Supercomputing for Accelerating Neutrino Physics

Elisabeth Petit - Bois

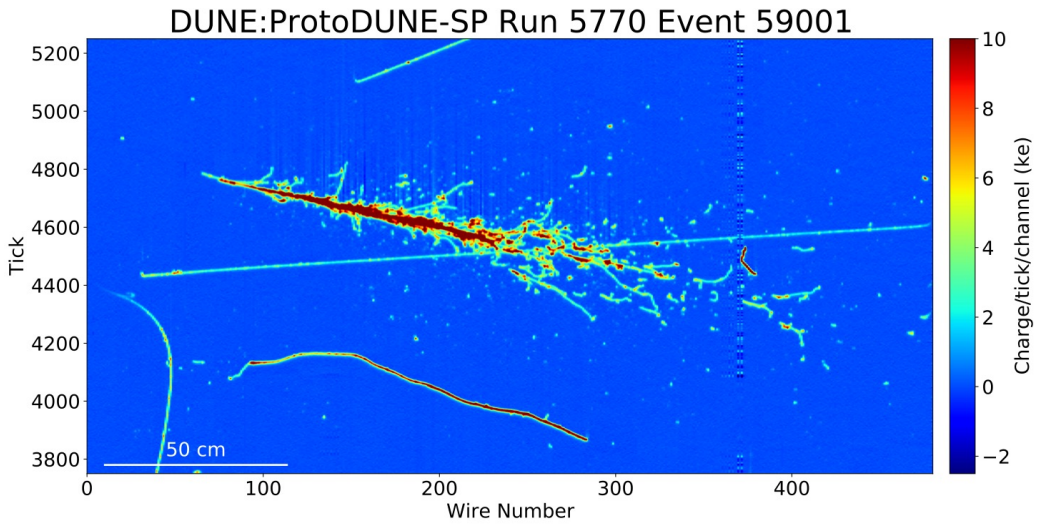
GEM Project Presentation

16 June 2021

Supervisors: Kenneth Herner, Michael Kirby, Andrew Norman

Background

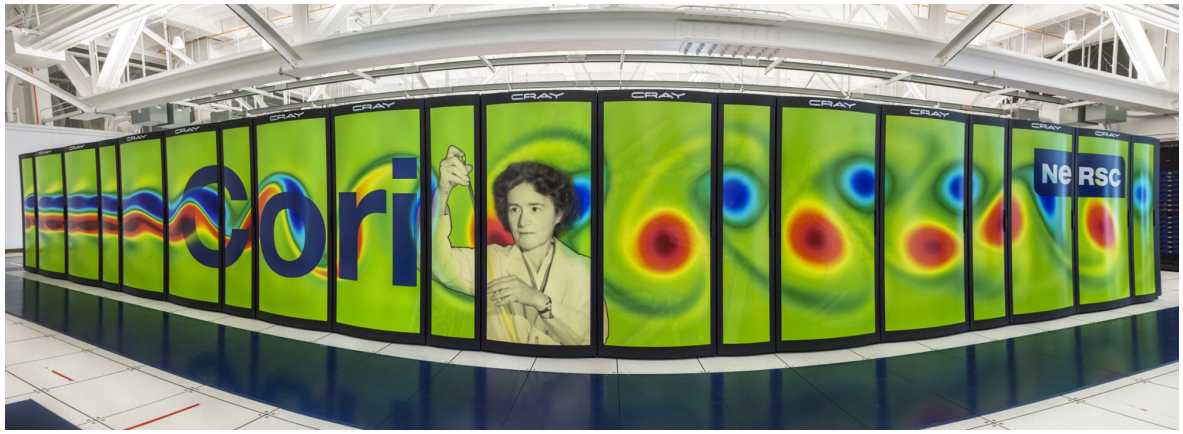
- The analysis of neutrino physics data requires application of advanced mathematical techniques.
- Activity within particle accelerators generates **millions** of events.
- These events undergo **machine learning (ML) inference** to determine the type of particle interaction that the detector recorded.
- This large number of events generates **terabytes upon petabytes** of data to process.



Fermilab

SONIC and GPUaaS

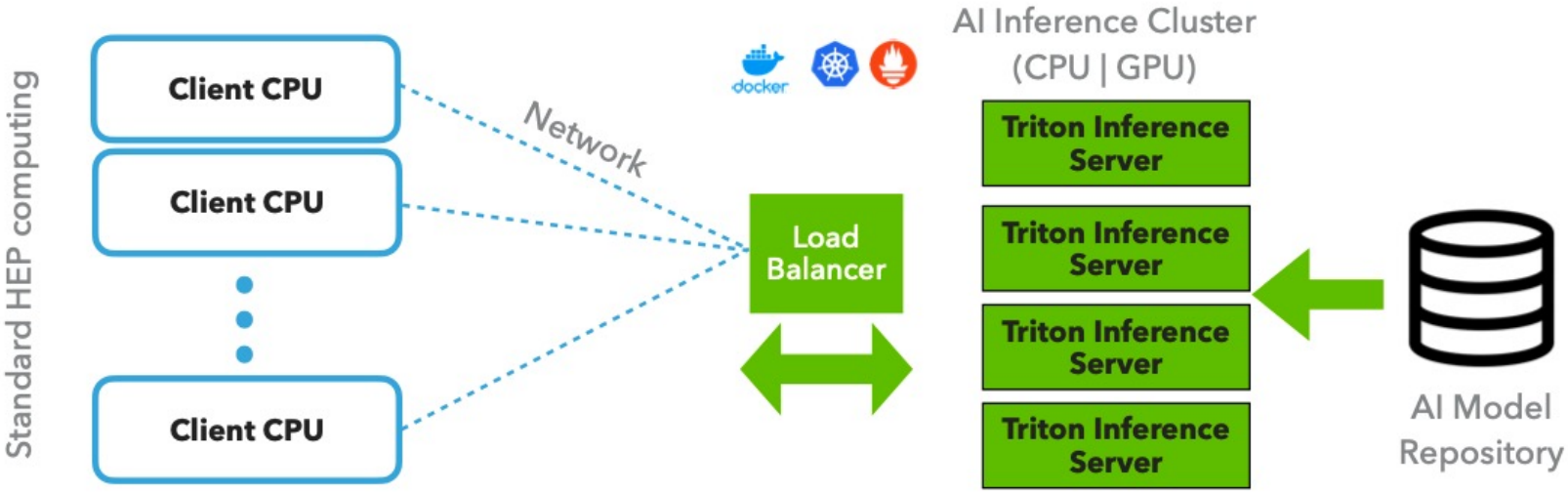
- Services for Optimized Network Inference on Coprocessors (SONIC)
  - A Fermilab-developed framework used to enable **heterogenous computing**.
  - Operates using a **client-server model** to integrate external computing resources.
- SONIC coupled with graphics processing units (GPUs) allows for a GPU-as-a-Service (GPUaaS) model for ML acceleration.



Cori Supercomputer at Lawrence Berkeley National Laboratory

Fermilab

How It Works



Results

- For ProtoDUNE event reconstruction using Triton + cloud:
  - Speeds up ML inference by a factor of 17
  - Reduces overall event processing time by a factor of 2.7 (330s to 123s)

- For ProtoDUNE event reconstruction using FermiGrid:

File	CPU avg event time	GPU avg event time	Ratio	CPU EmTrkMichelId time	GPU EmTrkMichelId time	Ratio
1	844.531	566.278	1.49	328.192	21.813	**15.05**
2	983.785	431.857	2.28	359.097	20.792	**17.27**
3	880.617	382.902	2.30	323.983	20.694	**15.66**
4	840.133	634.182	1.32	306.585	24.325	**12.60**
5	527.301	404.612	1.30	188.112	20.279	**9.28**

Objective

- This project will work to adapt the SONIC system to run on the NERSC Cori supercomputer system.
- The DUNE experiment plans to use a combination of supercomputing facilities along with GPU accelerators for neutrino analysis.
- One of the first high energy physics (HEP) demonstrations of a hybrid supercomputing environment that can partition computations between a high performance CPU environment and a GPU accelerated environment.

References

- Wang, Michael, Yang, Tingjun, Acosta Flechas, Maria, Harris, Philip, Hawks, Benjamin, Holzman, Burt, Knoepfel, Kyle, Krupa, Jeffrey, Pedro, Kevin, & Tran, Nhan. *GPU-accelerated machine learning inference as a service for computing in neutrino experiments*. United States.
  - https://arxiv.org/abs/2009.04509
- https://www.nersc.gov/about/