A Coordinated Ecosystem for HL-LHC Computing R&D

Workshop Notes

- Indico page.
- Meeting participants.

Wednesday, Oct 23

Workshop Introduction - Peter Elmer

- Focus is on how to build partnerships between various efforts in the community (now-funded IRIS-HEP, CCE, SciDAC, Ops programs, LDRD, other NSF projects).
- Questions we should address:
    1. How does the ensemble of efforts fit together? What are the gaps?
    2. How do the US efforts fit into the international context?
    3. How should the US R&D efforts be structured to impact planned updates to the HSF CWP and the S&C part of the US Snowmass process / HL-LHC TDRs?
- Questions:
    1. Can we include some discussion / overview of the AI program that DOE is starting to put together? A: We should; will try to find 5 minutes later today.

US ATLAS - Torre; supplemental slides (BNL activities not covered in other talks).

- S&C challenges:
    - CPU: GEANT4 SIM time; generator speedup needed; utilization of accelerators.
        - ATLAS assumes that there will be a generator speed up *and* an increase in precision
    - Disk: 2028 disk usage plans very efficient usage; no “opportunistic storage”; greater use of cheap(er) cold storage; rely on dynamic network usage modes.
        - Question (LatB): Do we understand why the ratio between SIM and RECO are so different between CMS & ATLAS? A: This ratio tends to fluctuate over time -- past RECO investments made % of SIM go up.
- R&D approach:
    - Factor in other elements: DOE HPC, NSF strategy, IRIS-HEP, etc.
    - Prioritized work program toward HL-LHC; orienting hiring accordingly.
- R&D WBS: Software reengineering, workflow porting to new platforms, distributed computing development.

- S/W re-eng: fast sim & fast chain; accelerator-enabled code; data layout & concurrent IO; event generation.
- Workflow porting: currently smallest area; current ERCAP allocation is relatively small. ML & FastChain are the target workflows now.
- Distributed computing dev: Data carousel, fine-grained event processing. Good example of collaboration with IRIS. Services needed to support ML.

US CMS - Oli

- Brian & Mike are the HL-LHC L2 managers for US CMS
- Ops program includes responsibilities for core fwk, workflow mgmt, submission infra, data mgmt; R&D tasks are somewhat "hidden" in there as part of continuous development.
- Working to establish a PostDoc program in 2020 - co-funded postdocs to carry out S&C R&D. Allows for succession planning and talent recruitment.
- Programmatic R&D efforts:
    - Efficiency of software: Multi-threaded framework (existing), Pileup premixing (existing), tracking on 'advanced architectures' (major area of activity).
    - Resource utilization & efficiency: Data streaming, analysis data formats, HEPCloud (portal to various CPU resource forms).
- CCE proposal: workflow, portable parallelization, generators, fine-grained IO.
- Things not well covered:
    - Analysis Facility Concepts
        - How do you enable a group of physicists to analyze a petabyte of physics data?
    - Storage Facility
        - Data lakes, data sinks, exabyte cold, warm and hot storage
    - Networking concepts and connected facilities

IRIS-HEP - Gordon

- Overview of the project, including governance and management.
- Question (LatB): Does the SSL include actual hardware? A: Yes - "adopted" the River cluster at UC.
- Comment (MikeS): Note there were two efforts that went in parallel prior to IRIS. CWP for the entire HEP community; S2I2 focused on the conceptualization of what a NSF-funded HEP software institute would look like (more focused than the entire community).

Other Contributions - Rob

- Emphasis on direct involvement & two-way communication (feeds into IRIS-HEP).
- (Whirlwind tour of 8 projects - see slides.)

HEP-CCE

- CCE - 4 labs (BNL, Berkeley, Fermilab, Agronne)

- Cover all three frontiers
- 4 orders of magnitude more resources available at the start of HL-LHC by the start of Run 4. If we look at previous usage patterns, one might think we get x100 more than we need, globally.
- We have to be ready for unpredictability
    - Hardware timescales: 2-3 years
    - Software timescales: 10-20 years
- Throughput in and out of the systems will be good - lots of people have a data intensive set of problems.
- Data structures must be built for parallel workflows
- 4 primary areas (pilot projects):
    - Pilot projects on concurrency/offloading (“Portable Parallelization Strategies”)
    - Data model/structure issues and IO/storage
    - Event Generation
    - Complex distributed workflows on HPC systems
- Questions:
    - Lothar - Is the scale of resources correct? Past experience (PIF) indicates:
        - PIF - $3M/year for HEP, was about the right amount
        - Here it is 6 experiments
        - Is the request big enough?
    - Mike S
        - Simulation (GEANT)?
        - We can look at this, but this is a very very hard problem to solve

ECP - Doug & Team

- (Long discussion - not captured in notes)
- ECP is a significant investment. Aims to “uplift” HPC application codes from current trendline to a new one.
    - Various codebases have agreed to deliver a 50x performance improvement.
- ECP is broad - encompasses software stacks
- Example of neutron

SLAC

- GEANT{4,V} support in the USA is going away? We should determine the exact/correct statement of support.

Thursday, Oct 24

Room 321:
https://docs.google.com/document/d/1tDXskA3VzLPJlhdNZc9cHfUg_i9MkMspdep8WxxImwM/edit?usp=sharing

Room 323:
https://docs.google.com/document/d/1WAhnxsezdYFAH5GT4VBfTxPxy0OP5wvPhz8M7_AX4aQ/edit?usp=sharing

Plenary:

Storage infrastructure and Facilities

- R&D Projects
    - Network caching infrastructures
        - Xcache - service? Or software?
    - Tape Buffer storage
        - Mas - warm storage in front of tape - remove the redundancy because you know it is on tape by definition
        - CMS has small single purpose cache in front of its tape archive
    - Tape Carousel
        - next step ; defining the tape bandwidth needed at the Tier1s
        - is this going to work for HL-LHC-ATLAS?
        - Have not done optimization of writing for reading yet
    - DOE Exascale storage - Storage roundtable to come from Ben and Hellen ; they want to broaden the conversation - within the next couple of months
        - What is the goal? - social engineering - exploration of possibilities - build the ecosystem - should we create plots
    - Storage hierarchy
        - Currently the experiments/community define the storage themselves. Is this what we're going to do in 5-10 years?
        - Can the labs take this on and figure out how to do this hierarchy?
        - Is this then passed down to vendors?
    - Storage scale
        - Number of tape drives for exabyte storage
        - Speed to recall exabytes of data
    - IRIS-HEP SSL, OSG-LHC, SLATE, SCAILFIN
        - Declarative, reproducible deployments - containerization of services
        - Federated facility operation, edge
        - On-prem to mulit-prem service mesh
        - Revisiting and re-engineering the WLCG/Tier2 fabric due to changing roles, analysis facilities, diverse services
    - Storage optimized for low-latency analysis platforms
    - WLCG QoS working group
        - Was meant to me data carousel working group
    - Access patterns vs. file optimizations

        - Spending file size for random access
        - Vs. putting it on tape
        - Transforming complete files?
    - Object stores
        - Improve processing
        - Currently loading of the storage is killer
        - CCE project
    - Storage for low latency scalable analysis systems
        - T2 centers
- Gap analysis
    -
- Opportunities
    - Get rid of the second copy of the tape for run4? Rely on transatlantic networking
- Milestones
    - Prepare use cases for DOE initiative
    - Small group to make cost analysis why we use tape?
    - Define APIs for interacting with cold storage that are needed to use this dynamically; more then “Is this on tape? , fetch it for me”
    - Requirements review, use cases and fundamental definition of our needs to give to facilities and vendors to come up with solutions ➜ baseline model

Data Analysis & Data/Analysis Preservation

- R&D Projects
    - IRIS-HEP
        - Data Query & Facility
        - Statistical Models (pyhf)
        - MC techniques
        - Fitting
        - Utilities (decay language, etc.)
        - Python ecosystem
        - (DIANA follow-up)
        - REANA
        - RECAST

- SCAILFIN
- Coffea
- HEP-Accelerate (CalTech)
- US-CMS: facilities and support for data analysis
    - spark
- DOE Lab SciDAC-4: HEP on HPC

    - Object stores, hdf5
    - Arrange an IRIS-Blueprint type workshop to engage therm?
- International:
    - ROOT
        - RDF
        - RNtuple
        - Combine the two?
    - CERN OpenLab: Spark exploration
    - SWAN
- End-to-end data challenges?
- Analysis Facilities:
    - Earliest milestone: blueprint workshop to sketch architecture(s) for such facilities
    - How to demonstrate multi-user capabilities?
        - Scalability?
        - Short term milestone for small-ish number of users
    - T3's (or T2s) with Kubernetes substrate
    - Impact of different storage architectures
    - Social engineering to generate user base?
        - Ease of use important
    - Where, how, how many?
    - who pays?
    - Distributed? Or centralized?

- Gaps:
    - Appropriate underlying hardware for analysis facilities
    - Low on effort
- Opportunities:
    - Expand user bases of prototypes to develop user communities
        - Generate feedback
    - Make sure appropriate data is available to attract people
    - Requires PI engagement and postdoc time
    - Connect with LCF program on analysis/viz facilities
    - Connect FNAL SciDAC effort with experiments
- Potential Milestones:
    - Schedule blueprint meeting
    - IRIS-HEP Data Challenge 2
    - CMS onboarding plan - mid 2020 have Spark facility running and have solved data problems
        - Do declaratively so it is portable

CCE Proposal Discussion

ECP engagement

R&D projects

- Fenced by project 413
- However all code generated here has to be deployed freely through Spack

Gaps

Opportunities: Software stack part of the project aims to have broader impacts

Potential Milestones

- Joint Atlas/CMS use cases and establish formal contact with ECP (what is the process?)
- Form a consortium on jointly supported products?

Prep work:

- Attend Annual meeting (needs invitation) -
- At what level of participation, should we make a statement as a community to be strategic? Community white paper? Need to develop use cases.
    - A mission for CCE?
    - An introduction meeting would be a first step - understand what they are already doing.
- Ask for a contact? Why isn't this CCE? Because you need the other side of the conversation that is not HEP. You need an IRIS-HEP contact as well. Needed for significant asks. First meeting to tour what technologies are available and then connect CCE with relevant smaller team leads.
- ECP software technologies are meant to be production hardened libraries for users
    - Help with using Tensor cores in Linear Algebra
    - Has ECP picked up HDF5? Exa-HDF5 is middleware
    - Compression algorithms - HEP not yet involved here.

https://drive.google.com/file/d/1HQWdqa9r6XNGBOtcL5SQXze4LRwcN2Da/view?usp=sharing

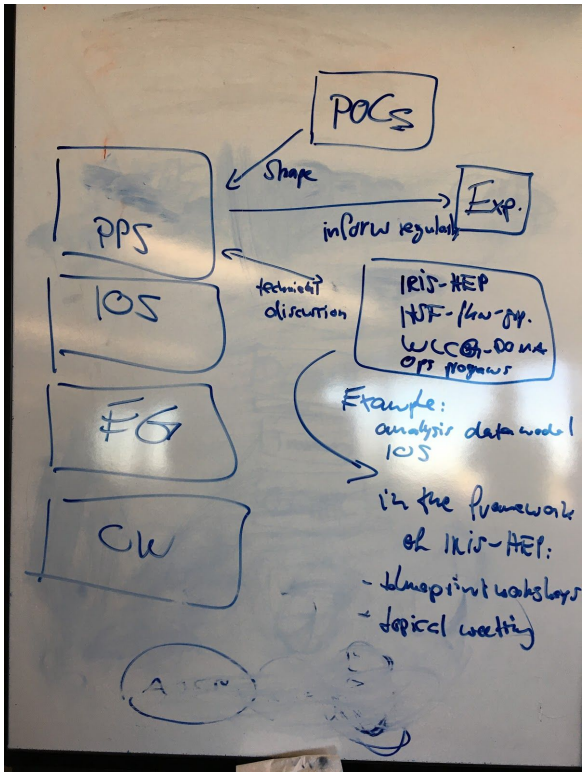
Milestones

- USATLAS and USCMS get together a group of people and preliminary list of areas of interest in ECP "products"
- ECP organizes more extensive briefing of ECP program
- USATLAS & USCMS define list of projects to work on
    - Which could result in bigger asks to ECP
- If approved, CCE could play a coordination role and could provide a home for discussions

- In parallel, explore on higher level possibility and process to get recognized by ECP, having an answer to funding agencies asking to get ECP effort working for HEP
- Yearly allocations do not work for HEP, we need programmatic allocations that also need to be higher to be impactful, discuss with ECP

CCE

- Once approved needs regular contact with experiments, and work can be reshaped depending on progress. Natural hook-in is in the framework groups. Where does it interface with IRISHEP?
- Debate on the value of fine-grained I/O - Deliverables are a set of guidelines.
- Clearly not enough effort for all areas - needs to be joint with research effort



Timeline

https://docs.google.com/presentation/d/1Gh0k7TwiZQHCPyu4q2YS31bKC-9tKCb6pzO7PQaJ_k0/edit#slide=id.g6f59d469dd_0_0

Friday, Oct 25

For posterity, here are the closeout slides from the 2017 meeting.

Some notes to help sort out the Weakness & Goals

- Data Analysis
  - * Appropriate underlying hardware for an analysis Facility undefined,
  - * Research vs Program effort
  - Integration with SSL-like substrate infrastructures
- Reco & Trigger
  - * Some R&D faces tension between Run 3 as testing ground and HPC facility timelines
  - * Involving subject matter experts in reengineering to ensure long term sustainability
  - * CWP focus areas un(der) covered - Real-time analysis beyond LHCb, and QA/QC (quality assurance / control) modernization
- DOMA
  - * Need an agreed-upon facilities model (esp. Integrating caching).
  - * Better define metrics to understand when we should transition to new models.
  - * Could use better alignment with SciDAC-4 & CCE plans.
- Storage Infrastructure
  - * Currently the experiments / community define storage hierarchies themselves. Is this what we're going to do in 5-10 years? Or are the facilities optimizing this for the experiments?
- Data Transfer and Networking
  - * Updating data transfer protocols ➜ coordinate OSG with ESnet studies
- Workflow and Resource Mgmt
  - * Integration of distributed resources with HPCs
  - * Usability of accelerators by HEP applications
  - -- Running data intensive applications on HPCs
- Event Processing Frameworks
  - * R&D Developers feel overly constrained by weight of existing frameworks and software
  - * Early choice forced by schedule constrains the field before sufficient R&D is done
- Physics Generators
  - -- These people's careers are judged on physics publications, not on how much they improve LHC generator efficiency
  - * No European FTE towards generator optimization

- Simulation
    - * Interactions with CERN and community on Geant are complex
        - CERN seems willing to let the US take the lead on Geant for GPUs
    - * Geant4 needs to be fully supported until GeantX is ready