Data preservation at the Fermilab Tevatron

S. Amerio^a, S. Behari^b, J. Boyd^b, M. Brochmann^c, R. Culbertson^b, M. Diesburg^b, J. Freeman^b, L. Garren^b, H. Greenlee^b, K. Herner^b, R. Illingworth^b, B. Jayatilaka^{b,*}, A. Jonckheere^b, Q. Li^b, S. Naymola^b, G. Oleynik^b, W. Sakumoto^{b,d}, E. Varnes^e, C. Vellidis^b, G. Watts^c, S. White^b

^aIstituto Nazionale di Fisica Nucleare, Sezione di Padova-Trento and University of Padova, I-35131 Padova, Italy

^bFermi National Accelerator Laboratory, Batavia, Illinois, 60510, USA

^cUniversity of Washington, Seattle, Washington, 98195, USA

^dUniversity of Rochester, Rochester, New York 14627, USA

^eUniversity of Arizona, Tuscon, Arizona, 85721, USA

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 approximately 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 and beyond. To achieve this goal, we have implemented a system that utilizes virtualization, automated validation, and migration to new standards in both software and data storage technology and leverages resources available from currently-running experiments at Fermilab. These efforts have also provided useful lessons in ensuring long-term data access for numerous experiments, and enable high-quality scientific output for years to come.

---

1. Introduction

The Tevatron was a proton-antiproton collider located at Fermi National Accelerator Laboratory (Fermilab). Run II of the Tevatron, occurring from 2001 to 2011 and having collisions with a center-of-mass energy of 1.96 TeV,

---

*Corresponding author. Tel: (630) 840-5014; E-mail address: boj@fnal.gov

17 saw the CDF and D0 collaborations [1–3] record datasets corresponding to an
18 integrated luminosity of approximately 11 fb^{-1} per experiment. These datasets
19 helped make groundbreaking contributions to high energy physics including the
20 most precise measurements of the W boson and top quark masses, observation
21 of electroweak production of top quarks, observation of B_s oscillations, and first
22 evidence of Higgs boson decay to fermions. The unique nature of the Tevatron’s
23 proton-antiproton collisions and large size of the datasets means that the CDF
24 and D0 data will retain their scientific value for years to come, both as a vehicle
25 to perform precision measurements as newer theoretical calculations appear,
26 and to potentially validate any new discoveries at the LHC.

27 The Fermilab Run II Data Preservation Project (R2DP) aims to ensure that
28 both experimental collaborations have the ability to perform complete physics
29 analyses on their full datasets through at least the year 2020. To retain full
30 analysis capability, the project must preserve not only the experimental data
31 themselves, but also their software and computing environments. This requires
32 ensuring that the data remain fully accessible in a cost-effective manner and that
33 experimental software and computing environments are supported on modern
34 hardware. Furthermore user jobs must be able to run at newer facilities when
35 dedicated computing resources are no longer available and job submission and
36 data movement to these new facilities must be accomplished within the familiar
37 software environment with a minimal amount of effort on the part of the end-
38 user. Documentation is also a critical component of R2DP and includes not
39 only the existing web pages, databases, internal documents, but also requires
40 writing clear, concise instructions detailing how users need to modify their usual
41 habits to work in the R2DP computing infrastructure.

42 **2. Dataset preservation**

43 *2.1. Collision data*

44 At the time of the Tevatron shutdown, the data for both CDF and D0
45 were stored on LTO4 tapes [4], which have a per-tape capacity of 800 GB.

46 An analysis of then-available tape technologies concluded that T10K tapes [5],
 47 with a capacity of up to 5 TB per tape, would be the near-term choice for
 48 archival storage at Fermilab. While it was theoretically possible to leave the
 49 CDF and D0 data on LTO4 storage, a decision was made to migrate these data
 50 to T10K storage for two reasons. First, if the Tevatron data were accessed
 51 for a long period of time after data taking ended, the LTO4 storage may be an
 52 unsupportable configuration; as LTO4 tapes decline in usage industry-wide there
 53 may not be replacement storage easily available. Second, as storage media and
 54 drives for older technology become scarce, their costs rise, potentially increasing
 55 the overall long-term cost of staying with LTO4 storage. Due to these concerns,
 56 the commitment was made to purchase T10K tapes and migrate all of the CDF
 57 and D0 data. It took approximately two years for the migration to be completed
 58 (Fig. 1) and the CDF and D0 data now share tape access resources with active
 59 Fermilab experiments. Both CDF and D0 migrated all data stored on tape,
 60 including raw detector data, reconstructed detector data and derived datasets,
 61 and simulation. CDF has also made an additional copy of its raw detector data
 62 at the National Centre for Research and Development in Information Technology
 63 (CNAF) in Italy [6]. Table 1 shows the amount of each type of data that CDF
 64 and D0 have stored over Run II.¹

65 *2.2. Non-statistical data*

66 Throughout the Tevatron run, both CDF and D0 used Oracle database soft-
 67 ware for non-statistical data, such as detector calibrations. The ongoing cost
 68 of maintaining Oracle licenses, which are not used by most current Fermilab
 69 experiments for scientific use, presented a long-term challenge. The database
 70 schema was heavily interwoven into the analysis software. As a result, convert-
 71 ing to a more economical open source database solution such as PostgreSQL
 72 would incur a prohibitive investment in human resources. Thus, both experi-

---

¹While not within the scope of R2DP, a copy of raw CDF data from Run 1 of the Tevatron (1992-1996) is also being made at CNAF.

Data type	CDF	D0
Raw collision	1.8	2.5
Reconstructed collision	3.5	3.4
Derived collision	1.5	1.4
Simulation	2.0	1.5
Total	8.8	8.8

Table 1: Types of data stored on tape for CDF and D0 along with approximate size in PB. The numbers do not include the additional 4 PB of CDF raw data stored at CNAF.

ments 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. The CDF and D0 Oracle database schemas currently contain 1.89×10^{10} rows and continue to be accessed by physicists generating simulated data for ongoing analyses. A breakdown of the main types and size of data in the CDF and D0 databases appears in Table 2. While the project decided to retain the Oracle databases, there was a risk that the physical hardware could fail before the end of the project lifetime. To mitigate that risk, the experiments moved the databases themselves to virtual machines. Most of the servers used for database hosting and caching at both experiments were transitioned to virtual hardware with no degradation in performance or uptime.

3. Software and environment preservation

Both CDF and D0 have complex software frameworks 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

Database	Number of rows (millions)	Size on disk (GB)
CDF SAM	1 015	201.6
CDF Runs	1 336	195.6
CDF subdetector calib.	643	187.7
CDF file catalog	112	24.7
D0 Luminosity	5 162	973.2
D0 Runs	85	21.8
D0 SAM	9 178	2 222
D0 subdetector calib.	901	88.5

Table 2: Summary of CDF and D0 databases. SAM stands for Sequential Access via Metadata (see Sec. 4.2). CDF and D0 internal database structures somewhat differ, leading to different sizes for the major database categories such as SAM and Runs.

93 architectures. During the operational period of the Tevatron, releases of the
94 software framework were maintained on dedicated storage elements that were
95 mounted on the experiments’ respective dedicated computing clusters. As these
96 dedicated resources are no longer maintained, CDF and D0 have migrated their
97 software releases to CERN Virtual Machine File System (CVMFS) repositories
98 [7]. As CVMFS has been widely adopted by current Fermilab experiments,
99 this move allows for maintaining CDF and D0 software releases for the fore-
100 seeable future without a significant investment in dedicated resources. Further-
101 more, as the Fermilab-based CVMFS repositories are distributed to a variety of
102 computing facilities away from Fermilab, this approach lends to CDF and D0
103 computing environments that exist on a variety of remote sites.

104 At the time of the Tevatron shutdown, both CDF and D0 were running
105 software releases that, while operational on Scientific Linux 5 (“SL5”), depended
106 on compatibility libraries that were built in previous OS releases dating back to
107 Scientific Linux 3. The two experiments chose different strategies to ensure the
108 functionality of their software releases throughout the data preservation period.

109 3.1. CDF software release preservation

110 At CDF, stable software releases were available under two flavors: one was
111 used in reconstruction of collision data and analysis and another for Monte Carlo
112 generation and simulation. To ensure the long-term viability of CDF analysis
113 capability, the CDF software team chose to prepare brand new “legacy” releases
114 of both flavors. These legacy releases were stripped of any long-obsolete pack-
115 ages that were no longer used for any analysis and also shed any compatibility
116 libraries built prior to SL5. Once validated and distributed, older releases of
117 CDF software which still depended on compatibility libraries were removed from
118 the CVMFS repository. This meant that usage of CDF code for analysis on any
119 centrally available resources was guaranteed to be fully buildable and executable
120 on SL5. Furthermore, it meant a relatively simple process for further ensuring
121 the legacy release is buildable and executable exclusively on Scientific Linux
122 6 (“SL6”), the target OS for R2DP. The total size of the CDF code base in
123 the legacy releases is 326 GB, which includes compiled code and most external
124 dependencies.

125 3.2. D0 software release preservation

126 After careful study, the D0 software team chose to stay with its software
127 releases that were current at the time of the Tevatron shutdown, but updated
128 common tools where possible, and also made sure that 32-bit compatibility
129 system libraries are installed on worker nodes at Fermilab, where D0 plans to
130 run jobs throughout the R2DP project lifetime. If D0 physicists should wish to
131 run analysis jobs outside of Fermilab in future years, they will need to ensure
132 that 32-bit versions of system libraries such as GLIBC are available at any
133 future remote sites. Resources at Fermilab, however, are sufficient to meet the
134 projected demand over the project lifetime. Required pre-SL6 compatibility
135 libraries can also be added to the CVMFS repository if needed in the future.
136 The total size of the D0 software repository in CVMFS, including code base,
137 executables, and external product dependencies, is currently 227 GB.

138 4. Job submission and data movement

139 4.1. Job submission

140 During Run II, CDF and D0 both had large dedicated analysis farms (CD-
141 FGrid and CAB, respectively) of several thousand CPU cores each. Since the
142 end of the run, these resources have been steadily diminishing as older nodes are
143 retired and some newer ones are repurposed. While the computing needs of the
144 experiments have declined over the years (Fig. 2), preserving full analysis capa-
145 bility requires that both experiments have access to opportunistic resources and
146 a way to submit jobs to them. The Fermilab General Purpose Grid (“GPGrid”),
147 used by numerous other experiments based at Fermilab, is a natural choice for
148 the Tevatron experiments. Both CDF and D0 have worked with the Fermilab
149 Scientific Computing Division to add the ability to run their jobs on GPGrid, by
150 adopting the Fermilab Jobsub product [8] used by other Fermilab experiments.
151 Having users submit their analysis jobs via Jobsub solves the issue of long-term
152 support, but introduces an additional complication for users who are unfamiliar
153 with the new system, or who may return to do a Tevatron analysis many years
154 from now and will not have time to learn an entirely new system. Thus, both
155 CDF and D0 have implemented wrappers around the Jobsub tool that emulate
156 job submission commands each experiment normally uses.

157 In the case of D0, users who wish to submit to GPGrid instead of CAB
158 (which will be required once CAB is retired) can simply do so by adding an
159 extra command line option. The D0 submission tools will then generate and
160 issue the appropriate Jobsub commands without any direct user intervention.
161 Users can switch to submitting jobs to GPGrid with a minimum amount of
162 effort, and future analyzers will not need to spend time learning an entirely
163 new system. We have successfully tested submission of all common job types
164 (simulation, reconstruction, user analysis) to GPGrid using the modified D0
165 submission tools.

166 In the case of CDF, the existing CDFGrid gateway was retired with all re-
167 maining hardware (worker nodes) absorbed into GPGrid. Analysis and Monte

168 Carlo generation jobs are now submitted entirely using Jobsub and a CDF-
 169 specific wrapper in front. This also allows for CDF analysis jobs to go to remote
 170 sites on the Open Science Grid, specifically to sites that previously had to oper-
 171 ate CDF-specific gateways. These sites can now support CDF computing either
 172 opportunistically or via dedicated quotas without the need to support a separate
 173 gateway. This transition was completed in early 2015 and CDF computing use
 174 did not diminish in 2015 as compared to 2014 despite moving to an environment
 175 with no dedicated computing nodes. CDF physicists consumed over 5 million
 176 CPU hours on GPGrid in the twelve months following the transition.

177 *4.2. Data management and file delivery*

178 Both CDF and D0 use the Sequential Access to Metadata (SAM) service [9]
 179 for data handling. Older versions of SAM used Oracle backends with a CORBA-
 180 based communication infrastructure, while more recent versions use a PostgreSQL-
 181 based backend, with communication over http. Throughout Run II CDF and
 182 D0 used CORBA-based versions of SAM, but can now also communicate with
 183 their existing Oracle databases using the new http interfaces as part of R2DP,
 184 eliminating the requirement of supporting the older CORBA-based communi-
 185 cation interface through the life of the project. While CDF and D0 code bases
 186 had to be modified to interface with these new communication interfaces, these
 187 changes are transparent to the end user.

188 For D0, part of the SAM infrastructure included dedicated cache disks on
 189 the D0 cluster worker nodes that allowed for rapid staging of input files to jobs.
 190 As files were requested through SAM, they would be copied in from one of the
 191 cache disks if they were present. If they were not already on one of the cache
 192 disks SAM would fetch them from tape. At its peak this cache space totaled
 193 approximately 1 PB, but this cache space was only available to D0 and would
 194 have been too costly to maintain over the life of the data preservation project.
 195 D0 has therefore deployed a 100 TB dCache [10] instance for staging input files
 196 to worker nodes, as CDF and numerous other Fermilab experiments are already
 197 doing. The test results showed no degradation in performance relative to the

198 dedicated SAM caches, and the D0 dCache instance has been in production for
199 approximately two years.

200 As CDF was already using dCache for tape-backed caching, once the neces-
201 sary code changes were made to use newer versions of SAM, data access contin-
202 ued to be possible with no hardware infrastructure changes needed.

203 **5. Documentation**

204 Preserving the experiments' institutional knowledge is a critically important
205 part of the project. Here we define this knowledge to be internal documentation
206 and notes, presentations in meetings, informational web pages and tutorials,
207 meeting agendas, and mailing list archives. The largest step in this part of
208 the project was transferring each experiment's internal documents to long-term
209 repositories. Both CDF and D0 have partnered with INSPIRE [11] to transfer
210 their internal notes to experiment-specific accounts on INSPIRE. For internal
211 meeting agendas, D0 has moved to an Indico instance hosted by the Fermilab
212 Scientific Computing Division, while CDF has virtualized their MySQL-based
213 system. Fermilab will see to it that archives of each experiment's mailing lists
214 are available through the life of the project, and Wiki/Twiki instances are be-
215 ing moved to static web pages to facilitate ease of movement to new servers if
216 needed in the future. In addition to moving previous documentation to modern
217 platforms, both experiments have written new documentation specifically de-
218 tailing the infrastructure changes that the R2DP project has made, along with
219 instructions for adapting legacy workflows to the new systems.

220 **6. Summary**

221 The Run II Data Preservation Project aims to enable full analysis capa-
222 bility for the CDF and D0 experiments through at least the year 2020. Both
223 experiments have modernized their software environments and job submission
224 procedures in order to be able to run jobs in current operating systems and to
225 take advantage of non-dedicated computing resources. Wherever possible they

226 have adopted elements of the computing infrastructure now in use by the ma-
227 jority of active Fermilab experiments. They have also made significant efforts
228 at preserving institutional knowledge by moving documentation to long-term
229 archives. Materials costs for the project were dominated by media for migrat-
230 ing the data to T10K tapes, while the vast majority of the project’s other costs
231 came from salaries. The implementation phase of the project is complete and
232 both experiments are actively using the R2DP infrastructure for their current
233 and future work.

234 Acknowledgments

235 The authors thank the computing and software teams of both CDF and D0
236 as well as the staff of the Fermilab Computing Sector that made this project
237 possible. Fermilab is operated by Fermi Research Alliance, LLC under Contract
238 number DE-AC02-07CH11359 with the United States Department of Energy.
239 The work was supported in part by Data And Software Preservation for Open
240 Science (DASPOS) (NSF-PHY-1247316).

241 References

- 242 [1] D. Acosta *et al.*, “Measurement of the J/ψ meson and b -hadron production
243 cross sections in $p\bar{p}$ collisions at $\sqrt{s} = 1960$ GeV”, Phys. Rev. D 71, 032001
244 (2005).
- 245 [2] A. Abulencia *et al.*, “Measurements of inclusive W and Z cross sections in
246 $p\bar{p}$ collisions at $\sqrt{s} = 1.96$ TeV”, Journal of Physics G: Nuclear and Particle
247 Physics, 34, Number 12 (2007).
- 248 [3] V. M. Abazov *et al.*, “The Upgraded D0 Detector”, Nucl. Instrum. Methods
249 in Phys. Res. Sect. A 565, 463 (2006).
- 250 [4] The Linear Tape Open Consortium, <http://www.lto.org>.

- 251 [5] Oracle, "Storagetek T10000 data cartridge family data sheet",
 252 [http://www.oracle.com/us/products/servers-storage/storage/tape-](http://www.oracle.com/us/products/servers-storage/storage/tape-storage/storagetek-t10000-t2-cartridge-296699.pdf)
 253 [storage/storagetek-t10000-t2-cartridge-296699.pdf](http://www.oracle.com/us/products/servers-storage/storage/tape-storage/storagetek-t10000-t2-cartridge-296699.pdf).
- 254 [6] S. Amerio *et al.*, "The Long Term Data Preservation (LTDP) project at
 255 INFN CNAF: CDF use case," J. Phys. Conf. Ser. 608 (2015) no.1, 012012.
- 256 [7] P. Buncic *et al.*, "CernVM - a virtual appliance for LHC applications."
 257 Proceedings of the XII. International Workshop on Advanced Comput-
 258 ing and Analysis Techniques in Physics Research (ACAT08), Erice, 2008
 259 PoS(ACAT08)012
- 260 [8] Dennis Box *et al.*, "Progress on the FabrIc for Frontier Experiments Project
 261 at Fermilab", Journal of Physics.: Conference Series, 664 062040 (2015).
- 262 [9] R. A. Illingworth, "A data handling system for modern and future Fermilab
 263 experiments", Journal of Physics: Conference Series, 513, 032045 (2014).
- 264 [10] M. Ernst *et al.*, "dCache, a distributed data storage caching system", In
 265 Computing in High Energy Physics 2001 (CHEP 2001), Beijing, China.
- 266 [11] <http://inspirehep.net>.

Figure 1: Tape migration status from LTO4 media to T10K media over the course of R2DP for CDF (top) and D0 (bottom).

Figure 2: Monthly utilization of computational resources (in core-hours) on the Open Science Grid by CDF (light yellow) and D0 (dark red) from October 2010 to October 2015. These resources include CDF and D0 dedicated resources prior to 2015 as well as GPGGrid resources at Fermilab. Since early 2015 the bulk of D0 computing has run on its dedicated CAB cluster, which is not shown here from 2015 onward.