1 Offsite Computing

1.1 History and Perspective

Since Run 1, CDF collaborating institutions have a history of using significant computing resources outside those provided by FNAL, either located at home sites or in the Fermilab trailers. This trend has continued into Run 2, and it had for many years escaped any programming and accounting, mostly because, as long as those resources have been exclusively used by single collaborating institutions, there was no way to track down their amount and effectiveness. Effectiveness of off site computing facilities had been limited in the past also by lack of automated data transfer (hand carried tapes only) and of tools for cross-site user's authentication and remote job submission. But this has started to change already and there has been a significant shift toward the deployment of significant off-site computing resources for MC production and data analysis dedicated to the CDF collaboration, following the last Fermilab Computing Review (June 2002). This has come about largely through the availability of large Beowulf clusters in several countries (Canada and the UK) with significant capacity dedicated to CDF, and the improved WAN bandwidths with consistently $>$ 100 Mbps throughput between the off-site centers and Fermilab. These has been possible because of the following several reasons. In no particular order:

- WAN speed has enormously increased in the last years and practically no data has been moved on tape to/from Fermilab during Run 2
- we have much friendlier and robust means to distribute CDF code that make it possible to insure reproducibility of results from the FNAL main platform to any remote laptop
- tools (like SAM) have become available to allow transparent access to remote data
- tools (like CDF CAF) have allowed already to build several computing farms that provide uniform access from any CDF user's location to any of those farms
- tools (like JIM) are becoming available to allow transparent automated brokering of user's job among many farms

- implementing and operating transparent access to FNAL based data and code repositories and computing resources for remote users, while it is working enormously better then in Run 1, is proving to be expensive enough human and money wise that it starts making sense to ask remote institution to repay this cost at least by allowing symmetrical access to their local computers

- driven by the political and technological developments in preparation for LHC era, most collaborators start to have access to large computer facilities that are built around the world and which may both have spare capacity in the near future and provide an easy path for the addition of CDF-owned resources

- several CDF institutions are already starting to open access to their local resources to all the collaboration, up to having formal MOU's for that, something that never happened in Run 1

Putting all the above together, both in terms of tools and of contribution/retribution, is the prerequisite for developing a carefully crafted political arrangement among the many CDF collaborators, which have widely different patterns of financing, contribution to CDF in other forms, and history in the collaboration.

We believe it will be more effective for CDF to build its "GRID" infrastructure using a bottom-up approach, were we start with our working system at Fermilab and add in remote resources a bit at a time, after having developed and demonstrated the effectiveness of the tools for the users, rather then to try to define a framework of "contributions" before knowing whether and how much those remote resources can be used. As we try to add value to our working environment, it is important to remember that any new tool or feature must give to the end user (the "famous thesis-hungry student") a significant improvement in terms of "time-to-publication", or no-one will use it.

Moreover the spread of new powerful computer centers and the affordability of computer farm components, is making possible for many CDF institutions to have private computing resources quite larger then they used to have in Run 1. Analysis and simulation work is moving away from FNAL to locally owned machines. This is also a welcome contribution as it reduces load on Fermilab systems, even in lack of a Grid environment. It is very difficult to quantify these things into "savings" or "contributions" in dollars.

Again a flexible incremental approach based on reality as we go on will be the best way to manage this process and incorporate the resources into a coherent scheme.

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

Step 1 is done, and currently in use, although transparent access to large CDF data samples via SAM is still in development/test phase. Step 2 has started. Organized Monte Carlo for Summer Conferences 2003 was generated entirely outside of FNAL, as discussed in detail in Section 1.3. Step 3 is the near term focus and we expect this to start happening by Spring of 2004. Step 4 well be a combination of success with SAM (in 1.) and 3. Step 5 Is at present an R&D project. We expect that Step 6 will not lead to significant decreases in the FNAL computing budget as described in this report until FY06 given that the budget is driven by costs associated with Step 4 and Step 5 type of services, as well as tape archiving.

To arrive at a "Virtual Center" accounting model by FY06 without sacrificing service levels, significant human resource effort has to be expended in FY04 and FY05. FY04 needs to be focused on technical aspects of providing the services, including initial prototype deployments. FY05 is then focused on operational aspects, and thus understanding of "pricing" that reflects services rendered in Step 4 and 5.

1.1.1 2004 update

The above section was written one year ago, and aside for details still stands. Of course during FY 2004 we have been moving ahead on the list of steps and the situation now is that we have tools in place for 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 (Canada is doing this already, e.g.), 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.

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.

Note that most 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.

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)	750	22	1	(2) (4)
Italy (CNAF)	950	30	1	(1) (5)
Japan (Tsukuba)	150	10	1	(1) (5)
Korea	500	1	0.1	(2) (5)
Spain (ICFA)	200	0.1	0.6	(2)
Spain (Cantabria)	50	2	0.6	(1) (5)
UK (ScotGrid+RAL)	230	5	1	(2)
UK	200	24	1	(1) (5)
Taiwan		3	0.5	(1) (5)
TexasTech	250	1	1	(2) (5)
Rutgers	110	15	0.2	(1) (5)
MIT	100	0.1	1	(1) (3) (4)
UCSD	300	5	1	(1) (4)
UCSD	200	5	1	(2) (4)
TOTAL	> 5000	> 90	-	

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

Several sites are building local resources mostly for lots of CPU, rather then storage, according to the general perception that the most effective way to use off site resources will be user Monte Carlo production and reiterated analysis of smallish data sub-samples, while the skimming of large inclusive samples will best be dealt with at Fermilab. Still 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, and $\geq$ 250 Million events will be the likely total by the end of this calendar year. 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.

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.

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 is being 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 mcRunJob SAM 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 on November 15th.

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 e.g. in Italy within the INFN-Grid project.

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 for FY07.

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.

Until our grid vision is fully implemented, and shown to be fully functional, we consider it prudent to require that baseline needs of CDF are fulfilled by on site computing at Fermilab.

2.2 Offsite Productivity Issues

A common theme among University based institutions in CDF is a concern about being able to attract and retain the best young graduate students during their first year when they cannot be at the lab. This has been identified as one of the most important challenges in our field by several of the US collaborators. The CDF software is perceived to be sufficiently complex, and frustrating that it makes it very hard to motivate young students. In many cases the students leave the field of HEP to work on something that has quicker entry into physics and is thus more attractive.

There is a clear consensus that a focus on user productivity is required for the upcoming years. It is however less clear who should bear the burden of providing the human resources required to make substantial progress. A compromise might include increased FNAL computing division support for software tools like debugger, profiler, memory leak checker, and code browser, in combination with collaborator support possibly via offline shifters for improved documentation, and an effort on making programming interfaces more consistent, as well as defining default analysis objects like tracks, photons, leptons, etc. which are well understood and documented.

A related set of concerns revolve around making the production output generally usable for the end-user, with well documented, and understood quality of the data.

It is also clear that improvements in usage effectiveness (i.e. removing the number of mistakes of novice users) may play a role in lowering the overall load on the CDF computing infrastructure, and thus help us to do more physics with the resources we have. While is very difficult to put this into quantitative terms, but CDF is tackling this seriously and effort is in place to improve the overall organization of the physics analysis community.