HEP Software Foundation 2022 and towards 2023

Graeme Stewart, for the HSF Coordinators and Working Groups

Software and Computing Roundtable, 2023-01-17

HSF Organisation and Role

- The HSF exists to catalyze and enable common software efforts across high energy and nuclear physics
  - We do not own or allocate resources, so all work that is discussed is owned by those experiments/projects/teams
  - We are hugely grateful that people take the time to contribute to HSF and other community events and to share and grow their work in the wider context!
- HSF Coordination
  - Provides oversight and drives overall engagement
- HSF Working Groups and Activity Areas
  - Organising in key, focused topic areas for the field
- The HSF's role is one of an information conduit and meeting point
  - Report on interesting and common work being done
  - Forum for technical comments and discussion
  - Encourage cooperation across experiments and regions
  - Motivate the publication of summary documents or papers for future use or reference

A More Normal Year...

- In 2022 we had a gradual easing of restrictions related to COVID-19
  - In person workshops and conferences could start to happen again from Spring 2022
    - E.g., ICHEP in July, ACAT in November
  - More people were able to travel and meet with colleagues
  - Experiments resumed many in-person events
- All of this helped re-introduce a much needed face-to-face / coffee time / beer dimension to activities
  - This is essential to the long term health of our community
  - Student activities are particularly valuable, e.g., the CERN summer student programme
- However, the world is not the same as it was before
  - Virtual participation is now accepted for almost all events
    - Balance costs vs. quality of interaction
  - We are aware of environmental costs of travel, so when we do fly, we should maximise the benefits
- The HSF (and partners) has always had a strong distributed dimension and we continue to benefit from that, backed up by in-person interaction

Community Advocacy

- We continued to advocate for software in the community, with several talks at conferences and events
  - The HEP Software Foundation, SMARTHEP kick-off workshop, 24 November 2022, Benedikt Hegner
  - Sustainability and future of software frameworks, JENA Symposium, 5 May 2022, Graeme A Stewart
  - HEP Software Foundation and Software Project R&D, SWIFT-HEP Meeting, 24 March 2022, Graeme A Stewart
  - Software and Computing R&D, 30th International Symposium on Lepton Photon Interactions at High Energies, 14 January 2022, Graeme A Stewart
- In addition, the HSF submitted several papers and LOIs to the US Snowmass process, particularly in the Computing Frontier
  - Many US HSF colleagues involved
- We regularly give input to the LHC Committee at CERN with WLCG (LHCC)

HSF Workshops and Events

We got started again with a rich programme of workshops in 2022, organised, in many cases, with other partners:

- HSF Detector Simulation on GPU Community Meeting
  - https://indico.cern.ch/event/1123314/
- Analysis Ecosystem II Workshop
  - https://indico.cern.ch/event/1123314/
- PyHEP 2022 Workshop
  - https://indico.cern.ch/event/1150631/
- HSF - IRIS-HEP Workshop on Software Citations
  - https://indico.cern.ch/event/1212344/
- Future Trends in Nuclear Physics Computing
  - https://indico.bnl.gov/event/15089/
- MC4EIC
  - https://indico.bnl.gov/event/17608/

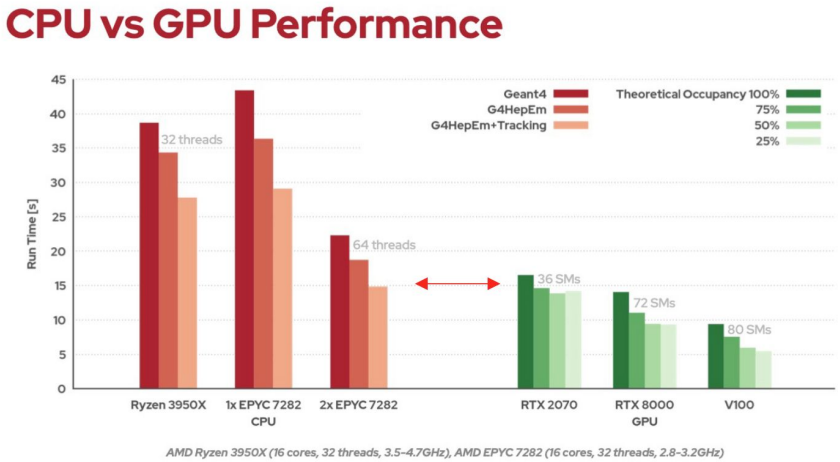
HSF Detector Simulation on GPU Community Meeting

- Increasing interest in GPUs for running HEP workloads
  - As these devices become more generally available at facilities can they be used for ‘generic’ HEP workloads?
  - Simulation an obvious candidate in terms of its huge resource consumption
- Two R&D projects in place
  - AdePT - CERN and UK SwiftHEP
  - Celeritas - ORNL
- Meeting organised by AdePT, Celeritas, Geant4 and HSF
  - Can we transform HEP particle transport to be efficient-enough on GPU?
  - How much effort will it take to create a production-level tool ?
  - What level of changes would be required to port key elements of the user code of production experiment simulation?

Challenges

- GPUs like homogeneous workloads
  - Particle tracking in inherently stochastic and divergent
- GPU memory accesses should be uniform for efficiency
  - HEP geometries are traditionally indirected and hierarchical
  - Particle creation and killing needs new memory and creates holes, respectively
- Both projects presented their status in terms of
  - Physics
    - Tension between capabilities and divergence
  - Geometry and Magnetic Field
    - Geometry is a particular bugbear
  - Integration
    - How to write scoring code
  - Prospects

Project Status



- AdePT performance is comparable on similar costing CPU and GPU
- Performance drops a lot on realistic geometries
  - New geometry code in development (surface based)
- Celeritas performance very similar
  - This translates into about x40 increase in events per second on very GOU heavy nodes (HPC centres)

- Many ideas shared and good communication between projects
  - No code shared yet, but could happen at a later stage
- Foresee another status update meeting this year

Note that other simulation R&D goes on as well and will hand an impact, e.g., the sub-event parallelism project that the eAST application (JLab/BNL) will rely on and helps with scheduling

HSF and IRIS-HEP Analysis Ecosystems Workshop

- Workshop held in hybrid mode at IJCLab
  - More than **70 people attended in person**
  - Held 5 years after the first workshop in Amsterdam
- Focused on 6 key topics for analysis
  - Analysis Facilities
  - ML tools and differentiable computing workflows
  - “Real-time” trigger-level analysis
  - Analysis User Experience and Declarative Languages
  - Analysis on reduced formats or specialist inputs
  - Metadata, bookkeeping and systematics handling
- HL-LHC was one focus, but not the only one
  - Run 3, Belle II, DUNE, ...

Topic Summaries

- **Analysis facility prototypes look fast enough now** (µs - ms per event)
  - AF focus now has to be on **ease of use** - for users and sites that deploy them
    - Many questions: scale-out, authentication, deployment complexity, user feedback, ...
    - Topics to be taken up in the HSF Analysis Facilities Forum
- **ML is much more widespread, becoming easier**, but still very dynamic; Autodiff is extremely interesting, but utility not yet established clearly
  - **Standard benchmarks** for performance will help
- User experience (UX) aims at reducing boilerplate and error prone/inefficient code
  - Do physicists need to do software engineering (and should they)? **There is training needed!**
  - **Bookkeeping and systematics remain pain points, as well as scale-out**
    - ROOT's .Vary() points the correct way
  - **Interoperability** between different ecosystem pieces is inconsistent
- **Reduced formats must to be used** to scale (NanoAOD, PHYSLITE)
  - Also need to **support the other analyses** - custom formats, dedicated skims?
  - Augmentation can be improved to only add for selected events
- Bookkeeping and systematics was discussed a lot in the UX context
  - Metadata paper reviewers suggested follow-ups, to be discussed in HSF
  - **Systematics challenge proposed**

Workshop Outcomes

- A few personal observations
  - Having an **in-person event was extremely productive**
    - Lots of opportunity for follow-on discussions and making contact with new people
  - We agreed that there is **one HEP analysis ecosystem**
    - ROOT and Scikit-HEP are both there and both highly engaged
- Workshop conclusions available on both Zenodo and arXiv [2212.04889]
  - Make columnar analysis easier with object facades
  - Tool interoperability should be strived for and used as a basis for training/onboarding
  - Open datasets are critical for performance evaluations (e.g. for ML models)
  - Metadata matters should be followed up at a dedicated workshop next year
  - Systematic uncertainties remain a major pain point for analysts - want common tools to make this easier and show how to use them in multiple experiments
  - Analysis facility work should continue, aiming to deliver an evaluation of solutions
    - **We have a very active group, The Analysis Facilities Forum, tracking specifically this topic and working closely with colleagues in, e.g., the IRIS-HEP Analysis Grand Challenge**

PyHEP 2022

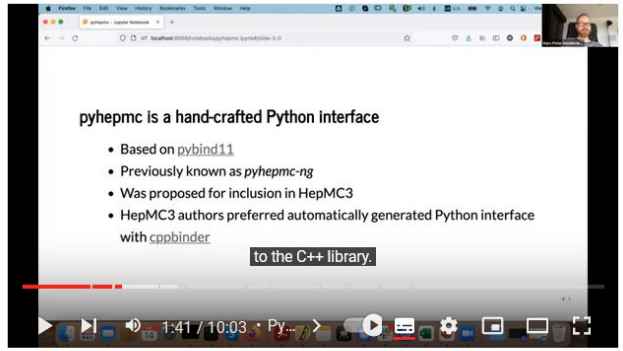
- PyHEP workshop series started 2018, bring together developers and users of Python packages in HEP
  - Recognising both the opportunity afforded by Python based data science tools and the needs to smoothly interface to create a coherent ecosystem (i.e. Analysis Ecosystem)
- First two workshops were in-person, ~70 people
- From 2020 the workshop, of necessity, went online
  - We hit an amazing vein of enthusiasm, with 1000 people registering
  - Many, many students, keen to learn and we introduced more didactic elements
- In 2022, again online: healthy integrated turnout of 420 though the week
  - But less than the 1000 who registered - free online events have low commitment, post-pandemic phase, so more in-person commitments for people?
  - Tried some novel forms of engagement: Remotely Green networking event and a ‘hackashop’ to encourage new developers

PyHEP2022 Highlights

PyHEP2022 Uproot Awkward Array hist Vector

Python usage in the basf2

- Steering file to arrange appropriate *modules* into a *path*, configure *modules'* options, and start data processing.
- Easy-understanding syntax for new users.
- basf2 *modules* can be written in C++ and Python.
  - Framework modules are written in compiled C++ codes.



PyHEP2022 pyhepmc a Pythonic interface to HepMC3

Recent Workshops

These workshops consist of modules that fall in the basics of the HSF Training Curriculum.

The topics covered during these are:

Software Basics Training
- Bash
- Git
- Python

Matplotlib
- Basics
- Styling
- HEP Specific

- Will hold the virtual PyHEP general workshop again this year
- However, want to have a **developer focused in-person** PyHEPdev event as well

HSF/IRIS-HEP Workshop: Software Citation and Recognition

- Workshop organised on Software Citation and Recognition
  - Review status of citation in HEP
  - Give credit to software developers and maintainers
  - Provide better and more sustainable software
  - Support for reproducibility
- Key principles developed by Force11 group
  - Importance, Credit and attribution, Unique identification, Persistence, Accessibility, Specicifity
  - Group then had task forces which helped to develop
    - Citation Format File standard (CITATION.cff)
    - CodeMeta
      - Metadata standard for software, a richer description of software

What to cite?

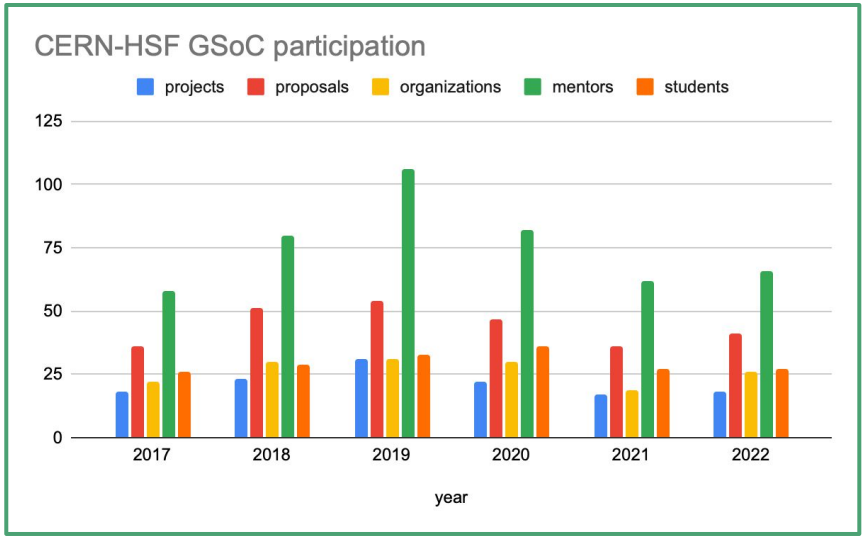
- An academic paper written about the software
  - This is the traditional approach, currently giving the most academic credit
    - Some feedback from RSEs - at least a subset don't like writing papers
  - There is a serious issue with ancestor papers picking up all citations
    - E.g., the 2003 Geant4 paper gets most citations - even though the code today is almost completely different and all the recent authors and contributors are missing
- The software itself
  - E.g., the Zenodo DOI
  - Not well rewarded academically
  - Does it describe why the software exists? The design choices?
- A combination of the two
  - E.g. the Journal of Open Source Software (JOSS)
  - Combining the code, plus a short paper describing the software
  - Code and repository is reviewed as well - has to meet best-practice standards like build instructions, basic tests, and user documentation

Citations and Recognition Outcomes

- Developers
  - If you want your software properly cited, put the citation everywhere...
    - In the README, in the documentation, on the distribution page (PyPI)
    - And make this a single source of truth!
  - Adopt a citation format file
    - CITATION.cff - first version can be easily generated via a webpage
  - Make sure you keep things up to date
- Experiments
  - Desire for consistency in citation recommendations - possibly curated by the HSF?
- Zenodo / Inspire
  - Better support for software citation coming this year, when automatically harvestable (e.g., from Zenodo INSPIRE HEP, CERN open data)
  - Will track citations ➜ credit for authors
- Workshop conclusions in preparation now

GSoC 2022

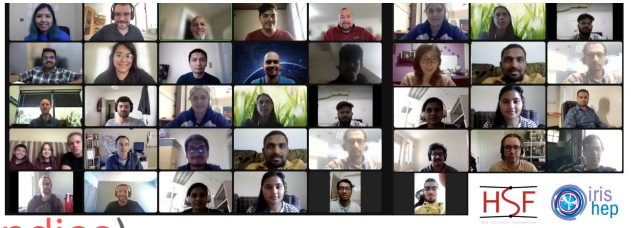
- We participated as an umbrella organisation again in GSoC
  - Programme improved last year: allowed to have short and long projects: 175 or 350 total coding hours
- Key numbers
  - 26 Organizations
  - 18 HSF projects
  - 27/41 proposals got a student (⅔)
  - 21/27 successful student projects
    - Record failure rate (22%)!
    - 2 withdrawn, 4 bad performance
    - Relaxed participation requirements (?)
- Student blogs available



GSoC 2023

- Same rules as for 2022, but even more open
  - Mix of medium (175 hours) and large (350 hours) projects
  - Flexible project duration
  - “Program open to students **and** to beginners in open source software development”
- Organization application deadline: February 7
  - We will need all project proposals by then!
  - A call for proposals will be made very soon
- CERN-HSF org admins
  - Benedikt Hegner + ?
    - Volunteers very welcome for this valuable task!

Training WG



Achievements unlocked:

- April 21-22: Matplotlib Training (indico)
- July 13-15: Software Carpentry Training (indico)
- July 25: Matplotlib Training Hackathon (indico)
- September 6: Containerization Training Hackathon (indico)
- Participation in ICHEP, PyHEP, Sustainable HEP conferences
- September 28-30: Software Carpentry Training (indico)
- October 11-13: Advanced C++ Training (indico)

Upcoming Quests:

- January 16-21: Analysis Preservation Training (indico)
- February 8-10: Software Carpentry training (indico)
- May 15-19: C++ Training - The American Edition @JLab (TBC)

Working Groups

- In addition to dedicated workshops and events we have many working groups and activity areas
  - Led by enthusiasts and advocates for common work and solutions
- These were active in 2022 and held many useful discussions in their field of expertise
  - See their Indico categories for more details on the meetings and topics covered

Working Groups

- Data Analysis
- Detector Simulation
- Frameworks
- Physics Generators
- PyHEP - Python in HEP
- Reconstruction and Software Triggers
- Software Developer Tools and Packaging
- HSF Training

Activity Areas

- Analysis Facilities Forum
- Conditions Databases
- Differentiable Computing
- Season of Docs
- Google Summer of Code
- intelligent Data Delivery Service
- Licensing
- Reviews
- Visualisation

Summary

HSF Website: https://hepsoftwarefoundation.org/
Main Forum Mailing List: HSF-Forum

- In 2022 HSF colleagues organised many events and discussions on important topics for High Energy and Nuclear Physics
  - Reinforcing the role of the HSF as a place where the community can gather for discussions and exchange of ideas
  - This feeds back into the work tackled by the software projects
  - We try to encourage diverse R&D, but also very practical solutions that deliver for the experiments
- In 2023 already look forward to another active year
  - A highlight will be the first in-person WLCG-HSF workshop since 2019, co-located with CHEP
  - Topics will be Analysis Facilities and Heterogeneous Computing