

USCMS Data Challenge 04 is Grid Based

The CMS experiment is being prepared to take data at the Large Hadron Collider (LHC) at CERN, near Geneva, Switzerland beginning in 2007. A central goal of the LHC program is the discovery and characterization of the Higgs boson, an as yet unseen elementary particle predicted to exist and necessary to the theory describing why particles have mass. The CMS collaboration consists of 2000+ scientists around the world with hundreds of scientists participating in the U.S. (USCMS).

The CMS Data Challenge DC04 is one of the milestones of the experiment scoped to ensure the experiment is ready with its global data distribution and analysis systems for the start of data taking. The performance metrics for DC04 were to provide a baseline to give the experiment input to the Physics and Computing Technical Design Reports in the next two years. These design reports will form the baseline to which the production data processing and analysis systems will be built and must perform. It is essential that these global data distribution and analysis systems work effectively in order to enable scientists and their students to be able to participate fully at their home institutions in addition to having available sufficient computing resources to accomplish the scientific mission of the experiment.

The goal of DC04 was to perform at 25% of the throughput needed at the start of data taking in 2007 (5% of the full LHC rate) including distribution of data from CERN to the primary regional centers in each country. Additional goals were to provide the software infrastructure to manage and distribute the data, and to provide an end to end demonstration of event reconstruction and analysis to show the state of readiness of the experiment's software infrastructure. This was to demonstrate the full chain of data processing and analysis to prove the experiment's software system. These goals were substantially met. About 1/4 of the globally produced 70M events were generated on the shared Grid3 environment in the U.S. (25 sites, 3000 CPU's) and achieved 50% greater throughput due to opportunistic use of resources than was possible with only those resources dedicated to CMS. These data were sent over the network to CERN for storage and processing. A typical data sample copied back over the network from CERN to Fermilab in Illinois is 5,200 gigabytes in 440,000 files. These processed data are now being analyzed by scientists in the U.S.

USCMS is participating in the Particle Physics Data Grid (PPDG) project, which together with the NSF funded GriPhyN and iVDGL projects, form the Trillium consortium. Trillium is a collaboration of several computer science and physics experiment groups that are enabling physicists, computer scientists and computational professionals to integrate, harden, deploy and run national and international scale data intensive distributed computing applications and systems. This work enhances and enables both the physical science and computer science goals of the participants, as well as a broader range scientists benefiting from improved software and the emerging shared persistent grid infrastructure.

URL's:

USCMS – uscms.fnal.gov

PPDG – www.ppdg.net

iVDGL – www.ivdgl.org

GriPhyN – www.griphyn.org