PATRIOT: Physics Archives and Tools Required to Investigate Our Theories

S. Mrenna,* FNAL, Batavia, IL 60510, USA

Abstract

The PATRIOT project aims to provide the best theoretical predictions of physics events for the high–p_T physics program of Run2 at Fermilab by combining several tools: (1) a new generation of Monte Carlo programs to calculate the hard structure of high–p_T events and event generators to make particle level predictions, (2) sophisticated queuing systems with job interdependencies for computations on farms, (3) an enstore or mass storage repository for datasets of events, and (4) the StdHep and MCFIO packages to define a common file format. These common files are processed through the detector and triggering simulations of the experimental collaborations CDF and DØ, and are being integrated into the common data handling system SAM.

MOTIVATION

The discovery of new physics at the Tevatron or the Large Hadron Collider (LHC) will likely require precise Standard Model predictions of event rates and kinematic shapes for high-p_T objects: leptons, jets, heavy quarks, etc. When it is practical, the experiments should have access to the *best* theoretical predictions possible to drive their analyses. When the experiments at Fermilab accumulate on the order of 1 fb^{-1} of data, the dominant systematic error for many measurements will come from uncertainty in our understanding of the theory of strong interactions (Quantum Chromodynamics, or QCD). Looking at past experience from Run1 at Fermilab, one learns that the desired result is not always easy to obtain. In that case, coding choices and event structures prevented the update of some generators within the experiments' software structure, so that many standard predictions were knowingly inaccurate. Furthermore, as Run1 evolved, a need developed for predictions from theoretical tools outside of the standard framework. Then, the question arose of how to pass these predictions through the trigger and detector simulations of the experiments.

These and other considerations led to the development of PATRIOT [1], the core of which is Monte Carlo datasets in a format that interfaces smoothly with the CDF and DØ software structure. There are many motivations for a database of Monte Carlo events (dbmc):

1. The generation of accurate samples of Monte Carlo events can be time-consuming and complicated. The dbmc avoids needless duplication of effort and prevents mistakes.
2. If both CDF and DØ use the same dbmc for their analyses, then it will be easier to combine their data since their theory uncertainties are the same.
3. Gets the best theoretical predictions to the experiments quickly, since the theorists can generate the particle-level events themselves.
4. A dbmc can be enable signature-based search strategies by providing a complete description of the SM event rates in any channel using *a priori* cuts.

* mrenna@fnal.gov

DESCRIPTION

PATRIOT exploits several tools to accomplish the full task of generating events and delivering them to the dbmc, for use by experiments or theorists. As new or better tools are developed, they can be easily incorporated into the general framework.

Theoretical Prediction

Several different tools now exist that automate the calculation of high-p_T events at the tree-level (i.e. at the lowest order in perturbation theory). While each of these tools rest on the same theory, the details of the calculation often differ, leading to somewhat different calculations. In principle, one would like datasets for the same physics process but using different tools to check for coding errors and to gauge the theoretical uncertainty in the calculations. For practical reasons, the calculations performed so far have relied on the generator MadEvent[2]. MadEvent uses a multi-channel integration scheme that is well suited to computational farms. In this scheme, the important regions of the Monte Carlo integrand are isolated and calculated independently. After each sub-integrand is know to a prescribed level of accuracy, the individual pieces are recombined to give the full integrand.

Farm Calculations

The Fixed Target Farm at Fermilab is used by several projects. It consists of about 200 worker nodes and two dedicated I/O nodes working within the fbsng (Fermilab Batch Server: Next Generation) framework [3]. The fbs queue system is driven by job description files, which allow a complete specification of the conditions for each job to begin. For example, the MadEvent calculation actually is performed in two steps. In the first step, the integrand is sampled coarsely to determine the largest integration regions. The results of the first step then determine the refined calculation done in the second one. The job description file can allow the second step to begin only after all the jobs in the first step have completed. Another dependency can be declared so that the final results are saved to a disk cache or a mass storage system when the second step has completed.

Format of the Results

The final result of the farm calculation (to be provided to the experiments) is a file of particle-level events. The `StdHEP` (Standard HEP) [4] tool was originally developed to provide a universal translation between several Monte Carlo programs in the form of the `HEPEVT Fortran` common block. It is also interfaced to `MCFIO`, a tool which writes the common block structure (or structures) to a file in a compressed and platform-independent format. Since CDF and DØ use `StdHEP` for their own Monte Carlo generation, this format is readable by both experiments.

Storage

`Enstore` is available for mass storage of datasets at Fermilab [5]. The `enstore` tools allow for a Unix-like listing of directories and files and the ability to `encp` to and from the enstore. The enstore is also expandable and can grow as more calculations are performed. `Enstore` is also used by CDF and DØ in specialized versions.

Metadata and `SAM`

As argued earlier, the datasets are valuable: valuable because of the resources used to generate them and valuable for how they can be used. The datasets lose their value, however, when the information describing them is inadequate. The choice of metadata is a serious issue, and it must contain the relevant information and should be organized for efficiency. `SAM`, or Sequential data Accessed via Metadata, is a dataset distribution tool that is being used by both CDF and DØ [6]. Several other presentations at CHEP 2004 describe the system in more detail. The primary characteristics that are relevant to `PATRIOT` are that the datasets are searchable on metadata and can be distributed globally to different analyses groups. The metadata for `PATRIOT` has been developed in a `mysql` database, which will be incorporated into `SAM` shortly. The metadata includes information about the mechanics of how the database was prepared, which tools were used, and which physics parameters where chosen.

FUTURE

Presently, the `PATRIOT` project has provided detailed calculations for backgrounds that are important for studying top quark pair production at the Tevatron. The theoretical details are described in Ref. [7]. Both experiments have processed these datasets through their full trigger and detector simulations, and analyses are under way. This is a success, but the distribution of the datasets and the information about them has been *ad hoc*. A serious use of this tool requires deployment into `SAM`. The further success of the project also will rely on the efforts of other theorists to make their calculations available.

REFERENCES

[1] `http://cepa.fnal.gov/patriot/`

[2] F. Maltoni and T. Stelzer, JHEP **0302**, 027 (2003) [arXiv:hep-ph/0208156].

[3] `http://www-isd.fnal.gov/fbsng/`

[4] `http://cepa.fnal.gov/psm/stdhep/`

[5] `http://www.fnal.gov/docs/products/enstore/`

[6] `http://d0db-prd.fnal.gov/sam/`

[7] S. Mrenna and P. Richardson, JHEP **0405**, 040 (2004) [arXiv:hep-ph/0312274].