SLAC-PUB-9182
August 2002

# Managing the BaBar Object Oriented Database

Adil Hasan, Artem Trunov
(Stanford Linear Accelerator Center, Menlo Park, Ca, USA)

**Abstract**

The BaBar experiment stores its data in an Object Oriented federated database supplied by Objectivity/DB(tm). This database is currently 350TB in size and is expected to increase considerably as the experiment matures. Management of this database requires careful planning and specialized tools in order to make the data available to physicists in an efficient and timely manner. We discuss the operational issues and management tools that were developed during the previous run to deal with this vast quantity of data at SLAC.

Keywords: BaBar, Database, Management

In BaBar the data is stored in Objectivity/DB (tm) [1] federated databases or federations. There are three types of federations used to store the different stages of data processing: online, for online data acquisition information, reconstruction, for the output of prompt reconstruction and analysis for data analysis. The three types allow for a seperation of federations where the data is predominantly read-only (such as the analysis federation) from those where the data is mainly write-only (such as the online and reconstruction federations). Each of these three types owns a disjoint sub-set of the 64k database id range.

The online federation serves the online data acquisition and stores the detector condition, configuration and ambient information. Access to the federation is restricted to sub-system experts and is controlled by a BaBar Database Authorization System (BDAS). The total size of the online federation is ∼100GB and all the data is disk-resident on one 4 cpu server. The data are regularly backed up to tape stored in the High Performance Storage System (HPSS) - the tertiary storage system used at SLAC.

The reconstruction federation stores the output of the prompt reconstruction. This includes: raw-, reco-, mini-, micro- and nano-level information. The federation is spread over six 4 cpu servers each with 1TB of disk cache. In addition two separate 1 cpu machines act as the lock and journal server and single 2 cpu server holds the federation catalog and metadata. This configuration has allowed prompt reconstruction to stay in step with the data acquisition. Like the online, the prompt reconstruction federation is a restricted access federation. Another reconstruction federation (the reprocessing federation) possessing a disjoint sub-set of database ids and sitting on a separate set of servers holds the reprocessed data. Each week ∼2TB of data is produced separately by prompt reconstruction and reprocessing.

The third type is the analysis federation. This federation holds a copy of the prompt reconstruction and reprocessing federation, where all the micro- and nano-level data are currently disk resident occupying a total of ∼15TB of disk space and using a total of 14 data and 1 lock and 1 journal servers. General access is allowed to this federation, but the data produced by prompt reconstruction and reprocessing is kept as read-only via the BDAS. Users are allowed to write to the database id range owned by the analysis federation. The reco- and raw-level data are mainly kept on tape, users are allowed to submit stage-in requests to have the data of interest staged onto disk from HPSS. The staging area for reco- and raw-level data occupies a total of 6.2TB of disk space.

For the simulation produced at SLAC only a generation and analysis federation exist, again each owning disjoint database id ranges. The generation federation consists of simulation-level as well as raw-, reco-, mini-, micro- and nano-level data. The generation federation, like the reconstruction federation is restricted access and is spread over three 4 cpu data servers each with 1TB of disk cache. A lock and journal server as well as a 2 cpu server holding the

Work supported by the Department of Energy contract DE-AC03-76SF00515.

Presented at the International Conference on Computing in High Energy and Nuclear Physics (CHEP '01), Beijing, China, 9/3/2001 - 9/7/2001

federation catalog and metadata complete the configuration. Simulation databases produced by off-site production are also registered in the SLAC generation federation as this is considered the master simulation federation. The analysis federation is general access and consists of a copy of the generation federation and is kept read-only via the BDAS. The micro- and nano-level data are kept disk-resident on the same data servers as used by the data analysis federation. The analysis and data analysis federations can support more than a total of 130 micro-level analysis jobs at the same time.

The reconstruction and analysis federations require detector condition and configuration information that is stored in the online federation. The condition and configuration databases from the online federation are automatically import into the prompt reconstruction federation when needed. The condition and configuration databases are copied from the reconstruction to analysis federation together with the reconstructed event data. The total size of the condition and configuration data is ∼16GB.

As previously mentioned, the reconstructed data is copied from the reconstruction to the analysis federation. The new databases are automatically registered in the analysis federation daily. Each week prompt-reconstruction stops processing data and the micro- and nano-level databases are physically copied from the reconstruction to the analysis federation. Unlike the condition and configuration databases, not all the micro- and nano-level data are on disk at the time of the outage. Some of the data has to be staged from HPSS onto disk and some copied directly from the reconstruction federation data servers. The micro- and nano-level databases are copied alongside previous versions in the analysis federation, typically ∼200GB of databases are copied. The whole process of copying the data requires an outage of ∼3-4 hours for prompt reconstruction. After the databases have been copied alongside older versions in the analysis federation, a shorter outage is required of the analysis federation in order to update existing copies of reconstruction micro-, nano-level, condition and configuration databases. The analysis federation outage is achieved by the presence of special semaphore files. User analysis applications use BaBar database and Objectivity code that recognizes the presence of these files causing the application to not to start a new transaction until these files have been removed, this minimizes the impact of the outage on user's applications.

The same procedure is used to copy simulation databases from the generation to the analysis federation. New simulation data produced by external sites (currently comprising ∼60% of the total simulation) is regularly transferred over the network to SLAC. This data is copied to a temporary location on a 4cpu import data server with 2.6TB of disk cache. This server also houses an import federation that acts primarily as a buffer between the external sites and the generation federation preventing corrupt databases from appearing in the generation or analysis federation. Automatic applications detect new imports by scanning the import directories for the presence of semaphore files denoting a complete transfer. These imports are registered in the import federation which, in turn, registers them in the generation federation if the the import is successful. Once a database is successfully registered in the import federation it is migrate to HPSS. Typically ∼ 300GB of simulation data is transferred to the import server every 2-3 days.

Bookkeeping is maintained by a set of scripts run automatically that load Oracle tables with information necessary to keep track of the imports, condition and configuration updates and data copied from reconstruction to analysis federation. A set of scripts run automatically are constantly monitoring the servers and the federations for dead objectivity servers or dead transactions. These monitoring tools use standard Objectivity/DB (tm) utilities in conjunction with standard UNIX tools and have greatly reduced the unscheduled outage time for the production federations. Many of the monitoring and data copying tools are written in Perl.

Based on current experience we have many plans and ideas that should allow us to take

a step closer to fully automating the process of moving data from reconstruction to analysis federations as well as reducing outage time (see [2]).

## References


[1] http://www.objectivity.com

[2] I. Gaponenko et al. "The BaBar Database: Challenges, Trends and Projections", CHEP 2001, Bejing.