SLAC-PUB-9181
September 2001

The BaBar Experiment's Distributed Computing Model

Dominique Boutigny LAPP-CNRS/IN2P3

on behalf of the
BABAR Collaboration's Computing Group

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

Stanford Linear Accelerator Center, Stanford University, Stanford, CA 94309

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

The BaBar experiment's distributed computing model

Dominique Boutigny LAPP-CNRS/IN2P3 on behalf of the BaBar Collaboration's Computing Group

Abstract

In order to face the expected increase in statistics between now and 2005, the BaBar experiment at SLAC is evolving its computing model toward a distributed multi-tier system. It is foreseen that data will be spread among Tier-A centers and deleted from the SLAC center. A uniform computing environment is being deployed in the centers, the network bandwidth is continuously increased and data distribution tools has been designed in order to reach a transfer rate of ~100 TB of data per year. In parallel, smaller Tier-B and C sites receive subsets of data, presently in Kanga-ROOT[1] format and later in Objectivity[2] format. GRID tools will be used for remote job submission.

1 Introduction

During its first run, the SLAC based BaBar experiment[3] has accumulated more than 22.10^6 Υ_{4S} disintegrations. The expected luminosity in the coming years will increase this number by a factor ~23 in 2005. This important step in statistics will have a very high cost in terms of computing.

The BaBar experiment recently decided to evolve toward a multi-tiers, distributed computing environment, where it will make use of the computing power and of the storage capacity available in national centers.

For its main storage, BaBar has adopted the Objectivity OO database technology. The events are stored in a hierarchical structure from the most detailed information (RAW and REC) up to the highest abstraction level suitable for physics filtering and analysis (TAG and MICRO). Recently, an intermediate MINI level has been added allowing to run detector studies or detailed analyses on a relatively compact format. Navigation structures allow to access any type of information from any level of the event store. The Objectivity database system is interfaced to the HPSS hierarchical storage system.

2 The multi-tiers model

In the BaBar terminology the Tier A sites together have a copy of all the data, some overlap may exist between the data-sets, especially for those accessed by many people or by the most urgent physics analyses. A Tier A should be able to host at least 30% of the total data sample. All the data format are available on disk or on mass storage, it should be possible to run high level physics analyses, as well as specific detector studies requiring RAW data.

Tier B serve a region and act as secondary data distribution centers. They host data in a format suitable for most of the analyses (Micro DST).

Tier C are typically individual institutes having access to a small sample of the data corresponding to the direct physics interest of the sites.

3 Data format

The Objectivity event store allows to easily create collections of filtered events (skims) with a little overhead as only the navigational information is stored. The collection of events has pointers to the actual event components. This is a very powerful tool for analysis, in sites hosting a large fraction of the total event store. Unfortunately, the exportation of such skimmed collections to smaller remote sites becomes difficult has they can potentially refer to a lot of databases. Moreover, due to the Objectivity limitation of 64K databases per federation, the size of individual databases should be large (2 to 10 GB) in order to keep a reasonable lifetime for a federation. The direct consequence is that the exportation of a single collection can result in the copy of hundreds of GB.

To solve this problem a new format (KanGA) has been introduced to store self-contained events in small ROOT files. This format allows only Micro DST level analyses and is suitable for exportation in remote sites. The skim collections contain a copy of real events.
A second approach is being developed now, consisting in supporting multiple Objectivity federations linked together with a single bridge federation[4]. This removes the 64K databases limitation and allows to decrease the size of individual databases. At the same time, the BaBar prompt reconstruction system has been improved in order to produce 20 self-contained physics streams constituting the units for data distribution in remote sites. It is foreseen to withdraw the KanGA/ ROOT format as soon the new system has been demonstrated to be suitable for remote analyses.

4 Data distribution

The distribution of KanGA/ROOT files is described in[5]. This section will only describe the Objectivity data distribution.
The following table shows the expected amount of real data (in TB) to be transferred to Tier A sites as a function of time. The last line of the table is the corresponding necessary network bandwidth. It is foreseen that transferring the Monte-Carlo sample will require to multiply these numbers by a factor of 2 to 3.

	2001	2002	2003	2004	2005
Tag	1,5	1,1	2,0	3,0	4,7
Micro	14,6	10,9	13,7	20,7	32,6
Mini	19,5	18,1	19,5	29,6	46,5
Reco	21,9	54,4	97,5	148,1	232,5
Raw	8,8	16,3	29,3	44,4	69,8
Sum	**66 TB**	**101 TB**	**162 TB**	**246 TB**	**386 TB**
Bandwidth (Mb / sec)	**54**	**82**	**131**	**199**	**312**

The data distribution system is using the BaBar/Objectivity C++ API to interact with the event store, and a collection of shell, tcl and perl scripts to deal with file manipulations and book-keeping. The sequence of operations is the following:

1) Dump the Objectivity catalog
2) Compare each entry in the catalog with a reference file and build a transfer description file (tdf) with those satisfying some selection criteria and corresponding to individual databases which have not been exported yet.
3) Extract the databases from the Objectivity servers or from HPSS and copy them on disk.
4) Update the reference file

This sequence is repeated for each active federation and for each site which subscribed to the distribution system (only one at the moment). Each site can have a different set of database selection criteria. For instance, it is possible for a site to receive only a limited number of physics streams.
The remote sites detect that a new export is ready by checking on a regular basis a list of completed database extractions. Then the copy over the Wide Area Network is done using an efficient multi-stream TCP based transfer tool: bbftp[6], databases are directly distributed among the various Objectivity servers in order to balance the load and are attached to the Objectivity federations.
The control of the data distribution servers at SLAC and in the remote sites is performed either by simple E-mails containing commands to be executed (Start a new database extraction, clean the disk etc…) or in automatic mode by cron jobs exchanging control files.
In order to speed up the Tier A ramp up, two data distribution systems are running in parallel on two dedicated machines. One is exporting the Micro DST event format and another one is exporting the detailed RAW, REC and Mini format.

4.1. Performance of the data distribution system

At the moment, Micro DST exports are run once a week just after the analysis event store has been updated with the latest reconstructed events. The size of an export at the Micro DST level varies from 200 up to 900 GB which are extracted from the event store and copied to disk at an average speed of 5 MB/s. During the 2001 data taking period 6.1 TB of data have been extracted, the overhead due to updated databases that should be re-exported represented 25%.
The detailed data (RAW etc…) are extracted almost continuously directly from HPSS by 200 GB chunks, each chunk is transferred to remote sites asynchronously, 3.4 TB have already been extracted with no overhead.

4.2. Network performance

Most of the BaBar data distribution traffic is going from SLAC to STARtap and then to IN2P3 through 2 different links. The detailed (RAW, REC and Mini) data are going through CERN with a 155 Mbit/s bandwidth and the Micro DST data are routed directly to IN2P3 through RENATER with a 34 Mbit/s link. With the current tools, the transfer speed via CERN is 70 Mbit/s in average with peaks at 90 Mbit/s, the speed through RENATER link is of the order of 30 Mbit/s. The total bandwidth used in the SLAC – ESNET portion is then 95-100 Mbit/s when both routes are used at the same time (see figure 1).

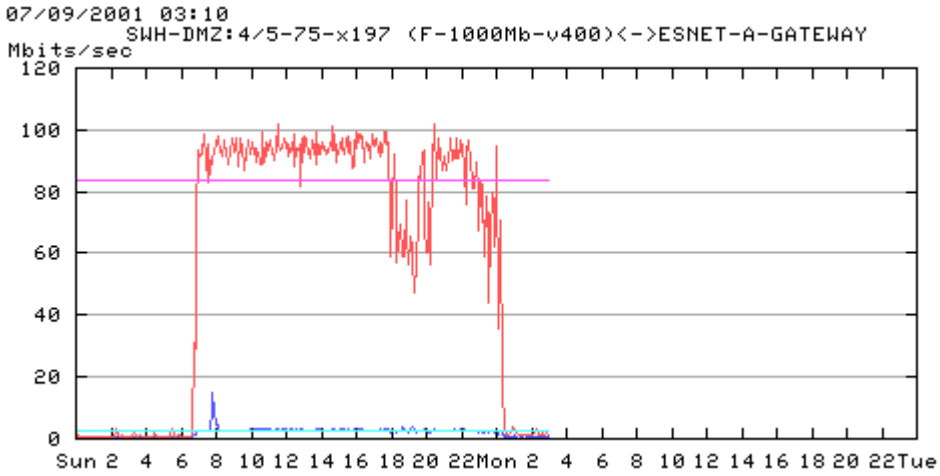

Figure 1: Bandwidth used between SLAC and ESNET as a function of time while transferring BaBar data to IN2P3

The reliability of the network as well as of the bbftp software with its automatic recovery feature has proven to be excellent. Improving the amount of bandwidth used during the transfer will require some experimentation and tuning.

5 The CCIN2P3 Tier A

Since January 2001 the IN2P3 computing center (CCIN2P3) is ramping up as a Tier A for the BaBar collaboration. Accounts are created on demand for every BaBar members willing to run some analysis or detector studies at the Tier A.
One of the major differences between the SLAC center and IN2P3 is related to the heavy use of HPSS at two different levels:

- HPSS provides a virtually unlimited storage for individual users, willing to store large HBOOK or ROOT files. Daily usage by BaBar and by other experiments, have proven this technique to be very reliable and efficient to analyze hundreds of GB of n-tuple like data.
- HPSS is also used to store Objectivity data, the automatic purging system has been activated in such a way that files are staged out when the disk space gets too small. The data are automatically recalled from tape when the Objectivity Advanced Multi-Threaded Server (AMS) detects that a database is missing from disk. This option is working well at IN2P3 where the number of users is relatively small; we should make sure to put the necessary protections to avoid to flush the HPSS cache in case of too many staging requests coming at the same time.

Full analyses requiring access to the 9 TB of 2000 data have been run on a regular basis at IN2P3, the CPU usage by local and non-local CCIN2P3 users is of the same order.
3.4 TB of RAW, REC and Mini data are already available for detailed studies.

6 Conclusion and prospects

The BaBar experiment is evolving toward a distributed multi-tier computing model. The SLAC center is the primary Tier A site, a second Tier A at IN2P3 is ramping up and is already providing a significant analysis resource to the whole collaboration. Data distribution tools have been setup to transfer the most recent data in a timely manner. The next step in the Tier A implementation is to stop the duplication between the 2 sites and to go to a model where part of the data sample is available in one site only.
Networks are critical for this model, we are now experimenting 155 Mbit/s links, faster links will be necessary in the future, but using efficiently OC12 links will require developments both in the data transfer tools and in the Objectivity data extraction software.
It is recognized that it is not practical to have to physically connect to different Tier A site in order to run an analysis on the whole data sample. An important effort is starting to develop remote job submission tools using the GRID technology.

References

[1] R. Brun and F. Rademakers, ROOT home page, http://root.cern.ch/
[2] Objectivity home page, http://www.objectivity.com
[3] BaBar Collaboration, B. Aubert *et al.* SLAC-PUB-8569, to appear in Nucl. Instr. And Methods
[4] I. Gaponenko *et al.* “The BaBar Database: Challenges, Trends and Projections” Abstract 4-013 at this conference
[5] T. Adye *et al.* “Distributing file-based data to remote sites within the BaBar Collaboration” Poster session at this conference – Abstract 4-023.
[6] G. Farrache, bbftp : see for instance: http://ccpntc3.in2p3.fr/bbftp/index.html