

Data Streaming Project

Project Definition, Schedule, and Budget

September 10, 2003

Introduction

US-CMS Software and Computing Project proposes to create a joint project with the CERN IT division to develop expertise in wide area data streaming in preparation for the start of the LHC experiments. The project will attempt to show the feasibility of selecting, transferring, and archiving streams of events in real time between the Tier0 center at CERN and remote Tier1 facilities. This project should serve as a pilot for additional collaborative projects between CERN and the US. It will strive to foster good working relationships that will benefit both projects in the very intense years that are sure to precede the start of the LHC analysis era.

The program of work will consist of development in reliable data transport and network optimization, archiving and storage, and in event selection and identification. The project will build on and leverage effort from storage, networking, and computing sub-projects already in progress including the Fermilab Network Laboratory Project, in cooperation with DataTag; the Storage Resource Management (SRM) collaboration; and the US-CMS Core Application Software Project.

The final goal of the project is to demonstrate that dedicated streams of data can be reliably and efficiently selected and transferred from the experiment site to a Tier1 center for reconstruction and analysis. This will allow additional computing resources at the remote centers to expand the physics potential of CMS by permitting dedicated analysis streams that might not ordinarily be processed, due to limited resources for reconstruction and storage at the Tier0. In addition, it will attempt to demonstrate that large streams of data can be reliably archived in mass storage at a Tier1 center. If successful, this would permit the second archived copy of the raw data to be distributed across the Tier1 facilities and remove this burden from the Tier0.

While the data archiving at the Tier1 centers and dedicated physics channel streaming concepts are relatively new, CMS has always foreseen using the Tier1 centers for re-reconstructing events. The transfer rates expected for re-reconstruction are the same as those for archiving, the difference is the real-time nature of the transfers and the need for archival storage. Much of the effort proposed in this project is needed for enabling re-reconstruction at the Tier1 centers, so this project is not seen as a significant increase in scope for the US-CMS User Facility Project. For archiving and dedicated analysis streams to be successful at the start of the experiment, there will need to be a commitment from the Tier1 centers for additional computing resources for analyzing the data streams and archival storage of the data transferred.

Project Goals

A data rate to mass storage of 100 megabytes per second is currently anticipated for the CMS detector, essentially from the start of running. In order to use Tier1 facilities as an archiving location for CMS raw data, as there are five planned, it would be necessary to achieve a sustained transfer rate of 20 megabytes per second during running conditions. The data requirements of a dedicated physics channel to the Tier1, in addition to archiving, depend on the type of channel investigated.

For this project we are attempting to demonstrate sustained average transfer rates of 30 megabytes per second with peak rates of 100 megabytes per second. For this project, simulated data will be streamed out of the CERN CASTOR system. We will take advantage of the large dataset being prepared for the CMS data challenge (DC04). The data will then be buffered in large data servers at CERN. A medium sized analysis cluster will process the data to select streams for transfer to Fermilab. The data will be transferred to Fermilab where it will enter deep buffers and be archived into the Fermilab Enstore system. A diagram of the components is shown in Figure 1.

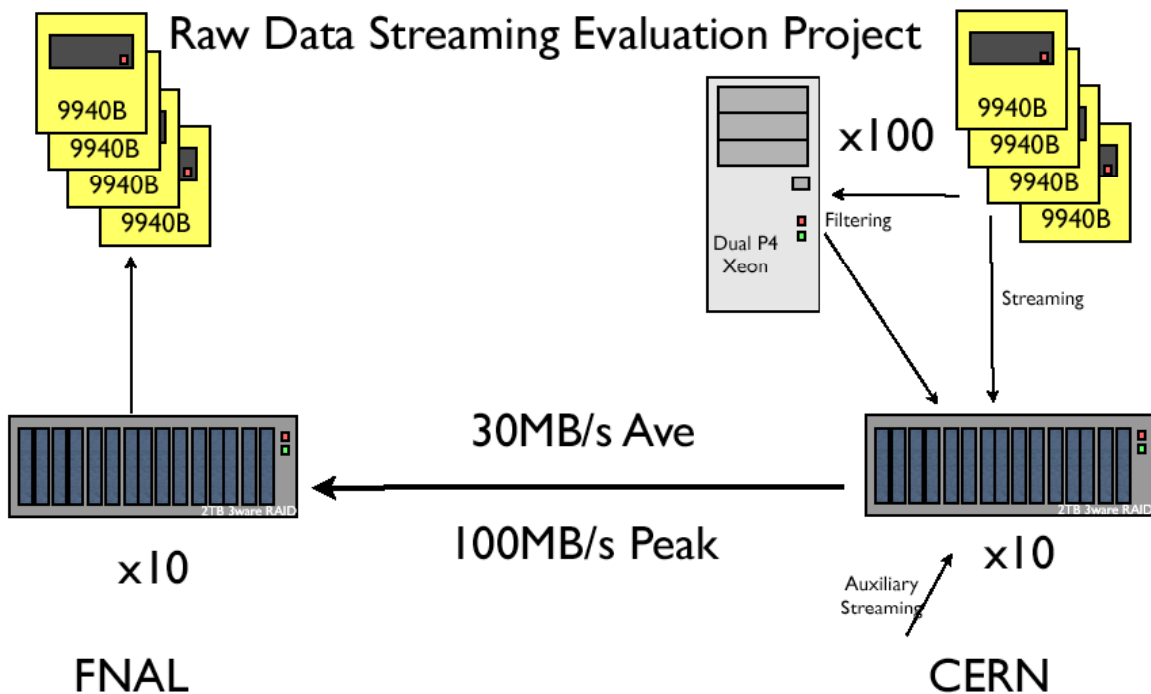


Figure 1: System components of data streaming evaluation project.

In order to show the robustness of the system, the data transfers will have to proceed for a believable period of time. For this exercise we propose to run for testing periods of a week. The project will also have to demonstrate robustness to faults. Both network and storage failures should be protected against to prevent the loss of data. During a week-long demonstration we will show robustness against faults of individual components: stopping the network and allowing the disk buffers to fill and recover, disabling individual storage elements, and protecting against data corruption.

Activities

The goals of this project rely in part on the success of several other activities within CMS and global grid projects. The manpower in this project is primarily assigned to integration, deployment, and testing, but the people assigned to work on this project will need to work closely with the other developers. The activities that this project relies on are networking research, mass storage interface research, and the CMS Computing and Core Software Project (CCS).

Networking:

Fermilab is currently working on a network lab project to evaluate low-level components for networking including kernel implementations of TCP, network buffer optimization, and network protocols. The lab is currently set up with network emulators designed to test wide area networking environments. The effort will concentrate on improving the network performance between the CMS computing facilities. The deliverable of this project will be a set of recommendations and instructions for optimizing network performance. We expect to use preliminary recommendations from the Network lab project to optimize the network performance for this project. The effort spent developing the recommendations will come from the lab project.

Interfacing to Storage:

Many computing centers are contributing to the development of a common protocol to interface storage devices. The Storage Resource Management project (SRM) has the features necessary in a common mass storage interface. We will work with the SRM collaboration and developers at Fermilab, if there are additional requirements for this project. Agreement will be needed within the project on a consistent version of the SRM protocol to ensure interoperability between the two sites.

Data Selecting:

The CMS CCS project has developed a reasonably advanced software framework. It has arrived at a sufficient level of completeness that it will be possible to select streams of events for transfer to Fermilab using a medium-sized cluster of systems located at CERN. The CMS DC04 data challenge will focus on reconstructing events at 20Hz using a large number of systems. In this demonstration we will attempt to select events without full reconstruction at a rate of approximately 10Hz on average. The creation of the application will require the help of the core software development team, but for this demonstration the application can be fairly simple.

Integration:

Once networking recommendations, storage interface protocols, and selection applications are delivered there is effort needed to integrate these into a data streaming demonstration. The majority of the project effort will be assigned to integration, deployment and evaluation.

Hardware Procurements:

This project requires substantial hardware procurements at both CERN and Fermilab, as can be roughly seen in Figure 1. We intend to procure a cluster of systems for data selection to be located at CERN. A farm of 100 dual CPU Xeon systems will be sufficient for the application proposed. 10 data servers will be procured at both locations to serve to stream data over the network and into mass storage as well as serve as deep buffers. At Fermilab 4 STK 9940B drives will be procured to augment the mass storage input devices dedicated to CMS. At CERN there are sufficient available tape resources for this project. CERN is establishing a connection from the computing resources to the 10Gbit research network. In order to succeed, the Fermilab fiber upgrade to Starlight needs to be in place.

Schedule

Below is a rough timeline of the project

Oct 15th 2003, finalize hardware choices and place purchase orders. Both centers are free to choose hardware that will integrate well with existing facilities.

Feb 1st 2003, hardware delivery to Fermilab and CERN

Feb 1st 2004, delivery of networking recommendations from Network Lab Project.

March 1st 2004, demonstrations of network performance between FNAL and CERN for verification of available performance. This test should run for a period of at least 48 hours and achieve the transfer goal of 30MB/s on average.

April 1st 2004, demonstrate event selection using the farm and CMS software on a raw sample of simulated events created for DC04.

May 1st 2004, perform the full data streaming demonstration from CERN to Fermilab for the period of one week. Data streamed from Castor should be analyzed on the selection farm, sent to Fermilab over the WAN at the goal rate, and archived into Enstore at Fermilab.

June 1st 2004, deliver final project report. Summarize the cost and effort required, the results of the demonstration, the feasibility for the start of the experiment, and recommendations for future work.

Budget

Hardware:

Data Servers (x20)	\$160,000
Compute Nodes (x100)	\$163,000
Tape Drives STK 9940B (x4)	\$120,000

Personnel:

2FTE for 0.5 years (\$70k per site)	\$ 70,000 (FNAL)
	\$ 70,000 (CERN)

FNAL Contribution	\$513,000
CERN Contribution	\$ 70,000