Fermilab and has served CDF well for the processing and analysis of data.

One of the interesting aspects of the job submission system is that it is not so clear what belongs to the job submission system and what is outside of the scope of job submission and is therefore part of other systems. There are many related hardware, middleware and software systems to consider when one is thinking about the scope and definition of a job submission system. In the current system and environment this can include the following:

* Job submission infrastructure
  - scripts
  - batch system
* Data handling system
* Data
* Databases
* Catalogs
* Grid interfaces
* Code
* Runtime environment
* Other

Some of the above can be thought of as part of the CDF code management or code system or are provided in some way by some other external entities. However, they will

have to be taken into account in order to properly understand the future development of the job submission system.

CDF and D0 built offline computing systems for Run 2 over many years. The initial implementation had separate systems for offline production computing on the CDF and D0 farms and other systems for the analysis computing systems. Many changes have been implemented, reflecting the long life of the experiment, the changing nature of computing, new techniques developed to reflect these changes, grid computing and the collaborative efforts required to utilize these systems, LHC computing and its impact on computing, among other things. Not all of these changes or updates are properly documented, but most of the current systems are documented reasonably well.

Figure 1 shows the current CDF offline system, as found on the web page www-cdf.fnal.gov/internal/ci/batch. The system consists of a set of front-end systems, worker nodes where the jobs run, and intermediate systems and middleware to provide access to the computing resources.

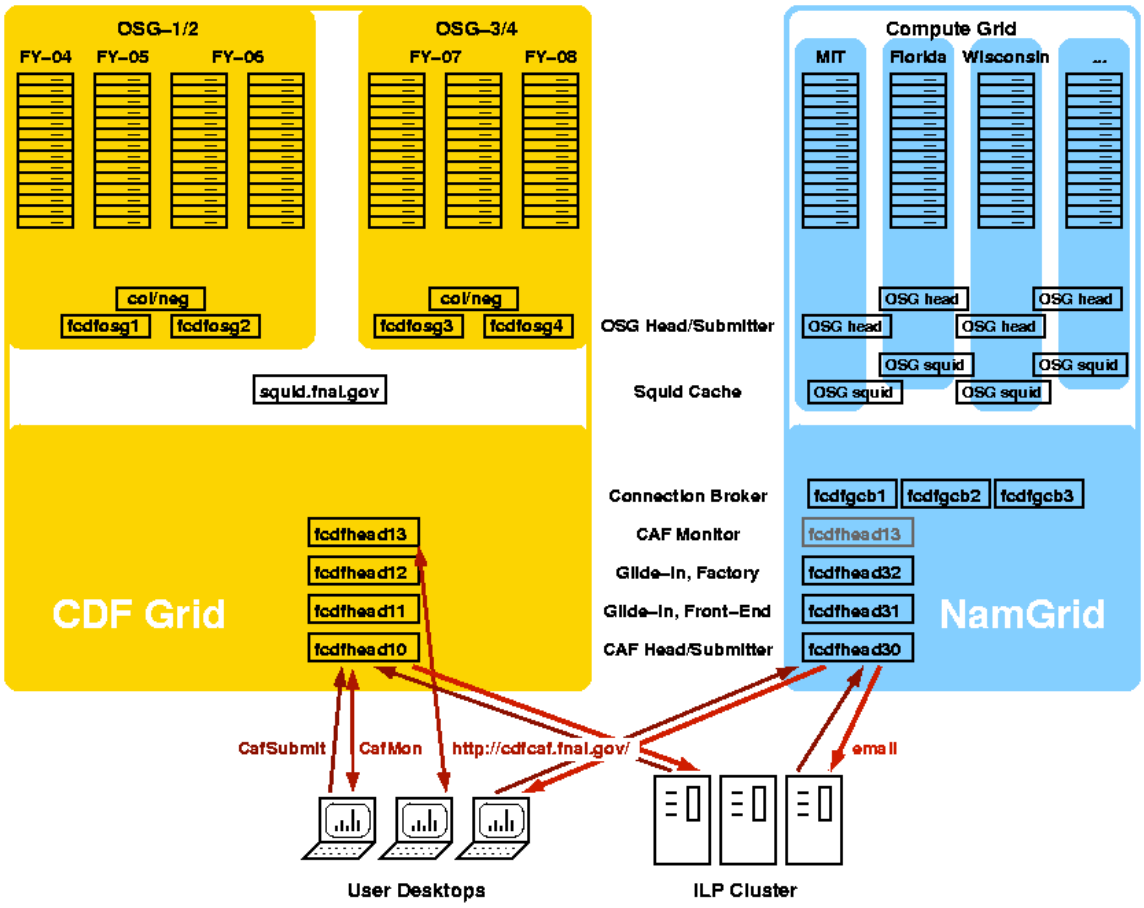

Figure 1: Current CDF offline system

It should be noted that many of the people who wrote and supported the systems

are no longer with CDF or the Fermilab Scientific Computing Division. In many cases they have moved on to other jobs or have retired. They may be working on other projects or other experiments and are not usually available for providing any effort. It may be possible to consult with some of them. In any case this has to be taken into account when looking at future plans for the job submission system.

7.1.1 CAF job submission and monitoring

Currently one person of the Fermilab Scientific Computing Division provides support for the CAF job submission system. The components of the system are mostly written in python. A job submitted to the CAF starts a SAM project to handle the delivery of data files, checks on things like the job length, and provides other general system activity. One interesting feature of the system is a custom kerberos module that is somewhat fragile and not well understood, but has been working for many years and has been upgraded when it has been necessary to do so.

The monitoring components are also written in python and use some daemons to provide information about jobs. There are various software packages and products that are used, including rrd and of course the batch system Condor. Again there has not been much maintenance required on this system in the recent past.

7.2 Future

The future of the job submission system and the offline computing system as a whole are tightly related. Nevertheless it is possible to provide some scenarios that may make sense for the future. One way to organize the thinking is to divide the future into near-term (4-5 years) and long-term ($>$ 5 years). This will be done in the sections that follow.

7.2.1 Near- term ($<$ 5 years)

In the near term it is likely that the current system can be maintained with nominal effort. The systems that are used will slowly upgrade and migrate as new software versions are installed, worker nodes and server computers are replaced or upgraded, batch systems and middleware are upgraded, etc. Much of this will happen in any case and has happened over the past years. There is experience with the changes and how they affect the CDF system and what the probable issues will be as they occur. Keeping up with and adjusting to the changes will take effort and possibly some money for hardware but they seem manageable. The REX department of the Scientific Computing Division has some plans for the budget for CDF computing for the next year and it will be examined, reviewed and revised as appropriate for that year and for upcoming years as part of a more general scientific computing portfolio management team activity.

7.2.2 Long-term (> 5 years)

Depending on how one counts the long-term begins in late 2016 or sometime in late 2017, as the run ended in late 2011 and one year has already passed since the shutdown. In any case the long-term is far enough into the future that there is time to develop a more detailed plan based on the experience of the past year and of the next one or two years with the CDF offline system, as well as watching the development of systems being used at the lab and elsewhere for the Intensity Frontier (IF) experiments and the LHC experiment, specifically the CMS experiment where Fermilab is hosting a large Tier-1 computing center.

There are two main options that should be considered for the long-term CDF job submission system. The first is to migrate the current system to the IF systems at Fermilab. This allows them to take full advantage of developments that will continue and have full support long into the future. The second option is to build a virtualized system that is "frozen" and can run in isolation and not be affected by changes in all of the packages and systems that the job submission system currently relies on. Obviously both of these options require effort and changes to hardware to produce a fully functional system.

The development of a system that aligns with and integrates with the IF systems is a mixture of developments that should be compatible with the evolution of CDF computing and changes that are required to fully integrate into the IF system. Many of the systems will be downsized and in some cases virtualized to adjust to the reduced demand and the ability to use virtual machines (VM's) to host these services. The CDFGrid and NamGrid systems can be combined. A project to produce a new job submission system for the IF experiments is being designed and, assuming that it goes forward, will be available in the next year or two. The CAFSubmit script/system can become a wrapper that connects to this new system and takes advantage of all of its features and support. Assuming that this transition is possible it will allow for very long-term support of the CDF job submission system. There are certainly some issues that will have to be explored and understood, not the least of which are operating system changes, Oracle interfaces, Root and its development over the years, SAM cache, dCache, Enstore, Condor, and other software systems that are not under CDF control in this scenario. Some of these will be handled automatically by the integration with the IF system (Condor, SAM cache, dCache, Enstore) and others are likely more problematic due to the fact that the CDF code will be affected by the changes.

The second option is to build a virtualized system, similar to the system designed and built by BaBar for long term data access (the Long Term Data Access (LTDA) system), see[3] section 4.1). It is not necessary that the system be identical to the BaBar system and in fact advances and changes in VM and Cloud technology may allow for variations of this system that might be more appropriate. Other options include (1) using CHROOT/JAIL to run jobs compiled on older OS such as SLFv5 to run on newer versions of SLF, (2) running virtual machines on the grid or the cloud and (3) maintain a system running old(er) operating systems but not in a virtualized

environment. All of the options will require further study.

7.3 Resources and timing

To maintain the current system of job submission is likely to require less than 1 FTE, as it does currently. That effort may increase as the system ages and is affected by changes in hardware, operating systems, middleware, and software products that are required to continue to run CDF jobs. Migrating to a new system requires an immediate effort to port to the new system, followed by a much smaller effort to maintain the system. This is especially true if the new system is essentially the same as the IF systems and if D0 makes the same transition. It is hard to make a good estimate of the effort required but the BaBar experience gives some idea of the order of magnitude. The BaBar effort estimates include more than just job submission but does not include some of the central computing activity such as batch systems, storage systems, hardware maintenance, database servers and so forth. The estimates given by Babar experiment are 10 FTE (2009), 7 FTE (2010), 4 FTE (2011), 2 FTE (2012) and 0.5 FTE (2013 and beyond). Estimates for the CDF case will only be possible when specific solutions are specified. As far as timing is concerned, in case we decide for the VM farm, the implementation should start as soon as possible, while CDF has significant computing resources to be dedicated to the new farm.

8 MonteCarlo production

8.1 Current system

The CDF Monte Carlo is composed by most of the basic components of the CDF offline software. It is written mainly in C++, although it includes several Fortran sub-programs, and it is supplemented with a significant number of shell, tcl and perl scripts for job control and module interfaces (see Fig. 2).

The master shell script running the CDF MC is MCProd. This script accepts a number of user-supplied arguments and options and reads in an input file with the general definitions of the MC dataset to be produced. Based on those it selects the running mode (e.g. event generation alone, generation + simulation + production), sets up the communication of the program with the CDF database, where various experimental parameters relevant to the simulation are stored, and initializes the random number sequences that will be used by the various program units. The core of the CDF MC is composed by three executables: cdfSim, TRGSim++ and ProductionExe. cdfSim does all of the physics simulation. It contains two main parts: i) a collection of event generators and decayers to generate the particles and ii) a set of methods using GEANT 3 and GFLASH routines and functions to model the various components of the CDF detector and to transfer the generated (HEPG) particles through them. The simulated responses of the detector components are digitized like the experimental data

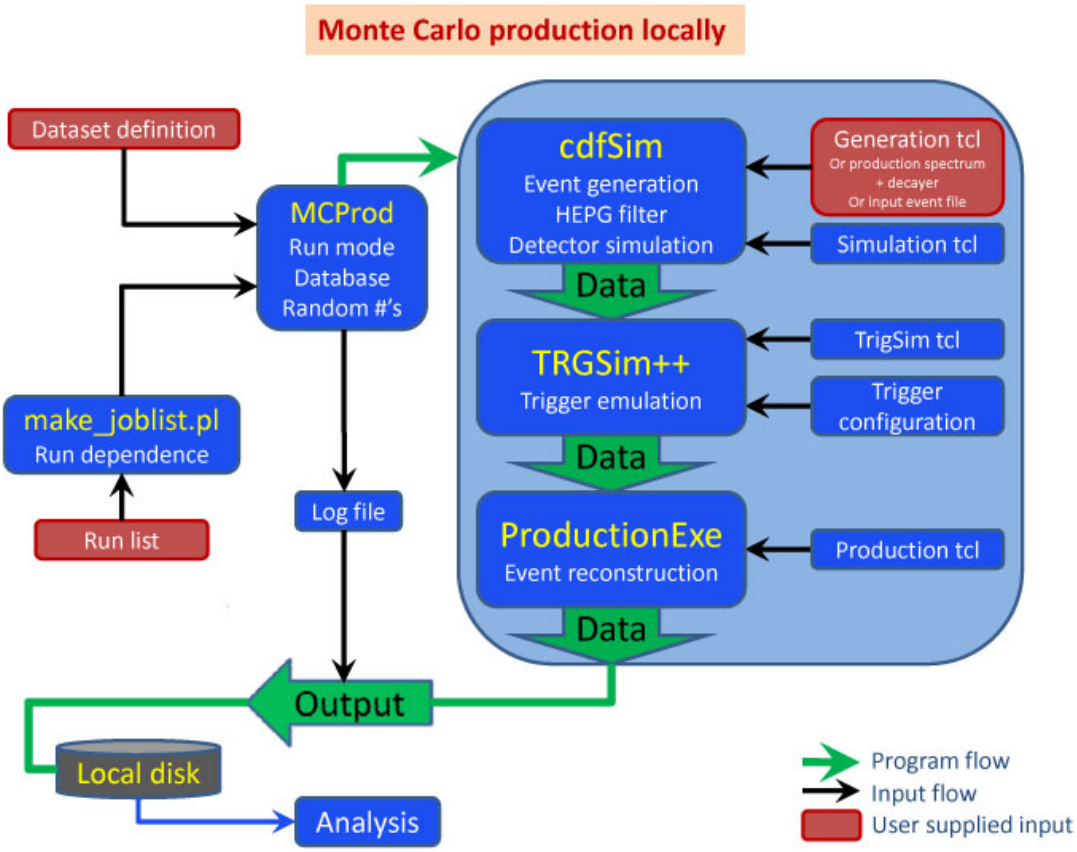

Figure 2: CDF MonteCarlo production flow.

and written out in a .root file. TRGSim++ is used optionally to emulate (deterministic) trigger signals. It reads in the output of cdfSim, fetches the trigger physics tables from the CDF database and evaluates the trigger bits for each event from the cdfSim digitized output data. The TRGSim++ output is appended to the cdfSim output into another .root file which is passed over to ProductionExe. This is the same executable used to process the real data from the digitized detector output units. It reads in the output file from TRGSim++ and, from the digitized detector data, it reconstructs the physics events in exactly the same way that the true physics events are reconstructed from the experimental detector data. ProductionExe writes out all results from cdfSim, from TRGSim++ and from the event reconstruction into a final .root file. The following generators are provided within cdfSim: Herwig, Pythia, Isajet, Bgen, HQGen, MBR, WGrad, Vecbos, WbbGen, Grappa, FakeEvent. Decay packages comprise EvtGen and Tauola. Alternatively, a StdHep-formatted input file containing physics events generated with some generator external to the program (MC@NLO, AlpGen, CompHep, Sherpa and MadGraph-MadEvent are currently available from the CDF MC Web pages) can be converted into AC++ framework-compatible format (HEPG)

using the hepevt2hepg executable, also available in the standatd CDF MC configuration, and then it can be supplied to the CDF MC core executables to run the events through detector and trigger simulation and through reconstruction. The standard CDF MC configuration allows for extension of this mode into a full run-dependent simulation, as with a generator built-in to cdfSim, using input events stored in many StdHep-formatted files.

The average time for the full chain of generation-simulation-production-output of a typical MC event to take place is a few seconds. Therefore, to produce a sample useful for physics analysis, typically of the order of some millions of events, the whole job must be divided into segments and these must be distributed as separate jobs over the grid. The challenges for this task are i) to ship to the nodes of a remote farm a self-consistent "tarball" (compressed archive including all files necessary for the run) that can run standalone while maintaining the communication with the home database to load the run-time needed calibration parameters and trigger tables and ii) to collect the output back to the home servers dedicated to temporary MC data storage ("data assembly servers"). The submission command of the MC tarball to a farm is executed by the script submit_MCProd (see Fig. 3). This script accepts a number of user-supplied arguments and options, some of which are passed over to the farm manager program (tarball name, farm name, job queue) and some others to the master shell script run1segment which controls the job. This in turn runs MCProd for each job segment, concatenates the section output files corresponding to different runs of the input run list into one production file for the whole job segment, gathers the segment's "metadata" (number of events, size of the output data file, check sum of the output data file, starting and ending time stamps of the segment job, generator used, user's kerberos principal etc.) which are written in a small text .ok file, and finally copies the output back to the home server. From there, after user's validation, the output files are sent to one of the two upload servers. On the upload servers a set of scripts run as Cron jobs perform further checks and trigger the automatic upload to the tape system. This process is split in three parts: i) the auto-import part copies the production .root and .ok files from the data assembly server to the upload server and checks the size of the production files, ii) the upload part checks the metadata and copies the metadata to the database and the files to the tape system, and iii) the auto-cleaner part deletes the files from the data assembly server after they are successfully uploaded. The auto-import and auto-cleaner parts are entirely written in python. The upload part is mostly written in python scripts along with some shell scripts. All three parts write out daily log files for monitoring, which are periodically read by shell scripts computing numbers and producing plots for the Data Handling Web pages.

To give a raw estimate of the resources currently needed to produce Monte Carlo events, to generate 3M of Pythia $H \to b\bar{b}$ events, about 700 segments on the grid (worker nodes) are needed, each producing a file with size of about 1.5 GB. Each segement lasts between 6 and 10 hours, depending on the working node CPU. To store the output, 1 TB of disk or tape is needed for the production files, and a slightly smaller amount for the analysis ntuples.

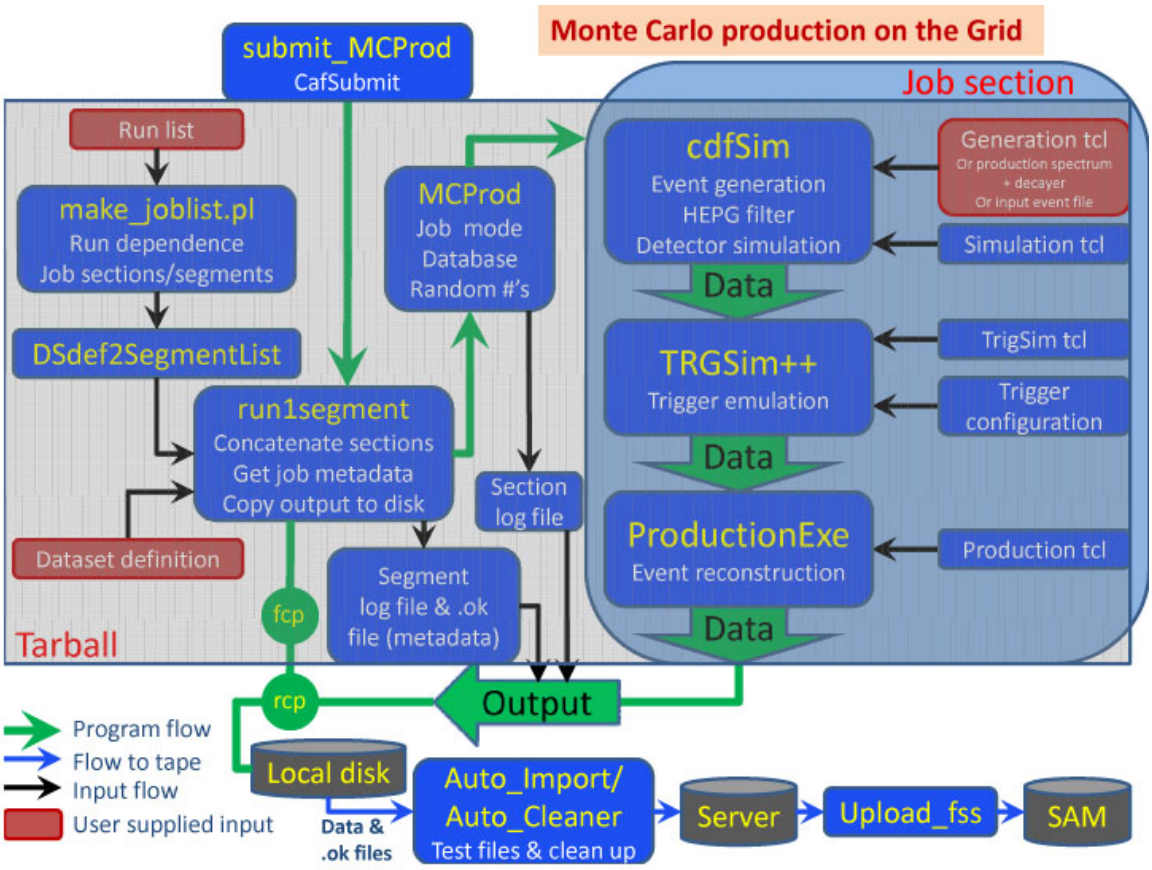

Figure 3: CDF MonteCarlo production flow.

The current Monte Carlo production system is based on the CDF organization in physics groups. Every group appoints a representative person responsible for producing samples upon requests from users, uploading the samples to tapes, and updating the MC page in the group's internal pages for the new samples. Each group is currently assigned half data assembly server with a local disk capacity of 11 TB. The two upload servers have a local disk capacity of 5.5 TB each. After uploading to tapes, the MC production output is converted to an analysis-level ntuple by the CDF Offline Operations team upon request from the MC representative. The ntuple files are also uploaded to tapes by the Offline team, using blank files for this purpose. An important step in ntuple file uploading is the renaming of the ntuple files, which are simply named by the ntuple code according to the number of the corresponding job section, to "data file catalogue" (DFC)-compatible names like the production files. This is done automatically by a script given the names of the original production files in the input file list of the ntupling job.

8.2 Future

In the remote future we need to assume the Monte Carlo production and ntupling effort to be transferred to the individual users, instead of the physics groups and the Offline group. At best, after the pages of the current physics groups will freeze, a person from Fermilab's Computing Sector could be charged with the responsibility to grant the future user, upon request and after authorization from a CDF or Fermilab delegate, with a temporary access to update a future Web page special for new samples and with a temporary access to a data assembly server to collect the output from the MC production and the ntupling jobs. That person could also monitor the user's jobs and provide a limited support for possible trouble-shooting. The minimal requirements necessary for the "future user" system are:

- The MC production and ntupling documentation pages must be available, with clear instructions for how to run each one of the two sequential jobs.
- The MC production code and the code to produce analysis-level ntuples must be operating. Each code should be available in the form of a Web-accessible self-contained tarball, for the future user to be able to download them on a local analysis machine, run a few tests to design his or her MC production job, and then submit first the MC job, upload the output, then submit the ntupling job, and then upload the ntuple.
- A farm where either tarball can run successfully must be available. This assumes a kerberos authorization for the user.
- The necessary information currently stored in the offline CDF database must be accessible from the farm nodes.
- A communication system (FroNtier or a system with the same functionality) between the database and the farm nodes, which queues the queries and caches locally the query results for times longer than the job time limits to avoid database overload, must be available.
- An utility for remote file copy to disk (fcp or similar), queuing copy requests on the working node to avoid copy time-outs, must be operating on the farm nodes.
- A data assembly server accessible from the farm nodes, with a local disk capacity of no less than 5 TB, must be available.
- An upload server with a similar local disk capacity and the auto-import, auto-cleaner and upload scripts installed and Cron jobs running to upload the files to tapes must be available. The Cron jobs can be run by the person monitoring the user's jobs only for the time that the user is uploading his or her files.
- The SAM database must be available for uploading files and cataloguing the new samples.

8.3 Resources and timing

The maintenance of basic resources like the CDF offline and the SAM databases, a farm supporting the CDF software etc. are discussed elsewhere in this document. Required resources special to MC production are a person from the Computing Sector to intermittently give access to users and monitor their jobs; a data assembly server, preferably with the up-to-date CDF software installed and operating, which can also host the new Web page with the description of the future datasets; and an upload server with the auto-import, auto-cleaner and upload scripts installed and operating. The time scale for the transition from the existing CDF MC production system to the "future user" system will depend on the life time of the current CDF organization in physics groups and the associated availability of persons to provide the MC representative service. This transition is not anticipated within less than about two years from now.

9 Preservation of documentation

The term *documentation* comprises a wide variety of information, from the technical details about data taking and maintenance to the instructions on how to run analysis code. The preservation of all the knowledge necessary to perform an analysis from scratch is a fundamental step of any data preservation project. In the following we will review the current organization of CDF documentation and then make a proposal for the future.

9.1 Current system

Extensive information about CDF experiment is maintained in the website `www-cdf.fnal.gov` hosted in the `fcdfweb` machine; this website is divided into a public section and a private one. The public one contains information that can be released to the general public, such as results made public by the collaboration, descriptions and events from the CDF detector. The private section is divided into four main areas: *online*, with information about data taking and details about CDF detector and trigger; *computing*, with information about all CDF computing facilities and tools - data handling, data production, code, job submission on the Grid; *physics*, with details about analysis (Bottom, Higgs, Exotics, Top, ElectroWeak and QCD) and common techniques and tools (High-pt b-tagging, Jet corrections, particle reconstructions, etc ...); *organization*, with information about CDF organization, meetings, talks, internal notes, publications, theses, etc..

The CDF internal notes and web-talks archives, accessible from the CDF webpage, contain detailed documentation about CDF analysis and should be indefinitely preserved. It has also to be mentioned that a lot of documentation is present on users private areas: it should be collected and properly organized.

Essential information is also contained in the CDF online web server http://www-cdfonline.fnal.gov/: data acquisition and sub-detector e-logs, details about detector systems, data acquisition and trigger operations are stored in these website.

The webpages are written mainly in htlm, but many twiki and tiki pages are also present. It has also to be noted that links to websites hosted on offsite servers, maintained by CDF institutions, are often present.

As far as maintenance is concerned, CDF website does not have a centralized organization. Different analysis or technical groups maintain their own pages.

9.2 Future

We will first highlight the requirements for analysis preservation in the different physics groups. Then we will summarize the proposal for a new website where all the necessary information to access and analyse data in the long term future will be organized. Finally, the usage of the digital library INSPIRE to preserve CDF internal notes will be described.

9.2.1 Top and exotic physics analysis preservation

The top and exotic physics group study the properties of the top quark and search for physics beyond the Standard Model. The common needs for data analysis, such as Monte Carlo simulation, data production are already standardized and all outputs are saved to SAM. A standard framework is developed for top analysis'. This package exists in the cvs and is well documented [25]. The essential tools for recontructing leptons, jets, missing transverse energy and finding b quark jets are standardized and already exist in the cvs. However there are individual codes, tools and data saved in local computers to be saved.

The top and exotic group are interested in analyzing events with different final states, such as events having single lepton and multiple jets (lepton plus jets) or events having two leptons and multiple jets (di-leptons), events having no lepton but only jets (all hadronic) and events having large missing transverse energy. Only the groups of lepton plus jets and di-leptons have common tools.

For the analysis using events with one lepton and multiple jets there is a generalized procedure for background estimation, the *Method II for You*. This is well documented in top physics web page [26]. To reconstruct top events a common tool, the top mass fitter, has been developed and should be preserved: it can measure the top mass but, given a top mass value it can also be used to identify the jet produced in association with the leptonically decaying W.

The majority of di-lepton analysis are based on the same code for for event selection. This code is currently saved on the web [27] and should be preserved in CVS instead.

The TopNtuples are skimmed for physics analysis using single lepton plus multiple jets and di-lepton plus multiple jets. All files are saved in SAM. Individual analysis might have data/MC files in their local area. These files have to be collected and saved.

Another common tool used in the top and exotic group is the to do the Kolmogorov-Smirnov test for histogram shape comparison. The code is documented in a web page([28]).

For the long term future it would be important to save and maintain the detailed analysis procedure together with the data. However due to limited manpower the analysis to be saved have to be prioritized. As an example, analysis specifically correlated to the proton-antiproton initial state should be preserved.

Here is a list of physics analysis which the top group aims at preserving:

- Production cross section
  - Measurements in the lepton plus jets, the di-lepton, all hadronic and missing Et plus multiple jets channels and their combination;
  - Production cross section of single top and combination from different channels.
- Top mass measurements
  - Measurements in the lepton plus jets channel, di-lepton channel, missing Et plus multiple jets and their combination.
- Forward and backward asymmetry of the top production
  - This analysis is particularly interesting due to its inconsistency with the theoretical predictions. It is also very specific for the Tevatron, because at the Tevatron the top is mainly produced via quark-antiquark annihilation , while at LHC top pairs are produced mainly via gluon-gluon fusion. This analysis is performed in the lepton plus jets and di-lepton channels; both should be preserved, as well as their combination.
- Measurements of top spin correlation
  - this is also a measurement specific to Tevatron. It is carried on both in the lepton plus jets and di-lepton channels. Currently no combination is planned.
- Analysis performed on the full dataset.
- Other analysis whose authors would like to preserve.

There should be a centralized area for all saved analysis. Each analysis area should contain a tarball and detailed instructions to run all the analysis scripts. These analysis areas should be linked from the data preservation web page.

9.2.2 Higgs physics analysis preservation

The LHC experiments have recently eclipsed the Tevatron sensitivity in all search channels for a standard model Higgs boson. Since then, 30 % more data has already been collected and a doubling of this data set is still expected before the end of this year. So, it is difficult to make a case for preserving the CDF SM Higgs searches.

That being said, the CDF Higgs search is already being preserved at the level of the SM Higgs boson search combination. All of the signal and background shapes templates, rate predictions, and systematic uncertainties used in the search a stored in a tarball along with the statistical analysis package Mclimit. So, for simple perturbations to the rates of signal or background processes, the effect on the result can be easily tested and results can be updated. Similarly, if someone completes or improves one of the SM searches it can be easily added to the combination. Studies of the affect of modifications to template shapes can also be studied, but rigorous studies of the affect of the the full analysis selection will not be easy to update.

In addition, preserving the CDF CVS repository will automatically preserve a large portion of the code used in the SM Higgs boson search. For example, two of the most sensitive low-mass channels (where the Tevatron is most competitive) use the same analysis package (WH analysis package, WHAM) and it is stored in CVS and well-documented on the cdf twiki (http://www-cdf.fnal.gov/htbin/twiki/bin/view/WHAM). There are several searches for Higgs boson beyond the standard model and the requirements for these exotic searches are included in Sec 9.2.1.

9.2.3 QCD and EWK physics analysis preservation

The SM group is characterized by the large variety of analyses that have been and still are carried on. The major analysis streams may be classified as:

* Jet studeis
* V+Jet studies (including Heavy Flavors)
* Photon studies
* Underlying Event studies
* Minimum Bias studies
* Diffractive studies
* Central Exclusive studies

Such studies make use of very different data and MC samples, and exploit different parts of the detector. For example diffractive and exclusive studies make large use of the forward Mini-Plugs, of the Roman Pot spectrometers placed along the beam line, and of the Beam Shower counters, while other typical measurements attain only the central detector. For this reason, although there are some common tools, many functions are tailored on a specific study; often also MC production is custom made and the samples are not present in SAM. Furthermore the datasets and triggers employed are many and with very different peculiarities. In general the analysis work require very different procedures, tools and expertise.

With these starting conditions it is clear that the goal of preserving completely a large number of analyses would require a manpower that may hardly be found:
- validation the the original code is very time-consuming;
- large part of the original code still reside only on personal desktop computers;
- full documentation is possible only for recent analyses or when the original authors are still active in CDF.

On the other hand, it may be possible to document a certain number of general tools and provide general examples to run a generic code. This would require some effort and dedicated manpower. Once this is granted, it is possible that a limited number of authors will dedicate some time to document in more detail their own analyses and eventually provide the missing data and the relative documentation.

With this goal in mind we point out some frequently used data formats, tools and necessary information without which no analysis may be appropriately carried on. Note that a large part of the planned work needs cooperation from group authors.

* The StNtuple framework is the most widely used within the group. It is documented in the wiki web pages [30]. We plan to provide examples of a generic code that runs on this data format, together with snippets of code to access specific data. Such code must be associated with user instructions and all examples should be previously validated.
* the two photon analysis packages, `diphoton` and `wgjj`, which are now in the CVS repository, should be preserved.
* The most relevant datasets should be documented in detail, while secondary datasets may eventually be described within their specific analysis (data [31] and MonteCarlo samples [32]).
* The most relevant triggers should be described and their efficiency documented. The trigger simulation should also be described (in particular when it does not reproduce the trigger as expected). An example may be found at this page [33] and links therein. Most of the documentation that is now available in the "Online" web page `http://www-cdfonline.fnal.gov/daq/` is to be preserved for this purpose.
* The tcl and log files for the generation of MC samples may be provided for the most common datasets together with their descriptions; this will be useful both for allowing generation of new datasets and as documentation of the existing ones. For older samples they are already available in this page [34].
* The jet energy correction package is already well documented [35] but it should be made clear that it is left to physicists to decide what correction level to apply depending on the analysis goal and that different corrections specific to HF jets have often been used.
* GoodRun lists should be provided and documented for various types of analyses, together with the corresponding trigger luminosities for the samples mentioned above. Note that many analyses may have actually used their own run list. The official QCD lists are available at this page [36].

* Jet clustering tools for StNtuples may be provided if still available [37].
* There is a quite complete documentation of the central and plug calorimeter and of the jet clustering packages that may be updated and reorganized [38]. The same is true for diffractive studies [39] but a large part will have to be recovered from off-site computers.
* The data from the low-energy-scan runs needs to be documented from scratch and validated.
* The (COT) track momentum scale should be described and the curvature corrections [40].
* The calorimeter energy scale for leptons should be described.
* The response and calibration of all calorimeters should be described.
* Baseline cuts and their efficiencies should be provided for:
  - leptons (e, mu) [41]
  - high and low pt tracks [42]
  - calorimeter tower thresholds
  - jets
  - photons
  - missing E_T
  - high pt b-tagging [43]
  - primary vertex position and quality [44]
  - diffractive event selection
  - exclusive event selection
  - underlying event selection

  for all these items it should be made clear to which study (or set of studies) they refer. Multiple versions of the same item may be acceptable for different analyses. Some of the items above are of interest to all the physics groups. Part of the information is already collected here [45].
* Validation plots should be provided whenever possible: some are already available on the web [46].

Part of the information to be provided has to be recovered from personal desktops and in the internal notes. Part is transmitted only in oral form: it is crucial that this information is collected from single authors and organized organically with the rest.

For what concerns the preservation of single analyses (assuming that someone is willing to do it), instead of revalidating all the code and spend time in documenting it, it may be possible in some cases to provide only working functions that reproduce the required functionalities. For example some dedicated techniques such as background subtraction for specific analyses may be preserved as a whole by providing an Ntuple file and a root script to read it. In general this method may be useful also for all those cases when the original analysis was done by reading data from a format different from StNtuple (like Production DSTs, TopNtuple etc).

There are two types of documentation to be provided. The first consists in the technical instructions on how to run the code and on what each peace of code is

intended for. The second consists in more general information concerning everything else is useful to know to complete a measurement. This latter part is more or less well documented in the internal notes, depending on the analysis. The less code, functions and procedures will be recovered, validated and preserved, the more becomes important to detail this part of the documentation. In order to provide an understandable level of information, ideally we should not limit ourselves to collect the available information in a single place (eg a web site). It would be much more useful to dedicate some time in skimming and comparing the documents, updating the information and organically reorganize everything. This process alone may well require several months of work by a dedicated person.

9.2.4 B physics analysis preservation

To achieve optimal sensitivity goals for a B_s mixing measurement, the CDF B group formed a task force in early 2004, known as the B Physics Analysis Kernel (BPAK) subgroup. The subgroup was charged with a multi-prong taskset to achieve maximal performance in:

- Track reconstruction: using L00 hits, track refitting with material integration options, increase acceptance with silicon standalone tracks.
- Vertex reconstruction: using event-by-event primary vertex and secondary vertex algorithms for low p_T b-tagging.
- Particle ID: using dE/dx and ToF in likelihood combinations
- B Triggers: incorporating instantaneous luminosity dependent requirements, to remain within B group trigger bandwidth budget
- Monte Carlo production: incorporating up-to-date generators / decayers, detector and trigger simulation software.

Most of these goals were met by 2006, resulting in a compilation of official tools in the `BottomMods` and `CharmMods` cdfsoft packages. A standardized `StNtuple` format for B group analyses, `BstNtuple`, was created. Following the B_s mixing discovery the BPAK group was dissolved.

Beyond 2006, the B physics analyses used the `Bottom/CharmMods` tools followed BPAK guidelines given in numerous B_s mixing analysis supporting notes. The `BstNtuple` format evolved in a semi-standardized way until Winter 2012 when an official version, **v90**, was released. This version is the most complete one, enhanced with a better tracking approved by the CDF Joint Physics group, track covariances and calorimeter jets, to name a few. An up-to-date twiki page, http://www-cdf.fnal.gov/tiki/tiki-index.php?page=BStntuple.Status, documents the details. The CDF Note 10771 serves as a very detailed `BstNtuple` manual containing code-snippet examples to access various information, e.g. GENP, PID, tagger, and perform vertex fits. It also

documents various procedures to create strips, mini-ntuples etc. and comes with an example tarball to perform a simple analysis.

The B Monte Carlo production framework evolved independently from the `mcProduction` framework used by the rest of CDF. It had several disadvantages, the most serious one being its dependence on cdfsoft due to which the B MC jobs could not be run on generic grid sites such as NAmGrid. The CDF Note 10307 documents the merger of the B MC framework into `mcProduction`. Implementation of an up-to-date version of heavy hadron decayer, `EvtGen`, is documented in CDF Note 10720. The B Monte Carlo webpage, http://www-cdf.fnal.gov/internal/physics/bottom/b-montecarlo/, provides instructions to create and validate B Monte Carlo samples. Most of the BMC components, e.g. runlists, SVT beamlines, input particle spectra, are frozen to their standard values. A large inclusive BMC sample is under preparation to facilitate background sample composition studies for future analyses.

The immediate goals for B group documentation is to add several more extensive BstNtuple analysis example codes to the current basic tarball. These codes will be accompanied with newcomer-friendly usage instructions. Similarly, more detailed examples of BMC production and usage instructions are to be provided on the BMC webpage.

In a longer term a more extensive compilation of analysis tools and their usage would be prepared. This task would be initiated by a group of current experts, e.g. analyzers who have used a certain tool recently. Starting from a set of supporting analysis notes information would be extracted into an analysis tools documentation. The tools to be documented are:

- Lepton ID: Up-to-date information on the usage of the dE/dx universal curves, ToF and the likelihood combination codes and their usage. These are already available in BstNtuple v90 but lack detailed usage and benchmark information.

- Taggers: The most recent knowledge of various components of the Opposite Side Taggers (OST) and the Same Side Kaon Tagger (SSKT). Several analysis supporting notes authored by the B_s Mixing Group have these information. However, information on codes and usage is largely unknown. The β_s analysis group has the current knowledge.

- Tracking / vertexing:

  - Document usage instructions for BstNtuple-level tracks, e.g. lower p_T tracks.
  - Provide instructions to calculate event-by-event PV and reconstructing Secondary Vertices with multi-step CTVMFT fits.

- Triggers: Ensure that the B trigger info. given in BstNtuple manual is complete. Document instructions to obtain historical information on paths.

- Monte Carlo:

  - More detailed examples with instruction sets. This should evolve into an automated validation suite.
  - Concise information on generators in use, input particle spectra, particle filter methods, manipulating EvtGen decay etc.
  - Information on SVT beamlines usage.

- Analysis tools:
  - Multivariate tools: B analyzers mostly use ROOT MLP neural net and BDT tools from the TMVA (http://tmva.sourceforge.net/) package. Some analyses also used the NeuroBayes NN package. Save and document one example code, e.g. $B \to \mu\mu$.
  - Likelihood fitting: B analyzers either developed their own fitter framework from scratch, i.e. using TMinuit, or used RooFit toolkit to develop custom fitters. Save a code from each type and document.
  - Statistical treatments: Several analyses, e.g. $\sin 2\beta_s$, employed very complex statistical methods based on pseudo-experiment techniques. Save a code with documentation.

Combine the documented tools into an analysis code to use for validation. This would use most of the important information in the BstNtuple worth validating. With help from the current analyzers prepare an official tarball with usage instructions.

9.2.5 New CDF website

Current CDF website contains everything a physicist has to know to understand the experiment and perform an analysis from scratch, but it is not easy to browse for a non-CDF user; moreover, in some cases information is outdated, and links - especially those pointing to pages hosted on offsite servers - are not working. For the long term future we propose to re-organize and complement the current CDF website; our target is a physicist, not necessarily from CDF collaboration, who aims at analyzing CDF data in the far future ($>$ 10 years from now), when likely CDF experts will not be available to guide through all the analysis steps.

In a first phase the new website can be divided into a public and a private section, as it is now. Many of the webpages of the current public section can be moved to the new page as they are: description of CDF detector for the general public, images and movies, public results. A new section about the archival mode should be added, with a description of the long term data preservation project and instructions on how to access CDF data. Here we should say the requirements to be eligible to analyze CDF data (now only CDF collaborators can access it, but in the far future we may want to allow a wider public) and how one can obtain an account to access data and computing resources at Fermilab. The public section should also contain information about CDF outreach program (see section 10).

The private section could be organized into the following main areas:

- detector: description of CDF sub-detectors; most of this information is already available in the current website, it has only to be re-organized.

- DAQ and trigger: details about trigger system; about data collection and organization in streams according to the trigger. A very useful information which is often difficult to access are the details about trigger selection. One can easily reach the trigger table of a specific run, and see the trigger cuts, but the exact meaning of the variables has to be retrieved from internal notes of from the code (for L2 and L3 triggers). Instead the meaning of the different trigger primitives could be collected on a dedicated page. This is a task which could be performed by a student, supervised by a senior CDF member.

- operations: all the content of the online web server (e-logs, operational details about detectors, daq and trigger, cvs repository of the online code, etc...) should be preserved.

- data

  - description of CDF data formats (raw, production, ntuples); sections dedicated to detailed description of the different flavours of ntuples. An essential information, which is often difficult to retrieve, is the meaning of the variables stored in the ntuples; this information should be collected on a dedicated page. As for the trigger, this task could be assigned to a student;
  - instructions for data access;
  - data quality.

- sections dedicated to the different analysis groups; for ease of browsing, these pages should have the same architecture; physics groups representatives should be responsible for these pages, with the help of students.

- code: description of CDF code; link to the CVS browser; instructions on how to download the code and build an executable; validation procedures.

- job submission: description of CDF computing resources; instructions to access data and to submit an analysis job;

- monte carlo production: the current MC page already contains detailed instructions for beginners; it should only be checked and updated.

- documentation: this section should contain links to internal notes and webtalks archive, as well as to public results; it should describe the procedure to create and upload a new analysis note. Analyzers should be encouraged to create a wiki page to document their analysis.

The re-organization of CDF documentation has high priority, given the natural reduction of the collaboration; besides physics groups, representatives for the different main areas highlighted above have to be found; participation of students is recommended: they will perform a fundamental service task for the collaboration, and learn at the same time.

9.2.6 INSPIRE for CDF

An essential component of Data Preservation for any experiment is to store and access all papers, notes, reports, written in the past as well as future documentation. Following the example of other experiments as BaBar, HERA experiments and D0, we are looking into porting all CDF Notes , currently stored on a dedicated Web Server at CDF, to INSPIRE, the SPIRE replacement for the 21st century, built on modern, open source "Invenio" software which allows to build a digital library or document repository on the Web and handle mySQL, Python and more. Four High Energy Physics laboratories (CERN, DESY, Fermilab and Slac) signed an agreement to use INSPIRE as documents storage system. Already more than a million records , such as papers, preprints, conference proceedings, reports and thesis, have been inserted. INSPIRE provides support for additional information tied to the original paper: for example, plots that did not appear in the publication and other support files and even extra data. INSPIRE also allows public and password protected level of access, so that internal notes can remain internal to the collaboration, if desired.

INSPIRE is ready to store and support the CDF publications.

The transfer of note into INSPIRE is labor intensive. We need to extract some metadata (title, abstract, author's list...) from our mySQL database and then convert the metadata into INSPIRE MARC (Machine Readable Catalog) format . INSPIRE does not support directory structure, so we'll need to tar up each note and its auxiliary material and send the tarball to the INSPIRE team that will them upload them into the official catalog. This whole process could be complicated by lack of uniformity in the metadata. CDF has around 10500 notes and some more in progress, due to current analysis.

9.3 Resources and timing

The preservation of the documentation is a responsibility of the CDF experiment and should have the highest priority in the long term preservation plan, given the natural reduction of the collaboration in the following years. The reorganization of all the necessary information in a new website will require at least 4 FTE for 1 year (0.5 FTE for each of the 4 physics groups representatives, plus 4 x 0.5 FTE for the other areas to be covered in the new webpage). These people should be chosen among CDF senior collaborators and should be supported by students. On the computing side, we will need long term support from the CS for the CDF webserver; moreover, as we are planning to create wiki pages for the new sections, we will need the CS to support

wikis in the long term future.

As far as the preservation of CDF notes in INSPIRE is concerned, since the CDF person works on this project part-time (0.20 FTE), we anticipate having a first pass of a subset of notes ready to be uploaded into INSPIRE, in about two months. The whole existing sample could be ready by the end of the year, if no major unforeseen issues arise. After all the existing CDF notes have been inserted, we might want to discuss the benefits and feasibility of inserting future CDF Notes directly into INSPIRE.

10 Outreach

After the CDF collaboration ceases to exist, and in the light of new competitive high-p_T data from LHC, our experiment will gradually become less sensitive about data ownership and its exclusive use. The public release of CDF data for scientific analysis – a long-standing request of the phenomenology community – will not necessarily be a desirable or easily achievable task, mainly because of the complexity of our data and the extensive corrections/calibrations/normalizations that take place at analysis time. Nevertheless, a small amount of CDF data can be released for educational and outreach purposes. The latter release of CDF data will be easier and faster because

1) the data doesn't have to be complete or properly documented
2) no claims of new discoveries or measurements can be made with the data
3) the format(s) of the data can be very simple
4) the data will not lead to any peer-reviewed publications

In terms of public service, the benefit for the society from the release of the data will be great. High-school or middle-school students will understand the field of experimental particle physics, and hopefully will choose a scientific and technological path (ideally physics and HEP) for their higher education. Moreover, triggering the interest of students in particle physics will educate them about the great advances of physics (like the recent Higgs discovery) and strengthen the public's support of future experimental facilities.

If we add some realistic complexity in the format of the released data, we can even create challenging projects for physics undergrads. This will help them understand what typical data-analysis challenges, and familiarize them with the typical work and lifestyle of a HEP experimentalist (collaborative work, delicate validation of data before use, long processing of data using grid computing, etc.) The interested students will be well informed, ideally attracted to the field, and follow a career in high-energy physics.

10.1 Types of data release

The CDF data cannot be released in the current complicated Stntuple (or Topntuple) format. There are two possible ways of data release: the implicit and the explicit one. With the implicit method, the data is released as part of a web-based application. The students don't get the actual list of events and cannot independently analyze the

data. With the explicit method, particle 4-vectors (maybe additional information) are released in a file format. The two methods are described below.

10.1.1 Implicit data release

With the implicit method, the data are not released in a list mode; instead it is presented with a web-based application. The advantage of the implicit release is that the data and its analysis are integrated. The user doesn't have to write his/her own code to analyze the data. Instead, the student will be directed by an application running on a web page, with the use of buttons, menus etc. on how the data will be analyzed. The result of the analysis will be automatically plotted. An example is CMS's e-lab [29]. A nice presentation of the field, the experiment, and the physics goals can accompany the actual data-analysis page. Interactive games can be performed online and make the learning experience more enjoyable (e.g., the event recognition game with CDF Run-I event-display lego plots [47]). The implicit release of data in principle includes the CDF event display (could be online or running on local Fermilab workstations), although our event display cannot compete with the CMS java-based one [48]. We can certainly release lego plots and ask the students to act as "triggers", perform visual pattern recognition, and categorize the events [47]. The education office of Fermilab was up to recently planning to integrate this game with an interactive multi-touch table [49] and turn the triggering game into a group activity for Fermilab visitors.

Disadvantage of the implicit released is that the student will not have the freedom to perform the statistical analysis he/she prefers. This is of course a problem for advanced students only.

10.1.2 Explicit release

CDF can also explicitly release files of data in list mode. In the past, the most popular file-based release for high-school students was in the form of spread-sheets [50]. Students fluent in Excel (or equivalent programs) were able to statistically analyze the data and make plots. In addition to spreadsheets, and especially for more advanced students and undergrads we could release lists of 4-vectors in ascii format along with the same information organized in a simple root ntuple or tree format. In principle more complicated information could also be released, but only if there is really demand for it.

For the release of the actual data files, code has to be written that takes the existing Stntuple files, applies trivial corrections and writes the data in a simple and understandable text format. Proper documentation should accompany the files. For a complete release, some statistical data-analysis techniques could be recommended, along with some step-by-step guidance. Special pages for teachers could be also written.

10.2 Challenging student projects

The support we offer to any educational use of CDF data is more important than the release of the data itself. We have to offer services complementary (maybe more interesting) to what other experiments already offer (e.g., through elabs [51] or quarknet [52]). Beyond the applications that will perform a trivial data analysis (e.g., detection of particle resonances), it would be desirable to create challenging problems for the students. For example, ask the students to discover a real (or MC) signal by removing background events through proper cleanup cuts. The understanding of the multi-dimensional nature of collider data and its analysis is critical and CDF can make a difference towards this goal. The design and maintenance of such a program, with the addition of new challenges, is a longer task. It would be nice to initiate competitions among schools for the discovery of a hidden real (or MC) signal in CDF data. Maybe some of the most challenging problems could be appropriate for undergrad university classes. We could envision that some of the more realistic projects could be used for training graduate students in experimental high-energy physics.

10.3 Resources

For the release of the data explicitly as files or/and implicitly as an integrated web-based application a 6-month 50% time of a physicist is required. The quality of the web-based release should be similar to quarknet or elab (actually it should be part of these programs). Beyond that, for maintaining the outreach program without any improvements, a 10% of a physicist is required for the duration of the program. If we want to collaborate with the education office and provide CDF data in any desirable (to them) form, a longer (and more intensive) engagement may be decided. If we want to maintain an active program of student challenges, then a 20-30% of a time of a physicist may be required.

The physicist will perform data skimming and trivial data analysis, minimal MC work, online applications development, web design, infrastructure support, communication with the education office. No interaction with students or teachers is assumed. The latter is possible for a more dedicated program. If the outreach will include Universities and challenging problems, and if the (online) communication with teachers/professors is desired, then this could be a long-term 50% time of a physicist, equivalent to traditional "service work" for the experiment (in this particular case, "service work" for society as well).

11 Summary

The CDF task force analysed CDF computing model to identify the minimum requirements to preserve the capability of performing a complete analysis on CDF data in the long term future ($>$ 10 years from now). They are summarized in the following.

- **Data preservation and access**
  We recommend that a copy of all CDF data and MC samples is preserved indefinitely at Fermilab. SAM has to be maintained as data access method; moving to a new system is not feasible given the reduced size of the collaboration. Instead of preserving the current SAM, we'd prefer to migrate to the future SAM which is being developed for the IF experiments; this solution would require immediate effort from the CS and the CDF collaboration to adapt current SAM commands to the new environment, but it will ensure long term support and ease of maintenance. CDF databases have to be preserved as well and kept accessible to users who need to re-process data or generate new MC samples. We think that any solution which implies the migration of our DBs from Oracle to another DB software would require too much effort and should be avoided, if possible. For data access preservation, we need significant support from CS. A CDF member (0.5 FTE x 2 years on CDF side) will support the CS expert.

  It would be preferable, though not mandatory, to have another copy of CDF data (or at least a subset of the data) stored offsite, at the CNAF or KISTI computing centers. A second copy of the data would ensure long term integrity against loss and a more distributed analysis capability. This option is being investigated with both computing centers to define the exact amount of needed resources.

- **Software preservation**
  As far as CDF software is concerned, we recommend that CDF adopt the porting strategy as far into the future as possible; to prepare for a time when this is not possible, alternative strategies should be investigated in collaboration with the CS. This effort will require at least 0.5 x 1 year FTE on the CDF side. A similar amount of effort will be required from CDF to work on the validation suites and documentation.

- **Job submission**
  In the long term future we should provide users with a mechanism to submit their analysis jobs on a suitable farm. Different possible solutions have been highlighted in the report and we need input from the CS to select the best in terms of ease of implementation and low cost and support in the future.

  Besides computing resources, users in the far future will need disk space to create their executables and store the results of their jobs. The exact amount cannot be estimated at this stage, it will depend on the number of analysis.

  For the development of the new job submission system CDF will need significant support from CS. On CDF side we estimate that 0.5 FTE x 2 years will be necessary to support this effort.

- **MonteCarlo production**
  It is fundamental to preserve in the long term future the capability to generate new MC samples. Proper solutions should be investigated to ensure new

generators can be interfaced with CDF detector simulation. Besides the MC production software and a job submission system, required resources special to MC production will be a data assembly server and an upload server.

- **Documentation**
  The preservation of the documentation is fundamental and it has to be one of the first priorities of the CDF collaboration. For the long term future, a new webpage and associated document storage and management are needed, where information contained in the current one can be re-organized and improved. INSPIRE has to be used to preserve CDF internal notes. CDF collaboration will be in charge for the preservation of the documentation. At least 4 FTE x 1 year will be needed to create the new webpage, plus 0.2 FTE x 1 year for the migration of internal notes to INSPIRE.

- **Outreach**
  Subsets of CDF data may be released for outreach purposes, in web-based application or in a simple text format. Outreach activities will be the responsibility of CDF members, in collaboration with the Fermilab education office, and with minimum support from the CS in case of need. A 0.5 FTE x 6 months will be necessary to create the framework and migrate CDF data to a simpler format; 0.2/0.3 FTE will be required in the future for maintenance. It has to be stressed that even if we consider outreach activities very important to promote HEP among students, we think the long term preservation of the current version of CDF data and the complete analysis capabilities should have the highest priority in CDF data preservation project.

All the requirements and the possible solutions described in this report need to be discussed with the CS, to assess their feasibility and the resources needed on the CS side. A strict collaboration with the D0 experiment has to be pursued, as common solutions should be investigated wherever possible. Together with the CS and D0 we will then proceed with the design of the final plan, which should be started to be implemented at the beginning of 2013.

12 Acknowledgments

We want to thank all CS experts who helped us to write this report and the D0 data preservation task force for the useful discussions.

13 References

References

[1] S. Bethke et al., *Determination of the strong coupling from hadronic event shapes and NNLO QCD predictions using JADE data*, Eur. Phys. J. C 64 (2009) 351 [arXiv:0810.1389].

[2] S. Schael et al. [ALEPH Collaboration], Search for neutral Higgs bosons decaying into four taus at LEP2, JHEP 1005 (2010) 049 [arXiv:1003.0705].

[3] *Status Report of the DPHEP Study Group: Towards a Global Effort for Sustainable Data Preservation in High Energy Physics*, DPHEP-2012-001 - May 21, 2012, arXiv:1205.4667 [hep-ex]

[4] The CDF Collaboration, Conference note 10807, http://www-cdf.fnal.gov/physics/new/top/2012/LepJet_AFB_Winter2012/CDF10807.pdf; V. M. Abazov et al. (D0 Collaboration), Phys. Rev. Lett. 84, 112005 (2011); arXiv:1107.4995.

[5] The CDF Collaboration, Conference note 10719, http://www-cdf.fnal.gov/physics/new/top/2011//SpinCorrDIL/SpinCorrDIL_Pub/spincorrPubnote.pdf; V. M. Abazov et al. (D0 Collaboration), arXiv:1110.4194 [hep-ex].

[6] The CDF Collaboration, Conference Note 10793, http://www-cdf.fnal.gov/physics/new/top/confNotes/cdf10793_SingleTop_7.5_public.pdf; V. M. Abazov et al. (D0 Collaboration), Phys. Rev. D 84, 112001 (2011).

[7] Tevatron Electroweak Working Group, CDF, D0 Collaborations, arXiv:1107.5255 [hep-ex].

[8] R.C.Lopes de Sa (on behalf of CDF and D0 collaborations), arXiv:1204.3260v2 [hep-ex].

[9] The CDF Collaboration, Conference Note 10784, http://www-cdf.fnal.gov/physics/new/bottom/120216.blessed-CPVcharm10fb/cdf10784.pdf

[10] http://www-ccf.fnal.gov/enstore/

[11] http://projects.fnal.gov/samgrid/WhatisSAM.html

[12] http://www.dcache.org/

[13] http://cd-docdb.fnal.gov/0047/004736/006/cheppaper.pdf

[14] http://cdcvs.fnal.gov/cgi-bin/public-cvs/cvsweb-public.cgi/?sortby=file&hideattic=1&logsort=date&f=h&hidenonreadable=1&cvsroot=cdfcvs

[15] http://www-cdf.fnal.gov/htbin/twiki/bin/view/CodeManagement/WebHome

[16] http://www-cdf.fnal.gov/htbin/twiki/bin/view/ProductionFarm/WebHome

[17] http://www-cdf.fnal.gov/internal/mcProduction/

[18] http://root.cern.ch/

[19] http://www-ekp.physik.uni-karlsruhe.de/~feindt/acat05-neurobayes.pdf

[20] http://www.fnal.gov/docs/products/ups/ReferenceManual/

[21] http://fermilinux.fnal.gov

[22] http://www-cdf.fnal.gov/htbin/twiki/bin/view/CodeManagement/OperatingSystemRequirements

[23] https://access.redhat.com/support/policy/updates/errata/

[24] inspirehep.net

[25] http://www-cdf.fnal.gov/~cplager/internal/Analysis/Tools/TopAnalysisTools/tools.html

[26] http://www-cdf.fnal.gov/internal/physics/top/m2fu.shtml.

[27] http://www-cdf.fnal.gov/~csmoon/internal/DIL/9.1fb/topcode/topana/.

[28] http://www-cdf.fnal.gov/internal/physics/top/RunIITopProp/tools/ks_tests.html.

[29] http://www18.i2u2.org/elab/cms/home/

[30] http://www-cdf.fnal.gov/tiki/tiki-index.php?page=OfflineStntuple.

[31] http://www-cdf.fnal.gov/tiki/tiki-index.php?page=Stntuple.Datasets.HighPtData

[32] http://www-cdf.fnal.gov/tiki/tiki-index.php?page=Montecarlo.QCDGroup.Main,http://www-cdf.fnal.gov/physics/ewk/mc_samples.html.

[33] http://www-cdf.fnal.gov/internal/people/links/MariaSorin/trigger/PHYS17/tabletrig.html

[34] http://www-cdf.fnal.gov/internal/physics/qcd/qcdmc6/

[35] http://www-cdf.fnal.gov/internal/physics/top/jets/corrections.html

[36] http://www-cdf.fnal.gov/internal/dqm/goodrun/good.html

[37] http://www-cdf.fnal.gov/~hatake/JetAlgoPage/JetAlgo.html

[38] http://cdfcodebrowser.fnal.gov/CdfCode/source/JetObjects/doc/JetPage.html, http://www-cdf.fnal.gov/CdfCode/source/JetMods/doc/JetModsDoc.html.

[39] http://www-cdf.fnal.gov/internal/people/links/KojiTerashi/

[40] http://www-cdf.fnal.gov/internal/physics/joint_physics/instructions/curvature_corrections_2005.html

[41] http://www-cdf.fnal.gov/internal/physics/joint_physics/instructions/electron_cuts_gen6.html, http://www-cdf.fnal.gov/internal/physics/top/r2leptons/etf/etf_main.html, http://www-cdf.fnal.gov/internal/physics/joint_physics/instructions/muon_cuts_gen6.html

[42] http://www-cdf.fnal.gov/internal/physics/joint_physics/instructions/gen5/COT_requirements.html

[43] http://www-cdf.fnal.gov/internal/physics/top/RunIIBtag/bTag.html

[44] http://www-cdf.fnal.gov/internal/physics/joint_physics/instructions/zvertex-efficiency.html

[45] http://www-cdf.fnal.gov/internal/physics/joint_physics/PerfIDia/PerfIDia.html

[46] http://www-cdf.fnal.gov/internal/physics/joint_physics/Validation.html

[47] http://ed.fnal.gov/projects/labyrinth/games/codecrackin/particle_graffiti/activity1.html?name=Your+Name

[48] http://www18.i2u2.org/elab/cms/event-display/

[49] http://ideum.com and http://openexhibits.com

[50] http://quarknet.fnal.gov/run2/run2_teacher.shtml

[51] http://www.i2u2.org/elab/list.html

[52] http://quarknet.fnal.gov/