



Distributed and/or Grid-Oriented Approach to BTeV Data Analysis

Joel N. Butler

Fermi National Accelerator Laboratory, Batavia, Illinois 60510-0500 USA

For the BTeV Collaboration

Abstract

The BTeV collaboration will record approximately 2 petabytes of raw data per year. It plans to analyze this data using the distributed resources of the collaboration as well as dedicated resources, primarily residing in the very large BTeV trigger farm, and resources accessible through the developing world-wide data grid. The data analysis system is being designed from the very start with this approach in mind. In particular, we plan a fully disk-based data storage system with multiple copies of the data distributed across the collaboration to provide redundancy and to optimize access. We will also position ourself to take maximum advantage of shared systems, as well as dedicated systems, at our collaborating institutions.

Introduction

BTeV [1] is an experiment to study CP violation and rare decays of particles containing b-quarks at the Fermilab Tevatron collider. The trigger is based on selecting events with evidence for detached vertices at the lowest level so that a very large fraction of all potentially interesting decays can be recorded without making rigid requirements on the exact nature of the final state. With this approach, over 1000 b-quark decays per second are recorded along with about 1000 directly produced charm decays, 1000 background events, and about 1000 calibration and alignment events. This results in a raw data set of about 1 petabyte per year. The data profile of BTeV is shown in Table 1. BTeV hopes to begin taking data in 2007.

Datasets of this size have typically been stored and processed at large centralized facilities, usually at the laboratories where the data was taken. After several stages of processing, small samples of highly selected events may be shipped off-site for the final stages of “physics analysis” on university facilities.

In recent years, technological advances have rendered this centralized model obsolete. CPU costs and more recently disk costs have fallen. University groups can now afford computing and data storage/access systems which can contribute meaningfully to the production processing of raw datasets. University departments are pushing for and receiving funding to provide excellent local resources to their researchers. This trend is driven by the development of computing codes in many areas of research which provide excellent simulations and computations that can replace many hours of trial and error in the laboratory. In fact, the computer is now becoming a virtual lab in which experiments can be simulated with great fidelity and unprecedented speed, convenience, and flexibility. Finally, wide area networking is improving very rapidly and it is now becoming possible to move large datasets across the network very quickly and reliably.

At the same time, national laboratories are feeling pressure on their budgets and are using an increasing share to support their unique facilities and experiments. They are finding it impossible to satisfy the ever-expanding needs of their experiments for computing resources and are more than happy to off-load parts of the task and cost to university groups. It will, therefore, be necessary for future experiments with large CPU and data storage/access needs to exploit whatever resources they have access to no matter where they are located.

Since BTeV is just beginning to develop its offline production system, it can design it to take advantage of these new trends and their extensions into the future. Below, we describe two possible approaches. The first, which we refer to as a “distributed hierarchical approach” is based on a system of dedicated “centers”. The second is an even less centralized model based on the emerging “computational datagrid”. In either case, BTeV data analysis will be highly distributed. We also describe an approach which starts with the first approach and gradually transforms itself into the second approach.

Data source	Event Size	Total Dataset Size
Based on 4×10^{10} events/year		
Raw data	50 kBytes	2 PBytes
Reconstructed Data	~ 50 kBytes	2 PBytes
Based on 1000 datasets of 10^7 events/each/year		
Summary Physics Objects	~ 10 kBytes	0.1 Pbytes
Condensed Summary physics data	2-5 KBytes	0.02-0.05 PBytes
Data Catalog Entry	~ 200 Bytes	2 TBytes

Table 1: BTeV data profile for one year

1 Approaches to Distributed Computing

1.1 A Hierarchical Distributed Model

In this model, computing resources are organized into a hierarchy of “centers” with different sizes and capabilities [2]. At the top of this hierarchy sits a large center which has the capability to do all analysis related functions but not the capacity. We refer to this as the “Tier0 center.” Below this sit a collection of large multi-service centers, each with a significant fraction, say 10% to 20%, of the Tier0 center. These are called “Tier1 Centers.” They are often seen as serving a geographical region and may also be called “Tier1 Regional Centers.” Below the Tier1 centers, there can be “Tier2 centers”, which provide services to a subregion and receive services from the Tier1 centers. Tier2 centers would typically be located at larger universities. This model can be continued, with “Tier3” representing typical university work groups and “Tier4” being the single user desktop. There may also be “special purpose” centers which carry out very specific functions, such as simulation only.

There are several important points to note about this approach. First, it will be easiest to implement if the resources at each Tier are dedicated to the specific experiment. That permits the experiment to set policy in an optimal way and removes the need for complicated software to set and enforce priorities among groups competing for resources. Second, it is clear that software and support are key issues for the implementation of such a strategy. Software will be necessary to locate, maintain, and optimally place the various datasets. Secure means of moving data around are required. It is essential that each center provide a critical mass of user support appropriate to its mission in the system. Third, the various centers must function over the lifetime of the data analysis, which could be a long period of time. This implies that each site must provide support for the full duration and for continual hardware evolution and even continual R&D. There will always be the risk of sites disappearing.

Examples of the kinds of software that are needed to implement this approach are management of large scale clusters of computers including features which provide reliability and fault tolerance; software installation and maintenance for systems of thousands

of computers; techniques for performance measurement and tuning; methods for dealing with problems of access control and distributed authority (security and policy enforcement); maintenance and tuning of interconnecting networks and hardware maintenance of computers and storage systems; and efficient protocols for migrating, replicating, and protecting data and metadata.

While these are difficult issues, the system has, at any time, a finite number of sites, with a well-defined architecture, a single mission, and clear lines of authority and policy control.

1.2 The Computational Datagrid

The idea of a computational datagrid [3] is to be able to use, in a transparent manner, distributed computing resources to solve a problem as if you were using a single workstation or PC. You are essentially able to assemble a virtual, distributed computing facility for each problem based on whatever resources happen to be available to you anywhere at the time. This, in turn, requires sophisticated software to evaluate the requirements of a computation, assemble the resources by exploring the network and discovering what is available, gain access to the resources, split the computation up among the computers, establish the appropriate environment, start the programs up, monitor and control their progress, recover the results, and maintain complete records of the task.

There are many issues that have to be addressed. Security, resource discovery, distributed ownership and local policy, preemption, system heterogeneity, validation, fault recovery, and many more. While many of these issues are addressed at some level in the implementation of the hierarchical model, the complexity of operating in such a distributed and heterogeneous environment and the degree of transparency needed to do this make this a much greater challenge. A great deal of money is being invested in doing R&D and eventually developing “grid-software” targeted at solving these problems.

It is clear that “grid-software”, if and when it exists, will be able to provide the software required to implement the hierarchical model. In fact, the earliest products of the grid software efforts address many of the issues required to support the hierarchical model. BTeV’s approach, which will be described below, will position itself to use emergent grid software initially to implement its hierarchical model. It will then expand its use of the full data-grid model as the software develops and as the needs of BTeV expand.

2 The BTeV Approach

In deciding on an approach to offline analysis, BTeV has to consider the following characteristics of its environment, which are already set:

- the existence of a very large Level2/3 trigger farm, consisting of over 2000 high speed processors, with significant excess capability. The trigger farm is only busy during accelerator operation and even during operation has “peaking capacity” well

above the average required capacity. It is reasonable to assume that about 2/3 of the capacity will be potentially be available for offline analysis.

- the existence and likely expansion of significant computing clusters at collaborating institutions. These clusters are in some cases dedicated to BTeV. However, several universities with BTeV collaborators are planning interdisciplinary clusters to be shared among several research groups in several fields. It is imperative that BTeV be able to take maximum advantage of this arrangement.
- the concept of the Level 4 (Level N) trigger. BTeV plans to perform additional “pruning” and “summarizing” of its dataset well after the fact of raw data recording. This would include summarizing various low physics interest samples, applying final calibrations and fixups to the main physics event samples and then deleting some of the raw data. These steps may occur days, weeks, or even months after the data are taken. With this approach, the dataset will continue to remain manageable even as more data is acquired every year. It will be possible to “go back” and look at data that is several years old because it will have been efficiently reduced in size.

One major obstacle in using trigger clusters for offline analysis has been the problem of getting the data back into the system from data tapes. This requires either large numbers of local tape drives and operator support or costly, expensive, and limited robotic mass storage systems. We propose to eliminate this issue by providing a multi-petabyte disk system as the main archival storage and data access system for BTeV. This system would be connected to the Level 2/3 trigger farm and would also be connected to the network for remote access. We expect to protect the data from destruction due to hardware failure or accidental deletion by having at least two more complete copies distributed throughout the collaboration on disk. We do not exclude the possibility of writing the raw data to tape but only for purposes of backup. All access to the data will be from disk.

Thus, the BTeV analysis architecture, viewed as a hierarchical system, will include a Tier0/1 level consisting of the Level 2/3 trigger farm. The next level of the hierarchy would be a series of Tier 1/2 centers – one at Fermilab and several at universities. These would be capable of performing event reconstruction, data selection (streaming, splitting, and stripping), simulation, and analysis. We envision between 5 and 10 such centers. We will require each center to have enough disk storage to keep a partial copy of the primary dataset to provide two complete copies of all essential data. Figure 1 shows a possible BTeV offline system architecture.

The use of an all-disk data storage/access system is key element of the strategy. Not only does it help BTeV utilize its trigger farm as an offline resource, but it eliminates a large, centrally located robotic tape-based mass storage system. The necessity to have such a system at the central site means that people who work on central machines have much better access to the data than people who want to use remote systems. This has tended to produce a concentration of analysis activities at the central site and has relegated the use of remote machines to only the final stages of the data analysis. Given the

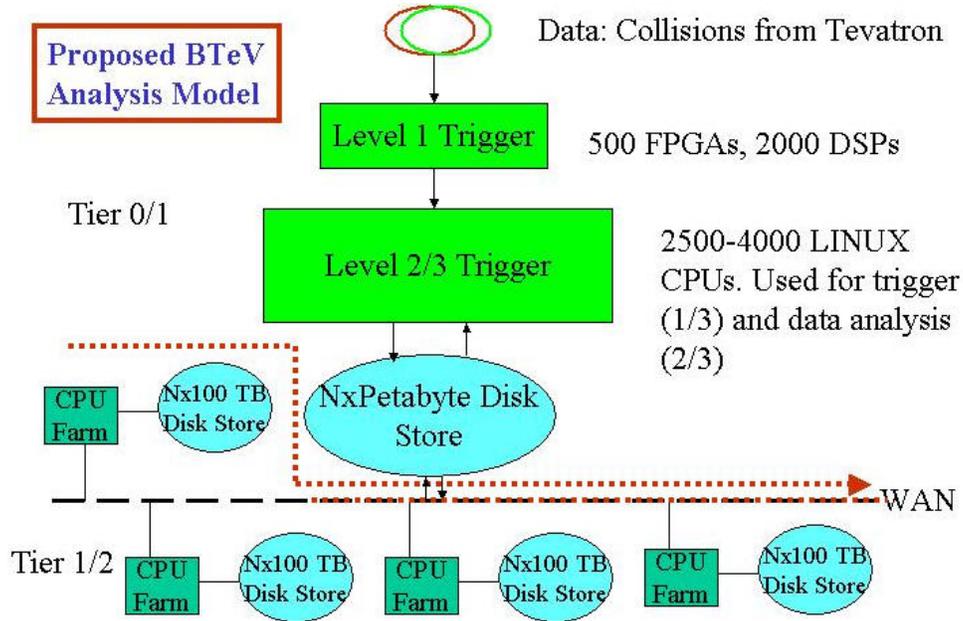


Figure 1: BTeV grid showing the Level 2/3 Trigger Farm, a dedicated Center at Fermilab, the large multi-petabyte disk store, and several centers at universities

availability of high speed networks and inexpensive disk, this approach is not necessary and should probably be avoided.

Drilling down further at the university level, we expect the university system to have (possibly) a dedicated cluster for BTeV and/or a shared “research cluster”. This means that BTeV will need to use resources that are not totally under its own control, even in the hierarchical model. There will be numerous policy issues, such as the application of priorities, fair share considerations, security, etc. As part of this organization, non-BTeV users will also have access to BTeV facilities. Figure 2 shows what a university cluster would look like.

Other resources, not shown but implied, include large facilities, such as supercomputing centers whose mission is to host user applications, and general sharing with other clusters on the grid.

Unique features of this approach are

- The blurring of the distinction between the trigger system and the analysis system. The former maybe used for analysis or simulation whenever there are available cycles, even during actual operation of the system as a trigger during data taking. The “Level 4 trigger”, and even conceivably parts of the Level 3 trigger, may well be done by resources usually called “offline” and may even be carried out on remote resources.
- The elimination of tape as a data access medium and total reliance on disk

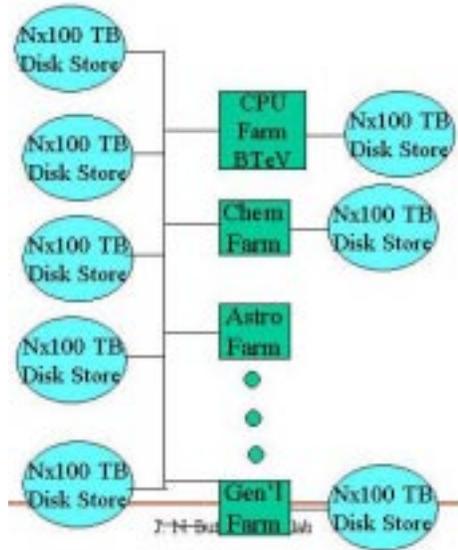


Figure 2: A typical BTeV university computing cluster showing both dedicated and shared resources

Obviously, significant software development is necessary to support this model. Some of this software will be written by the grid developers. However, additional software will need to be developed in several areas. These include policy modules for managing shared resources, preemption and priority schemes, data recovery and migration strategies, security (protection against accidental and malicious deletion of data), user interface, etc. As we develop our analysis framework, we are keeping the eventual implementation of this model in mind and providing in advance the necessary hooks to take maximum advantage of all the resources we have available to us.

Acknowledgement

This work was supported in part by Fermilab, which is operated by Universities Research Association Inc. under Contract No. DE-AC02-76CH03000 with the United States Department of Energy.

References

- [1] see <http://www-btev.fnal.gov/public/hep/general/proposla/index.shtml>, “BTeV Proposal” and “BTeV Proposal Update”
- [2] see, for example, the MONARC Project at CERN, http://monarc.web.cern.ch/MONARC/docs/monarc_docs.html, “Regional Centers for LHC Computing”

- [3] *The Grid: A New Infrastructure for 21st Century Science*, Ian Foster, Physics Today, February 2002