CDF/DOC/COMP_UPG/5858
February 17, 2002
Version 0.1

The CDF/D0 UK GridPP Project[1]

B.Todd Huffman [2], Ronan McNulty [3], Tara Shears [4], Richard St. Denis [5], David Waters [6]

1 Introduction

Run II of the Tevatron began in March 2001, and is foreseen to continue until the start of LHC. CDF is expected to store over 1 Petabyte of data on tape in Run II. Data handling in CDF has become an important issue as we devise schemes to reduce the data to a manageable size and proceed with physics analyses. Of particular interest to the UK are the problems of data access and resource sharing, which we envisage solving within the framework of the GRID.

The first part of the document, Sections 2 to 8, outlines the CDF requirements for a UK GRID. We start by summarizing the data model for the UK institutes on CDF. An evaluation of the requirements with regard to how GRID can provide solutions follows. Detailed use cases have been worked out to provide the rationale for the requirements. Following this, the UK CDF group describes the expertise and resources it has to offer. Timescales and goals for the project are given. Potential areas of collaboration and manpower requirements are outlined before concluding. Background information on the CDF Data File Catalogue is provided in an appendix.

Section 9 of the document follows with additional information summarizing the main points of the CDF proposal to work on grid. Following this in Section 10 is the executive sumary of the CDF/D0 joint project. Section 11 describes the portion of project to be executed by the Glasgow group and Section 12 the responsibilities of the Oxford and UCL groups.

2 UK CDF Data Model

We begin with a description of the data samples and analysis model. This is followed by a summary of dataset size, simulation needs and the CDF software framework. These subsections set the stage for the requirements stated in the next section.

2.1 Description of Data Samples

The data sets at CDF are organised by trigger type. The UK institutes intend to perform many physics analyses in the fields of B physics, electroweak measurements and searches for new physics. The relevant data samples are summarised Table 1 and have been based on CDF4718 [1].

[1]United Kingdom, Particle Physics Data Grid
[2]Oxford University
[3]Liverpool University
[4]Liverpool University
[5]Glasgow University
[6]University College London

Physics	Physics Trigger Set	No. Evts
B lifetime B oscillations CP violation	Central J/psi Displaced vertex	12 M 1000M
W/Z Higgs SUSY B physics	High Pt leptons Inclusive electrons Inclusive muons	2M 50 M 14 M
SUSY Calibrations	High E_t photons	58 M
Higgs	$Z^0 \to b\bar{b}$	6M

Table 1: Datasets that UK institutes wish to access.

2.2 CDF Analysis Model

Raw data are processed at Fermilab on a production farm of Linux PCs to produce **Secondary Data Sets,** which consist of higher level objects (e.g. tracks, jets, missing energy etc.) in addition to the raw data. UK physics analyses will proceed by filtering the Secondary Data Set to produce a **Tertiary Data Set**, which is a subset of the Secondary and may store less information per event. We will use the Tertiary Data Set to produce **ntuples** of the salient information for a given analysis, which can be analysed on local machines in conjunction with ntuples of simulation information.

The following operations on the data may need to be performed by the UK computing resources:

- **Reprocessing of the data:** If better detector constants are produced, secondary data sets containing sufficient information can be processed to make new Secondary Data Sets. These could be exported to FNAL.
- **Filtering:** Strip the Secondary Data Sets to produce Tertiary Sets. Each selected event could either be fully copied or a set of pointers to individual events could be stored.
- **Analysis:** Perform analysis operations on Secondary or Tertiary Data Sets and store the additional information along with the event. Analysis and filtering are often performed together.
- **Ntuple Creation:** From the Secondary or Tertiary Data Sets, store salient information in ntuples. Analysis and filtering may be performed at the same time.
- **Ntuple Analysis:** Analyse the ntuple, usually on a local machine.
- **Simulation:** Run a physics generator and a detector simulation to produce Secondary Data Sets that look like 'real' data, though they usually contain additional 'truth' banks.
- **Fast simulation:** Run a physics generator and a fast simulation (often just 4vector smearing) to access large statistics.

Data Sample	Secondary(Gb)	Tertiary(Gb)	Ntuple(Gb)
Central J/ψ	1200	120	12
Displaced vertex trigger	100000	1000	100
High P_t leptons	200	100	20
Inclusive electrons	5000	200	50
Inclusive muons	1400	56	14
High E_t photons	5800	580	58
$Z^0 \to b\bar{b}$	600	300	60

Table 2: Estimates of the Sizes of Datasets based on CDF note.

The facilities at Fermilab are not sufficient to allow us achieve this in a timely fashion, either for data analysis or for simulation (for which no central production is envisaged). We are looking to implement a GRIDlike solution, with shared processing power and the CDF data sets distributed about the disk storage in the UK.

2.3 Summary of Data Size

Table 2 summarises the size of the datasets at each stage of the analysis model.

2.4 Simulation

There is little centralized production of simulation at CDF. Thus the UK will be responsible for making simulated data samples for each of the analyses we are interested in. We will have two classes of simulation generation:

- Full simulation to optimise data selection in an analysis. This will result in tertiary data samples of about 50G which must be stored, and from which ntuples are made.
- Fast simulation to test biases or systematic errors in an analysis. Up to 10^{12} events may need to be generated, although only a small amount of disk space is required since the whole event need not be stored. This is consequently a CPU limited task.

2.5 CDF Analysis Framework

CDF reconstruction and analysis takes place in a software framework called "Analysis Control++" (AC++), similar in conception to the BaBar software framework with which it shares a great deal of low level code. The framework code, like the vast majority of the CDF software, is written in C++ and employs numerous objectoriented programming techniques to provide the most convenient API. The only significant remaining use of Fortran is in the Monte Carlo generators.

The framework is fundamentally modular, with users specifying sequences of modules in order to perform reconstruction or analysis tasks. User modules are easily compiled and linked into binaries containing standard CDF modules. AC++ jobs are controlled through a script written in Tcl/Tk, provided as an argument to the binary. The sequences of modules, together with the sets of parameters that determine the operations of the modules, are specified in this

Tcl file, along with Framework parameters such as the number of events that are to be processed. All module sequences must contain an input module. In the case of simulation this can be a Monte Carlo event generator; for data jobs it would usually be the data handling input module in which standard datasets or else specific run and event ranges are specified.

The CDF software framework reads in and writes out events in sequential ROOT format. The Event Data Model (EDM) is the CDF structure for arranging both the persistent (i.e. disk resident) and the transient (i.e. memory resident) representations of the events as collections of objects with various common properties, connecting links, etc. An important nearfuture upgrade of the CDF software is the implementation of multibranch ROOT I/O. Different parts of the event data will live in different branches and will only be physically readin as required, thus greatly reducing the I/O related overheads in many applications.

Databases are a crucial ingredient of CDF operations. Trigger, calibration, run conditions and data handling information are all stored in a database that, typically, every user application would need to access at least once per job. Within Fermilab CDF has decided to use Oracle, with which any future extensions of the CDF software must also be compatible. It is likely that, in the near future, a freeware database such as MySQL or PostgreSQL will be supported for offsite users such as those at UK institutions. A heterogeneous database environment must therefore be anticipated and planned for. Details of the CDF Data File Catalogue can be found in the appendix.

In the first instance it can be assumed that a CDF software installation exists on all machines available on the GRID to run CDF software tasks. These installations can be made to contain standard executables, such as those required to run standard reconstruction or make standard ntuples. However, alternative schemes are being developed for distributed Monte Carlo production whereby all necessary shared libraries and associated files are packaged together with the specified binary, such that it can be run on any machine of a given architecture and operating system version.

3 CDF Requirements

3.1 Hardware Requirements

- **HW1:** Adequate bandwidth between the UK and Fermilab to repopulate the disks with newly available datasets (following, for example, a full reprocessing of the data on the CDF production farms) in a period of time that does not reasonably interfere with analysis: 3 TB in 2 weeks. This corresponds to a sustained rate of approximately 20 Mb/s.

- **HW2:** Adequate bandwidth between UK sites for the transfer of secondary datasets (following, for example, reprocessing by a UK group) as they are created. This should be more than the bandwidth of HW1, and less than the limit imposed by disk access (i.e. between 20 Mb/s and 100 Mb/s).

3.2 Middleware Requirements

- **MW1:** A webbased browser tool that generates reports containing parameters that describe datasets. It must be possible to impose requirements on the parameters (cuts, ranges, lists, ...)

- **MW2:** Tools to add new datasets to the Data File Catalogue.

- **MW3:** A mechanism to copy data to and from a given site, via network or tape transport. Appropriate data integrity checks will be made and, if successful, appropriate metadata updates will be performed.

- **MW4:** A database system for tracking the data held at a site and available at other sites. This system has the following requirements:

  - Databases may be local, publishing their contents for use at remote sites.
  - In the first instance this can be a central database which is exported periodically (at least nightly) to all remote sites.
  - The database system must function with both Oracle and freeware databases. This probably requires the use of standard SQL.
  - The system may use proprietary SQL for read operations only when a significant demonstration of manpower savings is demonstrated.
  - The system may use proprietary SQL for write operations.
  - The database system must follow the standards described by the Fermilab database (ODSDBA) group.

- **MW5:** Sites must make information about system resources (including CPU, disk space, tape robot space and relevant network capabilities), available to remote users.

- **MW6:** The resource availability information must be translated into costs that are relevant to the enduser. The most obvious cost is the time taken for a specified file transfer or analysis task to complete.

- **MW7:** A mechanism to allow jobs from participating remote sites to be run locally. Amongst other things, this must involve the ability to appropriately prioritize different classes of job. For example, in the case of systems being used for both interactive work and batch jobs, priority must be given to local interactive users.

- **MW8:** Priority must be determined from a match of resources needed on a given computing node to resources available.

- **MW9:** Job status monitoring. Convenient tools must be made available to endusers to evaluate the status of their jobs, wherever they are running.

- **MW10:** All CDF jobs must see one of the following standard CDF environments:

  - Standard development: access to code for building and database server
  - Level 3 – like environment: database, executable and ancillary files exported as tar balls.
  - Farmlike environment: executable and ancillary files exported as tar balls, database server accessible across network.

- **MW11:** A method for deducing the most costeffective manner to store data, potentially across multiple sites or constrained to a specific site (e.g. MAP or a local machine).

- **MW12:** Resource availability must include read, write and execute permissions on all disk storage areas.

- **MW13:** Quality of service guarantees that distinguish amongst (in order of priority):
  1. Video conferencing
  2. Distributed collaborative work
  3. Metadata transfer
  4. Data transfer.
- **MW14:** A data entry method for simulation parameters that works with the database browser, so that simulations may be made starting either fresh or with parameters from a previous simulation.
- **MW15:** There must be a method to obtain random number seeds for simulation.
- **MW16:** GRID middleware must be compatible with Fermilab security policy.

3.3 Data File Catalogue Requirements

In addition to the information already contained in the CDF Data File Catalogue (for example, pertaining to the dataset name, run sections etc.), it is anticipated that the following information might be required of data in order for it to be used by GRID applications:

- **DFC1:** Location. In addition to geographical information, the metadata could contain a link to sitespecific information such as network connectivity and bandwidths, available CPU etc.
- **DFC2:** Persistency state. Are the data:
  - Disk resident,
  - Tape robot resident,
  - Available on tape for manual staging.
- **DFC3:** Reprocessing flags, which indicate the time and type of reprocessing that has been performed on the data. For example, the type of reprocessing will include the version of the CDF software used.
- **DFC4:** Parent/child links. For example, Standard Ntuple data must point back to the secondary dataset from which it was made. Enough information must be provided to ensure reproducibility of results.
- **DFC5:** Must accommodate storage of parameters for all HEP generators.

4 CDF Use Cases

This section outlines a number of sample CDF use cases for data analysis. Some use cases include others. These references are indicated by boldface type.

4.1 Identify a dataset

1. *Actions:*
   - User runs web browser to determine what metadata are available to describe a dataset.
   - User specifies metadata parameter cuts.
   - User submits query to the database using the web browser and obtains a list of datasets.
2. *Relevant Requirements:* MW1,DFC1, DFC2, DFC3, DFC4

4.2 Specify a new dataset

1. *Actions:*
   - User runs Data File Catalogue tools to describe new data sample.
2. *Relevant Requirements:* MW2,DFC1,DFC2,DFC3,DFC4

4.3 Submit a job

1. *Actions:*
   - User specifies resources needed, and optionally:
     - an input dataset
     - an output dataset
     - an input set of simulation parameters
     - The cpu and data requirements are matched.
2. *Relevant Requirements:* DFC2,MW1,MW4,MW5,MW6,MW7,MW8,MW10

4.4 Specify simulation parameters

1. *Actions:*
   - User runs simulation parameter GUI to describe new data sample
2. *Relevant Requirements:*MW0,MW13,MW14

4.5 Populate UK disks with secondary data set

1. *Actions:*
   - User **identifies secondary datasets** to be transferred to the UK.
   - User submits a list to an application which
     (a) locates data at FNAL,
     (b) stages data if needed,
     (c) copies data to appropriate mass storage in the UK.

- Once data is copied, the relevant bookkeeping database entries are updated.
- The status of the transfer is returned to the user.

2. *Relevant Requirements:*MW11,HW1,MW3,MW5,MW9

4.6 Reprocess secondary data set

1. *Actions:*

- User prepares revised reprocessing executable on own machine. Revised executable could involve addition of extra information (eg. secondary verticies) to event, or the improvement of existing reprocessing code.
- User **identifies secondary datasets** to reprocess.
- User submits executable and dataset specification script to application which directs job to where data resides.
- Data are written to disk as reprocessing proceeds. Location of dataset is updated concurrently in bookkeeping database.
- The status of the job is returned to the user.
- If desired (eg. if reprocessing motivated by problem in original secondary dataset), data can also be transferred to FNAL. Dataset description is updated in bookkeeping database.

2. *Relevant Requirements:*MW1,MW4,MW10,MW11,MW12

4.7 Skim to create compressed sample

1. *Actions:*

- User prepares skimming executable on own machine.
- User **identifies secondary datasets** to reprocess.
- User **specifies a new dataset** for output of skim.
- The output skimmed data are either written to disk or stored as pointers to run/event numbers. Location and description of dataset is updated concurrently in data file catalogue.
- The status of the job is returned to the user.

2. *Relevant Requirements:*MW1,MW2,MW4,MW10,MW11

4.8 Compression of sample into Standard Ntuple

1. *Actions:*

- User **identifies secondary datasets** to process.
- User **specifies a new dataset** for output Ntuple.
- User submits script to an application which directs job to where data resides.

- Location and description of output ntuples are stored in bookkeeping database.
- The status of the job is returned to the user.

2. *Relevant Requirements:*MW1,MW2,MW4,MW10,MW11

4.9 Use Ntuples

1. *Actions:*
   - User **identifies secondary datasets** to retrieve Ntuple to process.
   - User submits a job with the requirement it run on his own machine to analyse Ntuples.
2. *Relevant Requirements:*MW11

4.10 Simulation of general signal and background samples

1. *Actions:*
   - User **specifies simulation parameters**. There is no need to prepare a specific simulation executable as this is part of the standard release of the CDF software.
   - User **specifies a new dataset** for output of simulation.
   - User **submits job**.
   - User monitors progress.
   - Output data is copied to mass storage and its location and description catalogued in the bookkeeping database.
   - The status of the job is returned to the user.
2. Relevant Requirements :MW1,MW2,MW4,MW10,MW11,DFC5

4.11 High statistics simulation for investigation of systematic effects

1. *Actions:*
   - User **simulates general signal and background samples** with the constraint that the output histograms, numbers and Ntuples are directed to his own machine and that no dataset be made.
2. *Relevant Requirements:* MW1,MW2,MW4,MW10,MW11,DFC5

5 CDFUK Resources

The CDFUK group has a variety of hardware resources, software resources and experience that can be used to meet the requirements stated. This section outlines the current state of CDF hardware available in the UK, the CDF "Middleware" tools, and experience gleaned from the CDF Level 3 trigger and database.

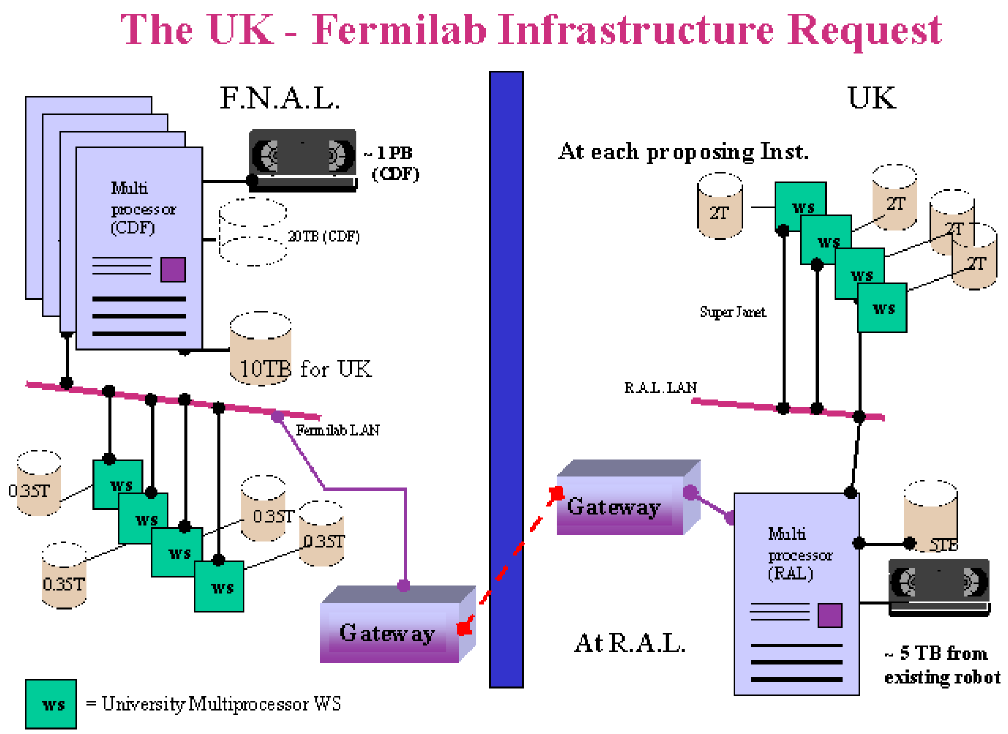

Figure 1: UK-Fermilab Infrastructure

5.1 Hardware Resources

CDF is expected to store over 1 Petabyte of data to tape in RunII. The UK CDF institutions have been awarded a 1.8 million pound grant from the Joint Infrastructure Fund (JIF) for computing equipment located in the UK and at the experimental site. About 4/5 of this award is being spent on highspeed highvolume disk storage with high bandwidth access to local CPUs while the remainder is allocated to lease a dedicated network link to Fermilab to improve interactive access to computing resources at the experiment and to provide internet video conferencing capabilities. A schematic of the infrastructure request is shown in Figure 1.

To date, a 10TB RAID has been provided to CDF and is being used to hold the express data stream. A similar sized contribution will be made to the experiment next year.

The JIF grant has supplied each of the four UK institutions on CDF (Glasgow, Liverpool, London, Oxford) with an IBM 8way SMP server connected via a Fibre Channel link to its own 1TB RAID. A similar system, twice as large, has also been provided to RAL for use by all the UK groups. Next year, the remainder of the grant will be used to approximately double the computing and storage capacity. UK institutes also have access to about 30 PCs at both Fermilab and the home institutions.

We intend to place the JIF computing and datastorage facilities on the UK GRID so that the CDF data and our combined computing power are seamlessly available, initially to all the UK groups, and later for access by the CDF collaboration. A further goal is to make our disk storage and CPU available to the whole particle physics community so that the CDF prototype GRID grows into, forms, and becomes part of the GRID for Particle Physics.

5.2 MAP Simulation Facility

The MAP facility at Liverpool is specifically designed for large scale simulation, and currently consists of 300 Linux boxes, each with local disk storage. Due to a recent SRIF award, MAP will be upgraded in December 2001 to 1500 processors. MAP has recently been GRIDenabled to allow remote job submission from authorized users, with job outputs returned to remote machines. MAP will satisfy all of the UK CDF simulation requirements, in addition to assisting with the LHCb and ATLAS data challenges (see respective proposals), provided sufficient human resources are available.

5.3 Current CDF 'Middleware' Tools

At present, we will proceed towards our physics goals in a completely manual and nonautomated fashion. Data sets are located at Fermilab and will be sent to (and around) the UK. Data sets in the UK will be tracked manually. All UK computers will be set up with the same directory structure and accounts and all UK CDF members have access to each computer. Local groups will have priority on their own machines. Large CPU usage will be scheduled by email or phone requests.

The following tools are at our disposal:

- ftp, bbftp and Gridftp for transferring the data. In addition there are automated scripts which search for new data, and transfer it when found.
- A 'Data File Catalogue Browser' which is a web based form to allow data of interest to be located.
- A staging facility named the 'Data Inventory Manager' that identifies whether data is on disk, and if not, stages it from tape.
- Standard data processing and simulation executables can be customised via tcl control commands.
- A mechanism by which the CDF database can be exported to remote computers.
- A system which schedules simulation jobs on MAP.

5.4 Level 3 Trigger

The UK CDF groups have the responsibility for creating the executables that are run on the CDF Level3 PC farm. This has many aspects that are of interest to GRID. For example, database bookkeeping methods have been devised for automatically tagging and keeping track of which code versions were used to build which executables in order to ensure strict reproducibility of the results. Methods have been devised for discovering and packaging up into tarballs all the dependencies (shared libraries, flat files etc.) of a given executable such that it can run on any machine with the correct OS installation. In addition, Level3 relies on the export into flatfile format of required database constants, again with the result that the executable can be run on machines without connections to the central CDF database servers.

5.5 CDF Database

The UK CDF groups have been major players in the design, implementation and management of the CDF database project, which is responsible for storing, cataloguing and retrieving calibrations, triggers, run conditions, hardware configurations and the data file catalogue. Experience has been gained in the following areas:

- Maintenance of the Oracle relational database management system: this includes defining the database architecture (ie. rollback, temporary and user segments, disk configuration including choice of RAID) and optimizing the parameters that control it (eg. system memory, simultaneous connections).
- Establishment of high availability (>99% up time) system due to online and offline requirements.
- Provision of a nonexpert environment to users.
- Provision of a webbased database browser for users.
- Simulation and monitoring of the database size, rate of growth and access patterns to ensure adequate performance.
- Replication. The online database is replicated offline, to protect online operations from offline use.
- API. Use of code generation to provide a standard C++ environment for read/write access.
- Export to remote institutes. Development of a prototype scheme for export of databases to remote sites.
- Evolution of schema to adapt to changing requirements.
- Manpower evaluation.

All of the items described above are relevant for the development and effective operation of the GRID.

6 Timescales and Goals

There are three distinct phases to our project. The first is a manual implementation and is currently underway. The second automates the handling of the data sets. The third automates the job submission and control.

6.1 Stage 1: Manual Implementation

- Arrange a common structure on all UK CDF machines.
- Make sure access is available to all UK CDF physicists. There may be some security issues here; at a minimum access from one UK CDF machine to another should be possible.
- Locate datasets at FNAL using the 'Data Catalogue Browser' and transfer to UK by ftp, possibly using automated scripts.

- Coordinate where datasets are located.
- Circulate this information by email, web pages, and word of mouth.
- Create executables by hand. Communicate with tcl file. Insert dataset name by hand.
- Use basic monitoring tools to show system load.
- Investigate a basic scheduling system (e.g. set the priority level so that local users have priority).
- Implement a basic sleep and kill facility so that local users have priority.
- Export the CDF database by hand to remote machines using existing tools.

These tasks are underway now. The timescale for completion is the end of 2001. They will be done irrespective of GRID developments and will use the existing UK physicist's effort on CDF and thus do not require additional manpower from GridPP.

6.2 Stage 2. Automatic handling of data sets

- Automatic registration and update of datasets: recording of movement, type, medium (disk, tape, box beside tape) and creation of datasets. Tools which might achieve this are LDAP, Globus, SAM etc.
- Monitor network traffic.
- Require a single method which returns the location of data, whether it be in the UK or at FNAL. (The FNAL end requires interfacing with existing CDF database.)
- Broadcast available CPU from each of the UK machines.
- Methods for the automatic distribution, suitability and performance of the CDF database (This is CDF specific.)
- More sophisticated monitoring: listing of jobs on each machine, with time elapsed and an estimate of time remaining.
- Interface of MAP simulation output with Data File Catalogue.

These tasks will automate the cataloging of the data sets. Jobs will still be submitted 'by hand'. For the most part, we will use common GRID tools similar to those we expect are required by the LHC experiments. There is some CDF specific implementation of them however, particularly in interfacing with the CDF databases. The timescale for completion is 'asap' since we are currently taking data. Realistically we suspect it will take one year. If it takes longer than 2004, then it is of little use to us. The tasks will not be achieved without new manpower.

6.3 Stage 3. Automated job submission

The job submission should be fully automated and seamless so that a user merely has to specify the executable and data. The decision on where to run the job takes account of available CPU, network conditions and disk contents. Possible tools are Globus, Condor etc. The timescale for this is 18 months to 2 years. Ideally it should be available by mid 2004 although it is still useful up until mid 2005.

7 Collaborative Effort and Manpower Needs

We expect that many of the tools we need are also required by the LHC experiments, BaBar and D0. Synergies with the SAM software of D0 clearly exist, and need to be examined to see if they satisfy our requirements. Hardware resources within the UK which can be pooled between experiments include the MAP simulation facility, and ScotGrid.

Contacts with Particle Physics Data Grid (PPDG) through Fermilab have been established and the Globus Toolkit will be installed on the CDF central systems and machines connected to the UK JIF network. Both Rutgers and the University of Chicago wish to collaborate with UK CDF on GRID issues. We have had discussions with Ruth Pordes, coleader of PPDG, who is anxious that a GRIDlike solution be found for data distribution at CDF. Currently, no one solution is envisaged for this; it thus presents an excellent opportunity to use Fermilab resources and for the UK to play a leading role in this extremely important initiative.

Under the assumption that GridPP will provide a common infrastructure for the UK physics experiments we estimate, based on our experience with databases at CDF, that about 2 FTE /year are required to interface the CDF specific elements to UK GRID. In addition, 0.5 FTE /year are required to develop and maintain the CDF interface to MAP.

8 Summary

CDF is in a position to offer concrete requirements and use cases based on the present experience of both commissioning and doing physics in a hadron collider environment.

Furthermore the CDF UK groups can offer the experience gained in their roles in design and coordination of the databases, the handling of data in level 3 and the operations of the processing farms. We believe that in addition to providing a testbed for data transfer and cataloguing for the GRID, CDF will help create, define and develop the GRID middleware tools for the future.

9 CDFUK Grid Addendum

This is an addendum to the rest of the document. This addendum outlines work that was continued after the original document was set down.

Since the submission of the requirements document given in Sections refsec:cdfuk-start to 8, work has continued on investigating different strategies for meeting those requirements. In particular, adapting for use by CDF the Sequential Access through Metadata (SAM) system developed by D0 continues to be an attractive way to proceed, given the relative maturity and degree of support for this product.

According to the best of our knowledge a SAM solution for CDF is feasible to implement, but we do need a "prototyping period" before this begins in earnest. During this period we would not only examine SAM but also look in detail at existing tools developed in the context of EU DataGrid and carefully assess the state and applicability to the CDFGRID of existing CDF middleware. Note that many of the Globus tools used by the EU DataGrid are likely to be integrated with SAM anyway, as part of the development of that product by Fermilab and D0. An important part of the "prototyping period" will be a detailed examination of existing and forecast usage patterns and user requirements. We want to know what people are doing now and how this can evolve into the use of a fullyfledged CDF GRID.

9.1 Project Overview

1. **Prototype phase:** We would initially take on the following items for evaluation within the CDF framework:

    - Usage patterns and user requirements,
    - EU DataGrid tools,
    - SAM.

    This evaluation would be completed within the first 3 months of the project.

2. **Implementation phase:** For the purposes of the rest of this document, we assume that we will adopt SAM. Furthermore the hardware will primarily be the multiple SMP servers and storage capacity purchased with Joint Infrastructure Funds (JIF) at the four universities and Rutherford Laboratory.

We emphasise however that should the framework or tools developed by EU Data grid prove more appropriate during the prototyping and evaluation phase, those tools could be adopted. To this end we need to pursue the following projects:

- Import, deployment and support of SAM stations in the UK.
- Adaptation of parts of SAM to function with CDF. This includes interfacing the CDF data handling tools so that they can translate cataloguing information to SAM. A key component of this task will be the implementation of Database schema compatible with SAM (or even in common with D0).
- Integration of the work package 2 tools for the replica catalogue, as we are required not only to use SAM but also to contribute to its further development into GRID. A separate development is required here for CDF because of differences that exist in the Event Data Model and how data files are catalogued between CDF and D0.
- Integration of the Liverpool MAP facility and JIF acquired PC farms into the SAMCDF framework.

These tasks take place after the pilot phase of the project.

We request the following resources:

- Two posts to work on adapting SAM for the CDF environment and for developing interfacing within SAM as is indicated during this adaptation exercise. We are basing this requirement on manpower estimates derived by Vicky White (Computing Division, F.N.A.L.) and presented at the CAF meeting. Consequently, two posts are required due to the large number of subtasks this adaptation requires. Among those subtasks are:
    - Set up SAM stations on the JIF equipment both in the UK and at F.N.A.L.
    - Interface SAM to the CDF framework
    - Interfacing the Liverpool MAP facility to SAMCDF
    - Maintenance of the middleware
- One post to develop the Grid User Interface Application and data catalogue browsing, aiding the current FNAL effort.

10 CDF and D0 Executive Summary

This section documents how the project outlined thusfar can be combined with the efforts of D0. This gives the full overview of the joint portion of the project. The Glasgow, Oxford and UCL specific responsibilities are given in Sections 11 and 12.

10.1 Introduction

CDF and D0 have many requirements in common in using Grid to access data and computing resources. We request posts from the GridPP project to work: firstly on the continued development and enhancement of SAM using standard and currently evolving Grid software; and secondly on experiment specific user applications .

10.2 Grid enhancements to SAM

The following outlines the programme of work to extend SAM by investigating, developing and incorporating other Grid tools. This work is common to both CDF and D0 and lasts throughout the two year timescale of the project.

- **1** Use of standard middleware to promote interoperability
  - **1.1** Use of Globus security infrastructure for Station to Station services, such as file transfer and job submission. Interoperability with Fermilab Kerberos security infrastructure and use of recognized Certificate Authorities on both sides of the Atlantic, in conjunction with work of other Grid projects.
  - **1.2** Use of GridFTP as one of the supported file transfer protocols
  - **1.3** Use of future enhanced Globus security infrastructure, including Community Authorisation Service
  - **1.4** Use of Globus as one supported system for job submission.
  - **1.5** Use of Condor and extensions as one supported system for job submission
  - **1.6** Publish availability and status of SAM station resources and services using standard Grid middleware components and services, when they emerge.
  - **1.7** Publish catalogue of data files and their replicas using standard or standards emerging from PPDG and DataGrid
- **2** Additional Grid functionality for Job specification, submission and tracking
  - **2.1** Use of full Condor services for migration and checkpointing of jobs – as much as is possible with D0 software and the D0 software framework. This may require work on both Condor and D0 software to achieve full functionality
  - **2.2** Building incrementally enhanced Job specification language and job submission services that ensure colocation of job execution and data files and reliably execute a chain of job processing, with dependencies between job steps. The first step in this is expected to be work in conjunction with the Condor team to provide for specification and execution of a Directed Acyclic Graph of jobs using an extended version of the DAGMAN product that CMS is testing for their MC job execution.

- **3** Enhancing Monitoring and Diagnostic capabilities
  - **3.1** Extensions to existing system of logging all activities in the system to both local and central log files as demanded by robustness and increased use of system.
  - **3.2** Incorporation of emerging Grid Monitoring Architecture and monitoring tools. Little exists on this at this point and this work will involve working with other Grid projects and participating in Global Grid Forum working groups.
- **4** Fabric Management: Optimisation of the CDF and D0 fabric which include the UK CDF JIF computers, the UK D0 JIF computers, the Lancaster farms, the Liverpool MAP facility, and future computing and data storage facilities such as Mosix or SMP clusters.

10.3 CDF Integration into SAM

CDF does not currently use SAM. Thus to contribute to this project, CDF must first be integrated into the existing SAM structure. The first step consists of a pilot project in which we will: set up the initial fabric; evaluate usage patterns and user requirements; and deploy SAM as well as standard Grid tools for evaluation. This pilot project will be completed within the first 3 months.

It will then be possible to begin integrating CDF into SAM. The following steps will be necessary and will take one year:

- Deployment and support of SAM stations in the UK
- Adapt CDF software to function with SAM. This includes interfacing the CDF data handling tools so that they can translate cataloguing information to SAM. A key component of this task will be the interfacing of the Database schema compatible with SAM (or even in common with D0).
- Adaptation of Data input modules to SAM. A separate development is required here for CDF because of differences that exist in the Event Data Model and how data files are catalogued between CDF and D0.

10.4 Maintenance and Support of SAM and Grid

Since CDF and D0 are running experiments and will continue to take data throughout the duration of the project, reliability and ease of usage are of paramount importance. We therefore identify projects, common to CDF and D0, to provide:

- user support, monitoring and diagnostic tools;
- fabric development, deployment and management;
- maintenance.

All of the above will ensure minimum disruption to the user, allowing development to proceed in parallel and at the same time easily incorporating improvements and advances in the system.These projects will be supported by manpower already working on the D0 and CDF projects and displaced effort.

10.5 Request for Posts

We request a total of 6 additional FTEs to work on the development of SAM and user applications of the experiments. The UK CDF and D0 groups are currently supplying approximately 23 FTEs to the development and maintenance of SAM within the context of the UK groups.

Four of the additional FTEs will work on integrating CORE Grid functionality into SAM as described in Section 10.2 . These positions would not be working on user applications. These FTE's would work in conjunction with the SAM project team at Fermilab, the PPDG and their collaborators. These posts will concentrate on the projects essential for meeting the proposed deliverables given in Section 10.6 and meeting the goal of having a functional GRID using modern grid tools working as proof that the GRID will work for future experiments.

The remaining two FTEs will work on developing user applications with one FTE per experiment.

D0 propose to have one FTE work on the developpment of its Monte Carlo Production system as the user application. This project is described in full in the separate D0 document.

CDF will allocate the user application FTE to work on the development of an interface for CDF user applications to SAM. The UK groups will supply an additional 1 FTE to work on the integration of CDF code with SAM. The FTEs working on Core projects will also offer support where required. Two FTEs are required over the first 15 months to adapt SAM for the CDF environment and for developing interfacing within SAM. We base this requirement on manpower estimates derived by Vicky White (Computing Division, F.N.A.L.) and presented at the CAF meeting.

10.6 Deliverables

The deliverables are specified in Table 3.

11 Glasgow CDF/D0 SAM/GridPP Integration Post

Primary investigator: Dr. R. St. Denis
Coinvestigator: Dr. A.S.Thompson, University of Glasgow

The text of this section is taken from a proposal that was successful. Thus, the work will be carried out as described in this bid.

The Glasgow group bids to GridPP for a post to work with D0UK and the Fermilab computing department to integrate the WP2 deliverables of GDMP and Spitfire into the SAM framework and SAM into the CDF framework. The implementation is through recruitment of new effort.

11.1 Deliverables and Milestones

1. Adaptation of CDF to SAM.

    - Interface the CDF data handling tools so that they can translate cataloguing information to SAM. A key component of this task will be the interfacing of the Database schema compatible with SAM (or even in common with D0).

Item	Time	Description
1	Month 3	CDF Pilot project ends with test of Globus tools and SAM. A deployed SAM station at one or more CDF institutes should be available to transfer files for data analysis.
2	Month 6	Integration of Globus Security infrastructure and GridFTP into SAM and deployment at several UK stations interoperating with Fermilab and other SAM stations.
3		First demonstration of MC production system using Request interface and automated job submission to one SAM station with limited intelligence in job distribution and load balancing.
4	Month 12	Fully commissioned MC Production System with reliable execution of jobs, splitting into subjobs as necessary, intelligent job distribution and load balancing, taking into account the economics of data movement versus job movement.
5	Month 24	A fully robust production quality system, excellent monitoring.
6		Interoperability with other Grid projects and EU DataGrid with sharing of some resources. Updated Fabric capable of handing the data on each experiment.

Table 3: Deliverables in the Joint D0/CDF GridPP project.

- Adaptation of Data input modules to SAM. Differences in the Event Data Model and how data files are catalogued between CDF and D0 require a separate CDF development.

2. Integration of Grid and SAM

   - **1.2** Use of GridFTP as one of the supported file transfer protocols.
   - **1.7** Publish catalogue of data files and their replicas using PPDG and DataGrid standards.
   - **2.2** Building incrementally enhanced Job specification language and job submission services that ensure colocation of job execution and data files and reliably execute a chain of job processing, with dependencies between job steps.

11.2 Work already being done

- Sam stations set up at Glasgow and UCL;
- Interface to AC++ to talk to SAM;
- Installation of Globus on CDFUKJIF machines;
- WP2 delivery of GDMP and Spitfire and participation in TestBed 1.

11.3 Support/infrastructure available locally

- WP2 center
- CDFUK Jif Glasgow Machines as described in detail in Section 5.

11.4 Match to internal priorities/plans of CDF/D0UK

- Operate in the context of the CDF Offline project's Data Handling Group to provide distribution of secondary data sets to the UK, using CDFJIF infrastructure.
- Ability to handle the data sets as outlined in detail in the main CDF submission to the GridPP.

11.5 Main objectives

In collaboration with other institutes, provide a job submission system available on user workstations that utilize CDF and D0 computing resources optimally for Monte Carlo production and data analysis of secondary and tertiary data sets as outlined in detail in use cases submitted to the GridPP. B) Provide experience for the Grid usage in preparation for LHC with data sets and physics analysis requirements that match those of the LHC era.

12 Oxford/UCL bid for a CDF GridPP Post

Dr. B. T. Huffman (Oxford), Dr. D. Waters (UCL)

12.1 Introduction

The text of this section is taken from a proposal that was successful. Thus, the work will be carried out as described in this bid.

The Oxford University and University College London (UCL) CDF groups request one post at the RS1A.3 level (National spine Point 6) to work in the area of the integration of the CDF and SAM analysis frameworks. The details of the overall work plan are outlined extensively in two previous sets of documentation by the CDFUK groups and also by the D0UK groups [both documents are on the web site: http://www.gridpp.ac.uk/eb/ExptBoard.html]. The post would be based at Oxford University but would work closely with the GridPP groups at University College London, Glasgow University, Lancaster University, and Imperial College in order to develop a common SAMCDF framework. This post is assigned the highest priority by the CDFUK groups.

The rest of this document follows the points spelled out in the request for tenders distributed by the Collaboration Board chairman.

12.2 Experiment

Tender is for CDF

12.3 Deliverables and Associated Milestones

- Month 12: The primary deliverable and objective in this project is the integration of the CDF data handling system into the SAM framework.
- Month 18: Comparison and test of SAM, Globus, and other Grid system tools to determine applicability and interoperability with the current CDF framework.
- Month 24: Design of framework interface to the appropriate Grid tool (for concreteness we will assume this is SAM, but could well be a different tool should the overall CDF/D0 group decide otherwise).

The items 1,3,4 and 5 of Table 3 are items to which this post should also contribute.

Additionally there is an aspect of Grid maintenance to the project. The post would be expected to contribute to the maintenance of the Globus/SAM framework.

This post will work closely with members of both the CDF and D0 groups here in the UK and at Fermi National Accelerator Laboratory in the USA. Therefore, in order to accomplish the deliverables in the stated times, money for travel to the laboratory and within the UK will be required.

12.4 Current effort

A web-based GUI has already been developed at Oxford using JAVA tools that allows remote database access in a relatively userfriendly manner. This work, using LDAP, was a first attempt at integrating the CDF analysis framework into the Grid with the Globus toolkit. (A working version of this interface was demonstrated during the regular meeting of DataGrid WP8 by Dr. Huffman at CERN in June 2001.) Dr. Huffman oversees the hardware supply of equipment to all four CDFUK institutions and Rutherford Laboratory and is the CDF representative on the GridPP Experiments Board. In conjunction with the Oxford Computing Laboratory (ComLab), a position exists which will be working on UML modelling of the Globus toolkit (and potentially other Grid frameworks) in association with the Oxford CDF group. In addition, Oxford is the location of one of the Escience centres under Prof. Paul Jeffreys. Oxford also has one LHCb position working on their framework on the Grid and will be requesting at least one other position to work on a GUI for the ATLAS/LHCb framework.

UniversityCollege London has a number of staff and students working on Grid and Grid related technologies. In particular it is heavily involved in network infrastructure and services under the auspices of WP7 of the EU DataGrid. As well as Grid work in relation to the CDF experiment, UCL has an interest in the development of a Grid for ATLAS. Furthermore, strong links exist between the High Energy Physics and Computer Science departments at UCL.

The CDF group at UCL has extensive experience of developing and using the CDF analysis software infrastructure, which will be invaluable in adapting it to use Grid tools, presumably in the context of SAM. UCL has also been active in other areas relevant to distributed computing such as the export of the CDF calibration database to remote sites.

12.5 Implementation of Deliverables

The deliverables would be implemented by advertising and filling a new position. There are several graduate students at Oxford University who are near graduation who have expressed an interest in this kind of position and they would be encouraged to apply. Dr. Huffman will oversee the position and take care of the administrative burden. Dr. Waters and Dr. Huffman would direct the work programme.

12.6 Locally available infrastructure and support.

Oxford University's Particle Physics subdepartment is the largest in the UK and also benefits from close proximity to Rutherford Appleton Laboratory. Oxford University is the principle budget holder of a Joint Infrastructure Fund Grant which supplies the equipment that will be used by the group in delivering the use cases outlined in the previous documentation (reference). Full office and IT services will be available to the holder of the post in accordance with an RS1A position. Easy access to train and bus services into London will further aid interactions with Dr. Waters at UCL and with the remaining groups in the UK. UCL benefits from close proximity to Imperial College London, whose D0 collaboration members are closely involved in work on SAM.

12.7 Match to the internal priorities and plans of the UK collaboration

The CDFUK collaboration fully supports this tender and has ranked it as the number one priority. The CDFUK group fully supports this request and agrees that it falls in line with the programme of work outlined in our previous documentation cited in the opening paragraph of this document.

12.8 Main Objectives

Listed are the main objectives of this position:

- Comparison and test of SAM, Globus, and other Grid system tools to determine applicability and interoperability with the current CDF framework.
- Design of framework interface to the appropriate Grid tool (for concreteness we will assume this is SAM, but could well be a different tool should the overall CDF/D0 group decide otherwise).
- Implement and maintain a CDF compatible interface to the SAM framework.
- Implement and maintain the necessary infrastructure for CDF data and metadata transport and remote job submission.
- Interact closely with the other CDF and D0 group members working on SAM and GridPP in the design of the interface.

13 Appendix: The Data File Catalogue

The Data File Catalogue (DFC) is the currently existing utility used to store and retrieve information about CDF datasets. It consists of a relational database (currently implemented in Oracle) together with an API to query the database from the CDF software. Datasets are described as collections of filesets, which in turn are described as collections of files. Files themselves contain run sections, which are the smallest units of data described in the DFC and correspond physically to approximately 30 seconds of datataking. The DFC also contains the mapping between filesets and tape partitions in the CDF tape store.

Datasets are specified to CDF applications in the fragment of the Tcl control file read by the Data Handling Input Module (DHInputModule). The DHInputModule gathers the tape partition numbers corresponding to that dataset's filesets by querying the DFC. The DHInputModule communicates these to a program called the Disk Inventory Manager (DIM) that contains a list of which tape partitions are currently staged to disk. The required filesets already present on disk are passed back to the DHInputModule to begin processing while the tape partitions that need to be staged are passed to a tape daemon to initiate the relevant tape mounting and staging operations. A key requirement in the overall design of these datahandling elements has been transparency to the enduser, in the sense that all the necessary tape and disk manipulations go on behind the scene once the datasets themselves have been specified.

The DFC and DIM together are similar in concept to a GRIDstyle replica catalogue. It is indeed forseen that DFC's will be installed to manage storage at remote sites as well as at Fermilab. Clearly though, there are many issues that need to be addressed in the context of GRID. For example:

- Event level metadata. The smallest unit of data is the run section, as described above. This has the consequence that "virtual" datasets formed, in the most nave sense, from lists of run and event numbers, cannot easily be accommodated in the current DFC scheme.
- Site specific information. The DFC and DIM have so far been developed and run only at Fermilab. Assumptions are made about local file structures, for example relating to the staging area path name. Therefore it is likely that some work would be needed to fully separate these utilities from site specific information that should be stored and accessed in appropriate database entries.

- Links to other metadata. Other metadata such as run conditions and trigger information is not directly linked to the DFC schema. Therefore the DFC as it stands is not a true metadata store.

14 Acknowledgements

We wish to thank the D0 SAM team for all their help and patience in describing the system and for collaborating with us on this project.

15 Bibliography

References

[1] CDF Note 4718:M.Albrow et al ., "Run-II Trigger Table and Datasets Plan"