FERMILAB-POSTER-21-050-STUDENT

Supercomputing for Accelerating Neutrino Physics

Elisabeth Petit – Bois, Georgia Institute of Technology – Kenneth Herner, Michael Kirby, Andrew Norman, Fermilab

Background

- 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 require ulti-petabyte storage systems.

Services for Optimized Network Inference on Coprocessors (SONIC)

- Heterogenous computing framework using a client-server model to integrate external computing resources.
- Uses graphics processing units (GPUs) for ML acceleration.
- SONIC proposal* reports ProtoDUNE event reconstruction using Triton + Google Cloud can:
  - Speed up ML inference by a factor of 17
  - Reduce overall event processing time by a factor of 2.7 (330s to 123s)

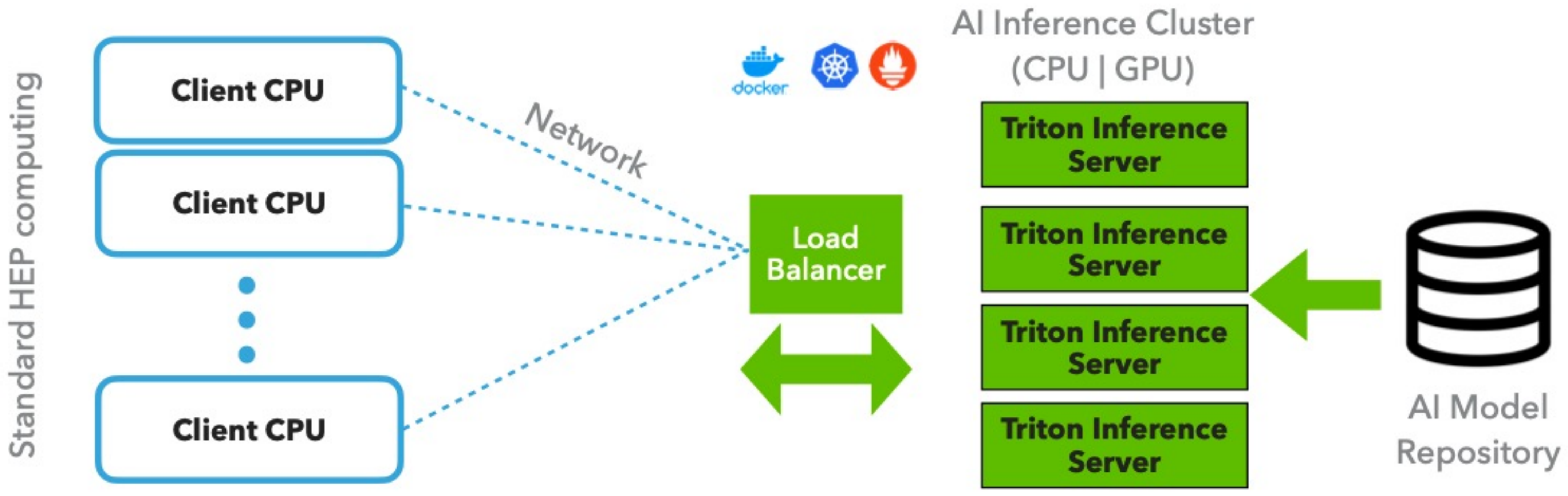

Figure 1: Diagram of SONIC Architecture

Objective

- Working to adapt the SONIC system to run on the NERSC Cori supercomputer.
- Prove SONIC proposal's metrics by showing overall event time and ML inference time are minimized using the framework.
- DUNE plans to use a combination of supercomputing facilities along with GPU accelerators for neutrino analysis.

*Reference: *GPU-accelerated machine learning inference as a service for computing in neutrino experiments"* https://arxiv.org/abs/2009.04509

ProtoDUNE event reconstruction using FermiGrid

File	CPU Total event time	GPU Total event time	Ratio	CPU ML inference time	GPU ML inference 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**

ProtoDUNE event reconstruction using Cori

File (Processor)**	CPU Total event time	GPU Total event time	Ratio	CPU ML inference time	GPU avg ML inference time	Ratio
1 (KNL)	1344.94	612.638	2.20	827.136	99.6944	**8.30**
1 (Haswell)	249.03	202.339	1.23	147.057	103.29	**1.42**
2 (KNL)	1118.84	456.305	2.45	717.792	63.5727	**11.29**
2 (Haswell)	203.805	264.244	0.771	127.486	186.394	**0.684**

**KNL is less-performant than Haswell

Acknowledgements

Thank you to my supervisors and the FastML team for creating a welcoming space for collaboration and innovation this summer. Also, thank you to The National GEM Consortium for their role in making this opportunity possible.

This manuscript has been authored by Fermi Research Alliance, LLC under Contract No. DE-AC02-07CH11359 with the U.S. Department of Energy, Office of Science, Office of High Energy Physics.