Developing an automated, customized backend for algorithms being loaded and executed on GPUs using SONIC

The CMS data acquisition and data processing challenge

The high luminosity Large Hadron Collider (HL-LHC) experiments will commence in a few years with a higher center-of-mass energy for proton-proton collisions and a higher intensity of protons in each proton bunch. The HL-LHC can probe a higher energy regime, making it exciting for numerous searches beyond the standard model and many measurements that requires high precision. However, It also poses challenges to data processing because more data will be delivered within the same time interval, requiring many algorithms increase their complexity to maintain optimal performance in reconstructing the particles and tracks generated from the collisions.

The CMS data-acquisition system comprises Level-1 trigger, HLT trigger and offline computation. Level-1 triggers promptly respond to an event's track or other prompt reconstruction objects, determining if the event is of physics interest and recording it. They are built on FPGAs to achieve prompt response. The HLT trigger is a CPU-based computing farm that further selects and categories events based on more properties of the prompt reconstruction objects. After the HLT, the data is taken offline and stored in Tier0 computing center at CERN, then distributed to Tier1 and Tier2 computing centers in many labs and institutes worldwide for full-intensity event reconstruction.

The currently CPU-based system for offline reconstruction will not be sufficient to process the data from HL-LHC. New approaches that leverage heterogeneous computing platforms, such as GPU-coprocessors, could provide solutions for CMS data offline computing in the HL-LHC era. The hardware deployment of this heterogeneous platform can be realized by locally connecting GPUs via PCI-express. A software project, 'Services for Optimized Network Inference on Co-processors' (SONIC), has been proposed to establish a stable and continuous service that optimizes computing efficiency and maximizes the utilization of the heterogeneous platform.

SONIC Project

SONIC manages experimental data processing requests, acting as a bridge between CPUs, designated as "clients," and the heterogeneous computing platform serving as the "server." It facilitates the transfer of processing tasks from the client CPUs, typically located in dataset processing centers, to the heterogeneous platform, which is not confined to specific locations. SONIC oversees the management of both client and server resources. For instance, it manages GPU servers using Kubernetes services and optimizes the CPU-to-GPU ratios based on the tasks at hand. Additionally, SONIC accounts for CPU fallback solutions in cases where server-side issues arise, or disruptions occur during data transfer between the client and server.

A prototype CMS workflow for processing datasets from the AOD data tier to the MiniAOD data tier via SONIC has been established on the Purdue Tier-2 cluster. Machine learning processes, such as B jet tagging and MET reconstruction, are delegated to the server for predictions. Both the "client" and "server" components are configured within the Purdue cluster. Performance evaluation of data processing using SONIC is conducted and compared against direct CPU-based inference, demonstrating increased data processing speed."

Automating customized backend for algorithms on SONIC

The CMS workflow encompasses various non-ML computations, some of which can benefit from GPU utilization. For instance, the Patatrack algorithm, originally a CPU-based track finding algorithm in CMS, has been implemented on GPUs with SONIC. This achievement signifies the technical feasibility of such adaptations. Our goal is to create automated tools that facilitate the loading of customized algorithm GPU kernels within the SONIC server and manage external work wrappers on the client side. This development will assist in recasting the CMS Run-3 GPU workflow to enable SONIC and test its performance in production environments, as well as establishing SONIC infrastructure to support customized GPU algorithms at HL-LHC.

The work plan, timeline, and milestones are demonstrated below.

2023/Q4: Automate the initial Patatrack-aaS workflow developed [10] by developing wrap- pers for inputs consumed by the algorithms and outputs from the customized backend kernel, as well as an automatic wrapper from the client side. Benchmark the performance in latency, throughput, and memory consumption against inferences on directly connected GPUs. The deliverables are the automation tools developed and the performance test results.

2024/Q1: Test the generalizability of the automation wrapper tools with other GPU re- construction algorithms. These can include although not limited to the algorithms in the current CMS Run-3 GPU workflow and algorithms targeting the HL-LHC such as the Line segment tracking. The deliverables are to identify shortcomings in the automation tool in handling algorithms consuming different data formats and further improvements to be integrated into the SONIC automation tools.

2024/Q2: This period is dedicated to developing and integrating the improvement in SONIC automation tools to successfully enable offloading of GPU algorithms in the CMS Run-3 workflow. The deliverables are improved support of customized GPU algorithms in SONIC on both the server and client side, as well as ingredients for developing a SONIC version of CMS Run 3 GPU workflow.

2024/Q3: A SONIC version of CMS Run 3 GPU workflow will be developed and fully tested at Purdue tier-2 data center. A prototype test in CMS production system will also be explored to understand the fallback behaviors illustrated in Figure 2. We will also start preparing a manuscript documenting the automation tool as well as performance test results.

Postdoc profile and mentorship plan

The proposed work will be conducted by Postdoc Yao Yao. Yao Yao earned her PhD from UC Davis and will join the Purdue CMS group as a post-doctoral researcher in July 2023. Yao will receive direct supervision and mentorship from PI Liu. She will also collaborate closely with the SONIC team and developers (Y. Feng, K. Pedro, N. Tran, P. MacCormack, P. Harris). This project integrates well into existing R&D efforts in SONIC Ku- bernetes server developments by computing professionals at Purdue (Dmitry Kondratyev, Stefan Piperov). Furthermore, Yao will also receive support from CMS software and track reconstruction expert Jan Schulte at Purdue. The CMS Tier-2/3 Center at Purdue, local computing resources including dedicated GPU clusters (Geddes and Gilbreth), and expertise available through the Information Technology at Purdue (ITaP) provide sufficient computing hardware, infrastructure, and technical support to ensure the success of the project.

Summary

The integration of a heterogeneous computing platform into CMS data processing presents a potential solution for the challenges posed by HL-LHC. The ongoing development of SONIC using CPU-GPU heterogeneous platforms indicates promising enhancements in offline data processing efficiency, particularly for tasks involving ML algorithms. To extend these efficiency gains beyond ML algorithms and enhance offline data processing, there is a necessity to develop tools that automate the creation of a customized backend for non-ML algorithms. This proposed initiative is timely and leverages existing investments in GPU software, hardware, SONIC, and portable software tools through the USCMS operations program. Adopting an 'as-a-service' approach will optimize hardware performance and flexibility while reducing software maintenance overhead for external heterogeneous resources in future CMS computing operations. Ultimately, it will significantly benefit the CMS physics program by enabling cost-effective and efficient processing of HL-LHC data and simulation samples.