CDF/DOC/COMP_UPG/PUBLIC/XXXX

Modifying the CDF production farms to use Enstore

Jaroslav Antos Yen-Chu Chen Miroslav Siket
Stephen Wolbers GP Yeh

March 8, 2002

1 Introduction

The original design of the CDF offline production farms uses the CDF Data Handling System [1] to provide input and output. Early in the project the I/O functions were split from the farms and the DH group was made responsible for the software, hardware and operation of the I/O for the farms. Well-defined interfaces were used to provide the input data for the farms and to write output to the tapes.

CDF is moving toward a new model of I/O based on the Computing Division written and supported package Enstore [2, 3]. It is wise to re-examine the I/O model of the CDF production farms and to explore modifications that makes the best use of experience gained up to now in Run 2 and to reflect the properties of the Enstore system.

This note will state the requirements of the CDF production farm in Run 2, give a short description of the DH-based I/O system, and finally will describe alternatives for using Enstore for the production farm I/O.

2 Requirements for CDF farms I/O

The CDF production farms are required to process raw data, reprocess raw data and secondary datasets, and generate and reconstruct simulated data. These requirements are laid out in more detail in [4, 5]. Briefly, the farms are required to handle 20 MB/s of input data in multiple (currently 8) streams and 10-20 MB/s of output data in multiple (currently 20) datasets for raw data processing. Reprocessing and Monte Carlo add to the requirement for I/O in the system.

The farms should reconstruct raw data as quickly as possible, ideally as soon as the data is available. In addition to the reconstruction of new data, the farms are required to reprocess old data with newer reconstruction code or with newer calibrations. Secondary datasets will also need to be reconstructed on the farms. Finally, Monte Carlo samples will be generated and processed on the farms.

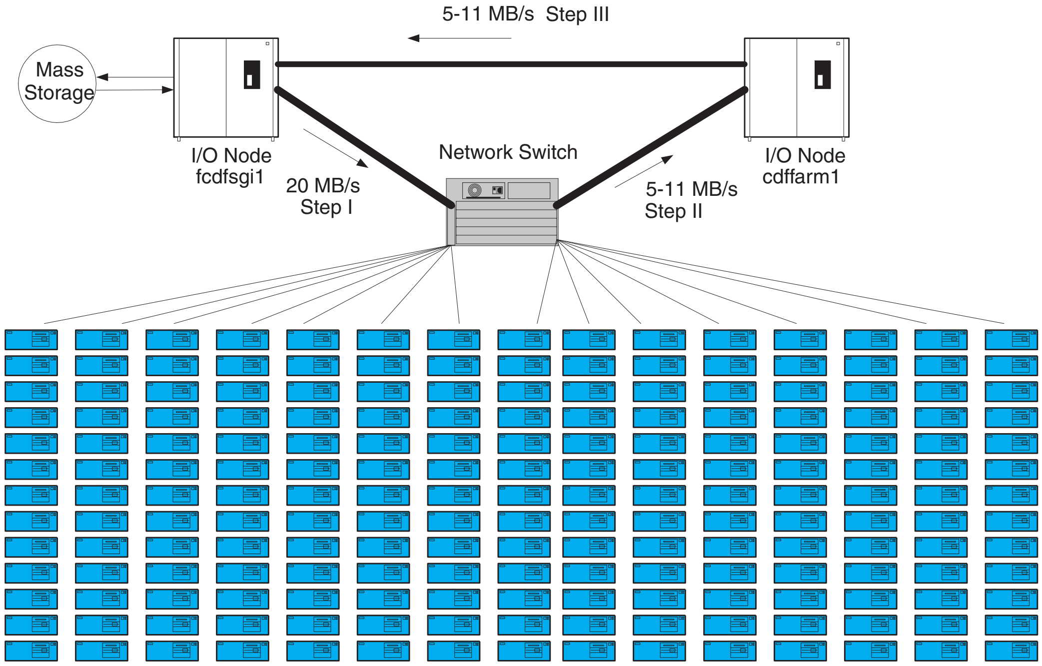
It is expected that the total of these requirements can be met in the 20 MB/s capability for input and approximately the same for output from the farms. Therefore, any system that is designed for the farms I/O should be able to achieve at least 20 MB/s to and from mass storage. A larger capability would be useful to allow the farms to catch up at times when a large backlog of processing exists. This can occur when the ProductionExe is not ready, crashes too often, has memory leaks, etc., or when the calibrations are not available. It can also occur when a significant amount of data is required to be reprocessed.

3 Current DH Farms I/O System

The current DH I/O system provides the input data to the farms and is the final repository of farm output data. A diagram that shows the original farms architecture can be seen in Figure 1.

Data is read from AIT-2 tape on tapedrives which are SCSI connected to the SGI O2000 node fcdfsgi1.fnal.gov. The maximum reading speed is 6 MB/s per tapedrive. The data is staged in filesets from tape to disk on fcdfsgi1. Once on disk, in DIM-managed space, the files can be copied to worker nodes for processing. There is a gigabit connection from fcdfsgi1 to the farms Cisco 6509 switch, giving the necessary bandwidith for the input files.

Figure 1: Run II Farms w/CDF DH system



Output from the farms is first copied to an SGI O2200 node cdffarm1. Approximately 10 Gbytes of output is copied from a single dataset. The files are run through the CDF program ac++dump to concatenate the small input files to output files with a target size of 1 Gbyte. There is a gigabit link from cdffarm1 to the CDF farm switch for the transfer.

The concatenated files are copied from cdffarm1 to fcdfsgi1 over a dedicated gigabit link. The files are stored on disk space on fcdfsgi1 and when sufficient files are saved a tape copy job is started by the DH system. The writing of partitioned AIT-2 tapes has an effective speed of 4 MB/s.

The writing of raw data from the CDF experiment occurs on fcdfsgi1. The node fcdfsgi1 has 8 AIT-2 tapedrives, 4 processors, 1 Gbyte of memory, and 4 Tbytes of disk.

4 Using Enstore for Farms I/O

Enstore is a mass storage system that provides direct reading and writing capability over the network using TCP/IP. Given this new approach the CDF farms group, working with the Fermilab Computing Division Enstore team and with the CDF DH group, and including discussions with the D0 farms group, has come up with options for integrating Enstore I/O into the offline farms processing. Figure 2 shows the architecture of the CDF production farms with Enstore.

Independent of the exact I/O mechanism, the farms group has been working on a new method for the concatenation of output files. This will be described first.

4.1 Concatenation on the Worker Nodes

Concatenation of the output files puts a relatively heavy burden on the cdffarm1 node. This limits the overall farm rate. One way to increase the capability of cdffarm1 is to increase its CPU power (faster processors or more processors). This is rather expensive, given the cost of R12000 or R14000 processor boards. Another approach is to use some of the farm worker nodes as concatenation nodes. The network switch is fast enough that files can be transferred easily and quickly between worker nodes, the worker nodes have sufficient CPU power, disk and memory, and more nodes can be procured cheaply. Therefore, the CDF farms group (Miroslav Siket) has written software that moves concatenation from cdffarm1 to a subset of worker nodes (4 nodes to start). The input for each concatenating node comes from the other worker nodes. The output is sent to fcdfsgi1 for logging to tape (using the current DH system). When using Enstore two different approaches for output are contemplated and they will be described below.

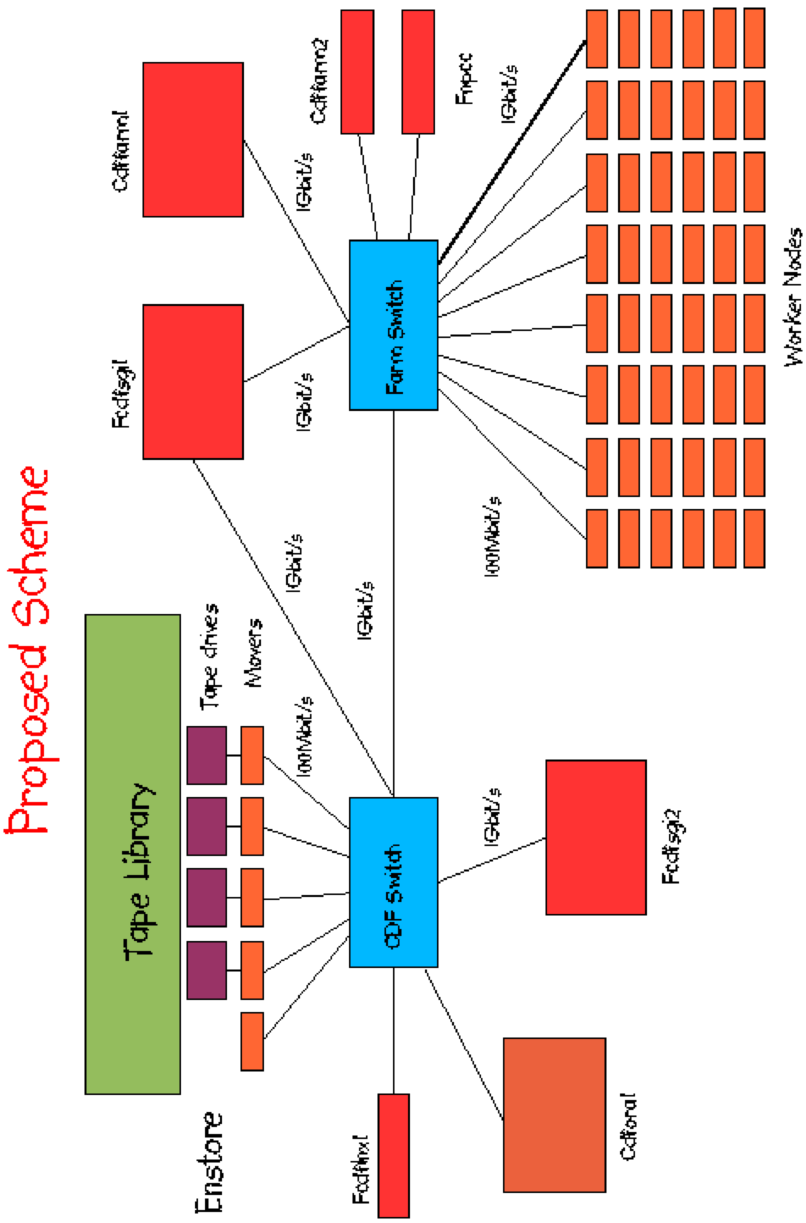
4.2 Output from the Farms Using Enstore

The choice of which approach to use for farms output depends on many factors. The final decisions will be made after different approaches are explored or when a satisfactory approach is determined.

4.2.1 Minimal Approach.

The output of concatenation can be sent directly to fcdfsgi1 and from there written to tape using Enstore. The changes required are not large, and this approach merely changes the low-level software that writes files (or filesets) to the tapes.

Figure 2:



Proposed Scheme
Tape Library
Enstore
Tape drives
Movers
100Mbit/s
CDF Switch
Farm Switch
Fcdfsgi1
Cdffarm1
Cdffarm2
Fnpcc
1Gbit/s
Worker Nodes
Fcdfsgi2
Cdfora1

4.2.2 Enstore Direct

To write data directly to Enstore requires that the Farms Processing System (fps) be modified to use encp from the worker nodes directly. This is basically the inverse of reading from Enstore tapes directly. The single write bandwidth should be 10 MB/s and multiple reads from different concatenation nodes should provide 10 MB/s per tapedrive up to 70-80 MB/s. The limitation is the Gbit link of the two switches.

4.3 Input to the farms using Enstore

The farms input when Enstore is used can be done in many ways. Two approaches are proposed. The first is the "Minimal Changes" approach. The second is to use Enstore directly.

4.3.1 Minimal Changes

The minimal approach modifies the low-level staging software on fcdfsgi1 to read data from tape (using Enstore's encp command). The files are copied to fcdfsgi1 disk and from there the files are copied to worker nodes. To the farms this system looks very similar to the current CDF DH system, and no hardware changes are needed. Very modest farms software changes would be needed to make this change. In essence fcdfsgi1 is a "cache" and a "buffer" for the data which is staged from tape. The data has to move over the network twice (once from tape to fcdfsgi1 and once to the worker nodes). This approach is straightforward and should be possible with a fairly small effort.

4.3.2 Enstore Direct

To use Enstore directly each worker node will connect to Enstore and copy data directly. The Enstore mover nodes are connected via 100 Mbit Ethernet to the CDF central analysis switch. The CDF central analysis switch is connected to the CDF farms switch with a Gbit uplink. An Enstore connection from a tape drive through a mover node and to a worker node should provide a full 10 MB/s of bandwidth. This has not yet been measured. In addition, multiple reads from multiple mover nodes to multiple worker nodes should provide 10 MB/s per tapedrive up to about 70-80 MB/s, limited by the Gbit link between the two switches. This also has not yet been measured.

This scheme requires changes to the CDF Farms Processing Software (fps) to coordinate the encp commands for best use of the tapedrives. Enstore itself will optimize and properly manage many of the requests for tape reads and writes. Given that the farms normally process data in a very systematic manner it should be straightforward to organize the data requests properly.

4.3.3 Caching Enstore

Instead of allowing each of over 100 (soon 200) nodes to read directly from the Enstore system one could instead use a subset of the farm to cache data from Enstore prior to needing the data for execution on the farm. Specifically, one could use some of the nodes to pre-fetch data files and use dfarm or some other mechanism to keep track of and move the data to the proper node when the job is started. Another possibility is using the disk cache system dcache to pre-fetch files prior to needing them on the farm. These ideas are attractive because they allow for the most efficient use of the tape drives/robot and the farm itself by reducing delays.

5 Plan

The farms have implemented concatenation on the worker nodes, using the current CDF DH I/O system. Given that this is in place the most natural approach is to use those worker/concatenation nodes as direct enstore writers. This avoids the fairly heavily loaded fcdfsgi1 and performs one less copy of the data. Gigabit connections to the worker/concatenation nodes will be necessary to effectively use these systems for Enstore writing.

For input the issues are not yet resolved. The three approaches each have benefits and possible drawbacks. It is likely that a buffer will be a necessary component of the input to the farm. fcdfsgi1 can server that role. But so can dcache or the worker nodes themselves. A decision will have to be based on which approach gives the greatest benefits.

6 Status

In addition to a change of access to mass storage CDF is making a change of tape technology from the 8mm SONY AIT-2 to the STK 9940. It is expected that the STK 9940 gives better performance with fewer problems than the

AIT-2. Experience at SLAC, BNL, Jefferson Lab, and at Fermilab with the 9940 has been excellent.

CDF will have 10 tapedrives for the experiment during the initial phase of Enstore use. This should be adequate, assuming the needs of the collaboration during this time. If not, the farms could potentially have to wait for input or output if tapedrives are heavily used and not available on demand. This could lead to a rethinking of strategy and may lead to more pre-staging than current thinking would require.

A gigabit connection between the CDF central analysis switch and the CDF farms switch has been installed. In addition, gigabit connections from the CDF farms switch and the worker/concatention nodes are being installed.

7 Acknowledgements

Miro and Steve would like to thank Don Petravick for early discussions of the Enstore system and how it might work with the CDF farms. Miro, Steve and Yen-Chu would like to thank Jon Bakken for useful discussions of Enstore. The CDF farms group thanks the CDF DH group over the years for helping to keep the data flowing and to making the data well-organized for the experiment.

References

[1] Rick Colombo, *et al.,* The CDF Computing and Analysis System: First Experience, FERMILAB-Conf-01/300-E, November, 2001

[2] http://www-isd.fnal.gov/enstore/index.html.

[3] Robert Harris, *et al.,* A Tape Handling System for CDF in Run 2, CDF note 5822, January 11, 2002.

[4] Jaroslav Antos, *et al.,* CDF Production for Run II, CDF Note 4810, November 15, 1998.

[5] Jaroslav Antos, *et al.,* The CDF Run 2 Offline Computer Farms, CDF Note 5702, August 8, 2001. FERMILAB-Conf-01/234-E, August, 2001.