PROCEEDINGS OF SCIENCE

# Building a distributed software environment for CDF within the ESLEA framework

**Valeria Bartsch**[*], **Mark Lancaster, Nicola Pezzi**
*University College London*
*E-mail:* bartsch@fnal.gov

A fast optical link (UKLight/StarLight) between UCL and the Fermi National Accelerator Laboratory (FNAL) in Chicago was established in 2004. It has been used by the CDF collaboration to send data between the USA and the UK at rates in excess of 500 Mb/sec and forms part of the CDF data-handling infrastructure. The issues involved in setting up the link and the CDF data handling system are described. The distributed grid environment and the problems encountered in establishing a data analysis environment utilising the link are also briefly described.

*Lighting the Blue Touchpaper for UK e-Science - Closing Conference of ESLEA Project*
*March 26-28, 2007*
*Edinburgh*

[*] Speaker.

## 1. Introduction

ESLEA is an EPSRC funded project which is seeking to establish a proof-of-principle demonstration of the utility of dedicated, guaranteed bandwidth light-paths in applications needing to transfer large volumes of data across Wide Area Networks (WANs). The ESLEA project utilises the UKLight switched circuit optical network to enable guaranteed high bandwidth network links for a range of eScience applications. The CDF experiment [1] is one such application endeavouring to exploit the potential of high speed optical network links.

CDF is a particle physics experiment trying to elucidate the fundamental nature of matter. It is presently taking data from proton anti-proton collisions at the Tevatron collider at FNAL in Chicago. Analysis of 2 Pb of raw data is underway by almost 800 physicists located at 61 institutions in 13 countries across 3 continents. The amount of raw data and the need to produce secondary reduced datasets have required new approaches to be developed in terms of distributed storage and analysis. Grid systems based on DCAF [2] and SAM [3] are being developed with the aim that 50% of CDF's CPU and storage requirements will be provided by institutions remote from FNAL. In order to effectively utilise this distributed computing network it is necessary to have high speed point-to-point connections, particularly to and from FNAL which have a bandwidth significantly higher than commonly available. To this end, as part of the ESLEA project, the use of a dedicated switched light path from FNAL to UCL was investigated. However in order to utilise the data sent to UCL from FNAL it was necessary to deploy the CDF data handling system at UCL and also make the CPU resources available to the rest of the experiment as part of a distributed grid. The issues associated with setting up this infrastructure, which necessarily formed a large part of the project, are also described.

## 2. CDF/ESLEA Objectives

Figure 1 shows the data flow of the CDF experiment. Raw data from the detector is accumulated at a rate of 2TB/day. This raw data is then *re-processed* which typically involves calibrating the data and applying more sophisticated algorithms to the data such that it can be utilised in physics analyses; this is termed *reconstructed* data. In general a given user then analyses a sub-sample of this reconstructed data and this analysis is facilitated by comparing the data to simulated or *Monte-Carlo (MC)* generated data. The generation of MC data is CPU intensive and typically performed on grid farms with the output data being stored centrally at FNAL. The re-processing of data for CDF takes place at FNAL. The user-defined datasets need to be made available to users across many institutes and in general they are distributed across different storage locations both within and outside FNAL. The two key objectives of the CDF/ESLEA project were:

- To generate a significant fraction ($\sim$ 10%) of CDF's MC need using grid farms at UCL and use UKLight/StarLight to send the generated MC data to FNAL.
- To make UKLight/StarLight available to users for the rapid transfer of user datasets from FNAL to the UK and then subsequently to German and Italian institutes.

To achieve these objectives, in addition to commissioning the UKLight/StarLight link, it was necessary to modify and port CDF's bespoke data-handling and grid-interface software to UCL so that it could be utilised within the context of the CERN LHC computing grid (LCG) [4] which remains the only supported particle physics grid system at UCL. In the following section, we briefly describe CDF's data-handling and grid interface infrastructure before describing the commissioning and usage we made of the UKLight/StarLight link.

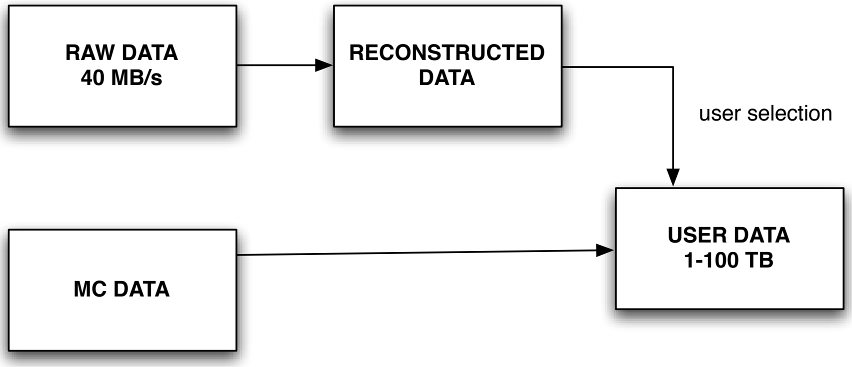

**Figure 1:** *The data flow and types of data utilised by the CDF experiment.*

## 3. CDF's Data Handling and Grid Analysis Systems

Transparent access to CDF's data is provided through a custom data-handling and cataloguing system called SAM [3]. SAM is used to store, manage, deliver and track the processing of all data. Each storage location is registered with a central SAM ORACLE database at FNAL and has an associated SAM server/station managing the local data. Users, through CORBA enabled SAM-clients, make requests for data and the data is delivered to a temporary local cache from the appropriate SAM server(s). In this way physicists have transparent access to the CDF data. At present $\sim$ 500 Tb of data is distributed across SAM servers in the US, Europe and Asia. For a given user it is most efficient if the majority of the data of interest to them resides at a local SAM station. The rapid population of a local SAM station with datasets of interest to UK physicists through UKLight/StarLight was one of the key aims of the project. Much of the development work to establish SAM as CDF's default data-handling system was carried out in the UK. Figure 2 shows the amount of data transferred through the SAM system per month for selected SAM stations, including the UCL station that was connected to FNAL via UKLight/StarLight.

Traditionally within CDF the analysis of data was done on dedicated resources of commodity nodes at FNAL managed as Condor [5] pools with a wrapper of CDF specific software around the batch system. This so-called CAF [2] system though has limitations within the context of farms at universities which are not easily customised and which, particularly in Europe, only have an LCG interface to the local batch system. However, the maintenance and monitoring of the bespoke scheduling and job submission software within the CAF system requires a significant manpower commitment which was not available within this project. Furthermore, the requirements of the CAF system that one node be outside of the firewall were not compatible with local internet security restrictions. It was therefore not possible to mimic the FNAL CAF system at UCL. We therefore devoted considerable time in evaluating and developing a more tractable solution that could make UCL CPU resources available via UKLight/StarLight utilising LCG and the SAM data handling system as opposed to deploying the bespoke CDF CAF system.

We successfully established user authentication with FNAL servers (via kerberos) and LCG servers (via grid certificates and a CDF VO) such that we could utilise the LCG resources at UCL using the gLite [6] interface. However within the time of the project we were unable to make the resources available for specific CDF analysis work owing to an instability in the software (PARROT/SQUID) serving the bespoke CDF analysis software to the batch nodes and the lack of an LCG compliant SRM interface within SAM. These issues are currently being worked on and expected to be resolved before the end of 2007.

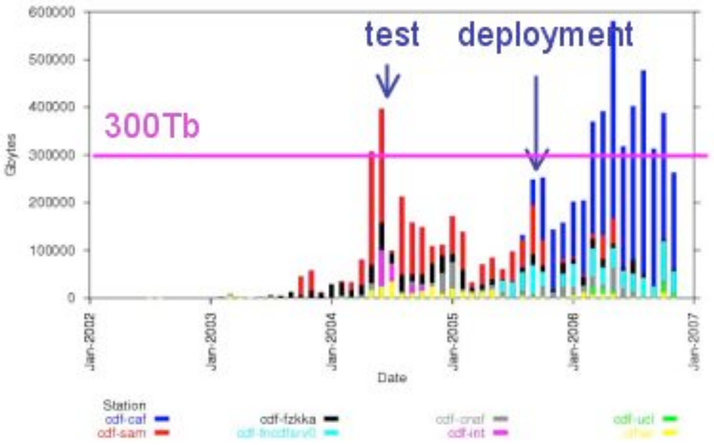

**Figure 2:** *Data read by the major SAM stations, cdf-caf and cdf-sam are located at FNAL, cdf-cnaf, cdf-ucl and cdf-fzkka are located in Europe.*

## 4. Data Transfers For CDF using UKLight/StarLight

Over the course of this project there has been a general shift of CPU resources to institutes remote from FNAL and now 50% of CPU resources are remote. The de-centralisation of the CPU has also required a similar migration of the data which in order to not hamper analysis progress has had to utilise the fastest available networks such as UKLight/StarLight. Typical CDF secondary datasets that are used for physics analyses are presently 1-100 Tb in size. Typical transfer rates from FNAL to Europe (UCL) using the standard network are approximately 25 Mb/sec (for multiple streams). A 50 Tb dataset would thus take approximately 6 months to copy from FNAL. This is comparable to the entire time that a CDF physicist would spend analysing the data in order to produce a publication. CDF produces in excess of 40 publications per annum. The datasets themselves are typically distributed in many files, each approximately 1 GB in size. To be useful to CDF, we required that UKLight/StarLight:

- Deliver a throughput of $>$ 500 Mb/sec such that typical datasets could be made available on a timescale of a week.
- Deliver files without corruption.

Since real-time analysis of a data-stream is not undertaken, a modest retransmission rate of files/packets at the 10% level was acceptable and the order in which files were received was not critical.

A dedicated 1 Gb/s circuit connecting UCL and FNAL utilising UKLight/StarLight infrastructure was setup late in 2004 and ultimately satisfied our requirements with transfer rates above 500 Mb/sec sustained over several days. However this was only achieved after incremental modifications to the hardware and software configurations that we briefly describe below.

In the initial period of the project, the link was unavailable for several months due to technical problems with switches and its prioritised use in demonstrations at conferences. We utilised this downtime to optimise the performance of the dedicated UCL PC connected to the optical network. It was clear from initial tests using "iperf" and "dd" that the PC was not configured in an optimal way for transferring large data volumes across the network and for writing this data to disk. The kernel TCP settings were modified [7], the file-system was re-formatted as an XFS file-system and parameters of the 7.5 Tb SCSI RAID-0 array were modified [8]. After these modifications we were able to make memory to memory transfers between PCs on the UKLight network at 950 Mb/sec and write to disk at 750 Mb/sec.

Ultimately we were interested in the disk to disk transfer rate via gridFTP [9] (as implemented in the SAM software package) from the CDF data pools at FNAL to the disk on the UCL UKLight PC. Initial tests could not deliver a throughput higher than 250 Mb/sec. This low rate was due to three factors:

- A switch within the CDF/FNAL network that was not appropriate for the network.
- Concurrent disk access from other CDF users at FNAL.
- Non-optimal gridFTP settings.

It also became clear that it was extremely important in diagnosing problems of the first type to have a good and regular communication channel with the FNAL experts and people controlling the FNAL hardware. Redundancy in the hardware and control over all steps of the network was vital in achieving a reasonable throughput. This control was ultimately only achieved by establishing a good working relation with the FNAL/StarLight experts over a period of time through regular presentations at their weekly network meetings. Redundancy at the UCL end and the ability to support multiple ftp streams was achieved by adding a second identically configured PC to the UKLight/StarLight network at UCL. Our highest throughput was achieved using 20-25 parallel gridFTP transfers concurrently to two UCL PCs. Typical transfer rates [10] are shown in Figure 3 which shows peak rates in excess of the required 500 Mb/s. The reductions are due to disk contention from other users accessing the same CDF/FNAL disks that we were copying data from.

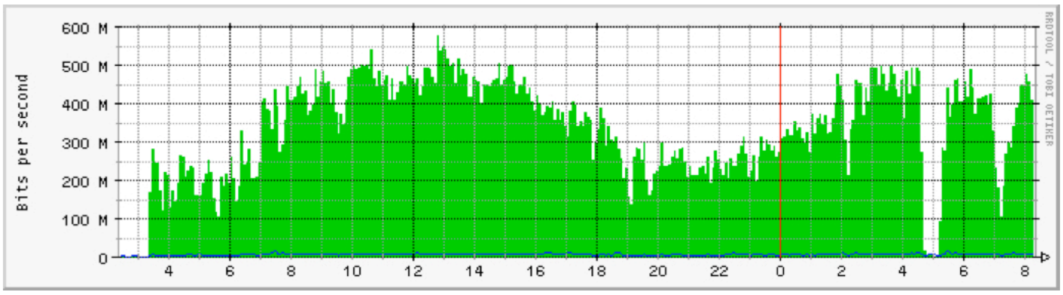

**Figure 3:** *Typical data transfer rates achieved across UKLight/StarLight over a 24-hour period between the CDF/FNAL storage and the UCL storage.*

## 5. Summary

The key objective to transfer files from CDF/FNAL to the UK over UKLight/StarLight at a throughput in excess of 500 Mb/s was achieved. It was however not possible to make the UCL CPU resources available to CDF via UKLight/StarLight due to problems encountered in deploying the CDF analysis and data-handling software within the supported LCG environment. As a proof-of-principle the project illustrated that a fast optical network can be deployed within the context of a running high energy physics experiment and deliver the required bandwidth. However the issues associated with trying to deploy highly specialised and custom software from a running experiment within a generic environment such as LCG need further work before the network could be used in a production capacity.

## References

[1] The CDF experiment, `http://www-cdf.fnal.gov`

[2] DCAF, Decentralised Analysis Farm for CDF, `http://cdfcaf.fnal.gov`

[3] SAM, Sequential Access to Metadata, see CHEP04 conference contribution, `http://projects.fnal.gov/samgrid/conferences/chep04/chep04.html`

[4] LCG, `http://lcg.web.cern.ch/LCG`

[5] Condor project homepage, `http://www.cs.wisc.edu/condor`

[6] gLite, Ligthweight Middleware for Grid Computing, `http://glite.web.cern.ch/glite`

[7] Modifications were made to 4 files.
/etc/sysctl.conf :

```
 kernel.core_uses_pid = 1
vm.max-readahead = 2048
vm.min-readahead = 1024
```

```
  kernel.shmmax = 1073741824
```

/proc/sys/net/core/wmem_max :

```
   8388608
```

/proc/sys/net/core/rmem_max :

```
   8388608
```

/proc/sys/net/ipv4/tcp_rmem:

```
   4096 87380 4194304
```

[8] The parameters of the RAID array were modified using "dellmgr" to:

```
   Stripe Size = 128kb
  Write Policy = WRBACK
  Read Policy = READAHEAD
  Cache Policy = CACHED_IO
```

[9] gridFTP, `http://www.globus.org/alliance/publications/papers/GFD-R.0201.pdf.`

[10] Network performance for the UKLight/StarLight was monitored from the URL: `http://tinyurl.com/2ejd2a`