

CDF's Utilisation of Switched Circuit Optical Networks

V. Bartsch^{b,c}, P. Clarke^d, **B. Garrett**^a, C. Greenwood^d, M. Lancaster^c, N. Pezzi^c

^a University of Edinburgh, Scotland, UK; ^b Fermilab National Accelerator Laboratory, Illinois, USA; ^c University College London, England, UK; ^d National E-Science Centre, Scotland.

Abstract

The ESLEA project utilises the UKLight switched circuit optical network to enable guaranteed high bandwidth network links for a range of eScience applications. The high energy particle physics (HEP) experiment, Collider Detector at Fermilab (CDF), is one such application endeavouring to exploit the potential of high speed optical network links. CDF produces a large amount of raw data (presently 1000 Tb per annum) which needs to be made available to 800 physicists who are analysing the data at 61 institutions in 13 countries across 3 continents. Presently the majority of the data analysis proceeds at the site of the experiment at Fermilab just outside of Chicago. However the recent commissioning of the optical networks : StarLight in the USA; UKLight in the UK; and NetherLight in the Netherlands, has opened up the possibility that European physicists could host a significant fraction of CDF's data and make available CPU resources for its analysis to CDF physicists worldwide. Initial transfers of data from Fermilab to University College London (UCL) in the UK have begun using the StarLight and UKLight networks and the Grid middleware software is presently being commissioned on a 200 PC Linux cluster at UCL.

1. Introduction

CDF[1] 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 world's highest energy colliding beam facility, the Tevatron at Fermilab, which is located just outside Chicago in the USA. The experiment currently produces approximately 1PB of raw data per year and will continue to do so until 2009. Analysis of this 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 Fermilab. In order to effectively utilise this distributed computing network it is necessary to have high speed point to point connections, particularly to and from Fermilab, which have a bandwidth significantly higher than commonly available. To this end, as part of the ESLEA[4] project, the use of a dedicated switched light path from Fermilab (using the US Starlight network) to UCL in the UK (using the UKLight network) is being investigated. This paper describes initial experiences of utilising Grid middleware in a switched light path environment for the CDF experiment and the transfer rates that have been achieved.

2. Data Transfer Needs

Typical CDF secondary datasets that are used for physics analyses are presently 1-50 Tb in size. The total size of these datasets is expected to quadruple in the next 2-3 years. The CPU resources required to repeatedly analyse such datasets will exceed those available at Fermilab and so the datasets need to be distributed to centres in Europe and Asia to facilitate the use of their significant CPU resources. Typical transfer rates from Fermilab to Europe (UCL) using the standard network are approximately 25 Mbit/sec (for multiple streams). A 50 Tb dataset would thus take approximately 6 months to copy from Fermilab. 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 20 publications per annum. Datasets therefore need to be made available to a physicist on the time-scale of days not months. A transfer rate of 700 Mbit/sec would allow 50 Tb datasets to be transferred to Europe in under a week, which is sufficiently quick to not significantly delay an analysis. The datasets themselves are typically distributed in many files, each approximately 1GByte in size. Each file must be received without corruption, but the order in which files are received is not critical. The requirements on data-loss/delays are thus not critical since there is no real-time analysis of the files as they are copied. It is only the integrity of the entire dataset that is necessary. Retransmissions of files and packets can thus be tolerated at the 10% level. The fast network needs to be made available to physicists on the time-scale of minutes, preferably using a

command-line request utility to the control plane software[5]. The link needs to be maintained for the duration of the large dataset transfers which typically would be from a few hours up to a week at a time. The use of the UKLight/StarLight network will expedite the distribution of datasets to a large number of CDF physicists and will ensure that CPU resources in Europe can be used to alleviate the load at Fermilab and thus allow the continued, timely completion of CDF physics publications in the next 1-3 years.

3. CDF's Data Handling and Analysis Systems

The analysis model of data in high energy physics is highly sequential and is generally carried out on dedicated Linux PC analysis farms. Typical analysis farms contain 50-500 Linux PCs and CDF presently has 5000 GHz of exclusive CPU. Presently such farms run custom CDF analysis environments (DCAF), which consist of a CAF cluster environment (see Figure 1), the CDF software, a SAM Station. Authentication is centrally controlled via Kerberos tickets authenticated at Fermilab.

To support the distributed storage and analysis of datasets CDF has developed a custom data handling and cataloguing system called SAM[6] (Sequential Access to Metadata). Each site hosting CDF data has an associated SAM server and physicists use SAM client software based on CORBA to request data. If the data is not available through a local SAM server the data is copied using gridftp from another server. The metadata pertaining to the files under the control of a given SAM server are stored centrally in an ORACLE database at Fermilab. In this way physicists have transparent access to the CDF data. At present 330 Tb of data is distributed across SAM servers in the US, Europe and Asia. Although SAM servers exist in the UK they have yet to be populated with significant amounts of data owing to the low bandwidth that has so far been available between the UK and the primary data storage locations at Fermilab. With the commissioning of UKLight/StarLight this will change and UK sites, starting with UCL, will be able to host large amounts of CDF data.

Ultimately it is the aim that such farms will use more generic grid middleware such as that being developed at Fermilab (SAMGrid[6]) and that being developed at CERN (LCG[7]) which interface to GLOBUS and CONDOR software. The JIM[8] (Job and Information Management) components are part of the SAMGrid software and complement the Data Handling system SAM, providing the user with transparent remote job submission, data processing and status monitoring. The logical entities of the SAMGrid

(see Figure 2) consist of multiple execution sites, which steer the jobs behaviour on the distributed analysis farms; a central resource selector, one or more job submission sites which take care of the authentication and the handling of the input tarball until it is transferred to the execution site; and multiple very light-weight user interfaces to the job submission handler. The execution sites have interfaces to various batch systems. Job handlers which are part of the execution site import the CDF software environment to the worker nodes of the analysis farms and control the job. Therefore job handlers for each generic type of job have to be created. This system is under test for CDF in the context of generating simulated datasets.

4. CDF's Utilisation of UK/StarLight

A dedicated circuit connecting UCL and Fermilab, utilising StarLight and UKLight infrastructure was setup late in 2004 and is presently undergoing integrity and bandwidth tests with a view to achieving a sustained rate of ~ 1Gbit/sec over several days. The CDF hardware at UCL comprises of a dedicated machine connected directly to UKLight, which is effectively on a private network and can only be reached from other machines e.g. at FNAL on the UKLight/StarLight network. From a public PC this node can only be reached from within the UCL HEP domain. Initial tests are proceeding with 7.5 Tb of RAID0 SCSI disk.

5. Results

Initial tests were performed using IPERF to gauge the capability of the network and the PCs to sustain high bandwidth transfers. IPERF does not store the data transmitted which allows measurements to be made independent of the disk IO on the systems involved. IPERF is capable of performing tests using both UDP and TCP. Tests with UDP have shown that while speeds of up to 700Mb/s can be obtained [9] speeds over this, result in packet loss and packet reordering is observed over 200Mb/s. Further tests with TCP indicate 190Mb/s is obtainable, which is consistent with the packet reordering observed during the UDP tests. Currently investigations are underway to identify the sections of network that are unable to support 1Gb/s and also to ensure that all disk IO is optimised by ensuring that RAID parameters are appropriately set (on SCSI disks) and that a fast file systems are used e.g. XFS or REISERFS4 on the critical machines.

6. Further Work

The DCache/Enstore[10] system at Fermilab which hosts all of the raw data is only able to sustain a throughput of 200 Mbit/sec. This is mainly influenced by the presence of gridftp-doors at Fermilab. A way around this bottleneck

is the storage resource manager (SRM)[11] concept. SRMs are middleware components whose function is to provide dynamic space allocation and file management on shared storage components on the Grid. An interesting feature of SRMs, for fast data transfer, is that the request for data transfer is sent to a SRM server which then negotiates with a DCache head node to determine to which pool the data needs to be sent. The data is then directly sent from the DCache pool node. This means that while individual nodes are not capable of maintaining Gigabit speeds, the aggregate rates of parallel transfers from several pool nodes would be able to (see figure 3). The interface of the SRM model to the SAM software still has to be updated and the effect on the UKLight link transfer rates will be studied.

Once the link speed and stability issues have been addressed the UKlight/Starlight link will be integrated into the existing grid architecture at UCL. A publicly accessible SAM station with 4.0 Tb reserved for permanent SAM storage and a further 2.5Tb of user storage will be maintained. This design will also be mirrored on the much larger UCL "CCC" system, which comprises of some 200+ Linux nodes and SAN storage, of which we hope to reserve ~ 10 Tb as a CDF SAM store (see figure 4).

7. References

- [1] The CDF experiment, <http://www-cdf.fnal.gov>
- [2] SAM, Sequential Access to Metadata, see CHEP04 conference contribution, <http://projects.fnal.gov/samgrid/conferences/chep04/chep04.html>
- [3] DCAF, De-centralised Analysis Farms for CDF, see <http://cdfcaf.fnal.gov>.
- [4] ESLEA, Exploitation of switched light paths for eScience Applications, see paper #474 in these proceedings.

- <http://www.eslea.uklight.ac.uk>
- [5] The ESLEA Control Plane Software, see paper #444 in these proceedings.
- [6] SAMGrid, <http://projects.fnal.gov/samgrid/> and <http://projects.fnal.gov/samgrid/conferences/chep04/chep04.html>
- [7] LCG, LHC Computing Grid, see <http://lcg.web.cern.ch/lcg/>
- [8] JIM (for CDF), see CHEP04 conference contribution, <http://projects.fnal.gov/samgrid/conferences/chep04/chep04.html>
- [9] Network performance for the Star/UKLight connected node: http://193.60.252.10/mrtg/195.194.14.1_8.html
- [10] CDF Run II dCache System Status <http://cdfdca.fnal.gov/>; CDFen Enstore System Status, <http://www-cdfen.fnal.gov/enstore/>
- [11] Fermilab Storage Resource Manager (SRM) Project: <http://www-isd.fnal.gov/srm/>

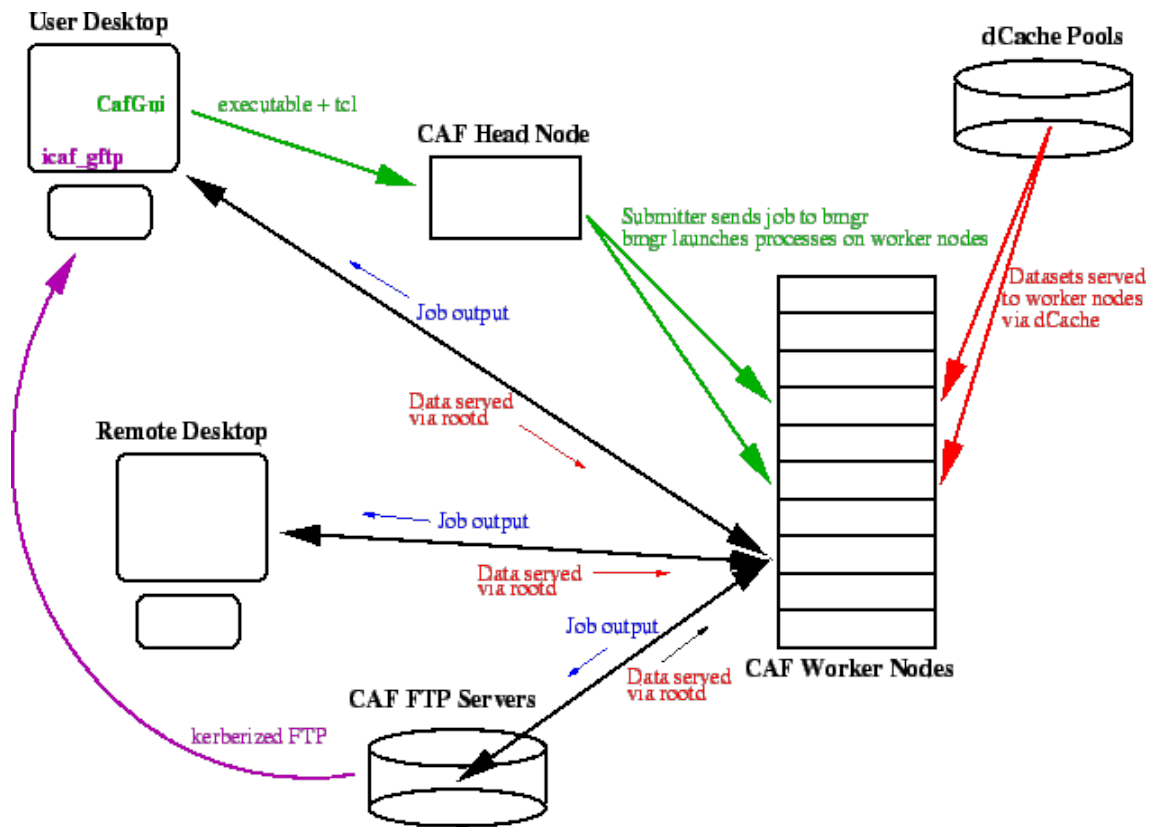


Figure 1: Basic CAF architecture.

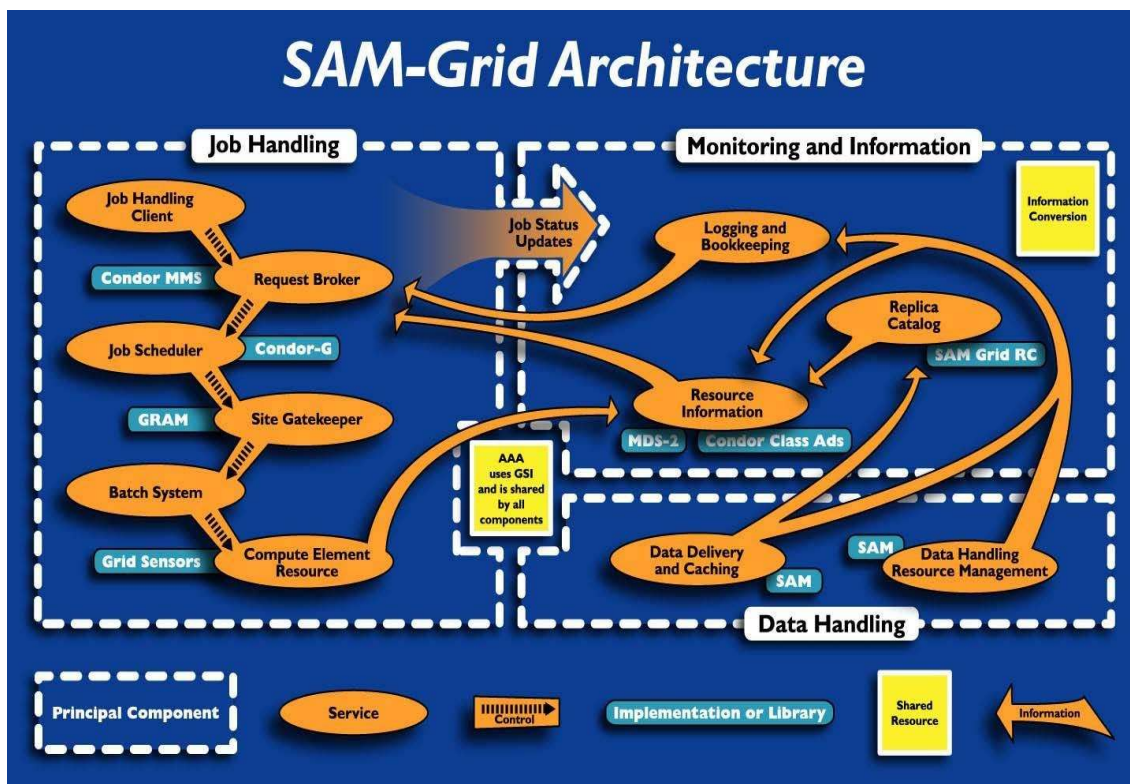
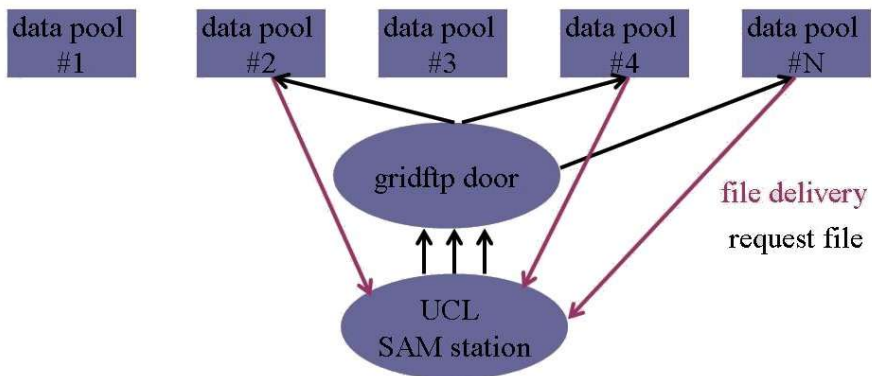


Figure 2: SAM-Grid Architecture.



**SRM: files transfers are redirected to data pools
 ⇒ disk IO and CPU not an issue any more for
 aggregated transfer rates**

Figure 3: SRM Model

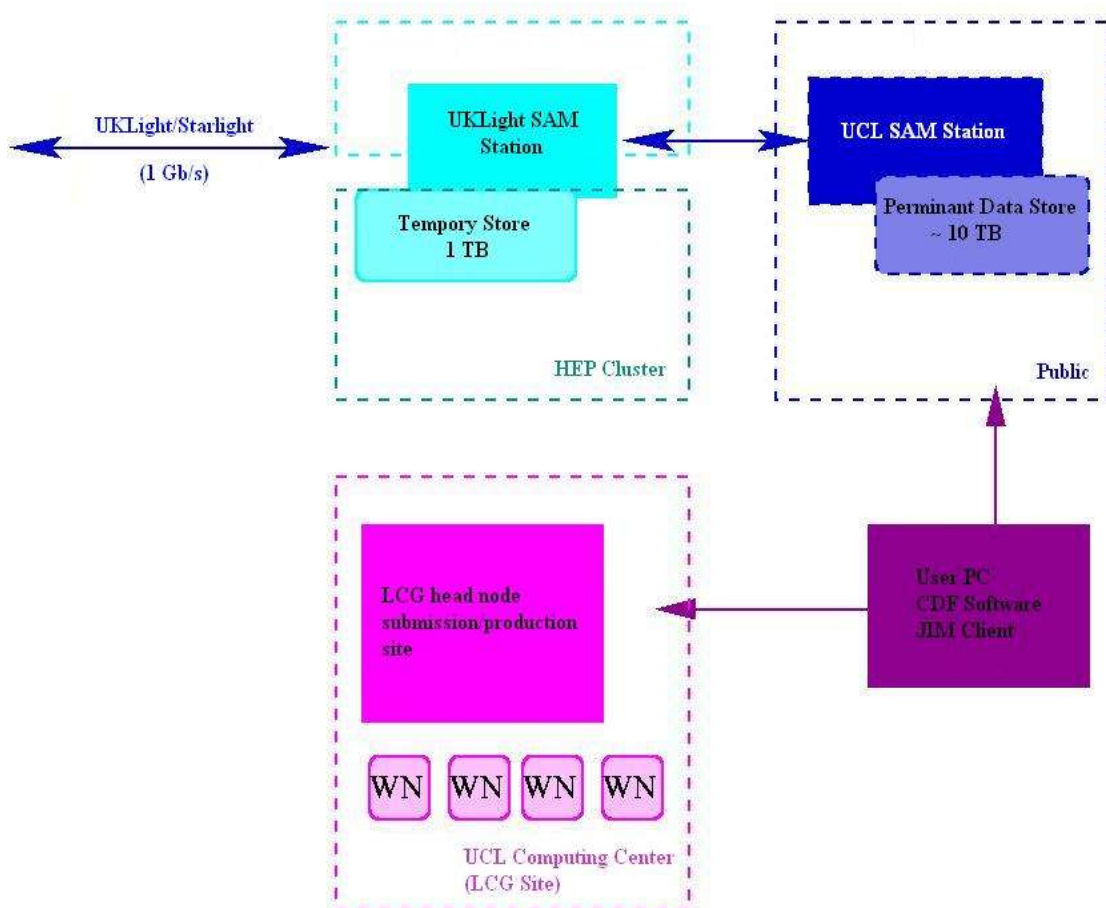


Figure 4: Evolution of UCL grid infrastructure.