Analyzing NOvA Neutrino Data with the Perlmutter Supercomputer

Norm Buchanan* *norm.buchanan@colostate.edu*
Steven Calvez* *steven.calvez@colostate.edu*
Derek Doyle* *derek.doyle@colostate.edu*
V Hewes‡ *vhewes@fnal.gov*

Alexander Himmel† *ahimmel@fnal.gov*
James Kowalkowski† *jbk@fnal.gov*
Andrew Norman† *anorman@fnal.gov*
Marc Paterno† *paterno@fnal.gov*

Tom Peterka§ *tpeterka@mcs.anl.gov*
Saba Sehrish† *ssehrish@fnal.gov*
Alexandre Sousa‡ *sousaae@ucmail.uc.edu*
Tarak Thakore‡ *tarak.thakore@uc.edu*

Orcun Yildiz§ *oyildiz@anl.gov*

*Department of Physics, Colorado State University, Fort Collins, CO USA
‡Department of Physics, University of Cincinnati, OH USA
†Fermi National Accelerator Laboratory, Batavia, IL USA
§Argonne National Laboratory, Argonne, IL USA

***Abstract*—NOvA is a world-leading neutrino physics experiment that is making measurements of fundamental neutrino physics parameters and performing searches for physics beyond the Standard Model. These measurements must leverage high performance computing facilities to perform data intensive computations and execute complex statistical analyses. We outline the NOvA analysis workflows we have implemented on NERSC Cori and Perlmutter systems. We have developed an implicitly-parallel data-filtering framework for high energy physics data based on pandas and HDF5. We demonstrate scalability of the framework and advantages of an aggregated monolithic dataset by using a realistic neutrino cross-section measurement. We also demonstrate the performance and scalability of the computationally intensive profiled Feldman-Cousins procedure for statistical analysis. This process performs statistical confidence interval construction based on non-parametric Monte Carlo simulation and was applied to the NOvA sterile neutrino search. We show the NERSC Perlmutter system provides an order of magnitude computing performance gain over Cori.**

***Index Terms*—NOvA, physics, neutrino, sterile neutrino, statistics, Monte Carlo, Perlmutter, Jupyter, NERSC**

I. Introduction

Neutrinos are the second most abundant particles in the universe and play key roles in the development of the early universe, stellar dynamics, and sub-nuclear interactions. Since their discovery in 1956, neutrinos have been shown to exhibit quantum-mechanical phenomenon known as flavor oscillation, which has vast implications for fundamental physical theories of the origin and evolution of the early universe. These discoveries are made computationally possible by large computing facilities that can analyze petabyte-scale data volumes and perform the computations needed to link experimental operations to fundamental theories. Recently, the NOvA experiment has employed the new Perlmutter system [1] at National Energy Research Scientific Computing (NERSC) to measure neutrino interaction rates and constrain parameters for the active-sterile neutrino mixing model using data from the NOvA detectors.

II. The NOvA Experiment

The NOvA (NuMI Off-axis ν_e Appearance) experiment consists of two massive detectors that perform high speed imaging of the interactions of neutrinos coming from the world's most intense neutrino beam, which is produced at Fermi National Accelerator Laboratory (Fermilab) in Bativia, IL [2]. The NOvA Near Detector (ND) is a 222 ton, $4\,\mathrm{m} \times 4\,\mathrm{m} \times 15\,\mathrm{m}$ liquid-scintillator-based detector positioned $100\,\mathrm{m}$ downstream from the primary target. The NOvA Far Detector (FD) is a 14 kTon, $15\,\mathrm{m} \times 15\,\mathrm{m} \times 60\,\mathrm{m}$ in size detector, positioned in Ash River, MN at a site $810\,\mathrm{km}$ away from the primary target.

These site choices optimize the ability of the experiment to observe the transition $\nu_\mu \rightarrow \nu_e$ and subsequently extract parameters of the Pontecorvo-Maki-Nakagawa-Sakata (PMNS) mixing matrix, as well as the extended (3+1) active-sterile mixing model [3]. Sterile neutrino models are currently one of the most intensely-pursued "beyond the Standard Model" physics scenarios being investigated in the field of neutrino science.

The NOvA near detector has to date collected on the order of 3-million images of candidate neutrino interactions. These data, or events, are characterized by nested trees of computed and derived values resulting from both classical and machine-learning algorithms. They are organized into tuples using the `ROOT` [4] analysis framework and data representation libraries. Recently, the data have been organized into an additional representation based on `HDF5` [5] tables. These data organizations support the `CAFAna` [6] and `PandAna` [7]

NOvA data analysis frameworks, respectively. These data form the basis of the scientific analysis work.

III. The Role of High Performance Computing

Computational data analysis demands are exceeding the traditional grid computing resources that are available to NOvA due to the growth in volume of NOvA data and computational sophistication of the analyses. In order to match these computational needs, we have developed a number of highly-scalable workflows on the National Energy Research Scientific Computing (NERSC) platforms. These workflows, which exhibit near perfect strong scaling, have dramatically improved our time-to-results and scientific exploration capabilities by allowing us to exploit the parallelism of near-exascale computing platforms.

A. Notebook Analysis Workflow

In the near future, we plan to transition to a workflow based on dynamically provisioned, shared pool of HPC resources with high speed connectivity to enable real-time exploration of our massive data sets. We have developed accompanying analysis tools to support such a workflow using the new Perlmutter system at NERSC and JupyterHub, including an implicitly parallel histogramming package, and a framework for measuring neutrino cross sections in the NOvA ND. We demonstrate scalability of this workflow with a realistic example measurement based on NOvA simulation.

B. Event Selection and Monolithic Datasets

Neutrino analysis begins with a filtering the full datasets to identify events that match analysis-specific criteria based on event attributes and/or metadata. Compared to legacy analysis techniques which perform this selection serially over blocks of data corresponding to files on the storage system, the `PandAna` framework utilizes a fully data parallel approach. `PandAna` evenly distributes work across an arbitrary number of MPI ranks where each rank processes a subset of the data including filtering and transformation. The events satisfying the analysis' selection criteria can then be further processed or aggregated depending on the application.

The NOvA experiment maintains petabytes of detector data and similar amounts of supporting simulation. This data is stored on Fermilab's hierarchical disk/tape archival storage systems. The data corresponding to a neutrino analysis are traditionally stored across $\mathcal{O}(10^4)$ files, each approximately $10\,\mathrm{MB}$ to $100\,\mathrm{MB}$ in size to facilitate data transfer to compute nodes when using distributed grid computing systems. We have translated and merged data into a monolithic `HDF5` format. This data organization achieved an 11x improvement in compression factor compared to the multi-file organization. This data representation, combined with the `PandAna` framework achieved a greater than 10x speedup in event selection when run on the NERSC Cori system.

C. Feldman-Cousins Procedure

Bounded parameters in the PMNS oscillation models paired with a low rate of data in the FD require confidence intervals to be empirically determined following the profiled Feldman-Cousins (FC) prescription [8]. The profiled FC method is a nonparametric Monte Carlo procedure where simulated FD data are statistically-fluctuated to generate millions of "pseudo-experiments". A log-likelihood minimization is performed with each pseudo-experiment, independently, to generate 1– and 2–dimensional p-value surfaces on regular grids of PMNS parameter values. Constructing a complete surface requires $\mathcal{O}(10^9)$ CPU hours due to hour-long fit-times. The FC procedure can be executed only after all previous stages of analysis are completed, often shortly before major conferences. Such stringent computational requirements can only be fulfilled on supercomputing platforms.

D. Feldman-Cousins Analysis Implementation on NERSC Supercomputers

NOvA members in collaboration with the SciDAC-4 project have built a framework based on the DIY [9] block-processing model to efficiently perform our profiled FC procedure on HPC. The DIY model maps data over available computational units known as blocks. Each CPU core is considered as a DIY block, over which a pre-determined number of pseudo-experiment fits spanning a grid of PMNS parameter values are assigned. Small-scale benchmark tests are performed to estimate average pseudo-experiment fit time. Large-scale jobs duration and node configurations are decided accordingly with static load-balancing.

The NOvA software environment required to run the DIY analysis code is containerized for cross-compatibility between HPC and Fermilab grid systems. The container also stores inputs to the FC procedure to leverage high-speed, simultaneous file access for thousands of MPI ranks.

We present performance results from our recent FC campaign to estimate confidence intervals on sterile neutrino oscillation parameters of the 3+1 PMNS oscillation model. The campaign was executed on both Cori-KNL [10] and Perlmutter systems to obtain p–value surfaces over the following parameter spaces: (1) $\sin^2\theta_{24} - \Delta m^2_{41}$ and, (2) $\sin^2\theta_{34} - \Delta m^2_{41}$. Resources used for each system are shown in Table I. Compared to Cori-KNL nodes, an overall performance gain of 13x on Perlmutter-CPU nodes is observed to complete same amount of computations. Such a improvements arise from the Perlmutter CPU architecture and increased core-count per nodes. The results [11] were presented at the Neutrino-2022 conference.

TABLE I

Configuration	Nodes	Ranks	CPU Hours	Parameter Space
Cori–KNL	1682	114 376	20 M	1
Perlmutter	200	25 600	1.5 M	2

ACKNOWLEDGMENTS

This material is based upon work supported by the U.S. Department of Energy, Office of Science, Office of Advanced Scientific Computing Research, Scientific Discovery through Advanced Computing (SciDAC) program, grant 1013935. This material is based upon work supported by the Fermilab Laboratory Directed Research and Development program, LDRD2016-010. This material is based upon work supported by the U.S. Department of Energy, Office of Science, Office of Workforce Development for Teachers and Scientists, Office of Science Graduate Student Research (SCGSR) program. The SCGSR program is administered by the Oak Ridge Institute for Science and Education for the DOE under contract number DE-SC0014664. This research used resources of the National Energy Research Scientific Computing Center (NERSC), a U.S. Department of Energy Office of Science User Facility located at Lawrence Berkeley National Laboratory.

REFERENCES

[1] Perlmutter system configuration. [Online]. Available: https://docs.nersc.gov/systems/perlmutter/system_details

[2] M. A. Acero *et al.*, "Improved measurement of neutrino oscillation parameters by the NOvA experiment," *Phys. Rev. D*, vol. 106, p. 032004, Aug 2022. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevD.106.032004

[3] ——, "Search for active-sterile antineutrino mixing using neutral-current interactions with the NOvA experiment, collaboration = The NOvA Collaboration, journal = Phys. Rev. Lett., volume = 127, issue = 20, pages = 201801, numpages = 8, year = 2021, month = Nov, publisher = American Physical Society, doi = 10.1103/PhysRevLett.127.201801, url = https://link.aps.org/doi/10.1103/PhysRevLett.127.201801."

[4] R. Brun *et al.*, "root-project/root: v6.18/02," Aug. 2019. [Online]. Available: https://doi.org/10.5281/zenodo.3895860

[5] The HDF Group. (2000-2010) Hierarchical data format version 5. [Online]. Available: http://www.hdfgroup.org/HDF5

[6] C. Backhouse, "The CAFAna framework for neutrino analysis," 2022. [Online]. Available: https://arxiv.org/abs/2203.13768

[7] Groh, Micah, Buchanan, Norman, Doyle, Derek, Kowalkowski, James B., Paterno, Marc, and Sehrish, Saba, "Pandana: A python analysis framework for scalable high performance computing in high energy physics," *EPJ Web Conf.*, vol. 251, p. 03033, 2021. [Online]. Available: https://doi.org/10.1051/epjconf/202125103033

[8] M. A. Acero *et al.*, "The Profiled Feldman-Cousins technique for confidence interval construction in the presence of nuisance parameters," 2022. [Online]. Available: https://arxiv.org/abs/2207.14353

[9] T. Peterka, R. Ross, W. Kendall, A. Gyulassy, V. Pascucci, H.-W. Shen, T.-Y. Lee, and A. Chaudhuri, "Scalable parallel building blocks for custom data analysis," in *Proceedings of Large Data Analysis and Visualization Symposium LDAV'11*, Providence, RI, 2011.

[10] Cori system configuration. [Online]. Available: https://docs.nersc.gov/systems/cori

[11] J. Hartnell, "New results from the NOvA experiment," 2022, Neutrino 2022. [Online]. Available: https://indico.kps.or.kr/event/30/contributions/880/attachments/172/367/Jeff%20Hartnell.pdf