Operational Experience with the *BABAR* Database

D. Quarrie[5], *T. Adye*[6], *A. Adesanya*[7], *J-N. Albert*[4], *J. Becla*[7], *D. Brown*[5], *C. Bulfon*[3], *I. Gaponenko*[5], *S. Gowdy*[5], *A. Hanushevsky*[7], *A. Hasan*[7], *Y. Kolomensky*[2], *S. Mukhortov*[1], *S. Patton*[5], *G. Svarovski*[1], *A. Trunov*[7], *G. Zioulas*[7]

[1] Budker Institute of Nuclear Physics, Russia
[2] California Institute of Technology, USA
[3] INFN, Rome, Italy
[4] Lab de l'Accelerateur Lineaire, France
[5] Lawrence Berkeley National Laboratory, USA
[6] Rutherford Appleton Laboratory, UK
[7] Stanford Linear Accelerator Center, USA

Abstract

The *BABAR* experiment at the PEP-II collider at SLAC is using an object database management system for its primary event store as well as for the storage of calibration and alignment information. We report on the experience gained since the experiment detected its first physics event in May 1999. This includes an overview of the software design, the hardware configuration, the data distribution strategy and a high level discussion of aspects that have proven to be important for performance and scaling. More detailed discussions of some of the technical details will be presented in other more focussed papers.

Keywords: *BABAR*, database, ODBMS, operations

1 Introduction

Since the detection of the first physics event in May 1999, the *BABAR* experiment has accumulated more than 30TB of data, mainly from the detector itself, but also as a result of detector simulations. The experiment is expected to accumulate about 300TB of information per year in the steady state. The design of the *BABAR* database has been detailed before, most recently in [1] and several other contributions to this conference describe aspects of the project in detail [2-10]. Here we discuss some operational aspects.

2 Production Servers and Federations

The following production database servers and federations have been setup at SLAC:

- *Developer test.* A dedicated database server with 500GB and dual lock servers are provided for developer test federations. The latter are necessary since we experienced saturation of the Objectivity transaction table with a single server. Any user can have several such test federations, typically corresponding to different *BABAR* software releases. They are allocated a range of 5 federation ids for this purpose.
- *Shared test.* Several developer communities (*e.g.* reconstruction) also have test federations for more integrated testing. These currently share the same servers as the developer test federations, but disk space is a problem here, so separate database servers are being installed.
- *Production releases*. These federations are used by the *BABAR* software release build process for each production release. These share the same machines as the developer test federations.
- *Online (IR2)*. This federation is used by the online system for calibrations, ambient (slow controls) and configuration (trigger settings, high voltage set points, *etc*.) data. The servers are physically located in the experiment control area rather than in the central computing building.
- *Online Prompt Reconstruction (OPR)*. This federation is used for the pseudo-online reconstruction of raw data from the experiment [10]. Initially OPR and the online shared a single federation running under two autonomous partitions, but some intermittent interference due to reading of the run information from the conditions database by OPR with simultaneous update

access from the online have caused us to split these two apart into separate federations until this issue is resolved. Data from the experiment is first spooled onto a disk subsystem in a non-Objectivity format in order to prevent possible problems in OPR and the main computing system from generating deadtime. This information is also written to tape to form an independent backup of the original experiment data.

OPR runs with 100 client machines, shortly to be upgraded to the design value of 200. Two database servers with a total of 2TB of disk will be augmented in Spring 2000 by another server with an additional 1TB of disk. This configuration will support the design 100Hz input rate. Output data is automatically migrated to a HPSS [11] Hierarchical Mass Store (HSM).

- *Physics Analysis*. This is the main federation for physics analysis activities.
- *Reprocessing*. We are about to embark on the first major reprocessing of the experiment data with improved algorithms. A clone of the OPR federation and servers is being established for this purpose using tapes written from the spool disk as input.
- *Simulation Production.* This federation is used for the bulk production of simulated information. A small farm of client machines (~30) is used for this purpose, augmented by a further set of machines at LLNL. Some further simulation production takes place off site at other Institutions and is imported into the Simulation Analysis federation.
- *Simulation Analysis.* This shares the same servers as the physics analysis federation but is a separate federation. This results in the limitation that a single job cannot simultaneously access both physics and simulated information, but increases the effective number of databases that can be supported.
- *Performance Testbed*. In order to understand performance issues, a separate testbed has been established using dedicated servers. This has been used to understanding performance scaling as a function of the number of clients, the number of server machines, filesystems per server and cpus per server, and many other configuration parameters. This work is continuing, but the results achieved so far are discussed in [3].

3 Integration with the Mass Store

The database server machines act as the primary interface to the mass store. In order to accommodate different staging and migration strategies, the file systems are divided into several regions:

- *Staged*. Databases in this region are managed by the staging/migration/purging service, which is configured appropriately for each server.
- *Resident*. Databases in this region are never purged, but can be migrated to tape.
- *Dynamic*. Databases in this region are neither staged nor migrated. Metadata databases such as the federation catalog and management databases that are frequently updated are located here. This prevents multiple instances of these databases being migrated to tape in an uncontrolled manner. Although such multiple migration only results in a single instance of the database in the HPSS namespace, it does consume space on tape. Explicit backups are taken of these databases during scheduled outages.
- *Test*. Databases in this region are not managed by the HSM. Test federations are used to test new applications and database configurations using the production server hardware prior to adoption into production.

The information from an event is spread across 8 databases that are split between the staged and resident areas. Bulk databases such as the raw data and full reconstructed information are staged, whereas more condensed summary information are resident. This subdivision, both into multiple databases, and between the staged and resident areas, is designed to improve performance to the most frequently accessed information.

The analysis federations are configured with two database servers dedicated to staging operations, each having about 1TB of disk space. One server is used for user requests where any user can request that a set of databases be staged from tape. Such staging is not automatic, but is supported by a set of procedures that must be performed prior to execution of users applications. The other server handles *kept* databases which are centrally managed to give access to the data from particular physics runs.

3.1 Movement of data between federations

Several of the production federations (*e.g.* Online, OPR, Physics Analysis) form coupled sets. Data, in the form of databases, must therefore propagate between members of such sets. A data distribution strategy [8] supports this activity, as well as the flow of data between SLAC and several regional centers and other institutions. Internally within SLAC this strategy takes advantage of the HPSS catalog in minimizing the actual copying of databases between federations. For example, once a database generated by OPR has been migrated to HPSS, the catalog for the downstream federation (Physics Analysis) can be updated, without the necessity for physically copying of the database between the appropriate servers. The staging procedures then support transfer of a database from tape to disk.

Such movement between federations relies on the avoidance of clashes in database identifiers for information created at the source and destination federations. The Physics federations subdivide the range into three primary regions, for OPR production, Analysis activities, and reprocessing.

Our original goal had been to make the movement of databases between federations transparent to the user community, and to minimize the latency such that newly acquired information was rapidly made available to the physicists. This has not yet been achieved reliably, and the mean latency is about 2 days between data being processed by OPR and it being made available within the physics analysis federation. Two outages of the physics analysis federations take place per week, causing a total downtime that we have recently reduced to about 10% of the total availability, but which we would like to reduce further. Other management activities take place during these outages, such as updating of the schema, and backing up of the dynamic databases. We are also still working on procedures to reduce the latency, although this will become less important as the event samples increase with more integrated luminosity.

4 Physicist access to information

Physicists access their desired event samples using event collections. Initially the physics analysis federation itself was used as the catalog of which collections were available. However, we underestimated how many such collections would be created and the scanning and manipulation of this catalog became a bottleneck that impacted other analysis activities. There are currently approximately 30,000 event collections, including individual node collections from OPR, summary collections, the results of selecting events by the various physics analysis groups, *etc*. We have therefore created an independent production catalog, based on Oracle, that can be queried to determine which event collections exist, and whether the component databases exist on tape and/or disk. The staging procedures use the information in this catalog to access the required databases, depending on the event collections and the information within the events that are requested.

5 Production Schema Management

It is crucial that new software releases are compatible with the information in the production federations. A *reference* schema version is therefore used to preload the release federation (as described previously) that is used for a software release. If the release build is successful on all

supported hardware platforms, and some QA tests are passed, the output schema of the release are adopted as the new reference schema. Since online and offline software release builds can potentially overlap a token-passing management scheme ensures a strictly sequential updating of this reference. The schema for the production federations are updated from the reference during scheduled outages.

6 Support Personnel

Two administrators support the daily database operations and develop management scripts and procedures. They act as the first level of user support, and also perform all the data distribution internal to SLAC. They are augmented by 2 people who support data distribution to and from external sites. Two people support HPSS operations, including software development activities. Five database developers act as the second tier for problems as well as ongoing development.

7 Acknowledgments

Members of the *BABAR* Computing Group have made significant contributions to the work reported here, in particular in the areas of detector-specific event store and conditions information. We also acknowledge the considerable support provided by members of SLAC Computing Services, particularly in providing and managing the computing hardware infrastructure that supports the *BABAR* datastore.

TABLE 1. *BABAR* Database Statistics (Jan 2000)

Number of licensed users	655	Number of test users	429
Peak simultaneous users	87	Number of persistent classes	513
Disk space	~10TB	Accumulated data	~33TB
Accumulated databases (OPR)	~14000	Accumulated collections	~28000
Primary Database Servers	12	Secondary Servers (catalog, journal, lock)	15

References

1 The RD45 Collaboration, “RD45 Status Report”, CERN/LHCC 99-28, Sep 1999

2 A. Adesanya, “An interactive browser for *BABAR* databases”, CHEP 2000, Padova, Spring 2000.

3 J. Becla, “Improving Performance of Object Oriented Databases, *BABAR* Case Studies”, CHEP 2000, Padova, Spring 2000.

4 I. Gaponenko, *et al.*, “An Overview of the *BABAR* Conditions Database”, CHEP 2000, Padova, Spring 2000.

5 T. Glanzman, *et al.*, “The *BABAR* Prompt Reconstruction System, or Getting the Results out Fast: an evaluation of nine months experience operating a near real-time bulk data production system”, CHEP 2000, Padova, Spring 2000.

6 A. Hanushevsky, “Practical Security In Large-Scale Distributed Object Oriented Databases”, CHEP 2000, Padova, Spring 2000.

7 A. Hanushevsky, “Disk Cache Management In Large-Scale Object Oriented Databases”, CHEP 2000, Padova, Spring 2000.

8 E. Leonardi, *et al.*, “Distributing Data around the *BABAR* collaboration’s Objectivity Federations”, CHEP 2000, Padova, Spring 2000.

9 S. Patton, *et al.*, “Schema migration for *BABAR* Objectivity Federations”, CHEP 2000, Padova, Spring 2000.

10 G. Zioulas *et al.*, “The *BABAR* Online Databases”, CHEP 2000, Padova, Spring 2000.

11 HPSS: http://www.sdsc.edu/hpss/