ENABLING DISCOVERY AT THE LARGE HADRON COLLIDER THROUGH DATA-INTENSIVE COMPUTATION AND MACHINE LEARNING

Allocation: Illinois/140 Knh
PI: Mark Neubauer[1]
Collaborators: Robert Gardner[2], David Lesny[1], Dewen Zhong[1]

[1]University of Illinois at Urbana-Champaign
[2]University of Chicago

EXECUTIVE SUMMARY

The Large Hadron Collider (LHC) is the world's most powerful particle accelerator, designed to study the fundamental nature of matter and the forces that govern its interactions by colliding beams of protons at the highest-available energies. We are using Blue Waters to process, simulate, and analyze high-energy proton–proton collision data produced by the ATLAS experiment at the LHC and to improve our sensitivity to new phenomena by developing novel approaches to identifying Higgs bosons produced with high momentum at the LHC by using machine learning techniques.

RESEARCH CHALLENGE

The goal of particle physics is to understand the universe at its most fundamental level, including the constituents of matter, their interactions, and the nature of space and time itself. This quest is one of the most ambitious and enduring of human endeavors.

The Standard Model (SM) of particle physics describes all known fundamental particles and their interactions, including the Higgs boson, which was discovered at the LHC [1,2] in 2012, with significant contributions by the Illinois (Neubauer) Group. The discovery led to François Englert and Peter W. Higgs receiving the 2013 Nobel Prize in Physics. The SM has withstood the last 40 years of experimental scrutiny, with important exceptions being neutrino mass, dark matter, and dark energy. Recent developments in particle physics and cosmology raise the exciting prospect that we are on the threshold of a major step forward in our understanding.

It is an enormous challenge to process, analyze, and share the 15 petabytes of data generated by the LHC experiments each year with thousands of physicists around the world. To translate the observed data into insights about fundamental physics, the important quantum mechanical processes, and the detector's responses to them, need to be simulated to a high level of detail and with a high degree of accuracy.

A key thrust of this project is to use the recently discovered Higgs boson to search for new physics in novel ways enabled by the Blue Waters supercomputer. The enormous energy available in proton–proton collisions at the LHC leads to the production of particles with very high velocity relative to the ATLAS detector (lab frame). Even massive particles like the Higgs boson can have a large momentum and, therefore, large Lorentz factor (γ) in the lab frame. When these "boosted" particles decay, their decay products are highly collimated and not easily distinguished in the detector instrumentation (e.g., by calorimeters). This limits the sensitivity of searches for new physics such as $X \to hh$, where X is a new massive (~TeV/c^2) particle.

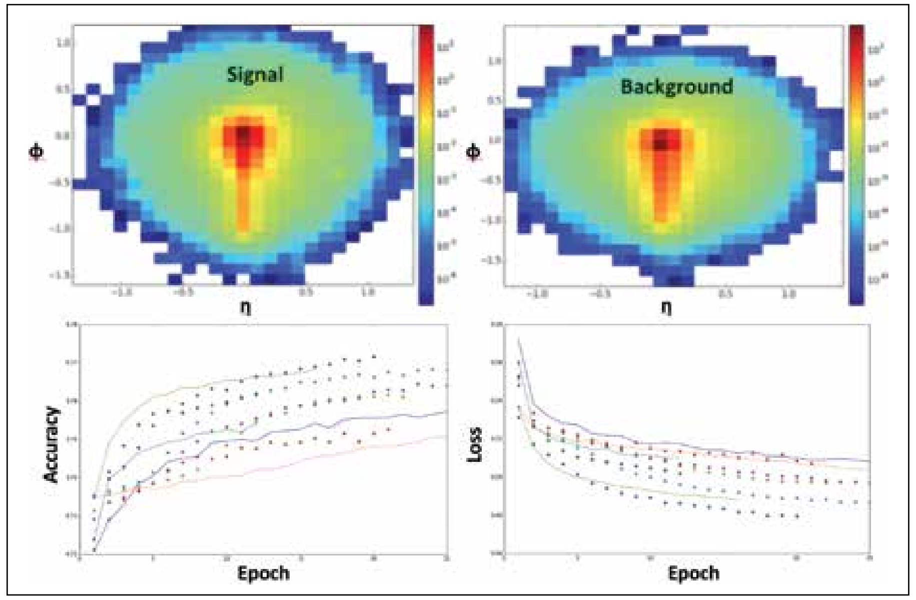

Figure 1: (Top) "Images" created by jets of charged particles in Higgs decay signal events and background events that are detected by the ATLAS tracking system. (Bottom) Higgs boson identification accuracy and signal loss as a function of the number of training epochs for a variety of CNN configurations and hyperparameter settings.

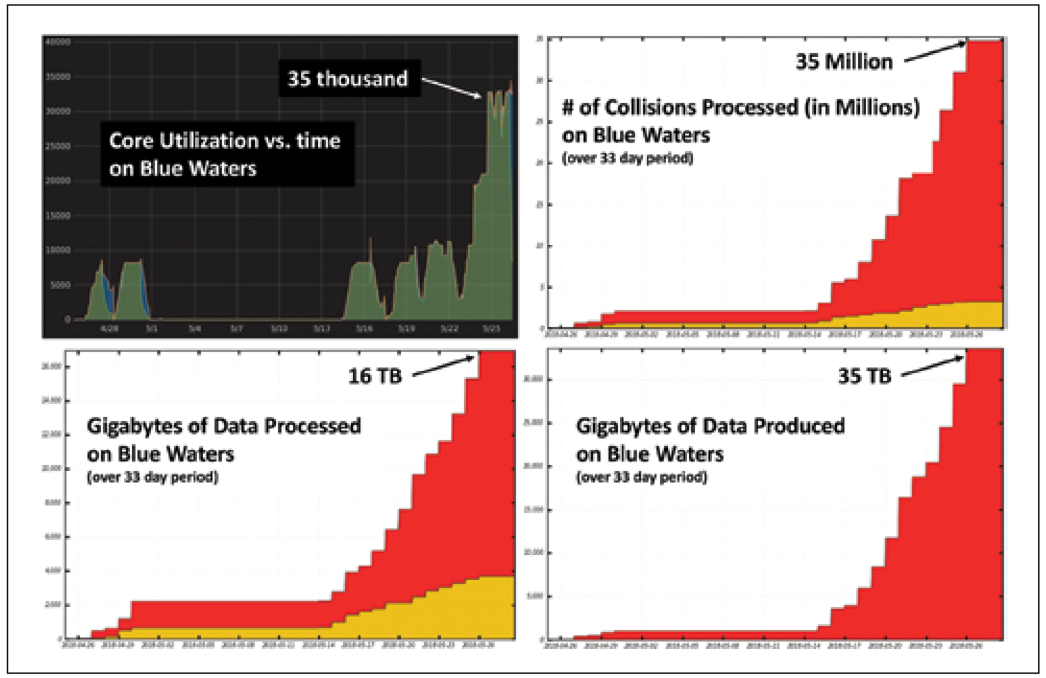

Figure 2: CPU core utilization, number of collision events processed, data consumed and simulated, and refined detector data generated during a 33-day period during the project.

METHODS & CODES

We have integrated Blue Waters into our production processing environment to simulate and analyze massive amounts of LHC data. Blue Waters resources are made available to the ATLAS computing fabric using a system called ATLAS Connect [3], which is a set of computing services designed to augment existing tools and resources used by the U.S. ATLAS physics community. Docker Images are delivered via Shifter to create an environment on Blue Waters' nodes that is compatible with the ATLAS job payload.

The approach we are currently taking to identify boosted Higgs bosons is to use a convolutional neural network (CNN) trained using "images" created by jets of charged particles in Higgs decay events, as shown in Fig. 1 (top). This work was done in collaboration with Indiana University and the University of Göttingen.

The images shown are the particles' angles in the detector as centered on the Higgs boson, with the color represented by the particles' momentum transverse to the proton beamline. We have successfully performed CNN training using GPUs on Blue Waters with these images via a sequential Keras model with a TensorFlow backend.

We have also studied deep neural networks, again using Keras with TensorFlow. We are developing an alternative approach that uses the particles' four-momentum and jet-clustering history rather than images fed into a recursive neural network, which draws inspiration from natural language processing.

RESULTS & IMPACT

Fig. 2 shows the CPU core utilization, data consumed and generated, and the number of collision events processed during a 33-day period of the project. The job output was made available to the rest of the ATLAS collaboration for use in analysis of the LHC data to improve measurements of the SM and to search for new physics. Fig. 1 (bottom) shows the Higgs boson identification accuracy and signal loss as a function of the number of training epochs for a variety of CNN configurations and hyperparameter settings. We are also using Hyperas, a convenience wrapper using Hyperopt with Keras models, on Blue Waters to automate the scanning of hyperparameters in a variety of machine and deep learning approaches to improve the Higgs boson identification over backgrounds. The techniques show promise in addressing the challenges of boosted Higgs boson identification and improving the sensitivity of new physics searches at the LHC.

WHY BLUE WATERS

Blue Waters, as a large CPU and GPU resource with high data-throughput capabilities, greatly facilitated our research. The strong support for containers allowed us to deploy our science application on Blue Waters nodes. Also, Blue Waters provided a means for a highly parallelized and automated scanning of free parameters in our machine learning configurations and, therefore, rapid optimization of our boosted Higgs boson identifier.