DØ Note ####

Status of Data Access and Databases

Lee Lueking, Vicky White

January 10,1999

Abstract

We describe the current status of the D0 Data Access and Databases efforts. This overview of the status of the requirements, specification, prototyping, coding, and testing efforts was prepared for the Von Ruüden Committee Review of Run II Computing at Fermilab, January 25–28, 1999.

Contents

1 Data Access

The Production and Data Access Group is responsible for delivering a data handling system which will provide rapid, convenient, and well tracked access to data. This must be well matched to the evolving priorities and phases of the experiment and must make optimal use of the finite and limited resources of tapes, tape drives, robot arm and shelf space, disk, cpu and network bandwidth. The primary objectives are to reliably store all raw data produced by the online system and by Monte Carlo simulations; and to provide access to the data for all subsequent stages of processing, such that physics results can be extracted as quickly and efficiently as possible.

The Computing Plan of January 1997 [?] identified the scope of the Reconstruction Processing, Analysis Processing, and Monte Carlo simulations which will be needed for Run II. The scale of the data management effort which accompanies these processing tasks is set both by the anticipated size of the dataset and the number and diversity of the expected analyses. The D0 data set is expected to total at least 560 TB. Of this, 300 TB will be RAW data from the DAQ system and the remainder needed for reconstructed and subsequent analysis stages. It also set the stage for a data handling system which will use a tape robot and tape storage for most data, caching frequently used data on disk, will utilize standard Farm systems for Production processing and a Central Analysis server.

1.1 The Sequential Access Model

During the early part of 1997 much work was done, largely in the context of a Run II Joint Computing Project working group, to further refine the computing needs for Run II. The Data Management Needs Assessment [1] report resulting from this study was published in June of 1997 and along with the basic assumptions from the D0 plan, outlined above, provided input to development of a D0 model for organizing and accessing data. At the summer workshop of 1997 the experiment decided to use Oracle as its central and primary database management system; the file and event catalog and production tracking database therefore were to be implemented using Oracle.

The D0 data access model, or data management strategy, was developed largely in the context of a Run II Joint project working group: one of 3 working groups convened in the area of Data Access. The model shares many features with the data management approaches used in previous runs and employs a file-based system of storage and book-keeping. For this reaason it has been called the "Sequential Access Model" or SAM. A detailed report on the Sequential Access Model was prepared for the October 1997 Von Ruden Review of Run II Computing [2]. The key features of this model are:

- Event Data is provided in files of event data units (EDUs) of some size and type. Data is expected to be accessed by reading sequentially through an entire file. The system is optimized for this type of access.
- An event is the logical storage unit which the overall data handling system catalogs and can arrange access to. Rarely analyses are expected to want access to one particular EDU in a file.
- Access to the data will be through various different modes of access. There are 5 basic types of delivery needs anticipated: 1) File and event database scans, 2) Experiment shared random access disk, called "Thumbnail" 3) Coordinated large dataset processing, referred to as "Freight train", 4) Random event picking and,

5) Individual user on-demand file access. These modes of data access each have specificly defined requirements for mass storage bandwidth, disk staging space, processing CPU and the number of concurrent users.

The goal of this model is to optimize the use of storage resources such as tape mounts and drive usage. In order to facilitate this goal, we have four primary objectives: 1) Clustering the data onto tertiary storage in a manner corresponding to access patterns, 2) Caching frequently accessed data on disk or tape, 3) Organizing file requests to minimize tape mounts, and 4) Estimating the resources required for file requests before they are submitted and, with this information, making administrative decisions concerning which data to deliver. In addition, it is desired to unload the burden of individual file tracking from the analysis physicists, and place it onto the data management system.

Clustering of data to optimize later access involves streaming the data directly from Level 3. Several possible streaming strategies were examined - exclusive streaming (an event is written to exactly one physical stream), inclusive streaming (some percentage of events, perhaps as much as 30

- Luminosity calculations
- Calibration data
- Providing Logical Streams of data to hide multiple physical data streams from the user

Streaming Strategies were studied and discussed in [6] and [7].

A series of detailed studies were made on data access flows based on various patterns and volumes of data access thought to be likely in the different stages of evolution of the detector and data processing. This document included detailed spreadsheets and calculations, based on historical patterns of access and information gathered from various physics groups and individuals. It is called “Flows and Controls of the Seasons of SAM” [3].

1.2 Data Handling Review

In March, 1998 the experiment held a Data Handling Workshop and Review to discuss and examine the proposed Sequential Access Model for data access. Several Physics groups were asked to carry out ’gedanken experiments’ to explore in detail how particular analyses might proceed using the exclusive data streaming proposed and the specific data tiers of event data under consideration.A series of questions were developed for Data Access method and streaming scenario proponents. The result of this workshop was an endorsement of the proposed Sequential Access Model [8]. Streaming based on level 3 was deemed to be reasonable, as well as the proposals for the 3 data tiers EDU250, EDU50 and EDU5, and the various modes of access proposed.

However, there was some concern about the following issues and they require ongoing consideration:-

- ramifications of exclusive streaming, including effects of data loss and complexity of luminosity calculations
- the physics content of EDU50 and EDU 5 remain to be finalized
- export of data to remote institutions

1.3 Luminosity Workshop

One of the more tricky issues associated with using an exclusive streaming method of data clustering is that of Luminosity tracking. The way in which streaming is to be done and the mechanisms for tracking data file event contents and luminosity have considerable effect on the Event/File Database design. In response to this Heidi Schellman and Rich Partridge organized a 2 day workshop on luminosity issues. A report from this workshop is available [9] as well as a more detailed discussion of a proposed method of normalization [10]. The responsibilities of the Online Data Logger to either record details about time-ordering of events, or possibly re-order events before writing to a Run Partition/Physical Data stream file is still under discussion.

1.4 Data Handling Layers Implementation Strategy

It was understood that the full data handling system necessary to store and manage the data would require several components working together to provide a functional system. A goal of the Run II Computing Joint projects was to identify and specify these components so that, as far as possible, common solutions could be developed for both D0 and CDF. With this in mind several working groups and projects were instituted to look into requirements, define boundaries and test possible solutions.

- RIP (Reconstruction Input Pipeline) to deal with the first phase of the data handling which involves moving data written to local experiment disk, by the Online system, to the Feynman center where it is to be archived and made available to the Farm Reconstruction system and to all subsequent stages of analysis and processing.
- Storage Management to provide for migration of data between disk and tape and for use of a tape robot.

- SAM Data Access to provide optimized and controlled access to data and to track the entire data creation, reconstruction and analysis process.
- Specification and development of a Production Management system for Run II, including Farm software and Batch system software.

Although D0 people are heavily involved in each of these projects and in particular in the successful integration and deployment of them at D0, they are all being carried out under the banner of the Run II Joint CDF/D0/CD Computing projects. CDF's data access model and proposed data handling system currently makes little use of the work of these projects (apart from the Farm system) although as recently as November 1998 a joint data handling workshop was held to examine the work of the prototype data access systems of both experiments in an attempt to find common ground.

1.5 Storage Management/ENSTORE

The Computing Division had gathered considerable experience in the use of the HPSS (High Performance Storage System) in the data center. As a result of this experience together with a detailed look at the requirements of Run II for storage management, including needs for reliability and placement of data on tape, an HPSS workshop was held in the Spring of 1998. Following this workshop a decision was made by the Run II Computing Joint Project Steering Committee to treat HPSS as a backup solution for Run to and to endorse the development of alternative storage management software [11].

The Run II Storage Management project is now focussed entirely on implementation of a Fermilab Storage management system called ENSTORE, [12] based on a DESY design. D0's requirements for Storage management are discussed in [13], in [4] and are further described in a presentation made at the November CDF/D0 data handling workshop [14] and an SAM interface document. These are being met by the ENSTORE system and testing of the Prototype SAM system involves use of, and integration with, the ENSTORE system. The ENSTORE system delivers files over an IP network from robot resident tape volumes. All details of tape allocation, queing, robot control and tape read/write error handling are managed by the ENSTORE system. Controls are provided to allow the Data Access layer to prioritize and re-prioritize the work in the ENSTORE queue, based on performance and operational statistics provided with each file transfer performed. The primary file/tape volume catalog is maintained by the ENSTORE system and is user visible via a pseudo NFS mounted pnfs file system. For performance reasons and to allow queries spanning multiple facets of file and event attributes a copy of the file/volume catalog is maintained within the SAM event and file database.

1.6 Data Access/SAM

The SAM Data Access project is building a prototype data handling system which stores and deliver data files; catalogs data files, data streams and events; and records all processing steps and analyses in an Oracle database. The Requirements document of June 1998 [4] remains the guiding document for the design and implementation of the SAM system. The Architecture document of June 1998 [5] also serves as the conceptual basis for the detailed design and coding work which is taking place as a Prototype SAM system is being built and tested. Additional detailed design work and specification of interfaces can be found in documentation and notes available from the SAM project website http://RunIIComputing.fnal.gov/sam [15]. This includes detailed design work on the entire database supporting SAM.

The recommendations of The D0 Global Computing Model Working group to adopt a 3-tier analysis system architecture further validated the design criteria of SAM - to build fully distributed system supporting multiple heterogeneous analysis and reconstruction systems via network-attached tape drives.

1.7 RIP

The RIP project, which will use ENSTORE to write data to network-connected tape drives over an IP network, has become an extension to the Online systems projects. The way in which the data written by RIP is to be cataloged in the SAM Data Catalog has been specified and understood. The RIP requirements specified in October of 1997 [16] remain the guiding requirements, including those which deal with total rate, glitch buffer and the placement of files on tape according to physical data stream. The decision to use ENSTORE and IP network-connected tape drives for this part of the data handling system means that Gigabit Ethernet technology can, and has, been chosen as the network technology for the interconnection between the D0 experiment and the Feynman Computing Center for Data Logging.

The Files written from the online system through RIP will be written using ENSTORE and also cataloged in the SAM event and file catalog. The online data logger will provide the meta-data required by SAM for each file written, including the event ranges/timestamps for events in a file. In addition the online system will provide information about each event in a file, including its trigger and filter bits, timestamp, and data stream. [17]. This information will be used to construct the Event and Data Stream parts of the SAM Event and File catalog.

1.8 Production Management/Farms/Batch

A Farms Batch system design [18]and a prototype system for use has been produced by the Production Management Run II Project. There are two possible models of use of this system under discussion which integrate D0 production code, SAM and the Farm system.

A Batch Working Group [19] has been setup up to examine the general requirements for a Batch system for Run II analysis processing. The interaction between the SAM data delivery system and a Batch system used to control such resources as CPU time and scratch disk, is being examined in the context of this working group. The working group will report by March and identify further work which needs to be done to either extend features of the LSF system, which is the primary contender for a Run II batch system, use its features in a appropriate manner, or in the worst (unlikely) case look for another batch system.

1.9 SAM and ENSTORE Prototype Data Handling System

A prototype SAM system is being constructed based on a suite of CORBA-based Servers which manage communication with the analysis/reconstruction programs and delivery of files to those programs. The Servers include:

- Project Master - to manage file delivery to multiple consumers for a single project
- Station Master - to manage the disk, file system and network resource available to a single 'Station" and to control startup and access to Projects
- Buffer Manager - to handle interactions with the Enstore system
- Optimizer - to control and optimize flow of file staging requests to Enstore
- Logger, Information Server and Database Server to manage internals of SAM and provide information

Client interfaces to Import files, specify file meta-data, specify an analysis project, start data delivery for an anlysis project and to consume data files all have version 0 implementations. CORBA IDL interfaces have been defined and code has been written in C++, Java and Python to provide the client functionality.

A full-featured database system has been designed to track events, files and all processing performed on them. Version 0 of this database has been implemented as an Oracle database.

A standalone test system has been put together to facilitate testing of SAM and ENSTORE and to simulate delivery of data to multiple network-connected 'stations' and to multiple projects each with many analysis user programs requesting access to the delivered data. The SAM test systems samson, delilah, sammy, samadams, samting, sameggs and samham are rack-mounted PCs running RedHat Linux and are connected to a similar PC ENSTORE teststand via a gigabit-capable switch connection. The node samson also serves as an ENSTORE system local test node to move data files between SAM and an ENSTORE storage system; in this case a disk based system capable of storing up to 18 GB of files. The gigabit connection between test systems has not been available until recently and so no high-rate tests between systems have yet been conducted. The ENSTORE system has demonstrated movement between sony AIT tape and client disk at full tape speed rates however.

Details of the design and work of the ongoing prototype SAM can be found on the SAM home page [15].

1.10 Milestones and Future Plans

October 11, 1998 - Functioning Prototype for Single Project This milestone was met and a test consumer process was able to obtain files from a SAM project,managed by a Project Master server process. The SAM servers were able to communicate with ENSTORE servers which were used to store and retrieve files from a psuedo-Robot archive. Shortly after this the prototype system was exercised with ENSTORE servers and movers connected to the STK robot and Redwood tape drives.

November 9, 1998 - Demo of SAM at Supercomputing 1998 in Orlando The SAM PC teststand was operational for this milestone and we were able to demonstrate running multiple different projects each fetching data files from an internal SAM Enstore system on node samson. Several 1GB data files were used as well as a number of few MB datafiles. Up to 30 different analysis programs simulataneously requested delivery of the files from a project - each simulating a wait to analyse the file and randomly creating error conditions whereby the file was not successfully marked as processed. Details of the project's definition and files and the consumption of these files by each analysis process were kept in the database.

November 30, 1998 - SAM Prototype connected to ENSTORE to Grau Robot This milestone was achieved by December 21st 1998, although only brief tests were conducted at that time.

December 30, 1998 - Documented Prototype for Monte Carlo file storage and retrieval This milestone is late although the system is actually capable of storing Monte Carlo files. Documentation exists for specifying the meta-data associated with a Monte Carlo file, including all its precedent files, parameter files.

January 31, 1999 - Version 0 of command line user interface and Farm System integration

The V0 interface is documented and largely implemented. Integration with the Farm system is underway.

February 3-5, 1999 - Performance tests/24 hour tests with ENSTORE. Demonstrate system at D0 collaboration meeting.

February 15, 1999 - Version 0 Production Event/File Database ready for use.

February 28, 1999 - Version 0 User Guide/File Import area on delilah

March 30, 1999 - Version 0 Integration with D0 Framework. This includes porting and testing of all user code to SGI machines and full integration with SoftRelTools.

April 15, 1999 - Version 1 of User guide

April 30, 1999 - Simple Pick Event capability, Tape Import capability

May 30, 1999 - Design of Station, Optimizer and Interface to Batch System complete

June 15, 1999 - V1 Event and File Database

June 15, 1999 - Begin Import Run I data into SAM

July 31, 1999 - V0 Implementation of Station, Resource Management and Batch Interface

September 15, 1999 - V2 Event and File Database

September 30, 1999 - V1 SAM Servers and Resource Management software complete

November 15, 1999 - V0 design for EDU5(Thumbnail) data access

December 15, 1999 - V2 SAM system including Pick Event and Thumbnail access complete

2 Databases

2.1 Databases Strategy

The Databases group (leaders: Vicky White and Lee Lueking) exists in order to coordinate the work on Databases, which will take place in the context of several other areas of D0 Run II Computing. The plan is to use a central Oracle database - one for the Online and one for the Offline system, and to provide for uploading and synchronization of data between them. The databases group tries to ensure that a common approach to designing the database and database applications is followed. They disseminate information and provide examples of ways to access the database. There is ongoing work in the area of the D0 persistency mechanism d0om. The need for part of the central Oracle database to be exportable to remote institutions, where Oracle may not be available is also understood and part of the charge of the group to propose the mechanisms for this.

The group provides the database infrastructure needed. An import aspect of this is the work carried out by the Support Databases Project, which is part of the Run II Joint CDF/D0/CD Computing effort.

A Sun 3500 and a Sun 4500 Server machine have been purchased and are in the process of being installed. These machines will serve as the production database machine and a development and test machine for the next year. Following extensive testing using the 4500 machine we will be in a position to specify the CPU capability and I/O capability needed for the offline database machine.

Meetings of the Databases group have been infrequent up to now - early meetings simply to establish the candidate database applications and timescales. In the past two months people have been starting to think about their database issues and regular monthly meetings of the Databases group may soon become appropriate.

An overview of the D0 Database Strategy was presented in a poster paper at CHEP 98 - http://d0db.fnal.gov/xxxx.

The group maintains a web page at http://www-d0.fnal.gov/ white/d0db and a users page at http://d0db.fnal.gov/d0db.

2.2 Databases Candidate Implementations

Seven different candidate areas which need a supporting database have been identified. These are:- File and Event Catalog, Run Conditions, Calibration (Online and Offline), Geometry, Alignment, Triggers,RCPs, Luminosity and Hardware Addresses.

Work to specify the nature of the database and the way in which it is to be used is taking place in some of these areas, as described below. In other areas this work has not yet begun.

2.3 File and Event Catalog /SAM database

This area has received most of the attention and therefore leads the way in providing expertise and tools for the other areas. Detailed analysis and design work has been done using the Oracle Designer tool to document the Entity/Relation diagram of the database and generate the tables. In addition a Database Server approach has been used, based on CORBA IDL interfaces and this also is forming the basisi for a core database server which may underly specific database servers in each of the other areas.

2.4 Support Databases Joint Project and Common Tools

All of the work which has taken place has been supported by and is in cooperation with the Support Databases Joint project. Members of the Support Databases project work closely with D0 people to specify the requirements and design the database and database applications.

A list of potential common tools, not only for D0 Database Applications, but also for CDF use, is being developed as part of that project and many may be used in support of the Database Applications and D0 Persistency package database back-end work.

2.5 Trigger Database

A database for tracking trigger configurations, statistics, and more, is being specified by Carmen xxx and Amber Boehnlein, working the Support Databases member Barbara Angelos. They are nearing the end of the design phase and will shortly have Version 0 of a Trigger Database.

2.6 Alignment Database

V0 of an Aligment Database has been developed by Al Ito, working with Support Databases member Julie Trumbo. Al is able to perform all uploads of alignment data from his NT PC desktop using Excel and Microsoft Access to front-end the Oracle database.

2.7 Hardware Database

Work is proceeding by Stan Krzyllllllll on the design of this database.

2.8 RCP Database

Marc Paterno and Harrison Prosper, working with Jim Kowalkowski, Vicky White and Steve White (from Support Databases project) are in the process of specifying and designing the RCP database.

2.9 Calibration Databases

Ela Barberis and Vicky White are working on an initial prototype implementation of an offline calibration manager and database. This exercise is also being used to understand which extensions to the d0om package to support calibration databases, are necessary or desirable.

2.10 Milestones and Future Plans

There are few specific milestones for this area since they are either related to the Support Databases Joint project, for database infrastructure needed, or they depend on manpower from other areas of D0 Run II Computing being ready to start work on their database problem. However, good progress is being made in the projects which are active and this should serve to lighten the work needed in other areas.

January 15, 1999 - V0 prototype for Offline Calibration Database and Offline Calibration Manager.

March 15, 1999 - V0 of RCP Database

April 15, 1999 - V0 of Trigger Database

Calibration Databases need to receive the most attention, after the Event/File Database for SAM. They will have to hold and provide access to significant amounts of data.

3 Data Access and Databases Open Issues

A number of issues remain open and finding solutions for them has not necessarily been included in the list of Milestones. Further experience with using the systems for Monte Carlo data and possibly Run I data may help to resolve some of these.

- Physics Content for EDU50 and EDU5. Decision on whether different streams may have different physics objects.

- File format for EDU5. Understanding of where ROOT file format/n-tuples begins and D0om file format ends
- Need for a freeware database and ability to extract part of the file/event catalog and suitable calibration, geometry and alignment data for a particular dataset. MSQL has been talked about and tried, but with the advent of Oracle on Linux PCs and the prevalence of Microsoft Access databases on NT machines MSQL may not be the obvious choice.
- Confirmation and testing that the Oracle file and event catalog can perform well with 2 billion events and several million files and that trigger and filter bits can be efficiently stored and selected.
- Further work on exclusive streaming with respect to data delivery and luminosity

References

[1] Drew Baden et. al.,"CDF/CD/D0 Data Management Needs Assessment", June 4, 1997. http://runiicomputing/runiidocs/NAG/cdrun2needs.ps

[2] Wolfgang Kowald, Lee Lueking, Kurt Ruthmansdorfer, Heidi Schellman, Vicky White "Sequential File Access Model Evaluation and Design for Run II", October, 1997. D0 Note xx. http://runiicomputing.fnal.gov/sam/doc/architecture/data_access.ps

[3] Mike Diesburg, Lee Lueking, Kurt Ruthmansdorfer, Heidi Schellman, Vicky White "Flows and Controls in the Seasons of SAM", December 4, 1997.D0 Note xx. http://runiicomputing.fnal.gov/sam/doc/requirements/five_seasons.ps

[4] Jon Bakken et. al. "Requirements for Sequential Access Model", June 1998. D0 Note 3465. http://runiicomputing/sam/doc/requirements/sam_vw_req_arch.ps

[5] Mike Diesburg et. al. "Architecture for Sequential Access Model", June 1998. D0 Note 3464. http://runiicomputing.fnal.gov/sam/doc/architecture/sam_architecture.ps

[6] Jeff Bantly, Amber Boehnlein, John Hobbs " New Phenomena Group Comments on Run II data access",July 10, 1997. http://fnpspa.fnal.gov/personal/cope/data_needs.ps

[7] Heidi Schellman and Iain Bertram "Study of possible Run II Streaming Scenarios using a Run I data sample", September 19, 1997. http://www-d0.fnal.gov/ schellma/upgrade_computing/bits_v3.ps

[8] http://www-d0.fnal.gov/newd0/d0atwork/dam_review.html

[9] Heidi Schellman and Rich Partridge, "Summary of Luminosity Workshop", http://www-d0.fnal.gov/ schellma/lumwork.ps

[10] Heidi Schellman,“Assurance of Data Integrity in Petabyte Data Samples”,December, 1998, Presented at CHEP 98, September, 1998. http://fnalpubs.fnal.gov/archive/1998/conf/Conf-98-335-E.html

[11] http://runiicomputing.fnal.gov/runIIdocs/runIIsteering/hpss/hpss.ps

[12] Don Petravick,“ENSTORE - An Alternate Data Storage System”, Presented at CHEP 98, September 1998.

[13] DØ Run II Computing and Planning Board, “MSS Software for Run II - DØ Evaluation”, June, 1998.

[14] http://runiicomputing.fnal.gov/sam/doc/talks and http://runiicomputing.fnal.gov/sam/doc/design/enstore-functional-components.html

[15] http://runIIcomputing.fnal.gov/sam

[16] http://runIIcomputing.fnal.gov/RIP

[17] http://runiicomputing.fnal.gov/sam/doc/design/RIP-interface.html

[18] http://www-hppc.fnal.gov/farms/r2farms.html

[19] http://www-hppc.fnal.gov/r2prodman/r2bwg/r2bwg.html