



HEP Software Foundation (R&D) activities

Eduardo Rodrigues - University of Cincinnati
On behalf of the HSF

The HSF and the Community White Paper



- The goal of the HEP Software Foundation (HSF) is to facilitate coordination and common efforts in software and computing across HEP in general
 - Our philosophy is bottom up, a.k.a. *Do-ocracy*
 - Also work in common with like-minded organisations in other science disciplines
- Founded in 2014, explicitly to address current and future computing & software challenges in common
- Finalised in Dec. 2017 a Community White Paper (CWP)
“A Roadmap for HEP Software and Computing R&D for the 2020s”
 - [arXiv:1712.06982](https://arxiv.org/abs/1712.06982)

“A Roadmap for HEP Software and Computing R&D for the 2020s”

HSF-CWP-2017-01
December 15, 2017

- 70 page document ([arXiv:1712.06982](https://arxiv.org/abs/1712.06982))
- 13 sections summarising R&D in a variety of technical areas for HEP Software and Computing
 - Almost all major domains of HEP Software and Computing are covered
- 1 section on Training and Careers
- Large support for the document from the community
 - > 300 authors from >120 institutions
- Journal submission to [Computing and Software for Big Science](#) has been made
- Attracted considerable attention
 - CWP article published in April in [CERN Courier](#)
 - Many presentations on it at major events

Contents

1	Introduction	2
2	Software and Computing Challenges	5
3	Programme of Work	11
3.1	Physics Generators	11
3.2	Detector Simulation	15
3.3	Software Trigger and Event Reconstruction	23
3.4	Data Analysis and Interpretation	27
3.5	Machine Learning	31
3.6	Data Organisation, Management and Access	36
3.7	Facilities and Distributed Computing	41
3.8	Data-Flow Processing Framework	44
3.9	Conditions Data	47
3.10	Visualisation	50
3.11	Software Development, Deployment, Validation and Verification	53
3.12	Data and Software Preservation	57
3.13	Security	60
4	Training and Careers	65
4.1	Training Challenges	65
4.2	Possible Directions for Training	66
4.3	Career Support and Recognition	68
5	Conclusions	68
	Appendix A List of Workshops	71
	Appendix B Glossary	73
	References	79

HSF post-CWP effort

- The Community White Paper laid down a roadmap for the future
 - Identifies main areas we need to invest in for future for our C&SW upgrade
- It was a major accomplishment made by the community, with HSF “coordination”
 - Process is concluded and has been a success
 - HSF, with its bottom-up approach, proved its worth in delivering this CWP
 - The process helped build a community (consensus) ... and the HSF itself as a recognised org.
- But the CWP is a milestone, not a final step
- HSF is far more than CWP & post-CWP activities to achieve goals set in CWP
- HSF activities reflect this ...
- ... and HSF collaborates or wants to collaborate with other initiatives
 - E.g., DIANA-HEP project around Data-Intensive ANALysis, CERN’s R&D programme on Experimental Technologies, etc.

The image features a large, stylized logo for 'HSF'. The letters 'H', 'S', and 'F' are rendered in a light pink, blocky font. A thick, flowing pink ribbon starts from the top of the 'S', loops around its right side, and then curves down to the bottom. The entire logo is set against a background of two horizontal grey lines that taper towards the center, with a grey circle at each taper point. The text 'HSF Events & Workshops' is overlaid in the center of the logo.

HSF Events & Workshops

HSF Events & Workshops

- Weekly coordination meetings
 - See <https://indico.cern.ch/category/7970/> for details
- Software Forum
 - Meetings to
 - Showcase common software projects
 - Introduce tools that help us face challenges like concurrency or vectorisation
 - Open dialogue with other like-minded communities
 - See <https://indico.cern.ch/category/10392/> for upcoming meetings
 - Suggestions for further meetings/topics? Email hsf-coordination@googlegroups.com
- HSF organises [workshops](#) and [WG meetings](#)
- Umbrella organisation for participation in Google Summer of Code programme



HSF/LPCC Generators Computing Workshop

- **Focus on the challenges for physics event generators**
that need to be faced in HEP in the next decade
 - Follow on from HSF Community White Paper
 - Need enhanced precision - generators which are both more compute intensive and have a larger spread of weights
 - Increasingly heterogeneous and concurrent processing landscape is a serious technical challenge
- Open to all interested parties, from the experiments and the theory community, but in particular to those who write and maintain the generator codes, from the major experiments and to experts in software engineering in HEP
- **November 26-27, 2018 @ CERN**
- You can [register](#) in [Indico](#) and [contact the organisers](#)

HSF/WLCG/OSG Workshop



- The next HEP Software Foundation Workshop will be **March 18-22, 2019 @ [Jefferson Lab](#)**
 - We will join with WLCG to tackle *software* and *computing* problems together
 - We will also co-host with Open Science Grid (OSG)
 - This presents some logistical challenges, but it's also an opportunity to join with another like-minded community
- More details to follow, but suggestions of good topics to cover are welcome
 - Parallel and plenary sessions are envisaged

Google Summer of Code



- Google-sponsored students working on open-source projects
 - Running since 2005
 - CERN EP-SFT active since 2011
- HSF acts as an umbrella organisation for many HEP institutes and projects
 - Makes it much easier for our institutes and software projects to get involved
 - Much lighter weight than going to Google directly
 - **This year we had 29 students accepted for projects** (amazing result, >2% of GSoC total)
 - ROOT, Geant4, Rucio, CVMFS, SixTrack, GoHEP, Falcon, YAMPL and many more
 - 26 students successfully completed their projects
 - Highlights:
 - Spark data analysis with PyROOT
 - Parallelised CNNs on GPUs
 - New JavaScript client for CVMFS

Copyright and Licensing






- We continue to work in this much **neglected area in HEP software**
 - Much code exists with no clear copyright or licence
 - The issues of large and deep stacks of experiments' software and license combinations were often neglected up to now
 - *Does impact on our ability to collaborate*
- LHC experiments continuing to be more open with their software
 - Goal is to maximise our useful user base and interactions with others
 - Good interactions outside LHC, e.g., with Belle II, who are also trying to resolve these matters
- **GPL licenses have become disfavoured** as they place obligations on any users
 - Can inhibit collaboration (e.g., industrial)
 - LHC experiments at large want non-GPL licences
 - Matches shifts at CERN, e.g., Indico moving from GPL to MIT
 - We made **significant progress** in moving packages like HepMC and DD4hep to *LGPL*
 - Widespread **use of GPL by theory community** still affects us greatly



HSF Working Groups

HSF Working Groups

- Data Analysis 
 - Detector Simulation 
 - Frameworks
 - Packaging
 - Reconstruction and Software Triggers 
 - Training
 - Visualization
-
- The 3 new WGs being set up are for the ‘mission critical’ areas
 - These groups should be helping to move the plans from the Community White Paper forward
 - Raise awareness of work being done in these areas in HEP
 - The mandates for the new WGs have been finalised
 - Want to nominate a convener? Email hsf-coordination@googlegroups.com by Oct. 29th ...

Working Groups ▾

Communication ▾

What are HSF working groups?

Data Analysis

Detector Simulation

Frameworks

Packaging

Reconstruction and Software Triggers

Training

Visualization

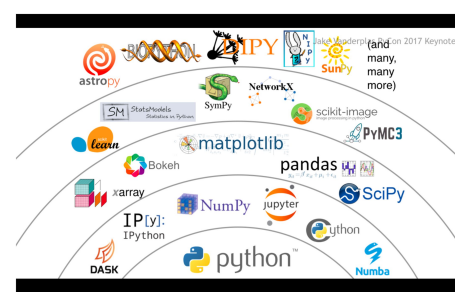
Data Analysis

- **Today we are dominated by many cycles of data reduction**

- Aim is to reduce the input to an analysis down to a manageable quantity
That can be cycled over quickly on ~laptop scale resources (when possible, which is not for all analysis cases)
- Key metric is 'time to insight'

- **Main R&D topics**

- How to **use the latest techniques** in data analysis that come from outside HEP?
 - Particularly from the Machine Learning and Data Science domains
 - Need ways to seamlessly interoperate between their data formats and ROOT
 - Python is the *lingua franca* here, thus guaranteeing our python/C++ bindings is critical
 - Functional / declarative expressions of analysis are more robust than imperative
- **New Analysis Facilities**
 - Skimming/slimming cycles consume large resources and can be inefficient
 - Grid resources not usually optimised for high I/O load analysis jobs
 - Dedicated analysis clusters may do better
 - Can **interactive data analysis clusters** be set up? SWAN, Spark, Dask interesting
 - Characterised by rapid column-wise access reads, with writes of new columns



R&D Outlook: needs coordinated work

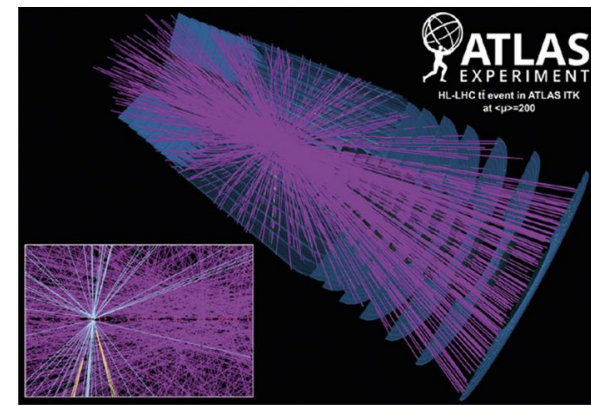
Detector Simulation

- **Simulating our detectors consumes huge resources**

- However, this is a recognised critical activity for LHC success
- Further gains needed for HL-LHC and intensity frontier experiments in particular

- **Main R&D topics**

- **Improved physics models** for higher precision at higher energies (HL-LHC and then FCC)
- Adapting to **new computing architectures**
 - Can a vectorised transport engine actually work in a realistic prototype (GeantV early releases)? How painful would evolution be (re-integration into Geant4)?
- **Faster simulation** - develop a common toolkit for tuning and validation of fast simulation
 - How can we best use **Machine Learning** profitably here? Multi-level approach, from *processes to entire events*
- **Geometry modelling**
 - Easier modelling of complex detectors, targeting new computing architectures



Data Management and Organisation



- **Data storage costs are a major driver for LHC physics today**

- HL-LHC will bring a step change in the quantity of data being acquired by ATLAS and CMS
- Notwithstanding **serious reductions** in the data stored by the experiments we need to optimise management and access

- **Main R&D topics**

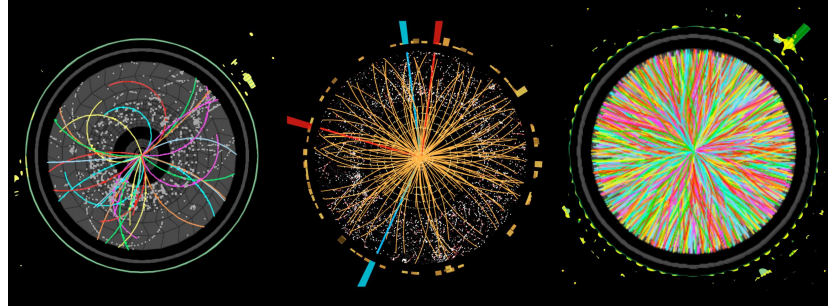
- Adapt to new needs driven by changing algorithms and data processing needs, e.g.,
 - The need for **fast access to training datasets** for Machine Learning
 - Supporting **high granularity access** to event data
 - **Rapid high throughput** access for a future analysis facility (previous slide on Data Analysis)
 - Efficient processing at sites with **small** amounts of **storage** (pre-stage buffer better than cache)
- **Consolidate** storage access interfaces and protocols
- Optimise storage with highly performant compression algorithms
- Support **efficient hierarchical access** to data, from high latency tape and medium latency network

Packaging



- One of the de facto areas of common interest between experiments
 - Building and deploying our software is a significant task and there is much duplicated effort
- WG decided to formalise the problem we are trying to solve
 - Use cases and test stack were established
 - Recognise that CVMFS and Containers simplified the problem a lot for us
 - Use cases can be enabled or become redundant as technology develops
 - We should be independent of site installed base OS
- Key drivers for the future:
 - Containers for deploying work at sites
 - CVMFS in *most* places (with work arounds where it's not)
- **R&D projects** looking at some of the directions for the future of packaging
 - Nix - pure functional package manager, build everything (really, even `libc`)
 - Portage - from the Gentoo Linux distribution
 - Spack - from LLNL, widely used scientific build orchestrator, very multi-version friendly

Event Reconstruction & Software Triggers

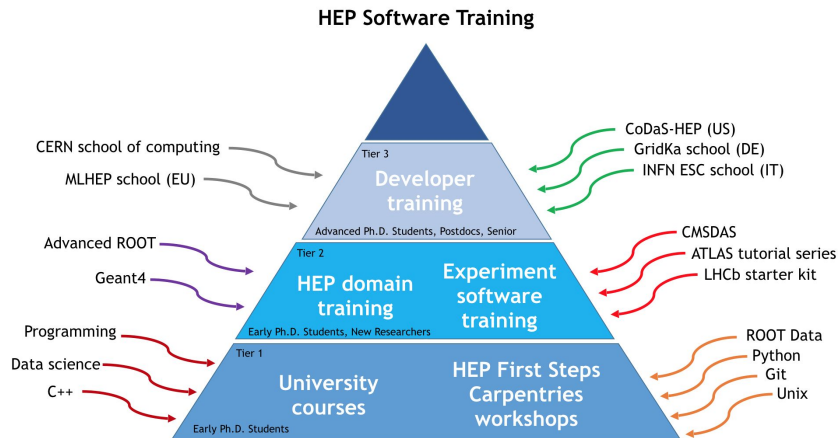


- **Move to software triggers is already a key part of the programme for LHCb and ALICE already in Run 3**
 - 'Real time analysis' increases signal rates and can make computing more efficient storage and CPU
- **Main R&D topics**
 - Controlling charged **particle tracking resource consumption** and maintaining performance
 - Do current algorithms' physics output hold up at pile-up of 200 (or 1000) and maintain low p_T sensitivity?
 - High granularity calorimeters bring a new type of detector into the tracking domain
 - Detector design itself has a big impact (e.g., timing detectors, track triggers)
 - Improved use of **new computing architectures**
 - Multi-threaded and vectorised CPU code, use of GPGPUs and possibly FPGAs
 - Robust **validation** techniques when information will be discarded
 - Using modern continuous integration, multiple architectures with reasonable turnaround times
 - **Reconstruction toolkits** can help adapt to experiment specificities: ACTS, TrickTrack, Matriplex
 - **Ideas and concepts can certainly be shared, code is a greater challenge**

R&D Outlook: A lot of projects in healthy states - keep up level of cooperation and sharing
(Connecting the Dots, Tracking Kaggle Challenge, etc.)

Training

- Recognition of **training ‘pyramid’**, from core skills to expert
- Organising some federation of training schools
- Working on a **curated set of training materials**
 - StarterKit organisers decided that the HSF is a good place to host common HEP training material - <https://github.com/hsf-training>
- Key point is to foster a community of trainers
 - We would like to establish closer links with the Software Carpentry community
 - Healthy interactions with many of the computing schools (Bertinoro, CERN School, GridKA) and with experiments (LHCb StarterKit)
- Will work with teams of 2 NSF projects recently funded
 - “Framework for Integrated Research Software Training in High Energy Physics (FIRST-HEP)”
3 year funding, <http://first-hep.org>
 - “Institute for Research and Innovation in Software in High Energy Physics (IRIS-HEP)”,
5 year funding, <http://iris-hep.org>



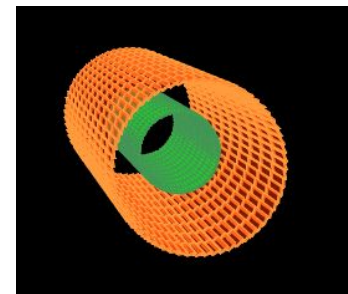
Other Working Groups

Software Development Tools

- Meeting on performance analysis software and how to share data
 - Common work on warehousing and visualisation possible
- Will also look at static analysers and grid tools

Visualisation

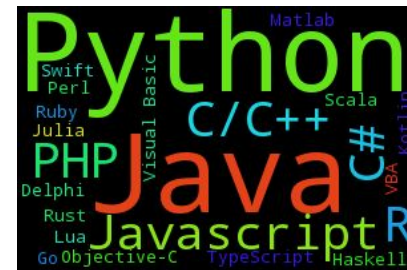
- Ed Moyses's WebGL event display now an HSF project (Phoenix)



Frameworks

- Pre-CHEP meeting focusing on data models (<https://indico.cern.ch/event/727646/>)
- Use of accelerators is a common concern; tools are of common interest
- Will continue to meet in context of Software Forum meetings

PyHEP - “Python in HEP”



- Python is a ‘first class’ language in HEP
- Traditionally an emphasis on developer productivity over code runtime
 - Popular in analysis and job configuration/steering
- Has become the lingua franca for data science and machine learning
 - Steering high performance backends gives excellent performance for the right problems
- HSF organised 1st event looking at the role of [Python in HEP](#)
 - Two-day [workshop](#) organised before CHEP, with 70 participants
 - Talks covered python ecosystem, LHC analysis, non-LHC experiments, C++ and ROOT bindings, distribution, education and training, core software, Python 3
 - Keynote from JupyterLab
- Post-workshop activities, building a community
 - Set up a [PyHEP Gitter channel](#) for informal exchanges
 - Created a community [inventory of packages, training materials](#) on Github



Thank you !