GLOBALLY DISTRIBUTED USER ANALYSIS COMPUTING AT CDF

A. Sill*, Dept. of Physics, Texas Tech University, Lubbock TX 79409-1051 USA
H.T. Lung, S. Hou, Academia Sinica, Taiwan
E. Lipeles, M. Neubauer, F. Würthwein, University of California, San Diego
A. Kreymer, Fermi National Accelerator Laboratory, Batavia IL 60510 USA
M. Burgon-Lyon, R. St. Denis, University of Glasgow
I. Sfiligoi, INFN, Frascati
A. Fella, INFN, Pisa
S. Belforte, INFN, Trieste
K. Cho, D.H. Han, H.W. Park, Kyungpook National University/KISTI
V. Bartsch, S. Stonjek, University of Oxford
F. Ratnikov, Rutgers University
L. Groer, R. Tafirout, University of Toronto
H. Matsunaga, University of Tsukuba

Abstract

To maximize the physics potential of the data currently being taken, the CDF collaboration at Fermi National Accelerator Laboratory has started to deploy user analysis computing facilities at several locations throughout the world. At the time of this writing, over 700 individual users are signed up and able to submit physics analysis and simulation applications directly from their desktop or laptop computers to these facilities. These resources consist of a mix of customized computing centers and a decentralized version of our Central Analysis Facility (CAF) initially used at Fermilab, which we have designated Decentralized CDF Analysis Facilities (DCAFs). The goals of this project are to reach a total of 25% of the experiment's overall computing through off-site resources by the end of 2004, and to expand this to 50% by the end of 2005 through grid-enabling these resources.

We report on experience gained during the initial deployment and use of these resources for the summer conference season 2004 that have allowed us to meet the 2004 goal. During this period, we allowed MC generation as well as data analysis of selected data samples at several globally distributed centers. In addition, we discuss our plans for developing a migration path from this first generation distributed computing infrastructure towards a more open implementation that will be interoperable with LCG, OSG and other general-purpose grid installations at the participating sites.

INTRODUCTION AND BACKGROUND

The CDF collaboration at Fermi National Accelerator Laboratory is currently engaged in analysis of physics data resulting from operation of the Tevatron collider. Like all operating experiments, the experiment faces a computing load that includes elements of initial processing of raw data, application of calibration and good run filtering information, splitting into physics data sets by candidate process selection, organizing and cataloging of the results, and further processing to produce ntuples used for detailed analysis by the physicists. In addition, Monte Carlo generation and production must take place in large scale for simulated data sets for each of the physics analyses to develop and refine analysis techniques, check backgrounds and acceptance, model detector response, etc. For CDF, the resulting computing load divides into the following steps in terms of analysis workflow:

- Initial production processing, calibration and physics dataset splitting is handled by a step called "production" on a dedicated processing farm at Fermilab.
- Monte Carlo generation and detector simulation allow simulated data to be prepared, which are then taken through a similar production processing path.
- The resulting data sets are catalogued and made available for processing into physics analysis ntuples by the physics groups.

During the past year, CDF has taken steps to move a large portion of the latter two categories of work above, which constitute the majority of our computational needs, into a state such that they can be conducted off site, both for convenience of the physicists and to lessen the load on CDF central computing resources. In addition, the development and widespread current adoption of grid computing methods throughout the world has created opportunities for CDF to position its workload and resources in a way that will enable us to take advantage of such resources in the future.

In this document, we discuss the current state of deployment of CDF global computing and the steps that have been required to reach this state. The total computational power

[*] Alan.Sill@ttu.edu

of the resources deployed so far represents 35% of the experiment's capacity in terms of compute cycles, placing us well in the target range of the desired goals. We also map out a short plan for grid-enabling these resources and merging this system with the developing worldwide high energy physics grid in the future.

CDF Analysis Farms

The bulk of CDF's physics processing load has been provided up to now by a central set of shared cpu and disk resources running Linux on Intel-compatible 32-bit hardware, using multiple RAID-5 file servers in a concentrated network environment [1]. The cpus are purchased in bulk and taken through an extensive burn-in process before acceptance. Transfers to and from the file servers take place for data using a combination of rootd, dcache and other secure transfer mechanisms, without explicit nfs mounts.

This architecture, called the CAF for "CDF Analysis Farm," uses Kerberos for user authentication and access, and is backed up by Enstore and dCache data handling systems [2, 3]. The result has been a highly successful user environment for shared resources, characterized by an access model in which universities or institutions that contribute resources are given priority usage on queues that correspond to their portion, but the entire cluster is open to opportunistic shared use by members of the collaboration at large. This model has produced a desirable combination of high utilization (essentially 100%) and manageable fair share guaranteed access for cluster contributors.

The challenge over the past year has been to extend the above successful model to the use of resources provided at the home institutions of our contributing members, located all over the world. CDF had already reached a state in which a large portion of the Monte Carlo production comprising almost all of the official samples is produced off-site, using clusters that were either dedicated to this task or available for this purpose, but not in general open to large-scale collaboration use on demand. It was our desire to extend the CAF model described above to the use of off-site resources, so that CDF members would be able to submit their jobs more freely to these resources for both Monte Carlo and real data analysis as desired.

DISTRIBUTED ANALYSIS

Physicists using the CAF system interact with the workflow of their jobs through a GUI or equivalent command line interface that is customized to this purpose. Users specify the number of instances of their job that they want to run, the directory are or tarball that contains the files needed to run the job, and the command that will start the analysis. Elements of the CAF workflow that we wanted to preserve included this easy-to-use interface as well as a simple method to specify the data set to be analyzed for jobs that require data input.

The SAM system [4] originally developed by Dzero was chosen for adoption as the data access model for off-site computing. This system provides an organized way to track the processing history of a physics dataset, as well as methods to manage the flow of data to distributed cache disks (and even tape), as required for off-site computing. To keep our data access model uniform, SAM was introduced for on-site computing at Fermilab as well. The CAF graphical user interface (GUI) and command set were altered to allow inclusion of a SAM dataset identifier, and the underlying CDF analysis framework and CAF software were extended to provide the functionality needed to manage SAM project workflow.

Basic elements

The resulting "DCAF" analysis environment, where DCAF stands for "Distributed CDF Analysis Farm," consists of the following basic elements:

1. The CAF cluster environment, consisting of a submitter, monitor, user analysis sandbox, and mailer to inform the user of the final status of the job.
2. The CDF software, organized by release and installed separately through our normal distribution methods.
3. A SAM station, required at each DCAF site.
4. Kerberos authentication, identical to the original CAF architecture and achieved through a special Kerberos cluster principal at each site.

To this we added Ganglia monitoring and batch system user monitoring, optional bandwidth monitoring between each site and Fermilab provided by the Fermilab network team, and a basic informational home page to describe contact information and details on special conditions that might exist at each of the DCAF clusters. When a data set is chosen, the software checks to make sure that it exists on the DCAF selected, and reports the appropriate details in terms of number of files and total data volume back to the user for confirmation before proceeding with the submission.

Table 1: DCAF Features as of Summer 2004

Feature	Status
Simple self-contained sandbox	Yes
Runs arbitrary user code	Yes
Automatic identity management	Yes
Network delivery of results	Yes
Input and output data handling	Yes
Batch system priority management	Yes
Automatic choice of farm	Not yet
Negotiation of resources	Not yet

Current status

Table 1 summarizes the basic features of the DCAF environment in terms of functionality commonly associated with large-scale grid computing. Although some desirable features are not yet in place in terms of automated job routing and automatic negotiation of resources, the resulting environment nonetheless provides a highly feature-rich and easy-to-use set of functionalities to CDF physics users.

Table 2: DCAF Capacity as of Summer 2004

Location	cpu (GHz)	Disk Space (TB)
Original FNAL CAF	1200	300
FNAL CondorCAF	2000	shared w/ above
CNAF (Italy)	300	7.5
KNU (Korea)	120	0.6
AS (Taiwan)	134	3.0
SDSC (USA)	280	4.0
Rutgers (USA)	100	4.0
Toronto (Canada)	576	10
Tsukuba (Japan)	152	5.0
Cantabria (Spain)	52	1.5
MIT (USA)	110	2.0
Totals:	5024	337.5

As of the end of the summer 2004 conference season, a total of 11 DCAF systems comprising two central CAFs and 9 off-site ones were in operation. Each of these facilities is open to all registered CDF physicists. Table 2 shows the cpu power and disk space online at each location, with cpu power measured in equivalent P4 GHz. Approximately 1.8 out of a total of 5.0 THz is now provided off-site, representing 35% of the CDF user analysis computing capability.

Disk space in the off-site institutions represents a much smaller portion, only about 15%, of the total available. This is partly due to the previous emphasis on the part of institutions and universities within the collaboration on providing storage at Fermilab to support analysis efforts. Approximately half of the FNAL-based storage is in the form of these university and institutional specific disk units. As the data handling system becomes more robust and the networks more mature, we expect this fraction of disk storage to grow over time at the remote DCAFs.

The SAMGrid data handling systems at these facilities, comprising the SAM project-oriented database combined with gridftp, have been operated in a mode that allows individual large-scale data sets to be pinned, i.e., transferred and locked, onto the remote storage on demand at any of the DCAFs. In addition, for the summer conference season we have supported the use of SAM to import and store remotely produced Monte Carlo results to make these available to the data handling system.

We are in the process of transition to a new release of the core SAM software with improved capabilities to handle transfer problems for large and complex data sets. In combination with other advances in storage deployment through the use of technologies like SRM-dCache [5], this should permit more general access pattern and use of the SAMGrid cache to analyze data without requiring the entire data set to be pinned remotely.

We are also implementing a new caching layer [6] to our calibration database access to reduce latency and to improve overall speed of access, both of which are major potential problems for high-intensity data processing at remote sites. This layer will be extended to other components of database access if it proves to be of value and necessary in field use.

UTILIZATION

Figure 1 shows the number of processes versus time on a scale that spans the past year for the central versus remote systems. As DCAFs have been commissioned, the overall usage has risen to peak values that are becoming proportional to the number of cpus deployed remotely. Overall utilization of the remote DCAFs has been more variable on the remote sites versus time and from site to site, but is rising will soon reach a steady high state. Up to now it has been necessary for users to select manually the DCAF to use for each submission. We are looking at various methods that can be used to distribute the jobs more automatically.

LESSONS LEARNED AND OUTLOOK

Methods used to achieve this deployment included a series of workshops to define and disseminate the required software set, creation of a rapidly updated “live” set of documentation to provide a cut-and-paste instructional reference for new site administrators, and a strong focus on operations, in which weekly operational summary and daily short coordination meetings were used to keep emerging problems under control. Deployment of a complex software stack was thus achieved with a small team for central coordination, allowing further development to proceed.

The long-term outlook for grid computing within CDF includes adoption of a variety of grid-enabled protocols beyond those we have adopted up to now. The strong focus on the use of a limited subset of these functionalities up to now, principally Condor, Kerberos, and SamGrid, has allowed deployment of a large set of computational resources for analysis in a form that is useful to the collaboration and that can be expanded in the future. In addition to the DCAF clusters described here, CDF has several institutional resources that can in principle be made available if adapted for access through appropriate interfaces. To achieve this, we are in the process of evaluating grid technologies including several reported at this conference. This evaluation process will be completed within the next few months and should result in further expansion of global user analysis capabilities for CDF in the future.

CONCLUSIONS

CDF has successfully deployed a global computing environment for user analysis based on a simple clustering and

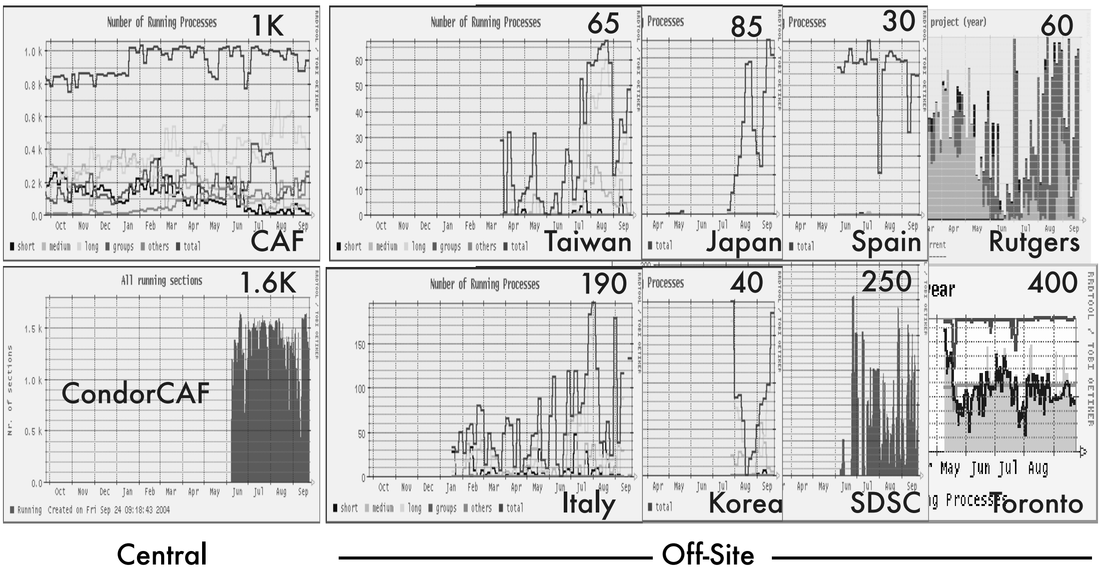

Figure 1: Utilization on central and remote DCAF systems, showing the number of processes active on each of these facilities versus time on the horizontal axis. The peak value on the vertical axis is shown for clarity in each case. As systems were commissioned, primarily in the latter portion of the past year, usage has grown to a bit less than 1/3rd of the total. (MIT, not shown, comprises approximately an additional 100 processes.)

submission protocol, with a large number of registered and active users. Starting from our all-central configuration in January of this year, a large portion, 35%, of the total cpu resources of the experiment are now provided offsite.

Aggressive trimming of the software suite has allowed us to bring up and use these computational resources in short order. Plans are in progress to build in more grid-like protocols to the overall system to provide access to other facilities and a bridge to the future.

ACKNOWLEDGEMENTS

We wish to thank the staff members of our respective institutes involved in system administration, database and data handling support, without which this work would have been impossible. Portions of this work were supported by the US Department of Energy, *(etc.)*.

REFERENCES

[1] M. Neubauer *et al.*, “Computing for Run II at CDF,” Proceedings of the VIII International Workshop on Advanced Computing and Analysis Techniques (ACAT 2002), Moscow, June 2002, Nuclear Instruments and Methods in Physics Research **A502** (2003) 386-390; T.H.Kim, M. Neubauer, I. Sfiligoi, L. Weems, and F. Würthwein, “The CDF Central Analysis Farm,” IEEE Transactions on Nuclear Science, June 2004, Vol 51, No 3, Part III, ISSN 0018-9499.

[2] http://www.fnal.gov/docs/products/enstore; see also J Bakken *et al.*, “The Status of the Fermilab Data Storage System,” these proceedings.

[3] http://www.dcache.org; see also P. Fuhrmann, “DCache, LCG Storage Element and Enhanced Use Cases,” these proceedings.

[4] http://projects.fnal.gov/samgrid ; see also R. St. Denis *et al.*, “Housing Metadata for the Common Physicist Using a Relational Database,” these proceedings.

[5] R. Kennedy, “SAMGrid Integration of SRMs,” these proceedings, and references contained therein.

[6] http://whcdf03.fnal.gov/ntier-wiki/FrontPage; see also S. Kosyakov *et al.*, “FroNtier: High Performance Database Access Using Standard Web Components in a Scalable Multi-tier Architecture,” these proceedings.