1 Offsite Computing

1.1 Status and Perspective

Offsite computing is by now a de-facto important reality of CDF computing environment. During FY04 we moved ahead along the plan outlined one year ago which still provides our blueprint for having a full grid-like environment by the end of FY05.

The motivations and history that led to that plan still stand as described in the 2003 version of this document and will not be repeated. We recall here the step-by-step procedure that we indicated then and use it as guideline for measuring development.

In practical terms, we expect CDF to extend from Fermilab to incorporating computing facilities outside the laboratory in a step-by-step process:

1. allow easy usage of remote facilities for code development, data analysis and MC generation by single institution

2. move off site (part of) organized MC production

3. move off site (part of) single user MC production spreading it uniformly across all institutions

4. move off site analysis of secondary and/or tertiary data sets, by duplicating CDF data at remote institutions and giving everybody access to it

5. exploit large off site CPU capability for interactive analysis and reprocessing

6. develop formal agreement and define the "price" of each service, only after it has been demonstrated to work

We focused FY04 efforts on technical aspects of making remote sites capable to provide the services. The main issue for FY05 will be operational aspects, and thus understanding of performance and usage to arrive at a "pricing" that reflects services rendered in Step 4 and 5 above.

Overall FY04 has been very successful and now the situation with respect to the 6 point list above is as follows (more details are given later on):

1. This has actually been in place since many years by means of a very successfully code distribution tool that still provides the foundation of our Grid effort. Since 2003 some institution have started to use SAM to transparently import data needed for analysis, and this is now the main operation mode of e.g. German CDF collaborators.

2. By now virtually all organized MC production is performed offsite. Automatizing of the process and of the data import is now in development/test phase, we plan to use SAM for data import back to Fermilab tape storage, file concatenation and bookkeeping. Already MC samples produced offsite are cataloged using SAM metadata. Use of RUNJOB tool for process control and of JIM for automatic brokering across several sites are options to be explored.

3. This is now reality. Since a few months several CDF users located e.g. at Fermilab have successfully used several offsite farms located at institutions other then their own, needing no additional help then a web page and an announcement "that they could".

4. This is also reality now. In the same situation as previous step, although still used on a smaller scale.

5. This is at present and R&D project.

6. Negotiations with CDF International Finance Committee on this point has started since more then one year. The politics is moving steadily in this direction and more details are given later. The missing technical step here is a uniform monitoring and accounting mechanism, that we are now developing.

We are at this point reasonably pleased with the status of steps 1.,2.,3. and 4. Those functionalities are now deployed at several sites and we have moved from development and beta test into production and operation phase, with weekly “offsite operation” meetings, transparent access for all CDF users to several offsite farms and few TB of data replicated offsite for local analysis. Large scale usage of those offsite farms is now limited not by the tools, but only by the very limited size of those installation (both as CPU and data disk) with respect to the central CAF, so that most users do not perceive and advantage with respect to using the FNAL machines. We expect this to change as soon as most popular datasets (e.g. inclusive hadronic B) are replicated offsite.

1.2 Status of off site resource as of Summer 2004

While some CDF collaborators own CDF-reserved computers, other share access to largish facilities with other experiments, this makes it almost impossible to tell a-priori e.g. how much CPU power CDF physicists can use off site. We expect that once a framework will have been clearly defined for usage of those facilities by all CDF collaborators, something like a minimum amount of CPU cycles available can be defined, while most likely accounting of off site facilities contribution will have to be done a-posteriori, based also on the actual efficiency and effectiveness of each single installation. Besides many institutions will keep priority of usage for their own members, so there will not be an a-priory guarantee on how much generic CDF users will obtain, but a-posteriori this can be a considerable fraction of those resources.

As we predicted one year ago, committing of computing resources by CDF institution for general access has been a slow process, which moved in some place faster, in some place slower then expected, due to local financial policies and pace of hardware acquisition. Nevertheless the following table, which is the snapshot of offsite hardware resources available for CDF by summer 2004, shows a clear picture of emerging large offsite facilities and confirms a growing trend with respect to last year document. In this table we only listed institutions that allow access to all CDF members, with the exception of the German GRIDKA site were this, while being the institution policy, is still technically problematic since access has to be done via LCG grid software.

Some sites are geared toward MC production especially as common CDF resources are concerned, so the local amount of disk is little informative, we write 0.1TB as storage in this case to mean that those sites do not envision allowing significant data access to CDF users at large, at least at present. Local access policies are constantly evolving and will possibly change in the future as a result of political negotiations, hardware additions and our attempt to enforce a common policy.

In addition to what is shown in the table, CDF groups in the UK at University of London and Liverpool and in Spain at University of Cantabria and Barcelona/ICFA are setting up local CDF farms that are expected to be accessible to CDF members by end of 2004. Very likley local users will

Institution	CPU (GHz)	Disk (TB)	Access Gbit/sec	notes
Canada	250	0.1	2	(1) (3) (4)
Canada	1500	0.1	2	(2) (3) (4)
Germany (GRIDKA)	1895	20	1	(2) (4) (6)
Italy (CNAF)	900	30	1	(1) (5)
Japan (Tsukuba)	150	10	1	(1) (5)
Korea	120	1	0.1	(1) (4)
Taiwan	135	3	0.5	(1) (5)
Rutgers	100	4	0.2	(1) (5)
MIT	115	1	1	(1) (3) (4)
UCSD	280	5	1	(1) (4)
UCSD	200	5	1	(2) (4)
TOTAL	> 4000	80	-	

Table 1: Computing resources available offsite to CDF users by summer 2004. Notes: (1) reserved for CDF group (2) shared with other experiments (3) dedicated to MC production (4) allows unrestricted equal access to all CDF members (5) gives priority of usage to local CDF members (6) CDF may use from 10% up of the CPU at GridKa, the indicated disk is CDF-own.

retain privilege, but we plan to have all those sites embedded in the same common framework for access, monitor and accounting.

Winter 2004 has seen the first deployments of large offsite data disk pools with local copies of selected datasets. Now UCSD, INFN, Canada, and Taiwan host several dataset replicas in the few TB range.

Most off site institutions have a high speed local connection to the Internet, so access to those facility can be highly efficient. Experience shows that effective throughput on WAN can be limited by many hard to find bottlenecks. Our best experience so far is with the Canada-FNAL link that has shown reliably and consistently $\sim$ 200Mbit/sec for production MC transfer. which seem to represent the maximum presently allowed by Fermilab-ESNET connection.

1.3 Offsite MC Production

At present, the primary use of off site computing resources for CDF as a whole is in the form of off site MC production, mostly in Toronto, UK, SanDiego and Italy. This is presently done by running MC jobs locally (i.e. no grid-like remote central control) and importing data back to FNAL by ftp to on site disks and hand write to Enstore. The Canadians have MOU responsibility for coordinating the MC production, as well as providing 1 Million events per day capacity in Canada; above the guaranteed minimum of 1 Million events/day, the Canadian cluster may be, and in past has been, able to produce significantly higher rate of MC production. Other institutions (UK e.g.) are considering the possibility to setup analogous MOU's for taking up responsibility to produce a fraction of the CDF needed MC events. The MC production group has already produced > 100 Million events in FY03, about 60% in Canada and 30% in UK. MC production proceed with increasing pace and more then 200M events have been produced in Canada and UK in FY2004. We expect this to increase more than linearly with luminosity as we improve the accuracy of the simulation, and physics analysis of Run 2 data matures. It is worth noting that as CDF reconstruction code evolves we do not only need to reprocess old data, but also to generate new MC samples with same reconstruction version as used in analysis.

All MC generated in this fashion is coordinated via the physics groups, but this is by no means all the MC generated by CDF. Most analysis works need very large MC productions that are tuned to that particular topic and can not be shard with others. These "single-user" MC samples are, and will be in the future, produced in a chaotic way by singles or small groups as needed and will not be managed by the physics groups.

The next steps for general CDF computing off site has been providing users the means to generate their private MC samples, e.g. exploiting CAF installations at remote institutions, which is now possible. While at present significant amounts of user level MC are still produced on the FNAL CAF, which is a poor usage of a system built for good data access and tightly coupled to the main CDF data repository, most of that work can be done easily off site and results copied back to FNAL using current CAF tools. Indeed this has started happening during winter 2004 and we are now in the process of establishing accounting tools to quantify this usage. Random sampling of offsite farms suggests that so far much more then 50% of their CPU time is being used for MC production.

Work is now undergoing to develop by end of Summer 2004 an "easy to use" users facility that will allow directly storing on FNAL's Enstore tape archive and cataloging into SAM the output of a MC run on a farm worker node running on any CAF around the world. Data will be moved using

GridFtp re-using technology already developed for the JIM project. We have identified the major issues there in the need of guaranteeing a minimum file size for efficient tape and metadata operation, so in some cases storage on intermediate disk buffer and file concatenation will be needed. We envisage to do this as well under SAM control, but plans are not final yet about the best way to organize the needed bookkeeping into SAM.

1.4 Offsite Data Analysis

In 2003 we presented the plan to expand, as our grid project mature, remote sites into more general user analysis centers off site. This is now a reality with several sites having an established procedure to import large data sets to local disk caches and make them available to CDF users for analysis, so relieving CPU load and data access congestion from FNAL's CAF. We have experienced that it is more effective to preload specific data sets, lock them on local cache disk, and advertise their availability to users, rather then import data on demand according to random analysis jobs and end with a lot of cache misses. At present about 10TB of data are replicated offsite and advertised to CDF users via web pages. Work is in progress for a more clear and unified presentation of this information. Those datasets range from J/Psi samples to Inclusive Hadronic B, to High Pt leptons. Since winter 2004 CDF has formalized this distributed computing effort with the creation of "GRID Operation" meetings and of a political body, the "CDF Computing Resources Coordination Board" where representatives from each institution that contributes to the CDF-Grid can coordinate their resource deployment.

2 Toward a CDF Grid

2.1 The Vision

The long term vision for CDF computing is that users develop and debug their application at their desktop somewhere in the world. They then submit their job to SAM-grid, specifying a dataset to analyze in addition to the usual CAF information. SAM-grid selects an execution site based on locally available data as well as CPU resources. The user job is queued at the local site, eventually instances of it start. SAM provides input data, and the user writes out their output into a local scratch area. After completion of an instance, the user may declare the files produced to the DH system for storage in a location where the user has sufficient quota to store the files on disk.

In addition, DH provides metadata catalogue services for the user's output data such that it can be reused as input for a future job by anybody in CDF. Data stored in this manner is not backed up to tape, and may be permanently erased by the user who owns the data. Tape archiving would require in general additional concatenation to achieve file sizes for efficient operations of tape archive resources, and we plan to use SAM to develop tools that allow users to do this effectively.

As organizational principle we expect to be guided by the notion of physics centers rather than regional centers. Assuming we can arrange policies of ownership for resources at the level of "virtual center" it is much more efficient to concentrate a given dataset and the CPU resources required for its analysis in a set of dedicated sites rather than spreading all datasets across all sites more or less evenly. Needless to say, this ideal will require some amount of negotiation and deliberation to be successful. On the other hand this is exactly how present offsite institutions are organizing their computer clusters in a spontaneous way.

2.1.1 The Tools

Technically, we will accomplish this initially via a combination of SAM, dCache, Enstore, CAF, and JIM. At present, dCache read operations, Enstore, and CAF are fully in production. SAM is routinely used offsite and has been deployed for usage also in central CAF. JIM is still in develop and test mode, but expected to ramp into production level support in FY05. In particular work is now in progress to use JIM and the RunJob utility to centrally manage distributed MC production across all CDF offsite farms. In addition to this batch based processing, we are also pursuing interactive computing, both in form of a central interactive platform at FNAL discussed in section **??**, as well as an interactive grid.

The interactive grid computing effort is a collaboration between UCSD, INFN, and MIT. The goal is to build an interactive grid computing system with response times to queries of order 10s for analyzing O(10GB) ntuples. The system is to be based on Root's PROOF tool by adding an interface to SAM metadata catalogue, and Condor/Globus grid middleware, a first prototype was developed for SC2003 last November.

As the CDF-Grid takes form and is deployed offsite, we will have to interface to and incorporate tools already in use in those sites, primarily the LCG/EGEE emerging LHC GRID, so that CDF farms in Europe, Taiwan and elsewhere may be run as a sub-sect of larger LHC clusters sharing manpower and resources. Work toward this integration is already started and constitutes next frontier of the CDF Analysis Farm (CAF) development which is by now a joint project of INFN and UCSD.

2.1.2 The Financial Side

The CDF International Finance Committee has been debating at length the formalization of foreign contribution to CDF costs. The current position of the Committee is that the CDF's plan to move 50% of analysis work offsite by 2006 is reasonable and it is a matter of fact that several countries have already contributed resources to that goals and are considering plans to increase their commitment up to the indicated level. On the other hand it is clear that such contributions will be on a voluntary basis, in a best effort spirit, and quite likely there will be no MOU-kind document. At the same time proper accounting of the usage of remote resources is perceived to be fundamental for a fruitful collaboration both as guideline for efficient usage and acknowledgment of the contribution. Assuming that the present developments are indeed successful it is conceivable that we start a more global computing cost accounting in FY05.

Up to today, even in lack of such a formalization, CDF has nevertheless received substantial financial contribution by foreign countries, Japan and Italy mainly, who have done so since many years, more recently also UK, Switzerland, Korea and Canada have contributed. In particular for what concerns computing, Canada has taken on a serious commitment to MC production and is providing very significant computing resources for it, and so is the UK and many US institutions who have produced Monte Carlo samples for CDF data analysis since roughly December 2002. It is CDF's first priority to preserve all positive sides of how things have been working till now, and therefore to be very cautious and careful in defining a brand new policy.