

--DRAFT--

# USCMS Plan for Meeting the DC04 Milestone

Lothar A.T. Bauerdick, Michael Ernst-Eschgarth, Ian Fisk, Greg Graham  
*Fermi National Accelerator Laboratory*

## Introduction to DC04

The CMS Data Challenge 2004 (DC04) is a major milestone on the road to a successful computing model for CMS. DC04 will involve crunching through an amount of data approximately equal to 5% of that expected during actual data taking in real time. This amounts to a 25 Hz event rate at the Tier-0 facility at CERN and distributing subsets of that data to the Tier-1 centers and some Tier-2 centers for analysis. Though it is not a hard and fast requirement that every aspect of the milestone be successful, it is imperative that we take this opportunity to learn as much as we can about the tools and architectures deployed in the service of the transportation and analysis of this data. In order to have an interesting data challenge, a pre-challenge production (PCP04) phase is in progress to generate and digitize some  $5E+07$  distinct Monte Carlo events amounting to about 100 TB of data CMS wide.

For the sake of planning, DC04 can be broken down into three concurrent phases:

1. Online data processing and streaming at the Tier-0 center
2. Transportation and Storage of data at the Tier-1 centers
3. Reprocessing of data at the Tier-1 centers and Tier-2 centers

This document will focus on planning the steps that need to be taken by USCMS at the Tier-1 site and Tier-2 sites in order to support the DC04 milestone.

## Online Data Processing and Streaming at the Tier-0 Center

Data will be retrieved from CASTOR and reconstructed on the farms at the Tier-0 site at a rate of 25 Hz, and the reconstructed data will be divided into streams. The mechanism for stream selection is not important to the milestone and will not be described here. Each Tier-1 center will be responsible for processing one or more streams. The streamed data will be kept in buffers at the Tier-0 site in the form of files not exceeding 2 GB in length. When a reconstruction program closes a file, the LFN of that file is registered in a replica catalog at the Tier-0 center. The catalog update is done by some agent; for DC04 the agent is planned to be a cron job.

At the Tier-0 center, the replica catalog will be an RLS catalog with the EDG extensions. The EDG-RLS catalog is required because it supports the concept of global unique ids (GUIDs). Furthermore, this catalog will interoperate with the POOL catalog at the Tier-0 center. In addition to the replica catalog, a database (called "the scoreboard") will keep tally of which files have been transferred offsite by the Tier-1 centers. Files that have been successfully moved will be removed from their respective buffers.

## Transportation and Storage of data at the Tier-1 Centers

The model of data movement to the Tier-1 centers is a pull model. An agent at the Tier-1 center, which again is a cron job in the case of DC04, will check the replica catalog at the Tier-0 site for newly created files (LFN) in the requested streams. A transfer of that file will then be initiated by the Tier-1 site. In order to keep track of these transfers from the Tier-1 side, a local replica catalog will also be maintained at the Tier-1 site. In addition, a database separate from the catalog will be maintained which has some state information included since the number of file transfers per day is so high. Upon successful receipt of the file at the Tier-1 center, it will be registered in the local replica catalog. If the file is intended to be archived forever and made available to other members of the CMS collaboration, a replica should also be registered in the Tier-0 replica catalog. At Fermilab, this will be the case since the data will be stored permanently in Enstore through the dCache interface. And in any case, the scoreboard at CERN will be notified that the file was successfully received so that it can be flushed from the output buffer there.

At Fermilab which relies on the Virtual Data Toolkit (VDT), the Globus RLS replica catalog is installed. This catalog does not at the current time interoperate with either the EDG extended RLS catalog or with POOL; however there are efforts underway in that direction. Therefore, the POOL MySQL based catalog will be used at the Tier-1 site. This will also be used as the local replica catalog and any local transfer state tables will be added there database as alluded to above.

### **Reprocessing of Data at the Tier-1 Centers and Tier-2 Centers**

In addition to being horizontally partitioned across multiple files, CMS events are also vertically partitioned (clustered) into multiple file categories, each category containing like subunits of each event. These include reconstructed event data, raw data, Monte Carlo data, CARF metadata, etc. A missing link in the above scenario is the consideration of how to group files from different categories together into an “execution unit” needed for processing of that data, and how to hold execution of processing until all necessary files are present. For DC04, we think that the execution unit may be the run, so a catalog that is able to produce the list of files needed for each run would be sufficient.

The simplest case will be if all file categories for reconstructed data are known beforehand and are all horizontally partitioned according to run number. In this case, since there is the additional partitioning into 2 GB files, all that may be required is a total file count with each run. We will recommend this case to the Tier-0 in advance of DC04.

Reprocessing at the Tier-1 center will take place after the data has been stored there. MCRunjob will periodically query the local replica catalog for files corresponding to new runs, create jobs for the USMOP environment, and keep the job processing tally locally. We assume a similar system could be in place for processing at the Tier-2 sites.

### **Analysis of Rates**

Rate of Data Into the Tier-1 Facility:

We assume that the Tier-1 center at Fermilab will consume about 20% of the total events produced during DC04. This corresponds to a 5 Hz rate at about 2.5 MB/event, or 1.1 TB per day. (A data transfer of this rate corresponds to about 16% of the

available instantaneous 622 Mbps trans-Atlantic bandwidth.) This corresponds to an average arrival rate of 22.5 2 GB files per hour. The dCache system backed by Enstore will be able to handle this rate.

Rate of Data Out of the Tier-1 Facility:

At this time, specific plans by the Tier-2 sites to analyze the DC04 data are unknown. However, given the above rates, it is not expected to be an unmanageable rate.

### **Schedule of Work and Effort**

In order to bring the above plan into fruit, short term effort must be applied by the Tier-1 facility into the following projects. (In the following, UF=User Facility and DAG=Distributed Applications Group.)

1. New File Discovery: This will be an agent by which the file catalog at CERN is queried for new files to transfer from the required output buffers at CERN. If the agent is a simple cron job, we estimate that the actual development will be about a day. This assumes that the catalog interface at the Tier-0 is known. 1.2 FTE-Week is budgeted here to be supplied evenly by the UF project and the DAG project.
2. Local Database Schema: We intend to use the schema proposed by Tim Barass in his note "Decomposing DC04 data distribution into modular agents with limited responsibility." A small amount of development to add the required table or tables to the local MySQL instance may be needed. Simple clients to access this database need to be developed. 1.2 FTE-Week is budgeted here from the DAG project.
3. POOL MySQL database: A pilot installation of the POOL MySQL database is to be installed at the Tier-1 facility at Fermilab. It must be checked against test production jobs and an operational plan developed for supporting the database throughout DC04. 1.2 FTE-Week is budgeted here to be supplied evenly by the UF project and the DAG project.
4. POOL MySQL compatibility with MOP: It is unlikely that a central POOL database located at the Tier-1 center will support distributed reprocessing of data using MOP. Some development to retrieve and load POOL XML fragments and distribute these with MOP jobs is needed. While it is not strictly necessary for the data challenge, effort can be estimated at 0.8 FTE-Week and assigned on contingency.
5. Training on POOL Tools: Estimated 5 people training for a day each.
6. New Configurators: MCRunjob development to focus on creating jobs from local POOL database information. Some communication with the Tier-0 is needed here to define precisely the proper execution unit of groups of files. Some work is needed in conjunction with the physicists at Fermilab to develop configurators to do the reprocessing. This is estimated at 1.2 FTE-week from DAG project.
7. Operational Support for File Transfers: UF 0.5 FTE-Week
8. Operational Support for Data Processing: DAG 0.5 FTE-week + UF 0.5 FTE-week.

	<b>Project</b>	<b>Estimated Effort (FTE-Week)</b>	<b>Completion Date</b>
1	Agent to Discover New Files for Transfer to Tier-1	1.2 (0.6 UF+0.6 DAG)	January 23, 2004
2	Database Schema to Keep Tally of File Transfer State	1.2 (1.2 DAG)	January 23, 2004
3	POOL MySQL Database Installation at Tier-1 with Tally Schema	2.0 (1.4 DAG + 0.6 UF)	January 30, 2004
4	MOP Compatibility (optional)	0.8 (0.8 DAG)	March 1, 2004
4	Training on Using POOL Tools to Get/Set XML Fragments for Repro Jobs	1.0 (0.6 DAG + 0.4 UF)	January 30, 2004
5	New Configurators to Support DC04 Job Creation	1.2 (1.2 DAG)	February 6, 2004
6	Support for DC04 Operations – File Transfers (Start March 1, 2004)	1.0 (1.0 UF)	March 31, 2004
7	Support for DC04 Operations – Repro (Start March 1, 2004)	1.0 (0.5 DAG+0.5 UF)	March 31, 2004

## References

Decomposing DC04 data distribution into modular agents with limited responsibility, Tim Barass, December 2003, Private Communication

Refining the DC04 Milestones, David Stickland, October 2003, Private Communication