

Proposed Research and Development Motivated by the BTeV Trigger System

We plan to research and develop methodologies and tools for designing and implementing very large-scale real-time embedded computer systems that

- achieve ultra high computational performance through use of parallel hardware architectures;
- achieve and maintain functional integrity via distributed, hierarchical monitoring and control;
- are required to be highly available; and
- are dynamically reconfigurable, maintainable, and evolvable.

The specific application that will drive this research and provide a test platform for it is the trigger and data acquisition system for BTeV¹, an accelerator-based High Energy Physics (HEP) experiment to study matter-antimatter asymmetries (also known as Charge-Parity violation) in the decays of particles containing the bottom quark. BTeV was recently approved by Fermilab² and will be constructed over the next 5-6 years to run in conjunction with the Fermilab Tevatron Collider. The experiment is expected to run for at least 5 years. It requires a massively parallel, heterogeneous cluster of computing elements to reconstruct 15 million particle interactions (events) per second and uses the reconstructions to decide which events to retain for further data analysis. **Creating usable software for this type of real-time embedded system will require research into solutions of general problems in the fields of computer science and engineering. We plan to approach these problems in a way that is general, and to produce methodologies and tools that can be applied to many scientific and commercial problems.**

The classes of systems targeted by this research include those embedded in environments, like BTeV, that produce very large streams of data which must be processed in real-time using data dependent computation strategies. Such systems are inextricably tied to the environment in which they must operate, and must perform complex computations within the timing constraints mandated by their environments. These systems require **ultra high performance** (on the order of 10^{12} operations per second). The level of performance requires **parallel hardware architectures**, which in the case of BTeV is composed of a mix of thousands of commodity processors, special purpose processors such as Digital Signal Processors (DSPs), and specialized hardware such as Field Programmable Gate Arrays (FPGAs), all connected by very high-speed networks. The systems must be **dynamically reconfigurable**, to allow a maximum amount of performance to be delivered from the available and potentially changing resources. The systems must be **highly available**, since the environments produce the data streams continuously over a long period of time, and interesting phenomena important to the analysis being done are rare and could occur in the data at any time. To achieve the high availability, the systems must be **fault tolerant, self-aware, and fault adaptive**, since any malfunction of processing elements, the interconnection switches, or the front-end sensors (which provide the input stream) can result in unrecoverable loss of data. Faults must be corrected in the shortest possible time, and corrected **semi-autonomously** (i.e. with as little human intervention as possible). Hence **distributed** and **hierarchical monitoring and control** are vital.

We believe that there are very significant advantages to connecting this research to the BTeV experiment. Not only will the software and methods produced by this research have significant impact on one of the most important areas of investigation in HEP, but the computer engineering research will also be directly applicable to a large class of similar real-time embedded computer systems. The BTeV trigger system hardware, which will be provided by Fermilab as part of the experiment, will supply an extremely important ingredient in this project: a large test-bed that represents millions of dollars of equipment and comes with a highly motivated set of users who will test the methodologies and tools developed in an extremely harsh environment over an extended period of time. The test-bed will be built gradually as the proposed research progresses, from a 5% system in 2002 to a full system in 2005-2006. It will therefore be possible for the software developers, aided and supported by the experimenters, to test and refine the software and strategies continuously and incrementally throughout the lifetime of this project. The close interdisciplinary contact between the experimenters and computer scientists will also help introduce important computer science research into the HEP community, which has not always been aware of work that has been done in this area and has not taken full advantage of it.

The team that has been assembled to carry out this research consists of the leaders of the BTeV trigger and data acquisition system development efforts and Computer Scientists who are experts in the field of embedded systems, real-time systems, and fault tolerant computing³. The Computer Scientists come from universities that are strongly involved in BTeV, and from Fermilab. The team is committed to carrying out the proposed R&D and implementing a series of systems of increasing size and complexity, using the experience gained at each stage to refine and improve the system until it is demonstrated to scale to the full BTeV system.

1. The BTeV Trigger System

This section has two parts. The first part describes the BTeV triggering and data acquisition system, in order to explain the problem that must be solved and the basic architecture and scale of the system that is planned to address it. The “trigger” or “event filter” algorithms that run on the resulting hardware platform are briefly described. These algorithms, which will largely be written by physicists from the BTeV experiment, are not, however, the thrust of the research proposed here. The second part of this section addresses the requirements of the infrastructure required to keep the trigger system operating, to assure that it is working correctly, and to detect and adapt to fault conditions both within the computing platform and within the experiment and machine environment. This has been called out⁴ as the major challenge in implementing the BTeV trigger:

“Regarding the robustness and integrity of the hardware and software design of the trigger system, these issues and concerns have only begun to be addressed at a conceptual level by BTeV proponents ... Given the very complex nature of this system where thousands of events are simultaneously and asynchronously cooking, issues of data integrity, robustness, and monitoring are critically important and have the capacity to cripple a design if not dealt with at the outset. It is simply a fact of life that processors and processes die and get corrupted, sometimes in subtle ways. BTeV has allocated some resources for control and monitoring, but our assessment is that the current allocation of resources will be insufficient to supply the necessary level of “self-awareness” in the trigger system... Without an increased pool of design skills and experience to draw from and thermalize with, the project will remain at risk. The exciting challenge of designing and building a real life pixel-based trigger system certainly has the potential to attract additional strong groups.”

The main thrust of the R&D proposed here is to address this key concern, which is a generic concern for highly-reliable embedded real-time systems of this scale and complexity.

1.1 The BTeV Trigger System Hardware and Filtering Algorithms

BTeV is a High Energy Physics (HEP) project that will carry out an ambitious experiment to search for and study differences between the decay of particles containing b-quarks and the corresponding decays of anti-particles containing b-antiquarks. The study of this asymmetry will shed light on the preponderance of matter over antimatter in the observable universe and is one of the intense areas of research in elementary particle physics⁵. BTeV will utilize the Fermilab Tevatron Collider for this research. The Tevatron will produce 15 million high-energy particle collisions (also referred to as interactions) per second at the center of the BTeV detector. Each one of these interactions typically creates between 10 to 100 subatomic particles that will travel through the detector, where they will be tracked and identified. The interactions that are likely to exhibit b-quark asymmetries are expected to occur once for every 1 million collisions in the BTeV detector. This means that for each year of operation BTeV must “filter” data from over one hundred trillion interactions to select a sample of 10 billion b-quark decays from which a few million will be used to reveal the mysteries of matter-antimatter asymmetries.

BTeV will generate an enormous amount of data, about 1.5 terabytes per second. The factors that contribute to this exceptionally large data rate are the interaction rate (15 million collisions/s), the large number of particles that are produced in the collisions, and the large number of electronic sensor channels (30 million channels in the current design) in the detector. Because recording all of the data on an archival medium for later analysis is simply impossible, the challenge for the BTeV trigger system is to analyze data from the detector in real-time and to select interesting b-quark interactions to write to permanent storage for subsequent offline analysis.

Electronic trigger systems are common in HEP experiments⁶. However, BTeV is pursuing an ambitious trigger strategy that is unique in HEP. Most other experiments use a simple “first level” trigger, based on dedicated hardware which makes relatively unsophisticated decisions based on the most obvious, but not necessarily the most fundamental, differences between the signal and the background events. Typically, these triggers accept only very specific event topologies that conform to the idea of what is important physics at that moment. This reduces the number of interactions that must be processed by subsequent trigger levels, but the simple first-level trigger limits physics analyses by rejecting potentially interesting data. For BTeV a conventional first-level trigger of this type would restrict the experiment to a limited selection of b-quark particle decays, and would thereby prevent us from pursuing the broad range of physics analyses, including some that are “off the beaten path”, that we feel are needed to study b-quark asymmetries. Consequently, the BTeV trigger must be considerably more complex than triggers used for other experiments. The design of the trigger is driven primarily by the physics goals of the experiment.

The BTeV strategy is to trigger on the most fundamental property that differentiates particles containing the b-quark from other types of particles. That property is the presence of an interaction vertex, where the B particle, the anti-B particle, and many other particles are produced, followed by vertices a few hundred microns away where the B particles decay. Detecting such small vertex separations requires the reconstruction of all vertices using a complex pattern recognition algorithm. The BTeV “vertex trigger” performs pattern recognition for every interaction in the detector (15 million /s)⁷. A WEB-based animation of the pattern recognition with explanatory text is available⁸.

An overview of the architecture of the BTeV trigger and data acquisition system is shown in Figure 1, and some significant numbers that characterize the system are given in Table 1.

Table 1: Sizing characteristics of the system scale

# of Gbyte/s data links	Buffer memory	Data rate to L1 buffers	Data rate to L2/L3 buffers	Data rate to archive	# of L1 DSPs	# of L2/L3 Processors
2500	1 Tbyte	1.5 Tbytes/s	25 Gbytes/s	200 Mbytes/s	> 2500	2500

The trigger system has three distinct levels, called Level 1(L1), Level 2(L2), and Level 3(L3) and will use massively parallel computation pipelines at each level. L1 includes the vertex trigger, and additional triggers are being considered. Figure 1 shows the buffers that receive data from the BTeV detector, an expanded view of the first-level vertex trigger, a Global Level 1 trigger manager that will process the results from all first-level triggers, a switch that will route data to a large computing farm called the Level 2/3(L2/3) cluster, and the buffers and processors that make up the L2/3 trigger. There is no distinction between L2 and L3 hardware, since the same processors will be used to execute the L2 and L3 algorithms. A processor that receives data for an interaction will process the data using the L2 algorithm. If the data satisfy the L2 selection requirements, the data are processed using the L3 algorithm. Data that fail L2 or L3 selection requirements will be dropped. Data for an interaction that satisfies the selection requirements will be written to archival storage at an average rate of 200 Mbytes/s.

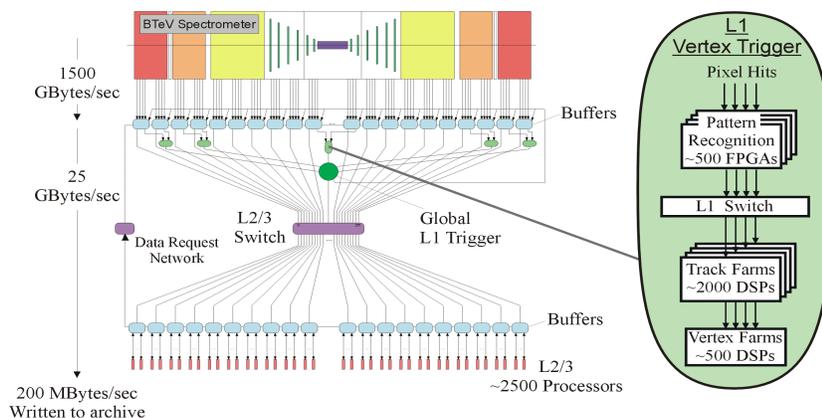


Figure 1: Schematic of the BTeV Trigger and Data Acquisition System showing (left side) the detector, buffer memories, L1, L2, and L3 clusters and their interconnects and (right side) a blowup of the L1 Vertex trigger

Detailed Monte Carlo simulations, data-flow simulations, and benchmarks for trigger algorithms have been performed to estimate the capabilities required of the BTeV trigger and data acquisition project. Here we describe the relevant results. As mentioned before, the data rate out of the BTeV detector and into the data acquisition system will be 1.5 Terabytes per second. All of the data for each interaction that occurs in the BTeV detector will be buffered long enough to give the L1 trigger time to decide if the data should be dropped or sent to an L2/3 processor. A subset of the data (the data from the silicon pixel vertex detector) will be sent to the L1 vertex trigger (inset shown in Figure 1). The L1 vertex trigger implements a pattern recognition algorithm using about 500 Field Programmable Gate Arrays (FPGAs) to find track segments in the pixel vertex detector (the first step in finding the particles that were created in an interaction). The track segments will be routed through a switch to a farm of Digital Signal Processors (DSPs) that reconstruct particle trajectories, followed by a second farm of DSPs that reconstruct vertices. Benchmarks of the reconstruction algorithms were performed for the Texas Instruments TMS320C67X DSP. The results from these benchmarks and estimates of the input data rate indicate that the L1 vertex trigger will require about 2500 DSPs to achieve the necessary processing power. This estimate does not include additional processors for fault tolerance or for other L1 algorithms, nor does it include the large number of support processors required to configure, monitor, and control the DSPs. The design goal for L1 is to reject 99% of the interactions that occur in the detector. The interactions that survive L1 will be passed to the L2/3 trigger. L2 will perform a more refined analysis of the data and impose more stringent selection criteria than L1. L3 will improve the analysis for each inter-

action that survives L2 even further by considering data from each detector subsystem and by performing a detailed physics-based analysis to identify interesting interactions. We have estimated that the L2/3 trigger cluster will require on the order of 2500 general-purpose computers, such as Intel Pentiums running the LINUX operating system. The L2/3 system will provide another factor of 20 in rejecting uninteresting events for a total rejection of 2000.

The requirements for the BTeV trigger are well understood, and the design is at an advanced conceptual stage. However, there is still enough flexibility in the design to adapt to new discoveries and different implementation strategies. For example, calculations that are currently done in DSPs may be migrated into FPGAs or we may use more DSPs and fewer PCs, or vice versa. The supporting software infrastructure must be flexible enough to handle variations in the hardware design. There will be variations due to the availability of new and different types of hardware, the addition of redundant hardware for reliability, elimination of superfluous hardware, or algorithm changes that require different hardware implementations. We must retain the ability to make design changes that permit the most cost effective use of the computing hardware. **The results from the proposed research will provide critical feedback to the designers of the BTeV system to achieve the required robustness and agility.**

1.2 IT Aspects of the BTeV Trigger and Data Acquisition System

While the hardware platform required to achieve the above-stated goals is extensive and complex and the trigger algorithms are quite challenging, an even greater challenge is to keep the system functioning and producing quality results over a period of several years. The system must serve well during the detector commissioning and debugging stage, during routine operations, troubleshooting, and calibration. Even during normal operations, it must operate in several modes and switch between these modes dynamically. It must continue to operate in spite of the failure of some of its components. It must adapt itself to varying conditions in the Tevatron accelerator and in the BTeV detector. It must evolve as hardware is replaced or as more powerful components are introduced to extend its capabilities, and it must accommodate changes in software as more is learned about the physics and as the detector itself evolves. In addition to computational performance, the key requirements for the BTeV trigger system include: (1) Dynamic reconfiguration and partitioning; (2) High availability, including introspective, self diagnosing, fault tolerant, fault adaptive capabilities; and (3) Life-cycle maintainability and evolvability.

1.2.1 Dynamic Reconfiguration and Partitioning

The system must be able to handle two aspects of dynamic reconfiguration: the ability to dynamically adjust parameters of the experiment, and the ability to reconfigure and repartition hardware used to analyze and filter data from the experiment. The latter will require some degree of dynamic reconfiguration of the computation network, so that the system can vary both the interconnectivity between different trigger levels and the number of processors assigned to different tasks.

In order to execute a number of different tasks (some of them simultaneously) the system must be able to operate in different modes. Examples are: 1) standard trigger operation; 2) special modes to support commissioning, debugging, and calibration of the detector; 3) verification of repairs, upgrades, and replacement of hardware and software components; 4) in situ (re)calibrations as the experiment proceeds; 5) use of parts of the system to test new trigger algorithms; 6) verification of detector alignment and calibration at the start of each physics run, which can occur a few times per day; and 7) introduction of special diagnostic packages to investigate problems in the detector, the trigger hardware, or software.

There is another significant area where dynamic reconfiguration is highly desirable. Physicists will use offline computing resources, requiring extensive processing power and I/O capabilities, for data analysis. One possible platform will be the L2/3 trigger system. The L2/3 trigger is not always fully saturated by real-time data acquisition, since the Tevatron intensity decreases during a run cycle⁹. Therefore, the L2/3 processors should be available to physicists when they are not needed for trigger tasks. Research on cluster-based services for the Internet¹⁰ has explored the use of *overflow pools* or sets of computers that can be used temporarily to handle prolonged bursts in demand for a service. We will explore a system that relinquishes resources during periods of low demand.

1.2.2 High Availability

The BTeV system must be available 24 hours/day, 7 days/week (24/7) to support all of the previously mentioned tasks. The highest possible data throughput and data quality must be maintained during normal trigger operations. Therefore, the system must be fault tolerant, introspective, self-diagnosing, and fault adaptive.

There are many ways that a complex heterogeneous system can malfunction, including the failure of individual processing elements or network switching elements. Moreover, a noisy accelerator environment or detector malfunctions could produce faulty data that impair the ability of the system to maintain data throughput or data quality. BTeV must be fault tolerant, but at an acceptable expense. For example, airline systems use two or three levels of redundancy in critical computer systems. This level of redundancy is costly, but the cost is warranted. In the case of the BTeV trigger, while it is unacceptable that the system loses large amounts of data, it is tolerable to lose a small fraction of the data for a short period of time. The key feature is that during fault conditions, the trigger system must continue to operate, possibly at decreased capacity (graceful degradation).

The system must be able to detect fault conditions, both locally and globally, and operator intervention should be kept to a minimum. Due to the system's complexity, it could take a very long time for an operator to recognize the existence of a problem, diagnose it, and remedy or mitigate it. All the while, valuable data from the experiment would be lost, or be of poor quality. The system must take the initiative to mitigate and adapt to faults.

One basic operation when adapting to faults is to remove the failed component from its partition, and restart its computation on a similar component. An example of this is a classic offline and batch processing system that relies on check-pointing. We will evaluate at which levels check-pointing is appropriate for high performance embedded systems, where loss of data, rather than compute time, may occur.

The proposed BTeV system has the capability to change many key operating parameters to repair or mitigate problems detected by its analysis programs, not only within the trigger and data acquisition system but also in the detector complex. For example, it can reduce the high voltage supplied to a noisy detector, or raise the threshold for deciding that it has valid information. Furthermore, there are typically several physics triggers and several calibration and monitoring triggers, which are "pre-scaled" to obtain an output data stream. The system can adjust these pre-scaled triggers or turn off lower priority triggers to preserve resources - computational, memory, or network bandwidth - for the highest priority data.

1.2.3 Life-cycle Maintainability and Evolvability

BTeV will be constructed over the next 5 years and will then run for at least 5 years after that. It will be essential to develop a trigger and data acquisition system that can be easily operated and maintained. It must be a system that can evolve as old hardware fails or becomes unsupported and as new hardware and software technologies that offer improved performance become available. In addition, the experiment itself will undoubtedly be modified to respond to new ideas and challenges as we learn more physics. The system must be able to adapt to support these changes. One has only to look back over the experience in HEP in the last decade to understand how difficult it will be to achieve this goal, and how necessary it is to address it from the earliest design.

2. Project Description, Goals, and Objectives

The BTeV trigger system is an example of a massively parallel, high performance, high reliability, computational system. The design and implementation of such systems cannot be achieved by the ad hoc approach of developing simple small-scale components and scaling them up into large-scale systems¹¹. Issues such as fault tolerance and performance must be explicitly addressed at multiple levels in the system design¹². We propose advances in system design methodology, tools and runtime infrastructure to facilitate these and more issues involved in developing such systems. We further propose to develop the software to accomplish the design and implementation of the system and to study its performance, utility, and scalability on the actual BTeV hardware as it grows over the construction phase of the experiment. The result of this research will be software, design methodologies, and the documented experience of the project.

3. Conclusion

Many real-time embedded systems applications require high computational performance and high availability. High Energy Physics is only one of many disciplines that must collect and analyze huge amounts of data in real time, and must continue operating under fault conditions. The following statement was taken from the President's Information Technology Advisory Committee Report (PITAC)¹³:

“The Nation needs robust systems, but the software our systems depend on is often fragile. Software fragility is its tendency not to work properly – or at all. Fragility is manifested as unreliability, lack of security, performance lapses, errors, and difficulty in upgrading. Examples can be found everywhere, from our huge information systems for air-traffic control to the personal computers on our desks, from the Pentagon to the Internal Revenue Service (IRS).”

This research proposes to develop new technologies for creating software for real-time embedded computer systems that must exhibit ultra high performance, must be highly available, and must be maintained and evolved easily over the system life-cycle. The results of the research will be a powerful software system that will support the BTeV experiment, as well as new engineering methodologies for developing software for large-scale embedded computer systems. These results could easily be applied to other such applications, including those referred to by PITAC. This research addresses the unreliability, performance lapses, errors, and difficulty of upgrading mentioned there.

This project will bring together a team of experts to develop new embedded systems technologies, which can be made available to a much wider range of applications and researchers. The team will prove the technologies by deploying them in support of the very demanding BTeV trigger and data acquisition system. By collaborating with an ongoing effort in experimental particle physics, the team will have access to a large-scale system for testing without needing to acquire all the hardware independently. Members of the team, which come from universities in the US and from Fermilab, are experienced in the development of this kind of software and will, through the support obtained from this project, make a breakthrough in software for embedded real time systems while at the same time advancing the investigation of a question of major significance in physics.

¹ The BTeV Proposal (May 2000) resides at:

http://www-btev.fnal.gov/public_documents/btev_proposal/index.html.

The BTeV Trigger is described in chapter 9 and the Data Acquisition System in chapter 10. These are located in Part 2 of the document. Trigger algorithm physics simulations are described in chapter 14, located in Part 3.

² Fermilab Director Michael Witherell's report on BTeV approval:

http://www-btev.fnal.gov/public_documents/Approval/index.html.

³ Links to the home pages of members of the collaboration may be found at:

http://www.hep.vanderbilt.edu/btev_rtes/.

⁴ Robert Tschirhart and Peter Wilson, private communication.

⁵ A general presentation on the fundamental objectives of High Energy Physics in the 21st century can be found at:

<http://www.fnal.gov/pub/inquiring/matter/future/index.html>.

⁶ Schaller, S.C., ed. 1999 *IEEE Conference on Real-Time Computer Applications in Nuclear Particle and Plasma Physics*, Santa Fe, June 1999, in *IEEE Trans. Nucl. Sci.*, Vol. 47, Issue 2, Part 1, April 2000.

⁷ Gottschalk, E. E., et al., "BTeV Detached Vertex Trigger," to be published in *Proceedings of the 9th International Workshop on Vertex Detectors (Vertex 2000)*, Homestead, MI, September 2000.

⁸ A WEB-based animation of the pattern recognition algorithm used in the BTeV trigger can be found at

http://www-btev.fnal.gov/public_documents/animations/Animated_Trigger/index.htm.

⁹ A run typically corresponds to all or part of a Tevatron colliding beam "store" which may last several hours. The collision rate decreases during the store as the particles in the beams are gradually used up by the collisions.

¹⁰ Fox, A., Gribble, S., Chawathe, Y., Brewer, E., and Gauthier, P., "Cluster-Based Scalable Network Services," *Symposium on Operating Systems Principles (SOSP-16)*, October 1997.

¹¹ Bapty, T., Sztipanovits, J., "Model-Based Engineering of Large-Scale Real-Time Systems," *Proceedings of the Engineering of Computer Based Systems (ECBS) Conference*, pp. 467-474, Monterey, CA, March 1997.

¹² Gartner, F., "Fundamentals of fault-tolerant distributed computing in asynchronous environments," *ACM Computing Surveys*, Vol. 31(1), 1999, pp. 1-26.

¹³ PITAC - Report to the President, President's Information Technology Advisory Committee, 2/24/1999.