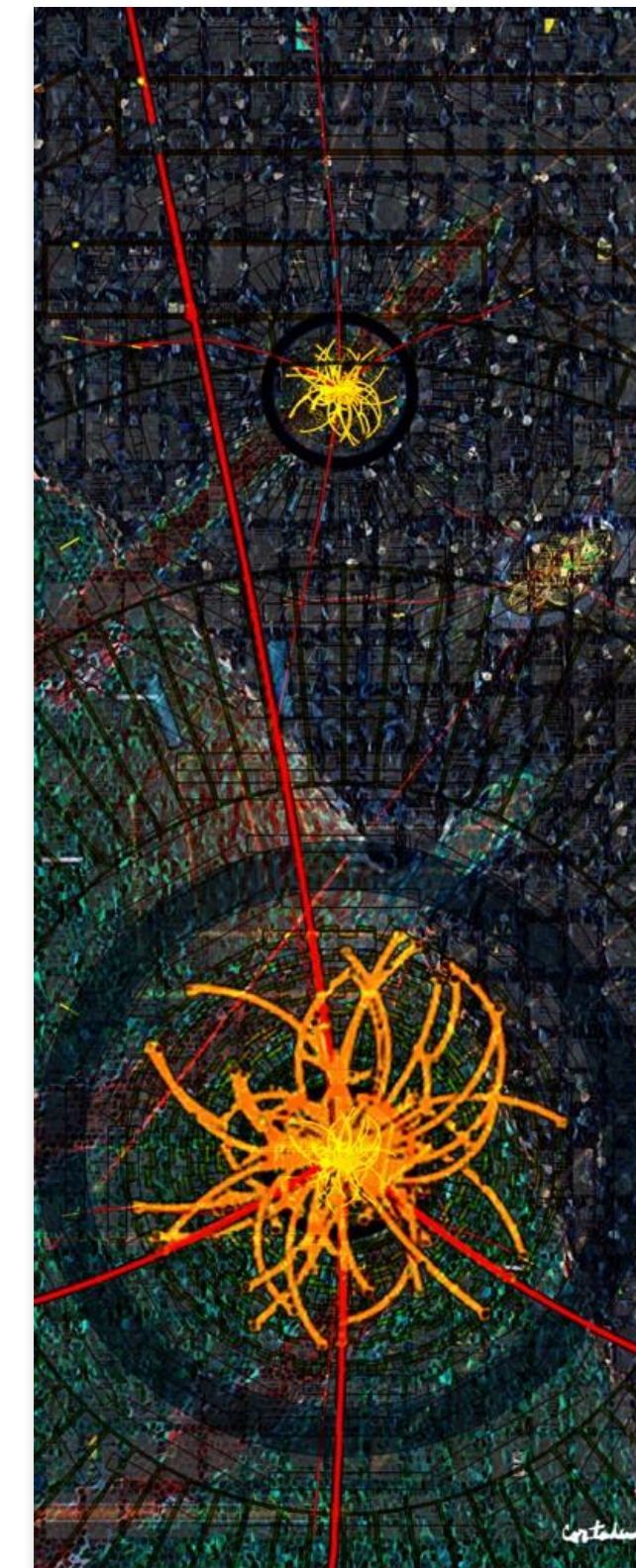
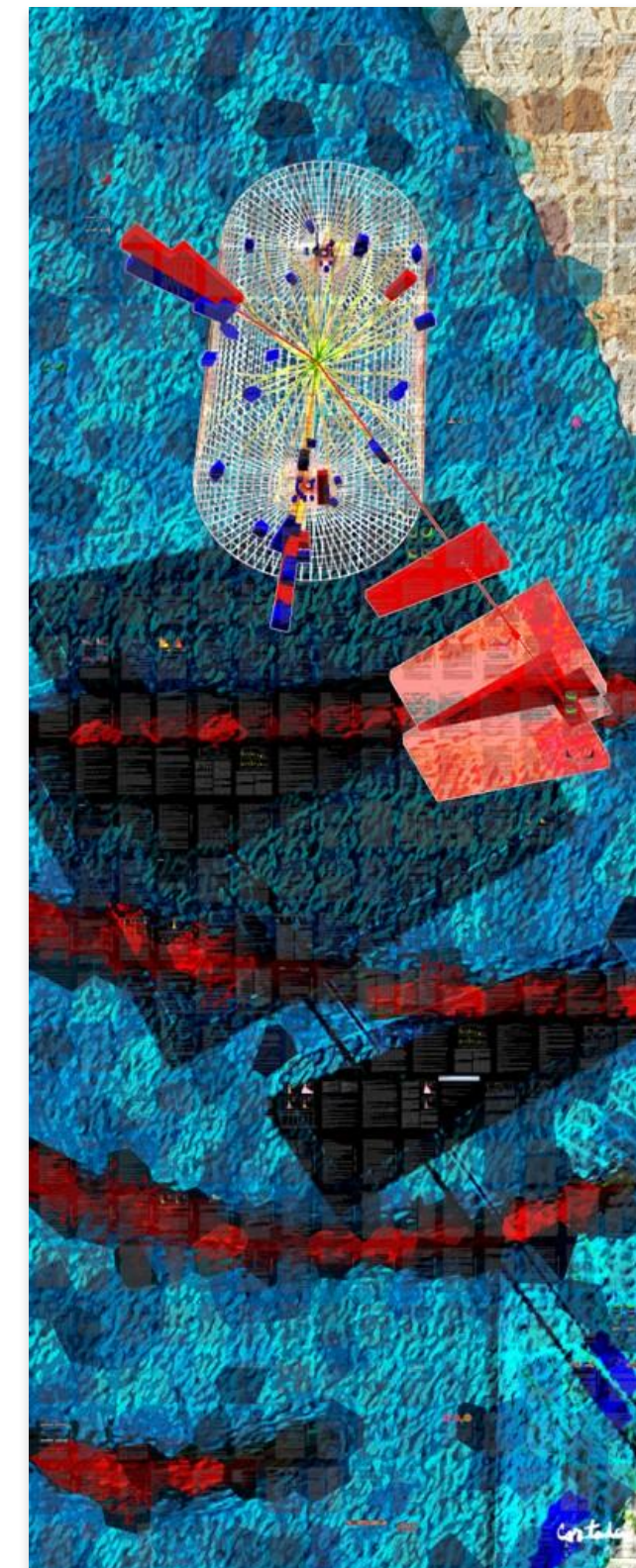
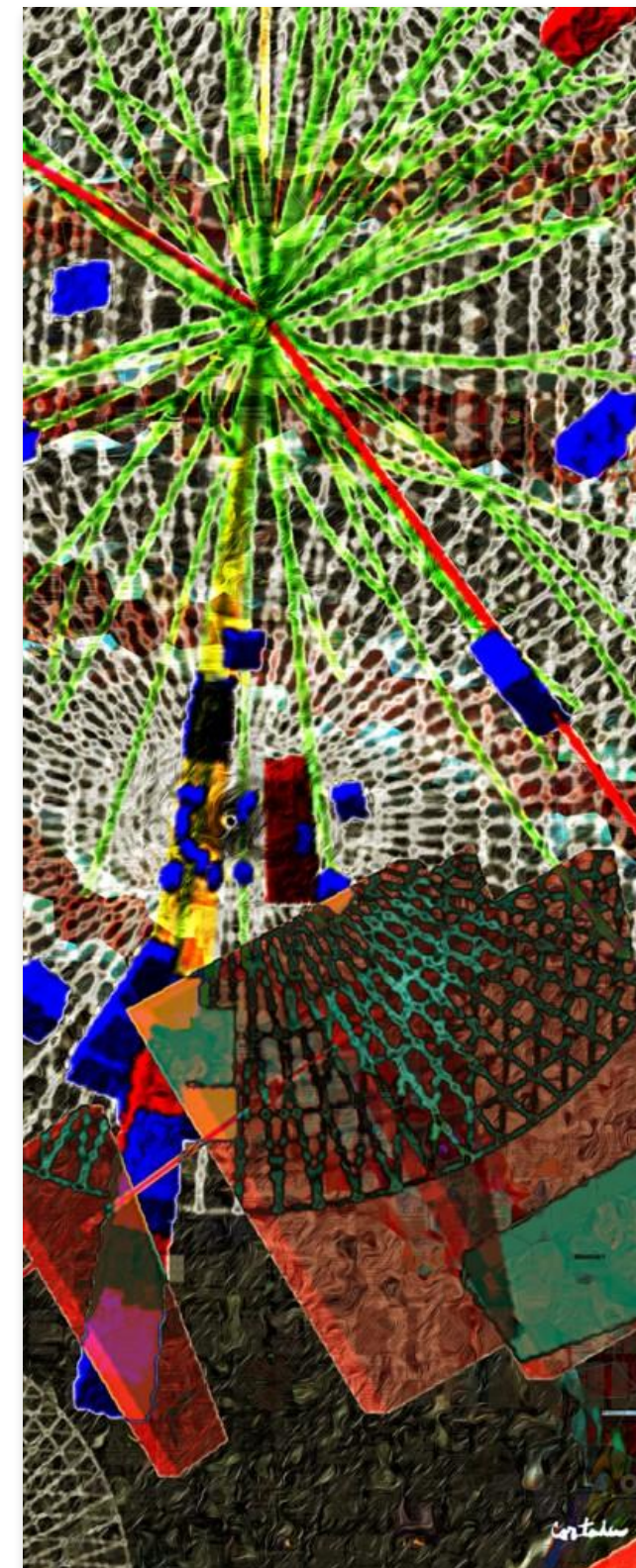
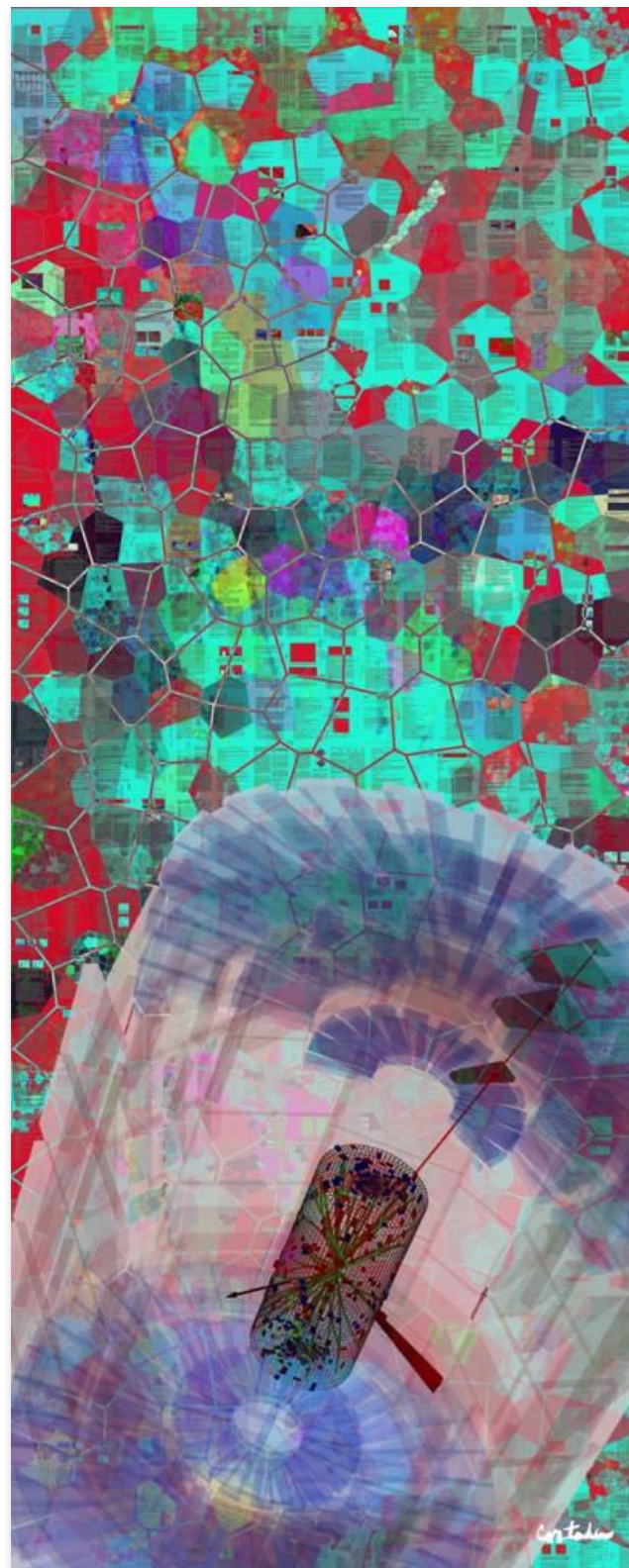


U.S. CMS Operations Program: Feedback

Oliver Gutsche for the U.S. CMS S&C Operations Program
IRIS-HEP Steering Board Meeting #13
15. February 2022



Disclaimer

- I am speaking for the U.S. CMS S&C Operations Program
- I cannot speak for the whole U.S. CMS Collaboration
 - The S&C Operations Program has in the past enabled and driven new capabilities and guided the Collaboration
 - IRIS-HEP is funding many projects that the U.S. CMS Collaboration is interested in
 - The S&C Operations Program will continue enabling the U.S. CMS Collaboration together with IRIS-HEP


Introduction & Outline

- We are happy with the collaboration with IRIS-HEP and benefit from IRIS-HEP projects
- The U.S. CMS S&C Operations Program has recently updated its HL-LHC R&D strategic plan for the agencies, along 4 grand challenges (term borrowed from IRIS-HEP)
 - GC1: Modernize physics software and improve algorithms
 - GC2: Build infrastructure for exabyte-scale datasets
 - GC3: Transform the scientific data analysis process
 - GC4: Transition from R&D to operations
- Go through the IRIS-HEP areas and show how IRIS-HEP projects integrate into the R&D strategic plan

- GC3: Transform the scientific data analysis process
 - We are integrating Analysis Facilities prototypes at multiple Universities and Fermilab
 - Goal is to have multi-user facilities in production during Run-3
 - With the new LHC schedule the data volume at the end of Run-3 might require these facilities to be available to efficiently perform analysis on Run-3 (and Run-2) data
 - IRIS-HEP projects are key ingredients to realize this goal
 - IRIS-HEP provides many foundation libraries like awkward-array, etc.
 - IRIS-HEP led projects fit community needs and are used widely in CMS
 - Popularization and effective use of columnar analysis tools through Coffea framework project
 - Lindsey Gray and Nick Smith leading development, deployment, and integration of user feedback
 - We are integrating with all the IRIS-HEP tools in our Analysis Facility Concepts
 - Coffea-Casa is widely used, enabled through IRIS-HEP SSL deployment
 - IRIS-HEP Analysis Grand Challenge instrumental on path to production quality facilities

- “GC2: Build infrastructure for exabyte-scale datasets” and “GC3: Transform the scientific data analysis process”
 - 3rd party copy efforts are doing well
 - Gridftp → HTTP-TPC completed in U.S., advanced in all of CMS
 - SRM/HTTP (initiating tape stage with SRM, move bytes with HTTP, last piece that requires sites to remove gridFTP and talk to tape) → next is a REST API (important ingredient)
 - XCache supports SoCal cache and Analysis Facility work
 - Next questions to answer: What is the long term vision for caches? How can we facilitate broader adoption?
 - Collaboration with ESnet projects for caches in the network?
 - Skyhook (“Arrow backend to CEPH”)
 - Lots of progress, we think we need to focus now on production-ready quality
 - Will require a lot of quality assurance to get it into production state
 - ServiceX supports MINIAOD and NANOAOB use cases
 - We will need stable releases for analysis facilities
 - Focus on building a user community with many users
 - General comments
 - Interested in adding columns from lower level data formats (MINIAOD) to higher level formats (NANOAOB) without the need to have an additional copy (of NANOAOB) (“joins in Skyhook”?)
 - But having users use the current ServiceX and Skyhook functionality is preferable before embarking on this
 - Data Lakes
 - Not part of DOMA/IRIS-HEP
 - Would be nice that this gets more visibility

- GC1: Modernize physics software and improve algorithms
 - mkFit big success → physics validation for Run-3 almost complete
 - Co-funded with IRIS-HEP
 - SegmentLinking is the next stage
 - Focussing on accelerators (mkFit focussed on vectorization for CPUs)
 - GNN tracking
 - Trying to focus on physics usefulness
 - Aiming for results that are physics interesting in ~6 months, looking at events with more tracks (lower pt threshold)
 - Had a successful hackathon in the CMS context
(<https://indico.cern.ch/event/1041335/> restricted to CMS)
 - Linking to SONIC efforts
 - Will connect FPGA efforts



CMS ML Hackathon: GNN-4-tracking

27 September 2021 to 1 October 2021
CERN
US/Central timezone

Overview

Timetable

Contribution List

Registration

Participant List

Join the #gnn-4-tracks Slack channel [here!](#)

The aim of this hackathon is to integrate graph neural nets (GNNs) for particle tracking into CMSSW.

The hackathon will make use of a GNN model reported by the paper [Charged particle tracking via edge-classifying interaction networks](#) by [Gage DeZoort](#), [Savannah Thais](#), et.al. They used a GNN to predict connections between detector pixel hits, and achieved accurate track building. They did this with the [TrackML](#) dataset, which uses a generic detector designed to be similar to CMS or ATLAS. Work is ongoing to apply this GNN approach to CMS data.

Tasks:

The hackathon aims to create a workflow that allows graph building and GNN inference within the framework of CMSSW. This would enable accurate testing of future GNN models and comparison to existing CMSSW track building methods. The hackathon will be divided into the following subtasks:

- **Task 1: Create a package for extracting graph features and building graphs in CMSSW**

Code will be provided from an EDAnalyzer that, given a miniAOD sample, extracts relevant features as flat ntuples to be used in graph building. There will be two subtasks;

 1. Re-factorise the "feature-extraction" logic in an existing EDAnalyzer to a CMSSW class interface
 2. Implement graph building in CMSSW

Code for the graph building will be provided in both Python and C++.

The intention of 1.1 is to create an interface that can be used in both the existing EDAnalyzer and directly with SONIC (see Task 2). Task 1.2 will enable graphs to be built within CMSSW which can then be sent to Sonic.
- **Task 2. GNN inference on Sonic servers**

The GNN uses Pytorch, which is not yet supported in CMSSW. We will therefore use SONIC (Services for Optimized Network Inference on Coprocessors). Pre-trained GNN models will be provided, which can be placed on the SONIC server. This task will aim to set up a workflow that sends data to SONIC, performs inference, and sends the predictions back into CMSSW.
- **Task 3: Track fitting after GNN track building**

The GNN does pattern recognition to reconstruct particle tracks using detector hits. In order to reconstruct track parameters, one needs to do the final track fitting with e.g. a Kalman filter. The track segments created by the GNN needs to be converted to a TrackCandidate that can be fed to the track fitter. There is already code in CMSSW for this that can be repurposed.
- **Task 4. Performance evaluation for the new track collection**

The fitted tracks will be compared to that of existing track methods. The commonly used MultiTrackValidator will be used for accessing the tracking physics performance. In addition, the workflow for evaluating the computing performance will be defined. This task will be the step for having a standard workflow for the comparison of the performance of new approaches, which will be developed in the future.

- “GC4: Transition from R&D to operations” the least established GC for U.S. CMS S&C
 - Very important for the success of HL-LHC
- OSG-LHC supports current operation of sites and services
 - We depend on it, working well
 - Highlight: token transition in OSG 3.6
 - Factories can send token-based pilots to OSG sites (within a pilot, X509 is still ok), can still send X509 based pilots to ARC CEs
 - Switch from gridFTP to HTTP-TPC (based on xrootd, and xrootd still can talk X509, so if non-OSG sites use HTTP-TPC, everything is ok)
- Scalable Systems Laboratory
 - Important ingredient in moving into operation
 - Good news: Nebraska Analysis Facility is part of SSL
 - In the future, would like to intensify usage of SSL concepts on Tier-2 level
- Training, Education and Outreach
 - Equally important to move R&D into operation, U.S. CMS community is participating well
 - Working on engaging undergraduates and graduates with U.S. CMS S&C projects
 - We want to collect projects and mentors within the program, and then enter them into the appropriate programs like the IRIS-HEP fellows program

Grand Challenges - Summary & Outlook

- Analysis Grand Challenge → works well
- Would be interested in 10 PB/day challenge
 - (serve 10 PB/day from infrastructure to HPC or remote processing resources)
 - Currently Morphed into participation in the WLCG data challenges
 - Do we want to accelerate some parts or do we want to do something US specific?
 - Could exploit the Fabric project and distributed caches, these would not be leveraged if WLCG only

Summary & Outlook

- We are happy with the collaboration with IRIS-HEP and benefit from IRIS-HEP projects
- U.S. CMS S&C organized R&D along 4 Grand Challenges, IRIS-HEP projects well aligned and key ingredients