Software and Computing for Small HEP Experiments

Maria Elena Monzani, Snowmass Seattle, July 23 2022

Workshop and Paper

Software and Computing for Small HEP Experiments

15-16 November 2021
US/Pacific timezone

https://indico.physics.lbl.gov/event/1756/

Software and Computing for Small HEP Experiments[1]

Dave Casper (editor),[1,a,b] Maria Elena Monzani (editor),[2,3,4,c,d] Benjamin Nachman (editor),[5,b,e] Costas Andreopoulos,[6,7,f,g] Stephen Bailey,[5,h] Deborah Bard,[5,8] Wahid Bhimji,[5,8] Giuseppe Cerati,[9,i] Grigorios Chachamis,[10] Jacob Daughhetee,[11,j] Miriam Diamond,[12,k] V. Daniel Elvira,[9,l] Alden Fan,[2,3,c] Krzysztof Genser,[9,m,n,o] Paolo Girotti,[13,p] Scott Kravitz,[5,c] Robert Kutschke,[9,m] Vincent R. Pascuzzi,[14] Gabriel N. Perdue,[9,q] Erica Snider,[9,i] Elizabeth Sexton-Kennedy,[9,n] Graeme Andrew Stewart,[15,b,n] Matthew Szydagis,[16,c] Eric Torrence,[17,a,b] Christopher Tunnell[18,r]

Contributed whitepaper:

https://arxiv.org/abs/2203.07645

Why a "small experiments" workshop?

Personal Perspective:

- I am the deputy operations manager (S&C) for LZ
- LXe-based direct dark matter experiment @ SURF
- LZ data is stored and processed at NERSC+UKDC

Data throughput (order of magnitude):

- Fermi-LAT (>2008): 0.3 PB/year
- LZ (2021-2026): 1-1.5 PB/year, 5 years
- ATLAS (>2010): 3.2 PB/year (raw)

Key Challenges for Dark Matter:

- No "tradition" for analyzing data at this scale
- Extreme "needle in a haystack" problem (10^{-9})

Why a "small experiments" workshop?

This workshop is centered around software and computing for the "small" experiments in our community. The mandate for this workshop is:

- Identify unique computational challenges of the "small" experiment community
- Gather input about what is needed in terms of computation for these experiments to be successful
- Connect members of the "small" experiment community to the computational frontier in Snowmass and encourage participation in topical groups
- Foster the development and re-use of open-source software, building on the work of the HEP Software Foundation and other collaborative efforts within the community

In order to be inclusive, we are not imposing a definition of "small" and have asked experiments to self-select.

Organizing Committee

Program committee:

- Dave Casper (FASER)
- Maria Elena Monzani (LZ)
- Benjamin Nachman (CompF)

Committee members:

- Matteo Agostini (LEGEND)
- Stephen Bailey (DESI)
- Wahid Bhimji (CompF4)
- Giuseppe Cerati (CompF1)
- Jacob Daughhetee (COHERENT)
- Mariam Diamond (SuperCDMS)
- Renat Dusaev (NA64)
- Daniel Elvira (CompF2)
- Jurgen Engelfried (NA62)
- Thomas Langford (PROSPECT)
- Gabriel Perdue (CompF6)
- Amy Roberts (CompF5)
- Daniel Whiteson (CompF3)

Agenda for the workshop

Day 1: Experimental Session

09:00	**Welcome and intro**	*Maria Elena Monzani*	09:00 - 09:10
	FASER	*Eric Torrence*	09:10 - 09:35
	MicroBooNE	*Erica Snider*	09:35 - 10:00
10:00	**g-2**	*Paolo Girotti*	10:00 - 10:25
	Break		10:25 - 10:35
	LZ	*Maria Elena Monzani*	10:35 - 11:00
11:00	**COHERENT**	*Jake Daughhetee*	11:00 - 11:25
	DESI	*Stephen Bailey*	11:25 - 11:50
	Wrap up	*Maria Elena Monzani*	11:50 - 12:00

Day 2: Session on "Tools"

09:00	**Brief Introduction**	*Dave Casper*	09:00 - 09:10
	HEP Software Foundation	*Elizabeth Sexton-Kennedy*	09:10 - 09:30
	HPC and small experiments	*Deborah Bard*	09:30 - 09:50
	Event processing frameworks	*Alden Fan*	09:50 - 10:10
10:00	**Reconstruction Frameworks (focusing on LArSoft)**	*Giuseppe Cerati*	10:10 - 10:30
	Break		10:30 - 10:40
	Generators (focusing on GENIE)	*Costas Andreopoulos*	10:40 - 11:00
11:00	**A perspective on detector simulations as done by (smaller) HEP experiments (and related challenges)**	*Krzysztof Genser*	11:00 - 11:20
	Machine Learning	*Christopher Tunnell*	11:20 - 11:40
	Wrap up	*Dave Casper*	11:40 - 11:50

Day 1 Summary and Wrap-up: Experiments

Maria Elena Monzani, D. Casper and B. Nachman

Common computing challenges across experiments

- **“Small experiment” ≠ “small data volume” or “small computing problem”** (Erica Snider)
- **Common tools are critical for success**: Have tried to avoid writing anything from scratch! (Eric Torrence)
  - Long-term support for common tools is unclear (MEM: Geant4 specifically)
- Continue building **dedicated computing expertise** in the community:
  - Small experiments will never have deep benches of experts (Erica Snider)
  - Too few production experts and too much data to process (Paolo Girotti)
  - Challenges with recruiting/retention in the NESAP program (ME Monzani)
  - Tough to stay ahead of problems without designated personnel (Jake D.)
  - Important that their performance/career is evaluated appropriately (S. Bailey)

The role of HPC in small experiments

- Access to HPC: Increasingly being told to use it (Erica Snider)
  - Experience so far has shown this to require significant work
  - **Using an HPC center isn't "free"** for experiments (or that center - S. Bailey)
  - Need established mechanisms for getting time available for the community
  - No control over architecture evolution. Fast pace evolution (Monzani/Bailey)
  - Downtime considerations when supporting running experiments at HPC

Benefits	Challenges
Scale to tens of thousands of cores when needed	Stability
One stop shopping for daily operations, big yearly reprocessing, science analyses	Queue wait time
Account management for hundreds of collaborators	Lack of control over configuration, upgrades, policies
Extra services: jupyter, docker, realtime queue, interactive queue, globus data sharing, databases, collaboration accounts for productions	Sharing resources with thousands of users on unrelated projects (esp. when those users are breaking the system)

Ideally, we would like common tools for data handling

- **Data movement** is a challenge given the size of the datasets (Miriam Diamond)
  - Authentication between different sites (federated identities?)
  - Also: collaboration accounts for productions, especially cross-site (S. Bailey)
- Common **book-keeping** needs: a few experiments looking at RUCIO
  - Does RUCIO support data movement? Does it support everything we need?
- Homegrown workflow / **job management** (not ideal, but didn't find alternative)
- **Long-term data storage**: All production data is backed up on tape (P. Girotti)
  - When files are needed, pre-staging process copies them to disks (dCache)
  - Experiments compete in a queue. System is not designed for peak loads
  - Space fills out quickly! Strategies for reducing data usage/confusion (JD)

Crucial discussion: public data, software, data access

- No consensus from LZ on **making the data and software public** (ME Monzani)
  - XENON thinks software should be public! (Chris Tunnel)
  - Reproducibility of results is essential! (Jacob Daughhetee)
  - Planning for data release forces use of best practices with version control
  - Provides a 'frozen' version of the analysis that can always be returned to
  - DESI: all open source at https://github.com/desihub (Stephen Bailey)
  - Significant benefit: external contributions vetting DESI code on real data
  - National lab challenge: conversations of open-source seem orthogonally at odds with security/export control concerns (Spencer Fretwell)
  - **Access to common data and tools for "sensitive countries"** (M. Diamond)
- Bringing everyone under one umbrella is difficult. Especially true with lack of continuity in organizational positions held by younger members (J. Daughhetee)

Day 2 Summary and Wrap-up: Tools

Maria Elena Monzani, D. Casper and B. Nachman

U.S. DEPARTMENT OF ENERGY | Stanford University

SLAC NATIONAL ACCELERATOR LABORATORY

Benefits of software re-uses across experiments

- **HEP Software Foundation successes** in a number of projects
- HPC “Superfacilities” to share development and support resources
- **Small experiments do not have smaller requirements and require the same full-featured frameworks as much larger experiments**
- Successful deployment of generalized tools like LArSoft and ACTS into production
- Generators are inherently experiment-agnostic; GENIE widely adopted in neutrino community
- Geant4, with curated physics lists
- Diffusion of machine learning methods to diverse applications
- Community should **strongly advocate open-source software in HEP**

People are essential

- HSF “do-ocracy” model
  - After common tools created, must be supported and/or adapted to new architectures
  - Benefits to engagement of users with improvements and development of new features
- Foster and **support “Research Software Engineers”** specialized in creating more robust and performant software
  - Also training in software development for physicists
- Funding **agency support for key infrastructure elements** like frameworks, Geant4 (and its models), and generators is vital
- HPC computing revolution is being led by a very small cadre of experts
- Generational C++ vs python cultural divide

Multiple revolutions are underway

- Migration from experiment-specific code to reusable community tools
- Migration from CPU to GPU
- Migration from C++ to python
- Domain-knowledge based algorithmic analysis to machine-learning

“Small experiments” paper

https://arxiv.org/abs/2203.07645

Paper Outline maps well onto CompF recommendations

Contents

Key recommendation: HPC support for experiments

SLAC

A pressing question for both experiment teams and HPC center staff is how to optimize the entire end-to-end workflow, not just individual applications. The NESAP program at NERSC (the National Energy Science Research Computing center) partners with application development teams and vendors to port and optimize codes to new architectures and platforms. Lessons learned from this process are shared with the wider NERSC community via documentation and training. NESAP is evolving with the types of systems we deploy, and will need to support workflow optimisation in the future.

Programs like NESAP are extremely valuable to small experiments and we advocate for a continuation and/or expansion of this program. However, we also heard about difficulties in recruiting NESAP postdocs. This is discussed in more detail in Sec. 2.6.

Key recommendation: experiment-agnostic G4 support

SLAC

In the US, the HEP support for Geant4 comes from the computing operations budgets of large experiments and intensity frontier operations at Fermilab. This leads to a stove-piping of Geant4 support such that issues specific to those experiments are addressed because they are needed for the operation of the experiment. Other experiments that use Geant4 are left out in the cold. In addition, Geant4's common software, such as physics models, no longer receives any US maintenance funding. The physics Generators used in the field (eg. Pythia, GENIE, Madgraph, Sherpa) also suffer from lack of stable funding in a similar way. This is not sustainable for the long term viability of small experiments. **Long term, experiment-agnostic Geant4 support is critical for the success of small experiments.** At the workshop, we also heard that many small experiments do not update Geant4 versions frequently (or ever!), and so they are unable to take advantage of new developments, whereas maintaining many old versions by the Geant4 collaboration is unsustainable. This means that small experiments often use no longer supported Geant4 versions and miss out on recent physics developments and computational innovations, such as multi-threading or the ongoing integration with Graphical Processing Units (GPUs).

Key recommendation: experiment-agnostic G4 support

SLAC

It is therefore essential that there is **funding for permanent software and computing experts**. The careers of these researchers should be evaluated appropriately with rewards to efficiency, stability, and robustness. A more transparent approach to software development and data sharing would go a long way towards improving the career prospects of software and computing experts, as it would allow individuals to claim credit for their work and be evaluated appropriately. There is little funding for software maintenance, even though this is critical. One possibility is the creation of research software engineer (RSE) positions that have long term funding independent of experiments, but for the support of existing and planned experiments. Another area of opportunity is the increase of joint particle physics and data science appointments at universities, which have become marginally more common over the last decade. We also need to increase the community investment in general software literacy of physicists, through (continuing) education initiatives and collaborations with industry, national labs and academia.