SciDAC: HEP Data Analytics on HPC

1. PIs: J. Kowalkowski (FNAL), N. Buchanan (Colorado State University), P. Calafiura (LBNL), R. Ross (ANL), A. Sousa (University of Cincinnati)
   Objective: Advance LHC and neutrino science by transforming data analysis applications, workflows, and data handling to effectively utilize resources available at HPC facilities

   Z. Marshall (LHC/ATLAS), S. Mrenna (LHC/CMS), A. Norman (NOvA/DUNE)
   S. Leyffer (ANL), J. Mueller (LBNL), H. Schulz (University of Cincinnati),
   M. Paterno (FNAL), T. Peterka (ANL), S. Sehrish (FNAL)
2. Met: Validation and extension of NOvA Feldman-Cousins 1D and 2D analysis procedures within HPC (Includes analysis of workflow and data needs), Generator HPC toolchain for tuning and optimization studies completed.
3. Near completion: A complete Pythia8 generator tuning study within HPC
4. References

   Acero, M. *et al* (NOvA Collaboration), Phys. Rev. D 98 032012, (2018)
   Acero, M. *et al* (NOvA Collaboration), to be submitted to Phys. Rev. Letters (2018) [antineutrinos]
   Three additional technical papers for 2018, see references on subsequent slides
5. Synergistic activity: SciDAC-HPC framework for event generation at colliders

   LHE generator data available to HPC workflows in HDF5.

NOvA Neutrino + Antineutrino Analysis

Scientific Achievement

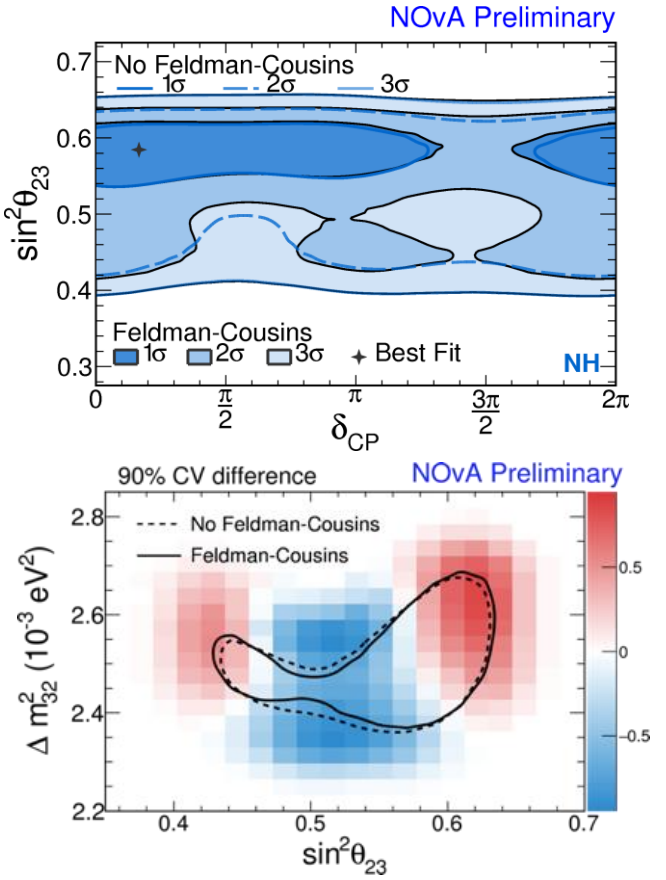
Large-scale analysis campaigns carried out at NERSC for the first time, in support of the first set of electron antineutrino appearance results shown June 4th at the Neutrino 2018 conference

Significance and Impact

Most precise measurement of antineutrino oscillations to date; Improved accuracy: 8x higher resolution than any prior NOvA result; 50x faster than any previous result: reviewed by collaboration in <24h

Research Details

- Comparing data with neutrino oscillation hypothesis to extract best-fit oscillation parameters and associated confidence intervals.
- Employs new fitting procedures, some of the most complicated currently used in neutrino physics.
- Requires advanced statistical treatment to account for non-gaussianity of errors in oscillation measurements due to: (1) Low statistics; and (2) parameters probed near physical boundaries
- Statistical treatment is extremely computationally-intensive, requiring billions of simultaneous multi-dimensional fits



Sensitivity contours under the Gaussian statistical assumptions compared to a Feldman-Cousins corrected computation. Corrected contours reveal large islands in parameter space where sensitivity is greatly improved.

A.Sousa. Presented at CHEP 2018, Sofia, Bulgaria. To be published in EPJ Web of Conferences (2019).
http://news.fnal.gov/2018/07/fermilab-computing-experts-bolster-nova-evidence-1-million-cores-consumed/

HEPnOS: Fast Event-Store for HEP on HPC

Scientific Achievement

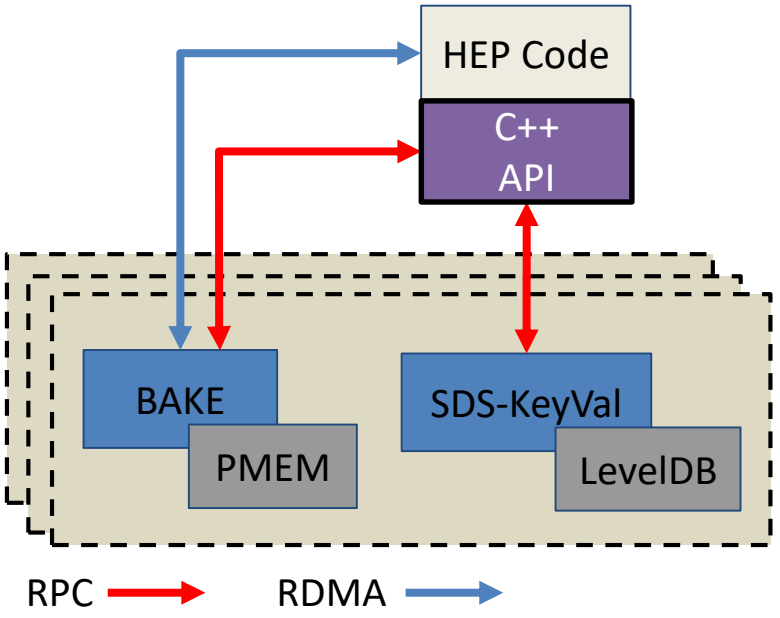
A custom data service for HEP, optimized for state-of-the-art HPC systems.

Significance and Impact

Extend the physics capability of HEP experiments by allowing HEP data analysis programs to harness the ever-increasing power of ASCR (and other) supercomputers.

Research Details

- Manage physics event data from simulation and experiment, through multiple phases of analysis
- Bypass file system to accelerate data access throughout analysis process
- Designed to seamlessly integrate into HEP software frameworks
- Leverage elements of ASCR *Mochi* project to rapidly develop and customize for HEP needs:
  - Physics object data stored in NVRAM, RAM, or SSD
  - Metadata stored in modern index (e.g., LevelDB)
  - RDMA used for client access to physics object data



The ASCR Mochi project is researching methods for rapid specialization of data services for SC mission needs.
https://xgitlab.cels.anl.gov/sds/HEPnOS/wikis/home

Tools for HPC-scale physics analysis

Scientific Achievement

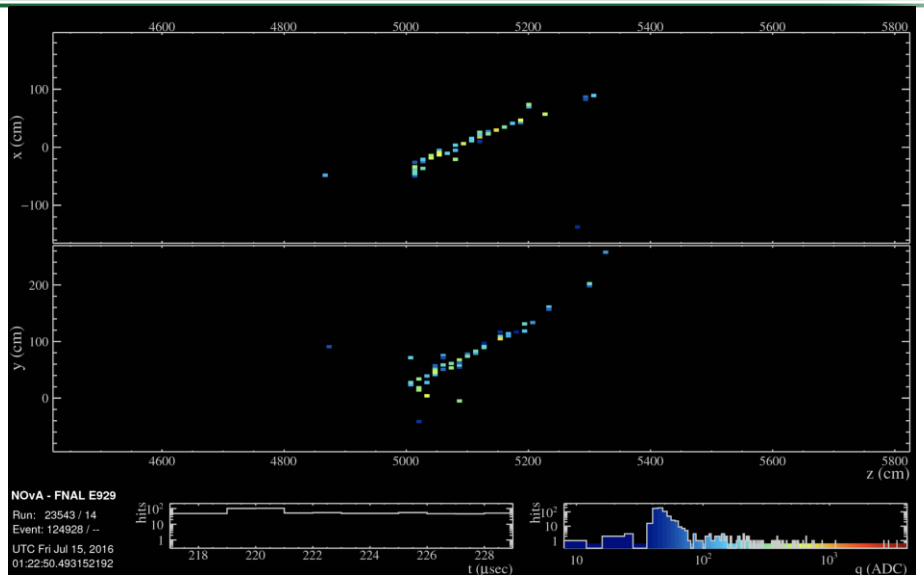
A parallel data storage and access library for multi-terabyte physics data sets for use in HPC environments.

Significance and Impact

Obtain better understanding of neutrino oscillations by allowing improved systematic study of event-selection criteria used in analysis.

Research Details

- Using hephpc::hdf5 C++ library for writing HDF5 files from traditional HEP analysis programs.
- Demonstrated parallel reading speed in Python prototype code: >40 TB read in <20 seconds, using >76,000 KNL cores on Cori at NERSC.
- Demonstrated conversion of more than 4 TB of analysis data from NOvA's HEP-traditional analysis data format to our HDF tabular organization.
- The NOvA collaboration has taken ownership of the *art* framework module developed for this.
- Demonstrated ease-of-use of efficient high-level libraries (Python *pandas*), to support implicitly-parallel analysis code.



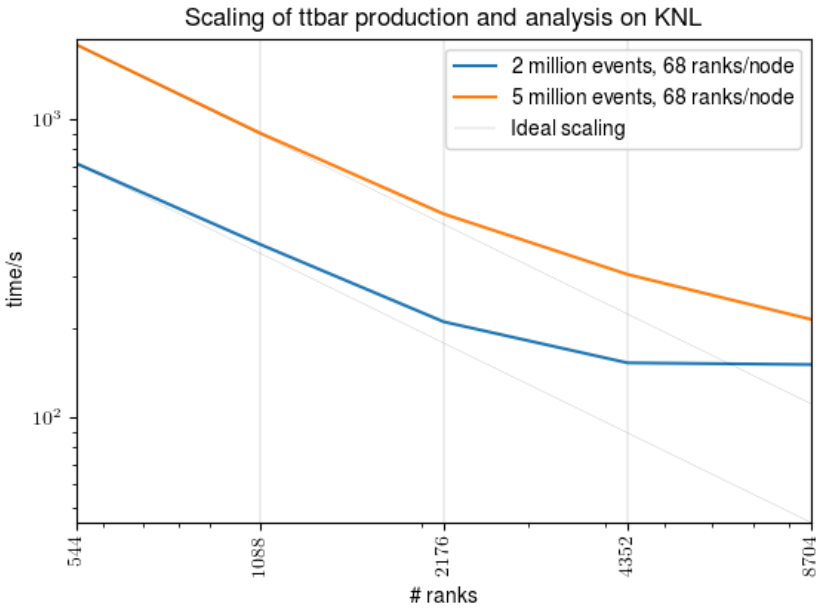
One of the 18 electron antineutrino appearance candidates selected by NOvA after analysis of 1.04 billion candidate interactions. NOvA observes a 4 sigma strong evidence for electron antineutrino appearance in a muon antineutrino beam

Currently published on BitBucket at https://bitbucket.org/fnalscdcomputationalscience/hep_hpc
art is described at http://art.fnal.gov
M.Paterno, *et al*. *Parallel Event Selection Performance on HPC Systems*. Paper presented at CHEP 2018, Sofia, Bulgaria. To be published in EPJ Web of Conferences (2019).

Parallel event generation with DIY on HPC

Scientific Achievement

Developed an application for generating and analyzing Monte-Carlo (MC) events on HPC architectures capable of running at a massively parallel scale.



Scaling behaviour when generating ttbar events with Pythia8 on HPC resources. The deviation from ideal scaling is due to the program overheads. With enough work assigned to each rank, we achieve perfect scaling. (Image Credit: Holger Schulz, U Cincinnati)

Significance and Impact

HEP phenomenology and all experiment simulation workflows require vast numbers of MC generator events. This application efficiently utilizes HPC resources and HEP community tools to accumulate events in parallel.

Research Details

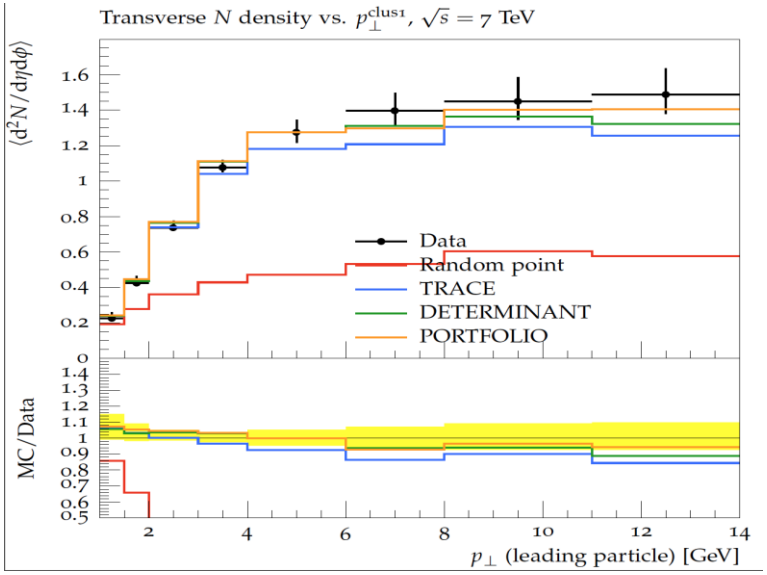
- Data parallelism with ASCR DIY library encapsulates all MPI communications into a block-processing program application.
- Implements full chain of event generation with Pythia8 and analysis with Rivet.
- Allows for extremely short turn-around time of large parameter space explorations in e.g. the field of generator tuning.
- Paves the way for new and advanced optimization algorithms that do not rely on surrogate models, e.g. for limit setting through reinterpretation of LHC search analyses.
- Can also be used to accelerate laptop analyses.

Currently published on BitBucket at https://bitbucket.org/iamholger/pythia8-diy/wiki/Home.

Automated physics event generator tuning

Scientific Achievement

Automated tuning of physics event generators: Formulation as bilevel optimization with equality constraints. Problem solved with surrogate-model optimization.



Optimization outcome: shown is the observation data (black) and the simulation model predictions (solid lines) when using three different outer optimization objective function models (portfolio, trace, determinant) and a randomly chosen solution (upper graphs). The lower graph shows how well the simulation predictions fall within the uncertainty band of the observations (ideally, the colored graphs lie within the yellow area). (Image Credit: Holger Schulz, U Cincinnati)

Significance and Impact

This approach allows for more robust theoretical predictions at the LHC. Current inefficiencies and potential biases in the treatment of observables addressed through bilevel optimization.

Research Details

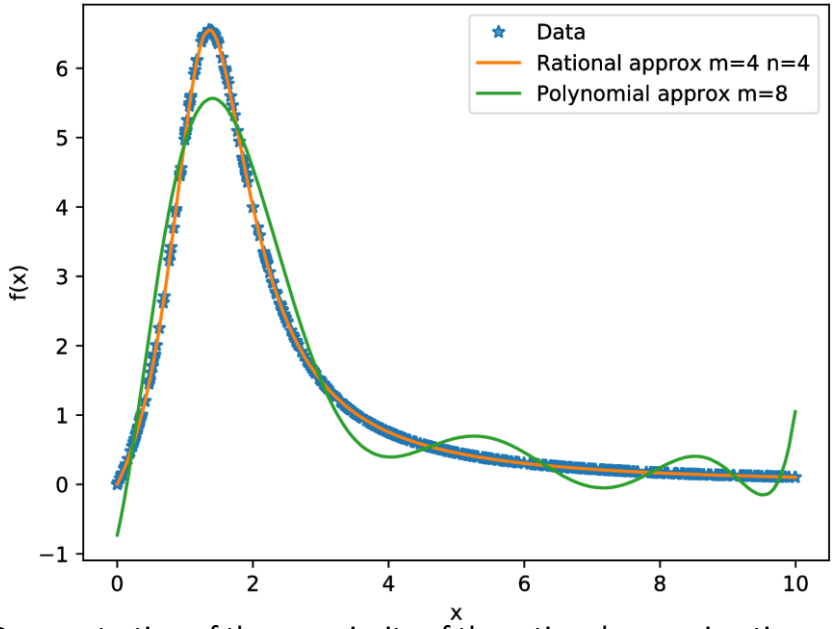
- Automatically adjusts the weights for each observable to influence the final tuning. Formulation of an outer optimization problem to *automatically* assign weights to observables used in the tune.
- Outer optimization by surrogate model approach with equality constraint to achieve good solutions efficiently.
- Modeling of the outer optimization problem with design of experiment (trace, determinant) and portfolio optimization (minimize the mean and variance of errors over all observables simultaneously) approaches.
- Inner optimization using HEP community tool PROFESSOR.
- Better tunes are obtained more efficiently.

H. Schulz, *et al*. *Teaching PROFESSOR new math*. Paper presented at CHEP 2018, Sofia, Bulgaria. To be published in EPJ Web of Conferences (2019).

Rational approximations for physics signal modeling

Scientific Achievement

Developed an algorithm that reliably calculates multivariate rational approximations.



Demonstration of the superiority of the rational approximation algorithm. We compare the input data (blue) with the predictions from a rational approximation (orange) and a polynomial approximation (green). The oscillatory behaviour as well as the unsatisfactory predictions at the edges of the interpolation domain so far restricted the usage of surrogates for applications in BSM studies where rational functions are frequently observed. (Image Credit: Holger Schulz, U Cincinnati)

Significance and Impact

Our algorithm allows the application of surrogate models in HEP to wider areas. For example, this enables new high-fidelity analyses to physics beyond the Standard Model, such as signal modeling for dark matter direct detection simulations.

Research Details

- The predictive power of polynomial approximations is limited if the data exhibits traits of rational functions.
- Simultaneous multivariate construction of numerator (order *m*) and denominator (order *n*) polynomials by means of singular value decomposition (SVD).
- Numerical stability improved through SVD and orthonormal bases.
- Automatic detection and rejection of solutions that have poles in the interpolation domain.

Generator tuning on unexploited data

Scientific Achievement

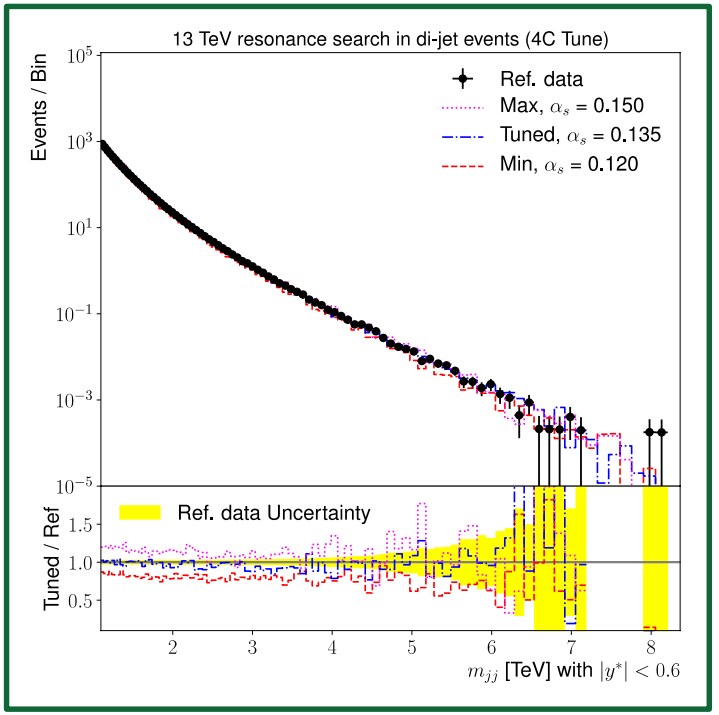
Developed novel automated HPC workflows to tune event generators with search data, expanding the data that can be used for tuning, allowing specialized tunings for new physics searches at the LHC.

Significance and Impact

Greater reach for new physics results because of improved background predictions.

Research Details

- Tuning will be possible on a wider kinematic region using search data.
- Use HPC to generate simulated data, use simple fast simulation to model detector effects, and tune directly to search data.
- The tuned results agree well with the search data (see the plot).
- Proof of principle with fast simulation: use the same HPC workflow to tune the simple fast simulation. A first step towards tuning of Geant4 simulation. Can be extended to provide an LHC search-data based fast simulation tune



Di-jet invariant mass distribution used for di-jet resonance search. Black dots are the observed ATLAS data. Blue line is simulation data after tuning strong coupling constant (αs), the other two lines are simulation data without tuning.