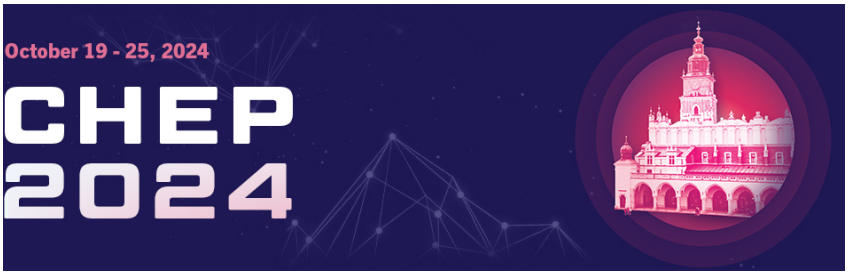

Contribution ID: **422** Contribution code: **WED 22** Type: **Poster**

IceSONIC - Network AI Inference on Coprocessors for IceCube Offline Processing

Wednesday 23 October 2024 16:00 (15 minutes)

An Artificial Intelligence (AI) model will spend "90% of its lifetime in inference."To fully utilize coprocessors, such as FPGAs or GPUs, for AI inference requires O(10) CPU cores to feed to work to the coprocessors. Traditional data analysis pipelines will not be able to effectively and efficiently use the coprocessors to their full potential. To allow for distributed access to coprocessors for AI inference workloads, the LHC's Compact Muon Solenoid (CMS) experiment has developed the concept of Services for Optimized Network Inference on Coprocessors (SONIC) using NVIDIA's Triton Inference Servers. We have extended this concept for the IceCube Neutrino Observatory by deploying NVIDIA's Triton Inference Servers in local and external Kubernetes clusters, integrating an NVIDIA Triton Client with IceCube's data analysis framework, and deploying an OAuth2-based HTTP authentication service in front of the Triton Inference Servers. We will describe the setup and our experience adding this to IceCube's offline processing system.

Primary authors: SHEPERD, Alec (University of Wisconsin-Madison); RIEDEL, Benedikt; SCHULTZ, David (University of Wisconsin-Madison)

Presenter: RIEDEL, Benedikt

Session Classification: Poster session

Track Classification: Track 4 - Distributed Computing