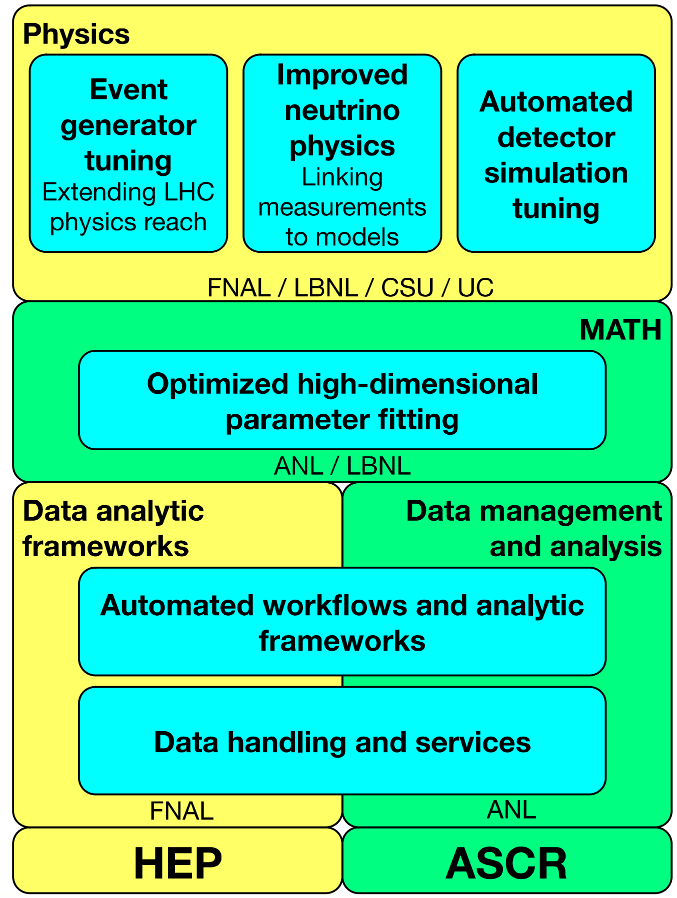
SciDAC-4: HEP Data Analytics on HPC

Project description and achievements

Jim Kowalkowski

Project Goals

- **Extend physics reach of LHC and neutrino experiments**
  - Event generator tuning (ATLAS/CMS)
  - Neutrino oscillations (NOvA/DUNE/SBN)
  - Neutrino cross-sections (NOvA/DUNE/SBN)
  - Detector simulation tuning (ATLAS)
- **Accelerate HEP analysis on HPC platforms, transforming how physics tasks are carried out in conjunction with ASCR math and data analytics**
  - High-dimensional parameter fitting,
  - Workflows supporting automated optimization
  - Distributed dataset management storage and access (*in situ*) for experiment data
  - Introduction of data-parallel programming within analysis procedures



Collaboration

- **PIs and senior researchers**
  - LHC and neutrino physics: **N. Buchanan** (Colorado State University, NOvA/DUNE), **P. Calafiura** (LBNL,ATLAS), Z. Marshall (LBNL,ATLAS), S. Mrenna (FNAL,CMS), A. Norman (FNAL, NOvA/DUNE), **A. Sousa** (University of Cincinnati, NOvA/DUNE)
  - Optimization: S. Leyffer (ANL), J. Mueller (LBNL), H. Schulz (University of Cincinnati),
  - Workflow, Data Modeling: M. Paterno (FNAL), T. Peterka (ANL), **R. Ross** (ANL), S. Sehrish (FNAL)
- **Research Associates and graduate students:**
  - A. Austin (ANL), S. Calvez (Colorado State University), P. Carns (ANL), P.F. Ding (FNAL), M. Dorier (ANL), D. Doyle (Colorado State University), X. Ju (LBNL), R. Latham (ANL), M. Wospakrik (FNAL), S. Snyder (ANL), J. Todd (University of Cincinnati),
- **J. Kowalkowski – PI (FNAL)**

http://computing.fnal.gov/hep-on-hpc/

NOvA Neutrino + Antineutrino Analysis

Scientific Achievement

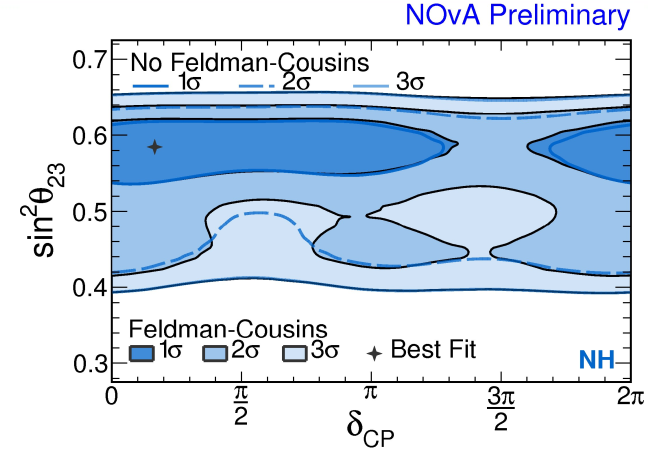
Large-scale analysis campaigns carried out at NERSC for the first time, in support of the first set of electron antineutrino appearance results shown June 4th at the Neutrino 2018 conference

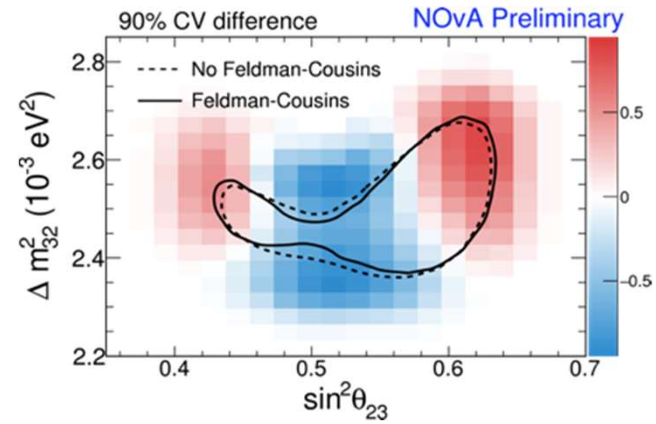
Significance and Impact

Most precise measurement of antineutrino oscillations to date. 8x higher resolution than any prior NOvA result. 50x faster than any previous result, reviewed by collaboration in <24h (not possible without SciDAC)

Research Details

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





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

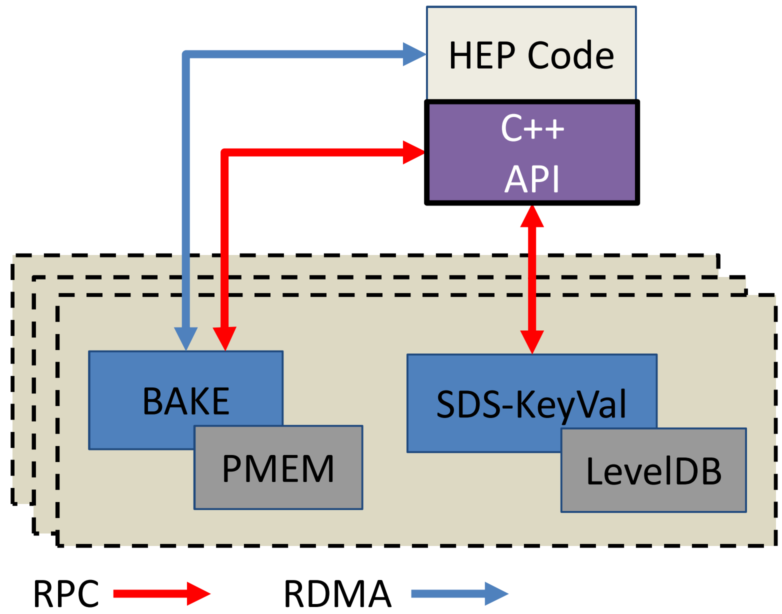
A custom data service for HEP, optimized for state-of-the-art HPC systems.

Significance and Impact

Bypasses file system to accelerate data access throughout analysis process (a major bottleneck). Extend the physics capability of HEP experiments by allowing fast data access through current HEP software frameworks.

Research Details

- Designed to seamlessly integrate into HEP software frameworks e.g. *art*
- 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
  - Optimized for current and future ASCR facilities
- Manage physics event data from simulation and experiment, through multiple phases of analysis



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

art is described at http://art.fnal.gov

U.S. DEPARTMENT OF ENERGY | Office of Science — Fermilab — Argonne National Laboratory — Lawrence Berkeley National Laboratory — University of Cincinnati — Colorado State University

Tools for HPC-scale physics analysis

NOvA/DUNE

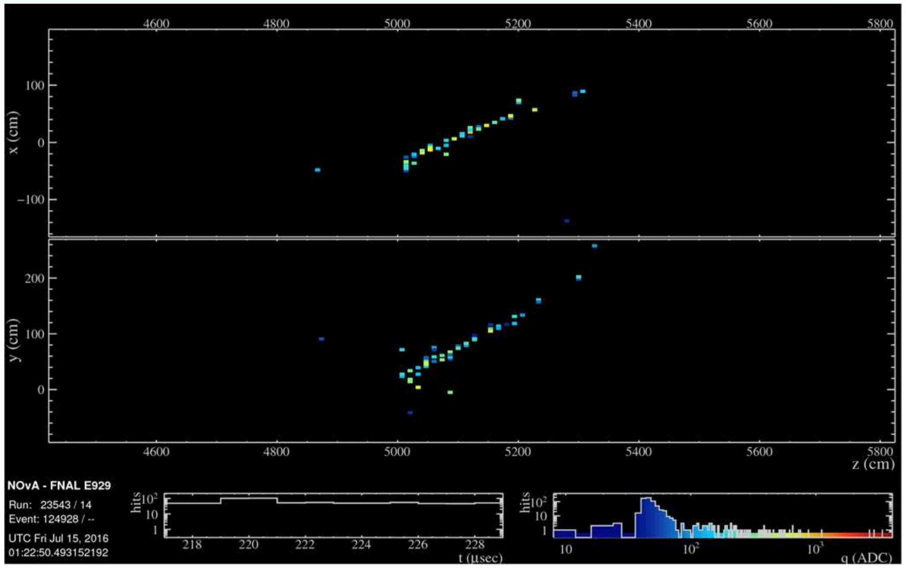
Scientific Achievement

We have developed 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 measurements by tuning of event-selection and evaluation of systematic errors.

The framework module and event selection code developed for this have been adopted by the NOvA collaboration.



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

Research Details

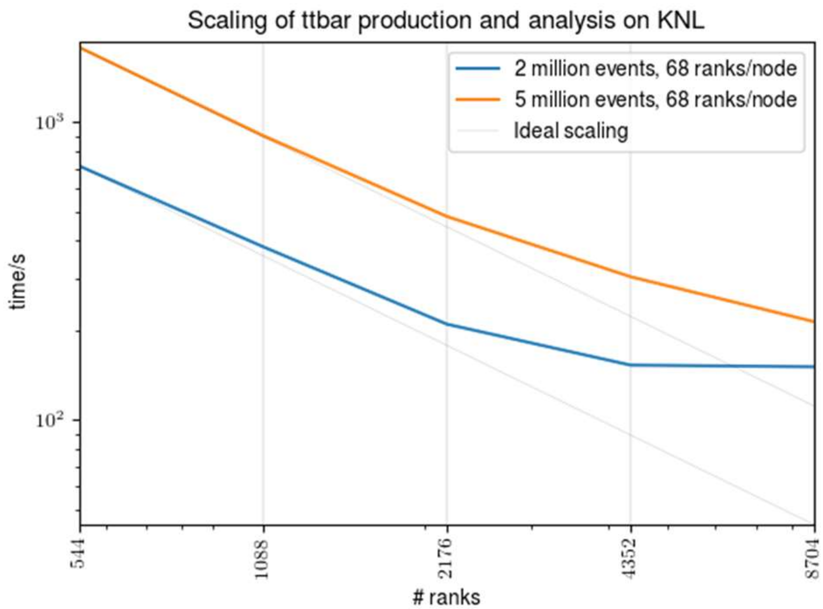
- Demonstrated parallel processing of liquid Argon data: 42 TB read in <20 seconds, using 76,800 KNL cores on Cori at NERSC, and three minutes to perform noise reduction stage on the entire LArIAT dataset.
- Demonstrated conversion of more than 4 TB of analysis data from NOvA's HEP-traditional analysis data format to our optimized HDF5 tabular organization (target is machine learning and event selection).
- Includes new tools and methods for writing analysis codes, aimed at easing development and maintenance, enabling advanced multivariable selection criteria not before possible.
- Using hephpc::hdf5 C++ library for writing HDF5 files from traditional HEP analysis programs.

https://bitbucket.org/fnalscdcomputationalscience/hep_hpc
M.Paterno, *et al*. 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 parallel, scalable application for generating and analyzing Monte-Carlo (MC) events on HPC architectures. This application efficiently utilizes HPC resources and integrates with HEP community tools.



Scaling behavior when generating $t\bar{t}$ 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.

Significance and Impact

Allows for extremely short turn-around time of large parameter space explorations, e.g. generator tuning.

Research Details

– Data parallelism with ASCR DIY library encapsulates all MPI communications.
– Implements full chain of event generation with Pythia8 and analysis with Rivet.
– Paves the way for new and advanced optimization algorithms.
– Speed is important: HEP phenomenology and all experiment simulation workflows require vast numbers of MC generator events.
– Can also be used transparently to accelerate analyses on laptops.

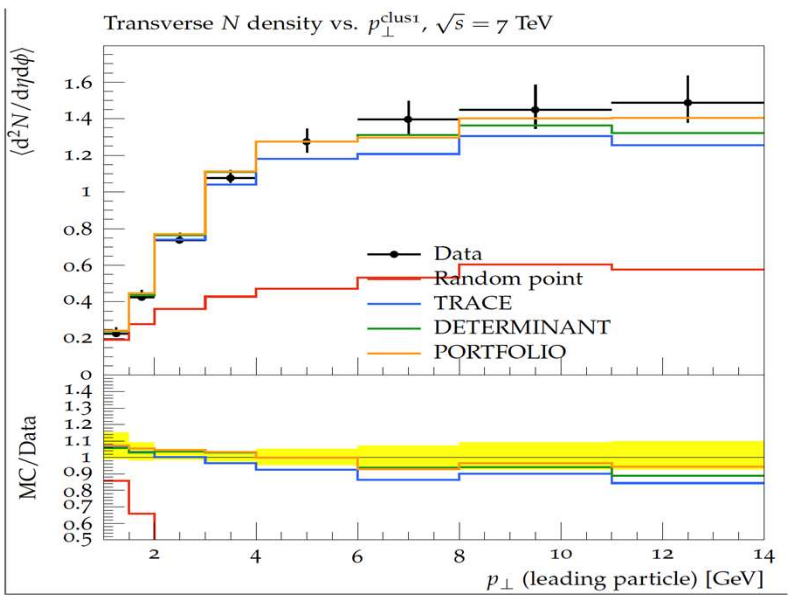
Currently published on BitBucket at https://bitbucket.org/iamholger/pythia8-diy/wiki/Home.

Automated physics event generator tuning

ATLAS/CMS

Scientific Achievement

We have developed an algorithm and implementation for the automated tuning of event generators.



Optimization outcome: shown is the observed data (black) and the simulation model predictions (solid lines) form three different outer optimization objective function models (portfolio, trace, determinant) and a randomly chosen solution (upper graphs). The lower graph shows how well the function predictions fall within the uncertainty band of the observations (the yellow area).

Significance and Impact

This approach allows for more robust theoretical predictions at the LHC. Better tunes are obtained more rapidly.

Research Details

- Current inefficiencies and potential biases in the treatment of observables is addressed through bilevel optimization.
- The *inner* optimization uses PROFESSOR, a HEP community tool, to optimize the event generator parameters.
- The *outer* optimization utilizes surrogate models to choose optimal weights to assign to the different experimental datasets.
- Several approaches are implemented: trace, determinant, and portfolio optimization.

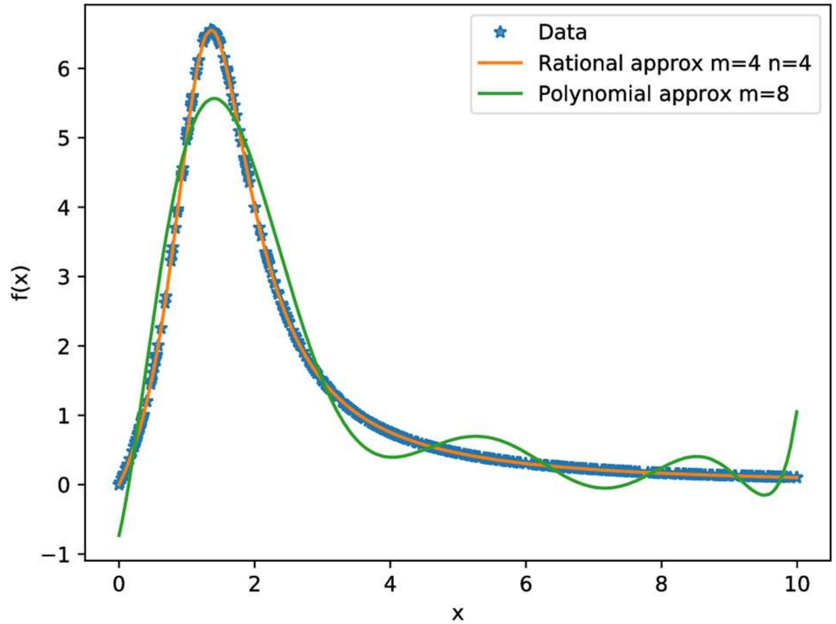
H. Schulz, *et al*. *Teaching PROFESSOR new math*. Paper presented at CHEP 2018, Sofia, Bulgaria. To be published in EPJ Web of Conferences (2019).

Rational approximations for physics signal modeling

ATLAS/CMS

Scientific Achievement

Developed an algorithm and implementation that reliably calculates multivariate rational approximations.



Demonstration of the superiority of the rational approximation algorithm. The input data (blue) is compared with the predictions from a rational approximation (orange) and a polynomial approximation (green). The oscillations as well as the unsatisfactory predictions at the edges for the polynomial interpolation has restricted the application of surrogates in some cases, such as BSM studies where rational functions are frequently observed.

Significance and Impact

Can now apply surrogate models to a wider variety of problems in HEP, e.g. signal modeling for dark matter direct-detection simulations. A model with far fewer parameter can be used to fit the data well.

Research Details

- Polynomial approximations break down if data exhibits traits of rational functions.
- Our algorithm calculates numerator (order *m*) and denominator (order *n*) polynomials simultaneously.
- Numerical stability achieved through SVD.
- Automatic detection and rejection of solutions with poles in the interpolation domain.

Generator tuning on unexploited data

ATLAS

Scientific Achievement

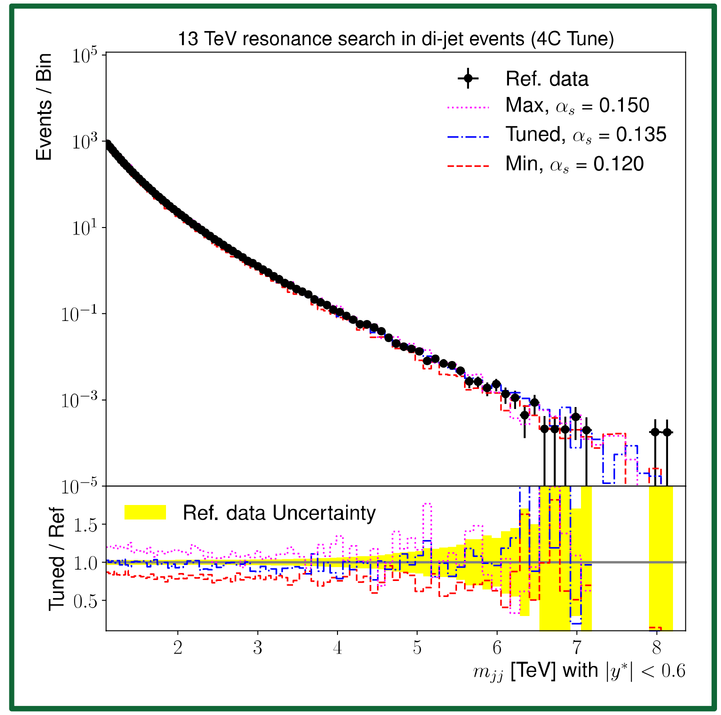
We developed novel, automated HPC workflows to tune event generators and fast detector simulations using the datasets used for BSM searches.

Significance and Impact

This yields improved background predictions and thus greater reach for new physics results.

Research Details

- Tuning directly on control regions of new physics searches removes the need for extrapolation.
- Expands the available data and allows specialized tunings for new physics searches at the LHC.
- Full chain of simulating data, approximating detector effects, and tuning directly to search data performed on HPC facilities.
- This is a proof of principle using fast simulation, a first step towards tuning of Geant4 simulation.



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.

Thank you

Accomplishments

- **First NOvA neutrino oscillation analysis using NERSC**
  - Time-to-result improved by 50x; first round completed within 16 hours
  - Used ~30M hours on Cori (and part of Edison) across two runs
- **Prototype event store (HEPnOS) built for serving data to HEP analysis codes**
- **Data-parallel NOvA pre-analysis event selection procedure**
  - NOvA accepted ownership of HDF conversion software for their data
- **Improved understanding of ATLAS and CMS data through generator tuning with Pythia, Rivet, and Professor**
  - Evolution of generator tuning algorithms, optimization of data selection, and development of DIY workflow
  - Generator tuning on unexploited LHC jet data and detector simulation tuning