Project Summary: Deploying GPU algorithms through SONIC

PI: Prof. Philip Chang (University of Florida)
Postdoc: Kelci Mohrman (University of Florida)
Project duration: Sep. 2023 - Aug. 2024

1 Overview

Starting with a brief overview, explain the goal of the project, the problem you aimed to address in your R&D project, and the method/process. What was the expected impact on the HL-LHC CMS S&C operations?

The goal of the project [1] was to implement a version of the Line Segment Tracking (LST) algorithm [2] with the SONIC framework [3] in order to enable flexible and efficient GPU usage. Because reconstruction tasks constitutes the largest fraction of CMS data processing, it is important to understand the resource requirements and to explore options for improving the efficiency of these steps. To this end, CMS is exploring reconstruction algorithms that are designed to make use of GPU resources. These include LST, which is a tracking algorithm that takes advantage of double-layer design of the HL-LHC outer tracker in order to perform hit correlations in a parallel way with GPUs. With more algorithms requiring GPU resources, it is important to understand the resource requirements and strategies for ensuring efficient deployment and usage. The SONIC framework provides the ability to make use of GPUs "as a service", enabling GPUs to be factored out of CPU machines. With this approach, the GPU-based servers may be remote from the CPU-based servers, allowing for more flexibility in the usage of GPU resources.

Our project thus aimed to demonstrate the successful implementation of the LST algorithm with SONIC, to demonstrate inter-site runs (with client CPUs and server GPUs at difference sites), and to explore the performance. If successful, this would help to improve the flexibility and efficiency of the resource utilization of the LST algorithm and contribute to a better understanding of the computational needs of the CMS experiment for the HL-LHC.

2 Techniques

What techniques did you use to address the problem? If more than one avenue were taken, what were the other techniques? And why were the chosen techniques expected to perform better than the alternative or baseline methods?

The SONIC framework was the main tool used in this project. We developed a custom LST backend for the LST algorithm, using the SONIC team's `TestIdentity` repository as an example setup [4]. We extracted relevant pieces of the LST `TrackLooper` repository [5] (at the `cuda_branch`, as described in Section 4) from its ROOT dependencies, and compiled these within the singularity environments provided in the `TestIdentity` setup.

On the client side, we created a modified version of the LST producer that passes the LST inputs to the server as a flat vector of doubles. On the server side, the inputs are decoded

and passed into the LST algorithm. The outputs of the LST algorithm are passed back to the client as a flat vector, which are decoded and saved to ROOT files (as in the standard non-SONIC LST setup). A summary of this workflow is presented in Figure 1.

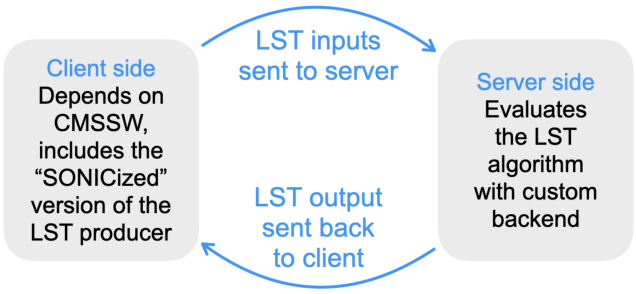

Figure 1: Schematic representation of the SONIC+LST workflow.

3 Outcome

Explain the outcome of the project, including alternative paths and their estimated impact as well. Was the outcome as expected as described in 3.? If not, why not?

In this project, we were able to successfully implement a version of LST with the SONIC framework, explore the performance, and demonstrate successful runs across multiple sites. The details of these main outcomes are summarized below:

- **Successful implementation of LST with SONIC:** As described in Section 2, a working version of the LST algorithm was implemented and run with the SONIC framework. Figure 2 shows an example efficiency plot produced with the SONIC implementation of LST. Instructions for setting up and running the SONIC+LST implementation may be found here [6].

- **Performance studies to explore scaling of SONIC+LST:** With the SONIC+LST setup, we tested running with multiple concurrent instances of LST `cmsRun` jobs and measured the runtime. For these trials, we were using the "step3" `cmsRun` configuration (as described in the LST readme [5]) and tested 1, 8, 16, 32, and 64 concurrent `cmsRun` jobs. We also tested with 4 threads (i.e. setting the `numberOfThreads` parameter in the `cmsRun` configuration file to 4); this is conceptually similar to simply launching more concurrent `cmsRun` jobs manually (except that in this case multiple threads would be able to share in one `cmsRun` and CMSSW handles the scheduling). For comparison, we also ran a similar configuration of the standard non-SONIC setup. However, it should be noted that this is not a fully equivalent comparison because, as explained in Section 4, the SONIC implementation is based upon an older version of the LST algorithm. As we scaled to larger numbers of concurrent jobs, we encountered crashes related to memory (as documented in the Sep 9, 2024 Tracking POG presentation [7]). The results of these timing runs are summarized in Figure 3. For these trials, there

was not an indication that the GPU is saturated (monitoring the `nvidia-smi` output during runs, we observed GPU usage from approximately 0-60%). It is thus likely that the scaling behavior observed in Figure 3 is influenced by other factors. Further studies would be required to fully understand this scaling behavior (as described in Section 5).

- **Inter-site runs:** The performance studies described above were run with both client and server at the Purdue Tier 2 site, but we were able to set up and run with the server and client at the University of Florida (UF) Tier 2 site as well. Furthermore, we were able to demonstrate successful runs with the client at Purdue the server at UF, as well as with the client at UF and the server at Purdue. It would be beneficial to explore the timing of these inter-site runs, as discussed in Section 5.

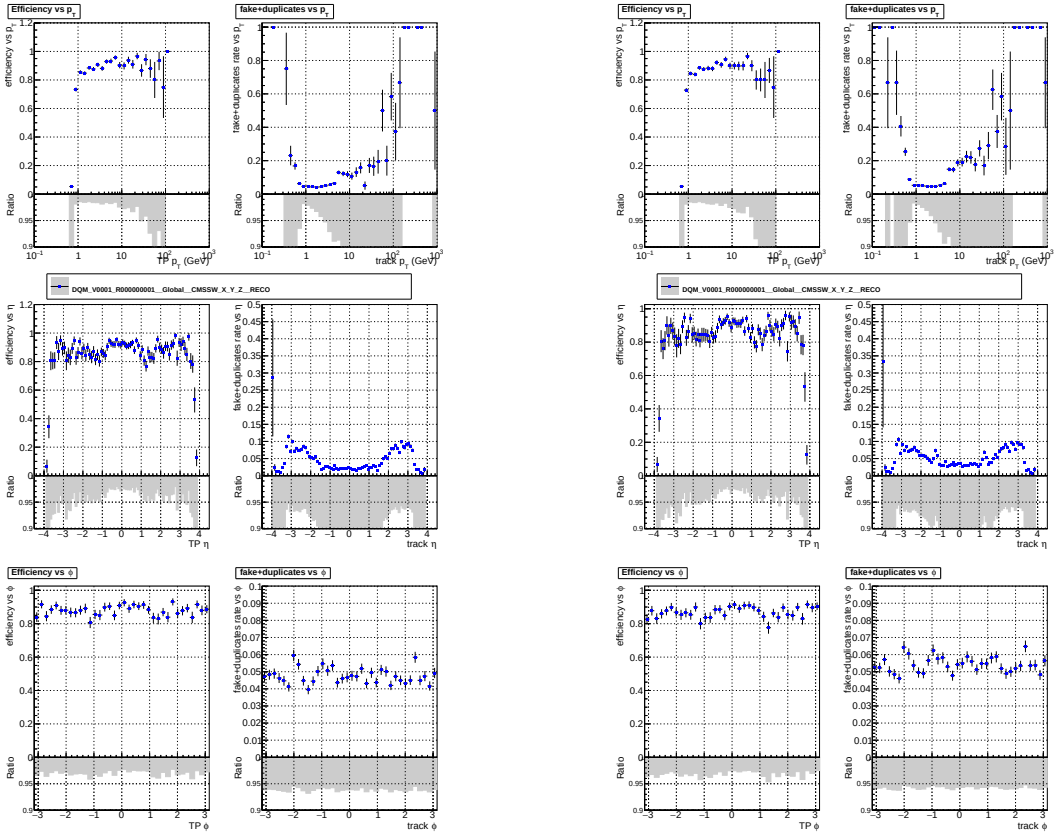

Figure 2: Efficiency plots (produced by the LST validation plotting script) for the standard non-SONIC run (left) and a run with the SONIC+LST workflow (run for 100 events). The non-SONIC setup is based on the master branch of LST (as of summer 2024, based on `CMSSW_14_1_0_pre3_LST_X`). As described in the text, the SONIC workflow is based on an older version of LST (from the `cuda_branch` branch). Although we do not expect identical results (because of the LST version differences), we can see there are no clear signs of significant qualitative differences in the validation plots.

4 Challenges

Was there a particularly difficult obstacle you encountered during this project? How did you navigate through it, and were there any pivotal breakthrough moments? What are some

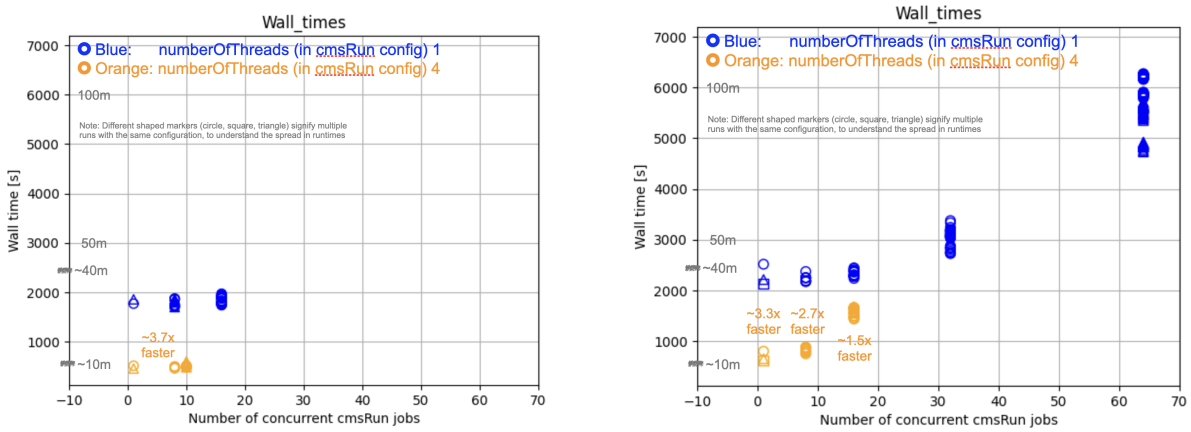

Figure 3: Timing runs with standard non-SONIC implementation of LST (left), and with the SONIC+LST workflow (right). The runs are processing 300 $t\bar{t}$ events. In blue, we show the results with a single thread (`numberOfThraeds` equal to one in the `cmsRun` configuration file), while the orange show the results with four threads. The different shaped markers (circle, triangle, and square) indicate different runs with the same configuration (to provide information about the spread of runtimes). These trials were run at the Purdue Tier 2 site, with the client and server at the same node. The node (with 128 hyper-threaded cores, and 512GB of RAM, and T4 GPU) was reserved for these trials (i.e. there would be no jobs from other users running on these nodes). For these runs, there were not indications that the GPU was saturated.

gained knowledge that others who may tackle similar problems in the future would want to know about beforehand?

One of the main challenges of this project was that the LST code was (and still continues to be) under active development, so the LST code changed significantly throughout the duration of this project. The most impactful of these changes was a move from `CUDA` to the `alpaka` library. However, `alpaka` is not available in the singularity environment used for the backend (provided in the example backend repository [4]). Because of this, the LST backend developed for this project is based on an outdated version of LST (from before the code moved away from `CUDA`). This makes validation of the SONIC implementation difficult, since the LST algorithm used in the SONIC setup is different from the current main branch of LST (which we were using for comparison).

Even though a full validation could not be performed, we were able to run the SONIC implementation through steps to create DQM validation plots, and compare these against plots created from the master LST code (as shown in Figure 2). While exact agreement is not expected (for the reasons explained above), we were able to observe that there were no clear signs of significant qualitative differences. However, in the future, would be important to work to update the LST backend to be in line with the current `alpaka` based master LST.

5 Next steps

What do you recommend for the future if the problem were to be tackled again, or to make improvements to your work?

The next next steps that could be investigated are summarized below:

- **Update the LST backend to be synchronized with the current LST main branch:** As introduced in Section 4, the LST backend was implemented with `CUDA`. The master version of LST has since been updated to use `alpaka`, so the SONIC LST backend should be updated to be synchronized with the updated LST algorithm. This will involve including `alpaka` in the server singularity environment.

- **Study timing and performance of the SONIC+LST implementation:** It would be important to better understand the timing studies summarized in Figure 3. For example, it would be good to understand what is causing the SONIC+LST runtime to increase with increasing number of concurrent `cmsRun` instances. One way to explore this would be to replace the LST backend with a simple pass-through (as opposed to the LST algorithm evaluationa0, and compare the runtime and scaling. Additionally, it could be useful to rerun the timing studies with a workflow that runs only the LST algorithm (instead of the more realistic "step3" workflow). Furthermore, it could be useful to explore tools for profiling the jobs in order to better understand what fraction of time is being spent on each step of the workflow. It would also be important to understand how these performance metrics compare between the SONIC and non-SONIC setups; however, in order to explore this aspect of the comparison, it would be useful to first update the SONIC implementation to the current master LST (based on `alpaka`), as described in the previous point.

- **Study the performance of inter-site runs:** While this project demonstrated successful inter-site runs, an important next step would be to explore the timing and performance of these runs. A preliminary observation (of a test run with the client at Purdue and the server at UF) indicated that the runtime is significantly longer (by about a factor of four) compared to a run in which the client and server were located on the same node at the same site. It would be important to perform repeated runs, carefully measure the timing, and explore the scaling with increasing concurrency of `cmsRun` jobs.

- **Integrating with main LST/CMSSW:** Moving forward, it would be important to explore options for integrating the SONIC+LST implementation into the main LST codebase.

6 Skills learned

What kinds of skills did you learn through the project? Do you think this experience has helped you understand how to develop and lead a research project?

This project allowed me to gain increased experience with a variety of general technical skills, including the usage of singularity environments, Triton servers, and GPU based workflows. Regarding experience and technical skills specific to CMS, I also gained more experience working with CMSSW and learned more about CMS tracking algorithms. Through frequent

discussions with mentors and PIs, I also had to opportunity to learn more about how to plan, organize, and execute technical research projects within CMS.

7 Presentations

How was your work presented to the wider CMS community? Could you provide a list of meetings, or conferences where the work was presented?

The work was presented primarily at CMS Tracking POG meetings:

- Sep. 9, 2024: 'LST with SONIC framework'" [7]
- Jun. 3, 2024: "Project status update: LST with the SONIC framework" [8]
- Feb. 12, 2024: "Project status update: LST with the SONIC framework" [9]
- Oct. 23, 2023: "Project introduction and plans: LST with the SONIC framework" [10]

References

[1] USCMS R&D Proposal "Deploying GPU algorithms through SONIC". https://uscms-software-and-computing.github.io/assets/pdfs/Kelci-Mohrman.pdf.

[2] Philip Chang et al. Segment Linking: A Highly Parallelizable Track Reconstruction Algtorithm for HL-LHC. *J. Phys. Conf. Ser.*, 2375(1):012005, 2022.

[3] Aram Hayrapetyan et al. Portable acceleration of CMS computing workflows with coprocessors as a service. 2 2024.

[4] "TritonCBE/TestIdentity/". https://github.com/yongbinfeng/TritonCBE/tree/main/TestIdentity#instructions-to-run-patatrack-aas-at-purdue.

[5] "SegmentLinking/TrackLooper". https://github.com/SegmentLinking/TrackLooper/tree/master?tab=readme-ov-file#build-tracklooper.

[6] "Setting up and running SONIC+LST". https://gist.github.com/kmohrman/015d7fa8f807099ff61e41e074e7124a.

[7] "LST with SONIC framework". https://indico.cern.ch/event/1443183/#50-update-on-soniclst-developm.

[8] "Project status update: LST with the SONIC framework". https://indico.cern.ch/event/1418266/#35-line-segment-tracking-using.

[9] "Project status update: LST with the SONIC framework". https://indico.cern.ch/event/1374894/#25-lst-running-on-gpus-as-a-se.

[10] "Project introduction and plans: LST with the SONIC framework". https://indico.cern.ch/event/1337451/#5-lst-running-on-gpus-as-a-ser.

Project Summary: Benchmarking current capabilities and exploring the acceleration of columnar processing via heterogeneous architectures

PI: Prof. Philip Chang (University of Florida)
Postdoc: Kelci Mohrman (University of Florida)
Project duration: Jan. 2025 - Dec. 2025

1 Overview

Starting with a brief overview, explain the goal of the project, the problem you aimed to address in your R&D project, and the method/process. What was the expected impact on the HL-LHC CMS S&C operations?

As we push to higher luminosities, the challenge of efficiently processing and analyzing our data will become increasingly important in the pursuit of new physics discoveries at the LHC. The goal of this project was to benchmark the performance of the step of end-user data analysis (operating on nanoAOD formatted data) in order to understand current capabilities, scaling, and bottlenecks for columnar analysis workflows and to explore the acceleration of columnar processing via GPU offloading. The results of these studies aim to help illuminate the challenges and opportunities that lie ahead as CMS pushes towards rapid and efficient turnarounds at HL-LHC scale, providing timely insights for USCMS S&C during this critical period in which our HL-LHC computing architectures are being determined

2 Techniques

What techniques did you use to address the problem? If more than one avenue were taken, what were the other techniques? And why were the chosen techniques expected to perform better than the alternative or baseline methods?

For the aspect of the project that aims to benchmark current capabilities of end-user columnar processing, we chose to focus on a skimming workflow, since this allows us to explore specifically I/O challenges. To this end, we developed a `coffea`-based skimming application (`cortado` [1]), making use of the `TaskVine` [2] framework for large-scale processing. This scheduler provides good performance and flexibility, and is maintained by developers who are familiar with the needs of HEP analyses. While working with the version of coffea that is deeply integrated with dask-task-graphs, we also made use of the Dynamic Data Reduction (DDR) tool [3]. This tool allowed us to handle the very large task graphs produced in our workflows. However, after the release of the modern virtual-array based `coffea` version, we have reimplemented the `cortado` workflow in a standalone way (without DDR) to study the performance in this simpler case.

For the GPU project, we made use of the CUDA backend of the `awkward` repository. This is currently the only GPU implementation of the `awkward` functions. During the course of the project, we have also begun iterating with the NVIDIA team to discuss the possibility of upgrading the CUDA backend to the CUDA Core Compute Libraries (CCCL), which may provide the ability to implement the `awkward` functionality in pure python without custom CUDA kernels [4].

3 Outcome

Explain the outcome of the project, including alternative paths and their estimated impact as well. Was the outcome as expected as described in 3.? If not, why not?

For the skimming project, we have implemented a `coffea`-based skimming application (`cortado`) in the latest virtual-array based `coffea` version. With this application, we tested the performance of skimming at the scale of 1.8B events. We have measured the performance of this workflow with varying numbers of CPU cores (the results of this are summarized in Fig. 1). The `TaskVine` tool provides diagnostics for each run (an example of which is shown in Fig. 1), indicating the activities the manager spends its time on. At a smaller scale (75M events), we have also tested alternative configurations (e.g., multiple concurrent instances of the `cortado` application, and alternate methods of file access). The results of these studies are summarized in [5].

For the `awkward` GPU project, we aimed to implement the set of ADL benchmark queries [6] on GPU. We were able to successfully implement all 8 queries on GPU (the code is available here [7]), and achieve qualitative agreement with the results of the CPU based queries (though some debugging remains in progress for Query 8, which is currently only runnable on small numbers of events). The comparison of the output histogram from the CPU and GPU implementations for an example query is shown in the left-hand slide of Figure 2, where the qualitative agreement is visible. We measured the total walltime of each query for GPU and CPU, which is reported in the middle plot in Figure 2. We further measured each step of each query (splitting up the query into read, load, compute, and histogramming steps); the timing for the compute step is shown in the right-hand side of Figure 2. From this timing comparison, we see that Query 7 shows a promising improvement on GPU (of a factor of about 40x) compared to the CPU time.

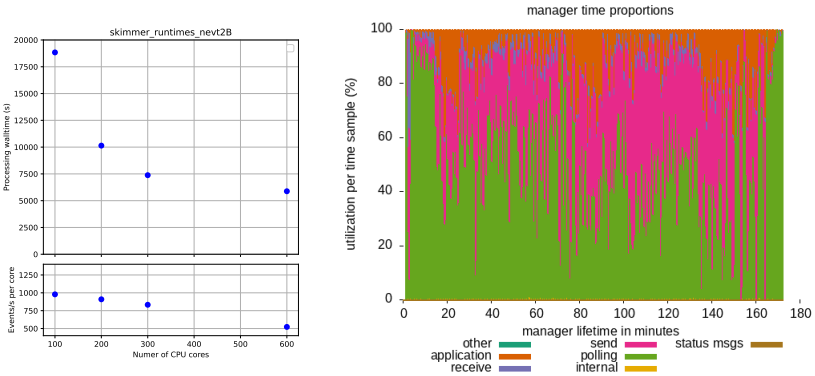

Figure 1: Runtimes for the `cortado` application with 1.8B events with varying numbers of CPU cores (left). Manager time breakdown for the 200 core run (right).

4 Challenges

Was there a particularly difficult obstacle you encountered during this project? How did you navigate through it, and were there any pivotal breakthrough moments? What are some

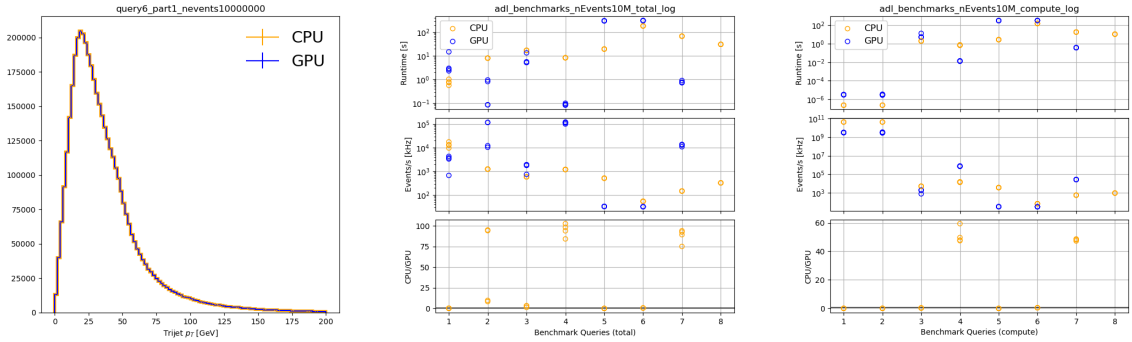

Figure 2: Comparison of histograms from GPU and CPU implementations of Query 6 (left). Walltime comparisons for the 8 queries (middle). Walltime comparisons of the compute step for the 8 queries (right).

gained knowledge that others who may tackle similar problems in the future would want to know about beforehand?

For the `coffea` skimming project, one challenge was that the `coffea` framework underwent a significant update over the course of this project. Because of this, the `coffea`-based skimming application we developed for the project (`cortado`) had to be implemented in multiple versions of `coffea`. In the future, the project should build upon the latest virtual-array based `coffea` version.

For the `awkward` GPU project, one of the main challenges was in validating the GPU results against the CPU results. While qualitative validation is straightforward via histogramming, event-by-event validation is more challenging because exact agreement is not expected between GPU and CPU (because of precision differences). This should be kept in mind for the future, as it can make validation challenging (especially when small differences in precision lead to different objects or events to be selected, which we encountered in our study).

5 Next steps

What do you recommend for the future if the problem were to be tackled again, or to make improvements to your work?

Several aspects of this work will be continued through my 2026 LPC DR project [8], so I will have the opportunity to follow up and extend this work myself.

For the skimming project, I aim to work towards larger-scale tests and developing the `cortado` package into a more usable framework. In order to achieve this, several challenges will need to be overcome:

- Failure handling and checkpointing (ideally the application should run to completion and provide a report to users with a trivial way for user to resubmit failed jobs).
- Understand and develop ways to avoid issues of hanging and long tails on final jobs.

- Technical improvements (e.g., including the Runs tree in the output skims).

For the `awkward` GPU project the next steps would include:

- Profiling the CPU and GPU queries to improve the understanding of the performance.
- Further studies of precision comparisons of GPU to CPU, to understand the differences (some of which are larger than would be naively expected for precision issues).
- Test a more realistic workflow on GPU (e.g., attempt to implement on GPU a `coffea`-based analysis such as SMP-24-015, benchmark any potential speedups, and/or identify any missing functionality).

6 Skills learned

What kinds of skills did you learn through the project? Do you think this experience has helped you understand how to develop and lead a research project?

Through this project, I have gained professional development experience through the process of writing the proposal and the opportunities to present the results at multiple conferences. I have also gained technical skills and experience, especially related to CUDA, GPU computing, profiling, benchmarking, and debugging large-scale processing. This project has indeed helped me to improve my understanding how to develop and execute a research program.

7 Presentations

How was your work presented to the wider CMS community? Could you provide a list of meetings, or conferences where the work was presented?

The work was presented at several meetings, both internal and external to CMS.

- May 21, 2025: "Towards rapid and efficient analyses at scale", USCMS Annual Meeting (poster) [9].
- July 14, 2025: "Towards rapid and efficient columnar-based analyses at scale", PyHEP developer workshop [10].
- July 23, 2025: "Benchmarking analysis workflows", CMS CAT General Meeting [11].
- August 28, 2025: "Towards rapid and efficient analyses at scale", Lepton Photon 2025 (poster) [12].
- December 8, 2025: "Towards rapid and efficient analyses at scale", CMS Week December 2025 (poster) [13].
- December 17, 2025: "Update on end-user analysis R&D and benchmarking studies", CMS CAT General Meeting [5].

References

[1] Cortado repository. https://github.com/kmohrman/cortado.

[2] TaskVine. https://ccl.cse.nd.edu/software/taskvine/.

[3] Dynamic Data Reduction, Ben Tovar. https://indico.cern.ch/event/1566263/contributions/6736112/.

[4] "Awkward Array: The Swiss Army Knife of Irregular Data (and Still a Little Awkward)", Ianna Osborne. https://indico.cern.ch/event/1515852/contributions/6522539/.

[5] "Update on end-user analysis R&D and benchmarking studies". https://indico.cern.ch/event/1615345/#2-updates-on-analysis-workflow.

[6] IRIS-HEP ADL benchmarks. https://iris-hep.org/projects/adl-benchmarks-index.html.

[7] Code for GPU-based ADL queries. https://github.com/kmohrman/columnar_gpu/blob/49b0860bbf24bc5863ea30231726f778043949ca/run_adl_queries.py.

[8] "Kelci Mohrman LPC DR 2026". https://lpc.fnal.gov/fellows/2026/Kelci_Mohrman.shtml.

[9] "Towards rapid and efficient analyses at scale". https://indico.cern.ch/event/1499327/contributions/6510006/.

[10] "Towards rapid and efficient columnar-based analyses at scale". https://indico.cern.ch/event/1515852/contributions/6591652/attachments/3103766/5500278/kmohrman_pyhepdev_jul2025.pdf.

[11] "Benchmarking analysis workflows", CMS CAT General Meeting". https://indico.cern.ch/event/1565053/#2-benchmarking-analysis-workfl.

[12] "Towards rapid and efficient analyses at scale". https://indico.cern.ch/event/1493037/timetable/?view=standard#347-towards-rapid-and-efficien.

[13] "Towards rapid and efficient analyses at scale". https://indico.cern.ch/event/1535613/page/40626-poster-session.