FERMILAB-CONF-15-219-CD

21st International Conference on Computing in High Energy and Nuclear Physics (CHEP2015) IOP Publishing
Journal of Physics: Conference Series **664** (2015) 032015 doi:10.1088/1742-6596/664/3/032015

Data preservation at the Fermilab Tevatron

J Boyd, K Herner, B Jayatilaka, R Roser and W Sakumoto

Fermi National Accelerator Laboratory, Batavia, IL, 60510 USA

E-mail: boj@fnal.gov

Abstract. The Fermilab Tevatron collider's data-taking run ended in September 2011, yielding a dataset with rich scientific potential. The CDF and D0 experiments each have nearly 9 PB of collider and simulated data stored on tape. A large computing infrastructure consisting of tape storage, disk cache, and distributed grid computing for physics analysis with the Tevatron data is present at Fermilab. The Fermilab Run II data preservation project intends to keep this analysis capability sustained through the year 2020 or beyond. To achieve this, we are implementing a system that utilizes virtualization, automated validation, and migration to new standards in both software and data storage technology as well as leveraging resources available from currently-running experiments at Fermilab. These efforts will provide useful lessons in ensuring long-term data access for numerous experiments throughout high-energy physics, and provide a roadmap for high-quality scientific output for years to come.

1. Introduction

The Tevatron was a 1.96 TeV proton-anitproton collider located at Fermi National Accelerator Laboratory (Fermilab). Run II of the Tevatron, lasting from 2001 to 2011, saw the CDF and D0 collaborations record datasets in excess of 10 fb^{-1} per experiment and make groundbreaking contributions to high energy physics including the most precise measurements of the W boson and top quark masses, observation of electroweak production of top quarks, observation of B_s oscillations, and first evidence of Higgs boson decay to fermions. After the Tevatron's shutdown, manpower and resources have steadily moved to the Large Hadron Collider (LHC) and to other areas of HEP such as neutrino and precision muon physics. The unique nature of the Tevatron's proton-antiproton collisions and large size of the datasets means that the CDF and D0 data will retain their scientific value for years to come, both as a vehicle to perform precision measurements as newer theoretical calculations appear, and to potentially validate any new discoveries at the LHC.

The Fermilab Run II Data Preservation Project (R2DP) aims to ensure that both experiments have the ability to perform a complete physics analysis on their full datasets through at least the year 2020. To retain full analysis capability, the project must preserve not only the experimental data themselves, but also the software and computing environments. This requires ensuring that the data remain fully accessible in a cost-effective manner and that experimental software and computing environments are supported on modern hardware. Furthermore user jobs must be able to run at newer facilities when dedicated computing resources are no longer available and job submission and data movement to these new facilities must be accomplished within the familiar software environment with a minimal amount of effort on the part of the end-user. Documentation is also a critical component of R2DP. Documentation preservation includes not

only the existing web pages, databases, internal documents, but also requires writing clear, concise instructions detailing how users need to modify their usual habits to work in the R2DP computing infrastructure.

2. Dataset preservation

2.1. Collision data

The data for both CDF and D0 were stored on LTO4 tapes at the end of the Tevatron run. An analysis of then-available tape technologies concluded that T10K tapes would be the near-term choice for archival storage at Fermilab. While it was theoretically possible to leave the CDF and D0 data on LTO4 storage, a decision was made to migrate these data to T10K storage for two reasons. First, if the Tevatron data were accessed for a long period of time after data taking ended, the LTO4 storage may be an unsupportable configuration; as LTO4 tape declines in usage industry-wide there may not be replacement storage easily available. Second, as storage media and drives for older technology become scarce, their costs rise, potentially increasing the overall long-term cost of staying with LTO4 storage. Due to these concerns, the commitment was made to purchase T10K tapes and migrate all of the CDF and D0 data. It took approximately a year for the migration to be completed and the Tevatron data now share all resources with active Fermilab experiments.

2.2. Non-statistical data

For non-statistical data, such as detector calibrations, Both CDF and D0 used Oracle database software extensively throughout the Tevatron run. The ongoing cost of maintaining Oracle licenses, which were not used by most current Fermilab experiments for scientific use, presented a long-term challenge. The database schema was heavily interwoven into the analysis software. As a result, converting to a more economical open source database solution would incur a considerable and prohibitive investment in human resources. Thus, both experiments decided to retain the Oracle database systems throughout the life of the data preservation period, following an upgrade to the most recent version of Oracle at the time of the Tevatron shutdown. Furthermore, as future upgrades to the Oracle database could potentially disrupt the existing schema, and thus the analysis software, a contingency plan was drawn up where the current version and schema could be frozen and run, in network isolation if necessary, in the future, even if support for that version had ceased.

3. Software and environment preservation

Both CDF and D0 have complex software releases to carry out simulation, reconstruction, calibration, and analysis. Most of the core software was developed in the early to mid 2000s on Scientific Linux running on 32-bit x86 architectures. During the operational period of the Tevatron, these software releases were maintained on dedicated storage elements that were mounted on the experiments' respective dedicated computing clusters. As these dedicated resources are no longer maintained, CDF and D0 have migrated their software releases to CERN Virtual Machine File System (CVMFS) repositories [1]. As CVMFS has been widely adopted by current Fermilab experiments, this move allows for maintaining CDF and D0 software releases for the foreseeable future without a significant investment in dedicated resources. Furthermore, as the Fermilab-based CVMFS repositories are distributed to a variety of computing facilities away from Fermilab, this approach lends to CDF and D0 computing environments being available on a variety of remote sites. At the time of the Tevatron shutdown, both CDF and D0 were running software releases that, while operational on Scientific Linux 5 ("SL5"), depended on compatibility libraries that were built in previous OS releases dating back to Scientific Linux 3. The two experiments chose different strategies to ensure the functionality of their software releases throughout the data preservation period.

3.1. CDF software release preservation

At CDF, stable software releases were available under two flavors: one was used in data reconstruction and analysis and another for Monte Carlo generation and simulation. To ensure the long-term viability of CDF analysis, the CDF software team chose to prepare brand new "legacy" releases of both flavors. These legacy releases were stripped of any long-obsolete packages that were no longer used for any analysis and also shed of any compatibility libraries built prior to SL5 Once validated and distributed, older releases of CDF software which still depended on compatibility libraries were removed from the CVMFS repository. This meant that usage of CDF code for analysis on any centrally available resources was guaranteed to be fully buildable and executable on SL5. Furthermore, it meant a relatively simple process for further ensuring the legacy release is buildable and executable on Scientific Linux 6 ("SL6"), the target OS for R2DP.

3.2. D0 software release preservation

D0 is staying with its existing software releases, but moving to updated versions of common tools where possible and making sure that 32-bit compatibility system libraries are installed on worker nodes where D0 plans to run jobs throughout the R2DP project lifetime. Practically this implies that D0 will likely be unable to run analysis jobs outside of Fermilab in future years, but resources at Fermilab are sufficient to meet D0's project demand in the future. Required compatibility libraries can also be added to the CVMFS repository if needed in the future.

4. Job submission and data movement

4.1. Job submission

During Run II, CDF and D0 both had large dedicated analysis farms (CDFGrid and CAB, respectively) of several thousand CPU cores each. Now that Run II has ended, these resources have been steadily diminishing as older nodes are retired and newer ones are repurposed. To preserve full analysis capability both experiments need access to opportunistic resources and a way to submit jobs to them. The Fermilab GPGrid, used by numerous other experiments based at Fermilab, is a natural choice for the Tevatron experiments. Both CDF and D0 have worked with the Fermilab Scientific Computing Division to add the ability to run their jobs on GPGrid, by adopting the Fermilab Jobsub product [2] used by other Fermilab experiments. Having users submit their analysis jobs via Jobsub solves the issue of long-term support, but introduces an additional complication for users who are unfamiliar with the new system, or who may return to do a Tevatron analysis many years from now and will not have sufficient time to learn an entirely new system. Thus, both CDF and D0 have implemented wrappers around the Jobsub tool that emulate job submission commands each experiment used previously

In the case of D0, users who wish to submit to GPGrid instead of CAB (which will be required once CAB is retired) they can simply do so by adding an extra command line option. The D0 submission tools will then generate and issue the Jobsub commands without any direct user intervention. In this way we accomplish the goal of submitting jobs to GPGrid with a minimum amount of effort by not requiring future analyzers to spend time learning an entirely new system. We have successfully tested submission of all common job types (simulation, reconstruction, user analysis) to GPGrid using the modified D0 submission tools.

In the case of CDF, the existing CDFGrid gateway was retired with all remaining hardware absorbed into GPGrid. Analysis and Monte Carlo generation jobs are now submitted entirely using Jobsub and a CDF-specific wrapper in front. This also allows for CDF analysis jobs to go to remote sites on the Open Science Grid, specifically to sites that previously had to operate CDF-specific gateways. These sites can now support CDF computing either opportunistically or via dedicated quotas without the need to support a separate gateway. CDF computing use has not diminished in 2015 despite moving to an environment with no dedicated computing nodes.

4.2. Data movement
Both CDF and D0 use the SAM service [3] for data handling. Older versions of SAM used a CORBA-based infrastructure, but recent versions use a Postgresql-based infrastructure, with communication over http. Throughout Run II CDF and D0 used CORBA-based versions of SAM but have switched to recent versions as part of R2DP in order to eliminate the future support requirements for the older versions, as other experiments at Fermilab are using more recent versions of SAM. There was work to modify the experiments' code bases to interface with these new versions, but the changes are transparent to the end user.

For D0, part of SAM included dedicated cache disks on the D0 cluster worker nodes that allowed for rapid staging of input files to jobs. As files were requested through SAM, they would be copied in from one of the cache disks if they were present, and if they were not already in one of the cache disks, then SAM would fetch them from tape. At its peak this cache space totaled approximately 1 PB, but this cache space was only available to D0 and would have been too costly to maintain over the life of the data preservation project. D0 has therefore switched to using a dCache [4] instance for staging input files to worker nodes, as CDF and numerous other Fermilab experiments are already doing. The test results showed no degradation in performance relative to the dedicated SAM caches, and the D0 dCache instance has been in production for approximately one year.

5. Documentation
Preserving the experiments' institutional knowledge is a critically important part of the project. Here we define this knowledge to be internal documentation and notes, presentations in meetings, informational web pages and tutorials, meeting agendas, and mailing list archives. The largest step in this part of the project was transferring each experiment's internal documents to long-term repositories. Both CDF and D0 have partnered with INSPIRE [5] to transfer their internal notes to experiment-specific accounts on INSPIRE. In D0's case most notes will eventually be made public. For both experiments there was a large fraction of older internal notes that existed only in paper form. There was a dedicated effort to scan these notes into the INSPIRE repository for each experiment. For internal meeting agendas, D0 has moved to an Indico instance hosted by the Fermilab Scientific Computing Division, while CDF has virtualized their mySQL-based system. Fermilab has agreed to maintain archives of each experiment's mailing lists through the life of the project, and Wiki/Twiki instances are being moved to static web pages to facilitate ease of movement to new servers if needed in the future.

6. Summary
The Run II Data Preservation Project aims to enable full analysis capability for the CDF and D0 experiments through at least the year 2020. Both experiments have modernized their software environments and job submission procedures in order to be able to run jobs in current operating systems and to take advantage of non-dedicated computing resources. Wherever possible they have adopted elements of the computing infrastructure now in use by the majority of active Fermilab experiments. They have also made significant efforts at preserving institutional knowledge by moving documentation to long-term archives. The implementation phase of the project is complete and both experiments are actively using the R2DP infrastructure for their current and future work.

Acknowledgments
The authors thank the computing and software teams of both CDF and D0 as well as the staff of the Fermilab Computing Sector that made this project possible. Fermilab is operated by Fermi Research Alliance, LLC under Contract number DE-AC02-07CH11359 with the United States Department of Energy.

References

[1] P. Buncic *et al.* "CernVM - a virtual appliance for LHC applications." Proceedings of the XII. International Workshop on Advanced Computing and Analysis Techniques in Physics Research (ACAT08), Erice, 2008 PoS(ACAT08)012
[2] Dennis Box *et al.* "Progress on the FabrIc for Frontier Experiments Project at Fermilab", In Computing in High Energy Physics 2015 (CHEP 2015), Okinawa, Japan
[3] R. A. Illingworth, "A data handling system for modern and future Fermilab experiments", Journal of Physics: Conference Series, 513, 032045 (2014)
[4] M. Ernst *et al.* "dCache, a distributed data storage caching system", In Computing in High Energy Physics 2001 (CHEP 2001), Beijing, China
[5] *http://inspirehep.net*