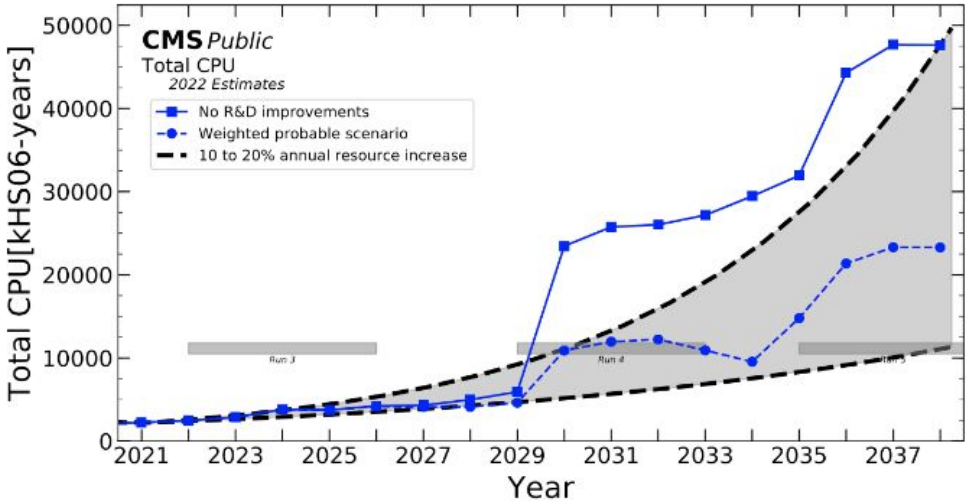
GPU Acceleration of Machine Learning Inference at CMS

Mentee: Hannah Green (Ohio Wesleyan University)
Mentors: Mia Liu, Dmitry Kondratyev (Purdue University)

CMS Computing

- HL-LHC will take 10x the data as LHC - how do we manage it?
- Computing resource usage is expected to spike as more data needs processing
- GPU efficiency at matrix multiplication makes for a quicker method to run parts of tasks currently ran on CPU (e.g. ML inference).
- Already free, pre-developed software for GPU network inference (unlike FPGAs, ASICs)



Portable Acceleration of CMS Production Workflow with Coprocessors as a Service (2023), unpublished

Main Workflow Software

SONIC - client/server interactions

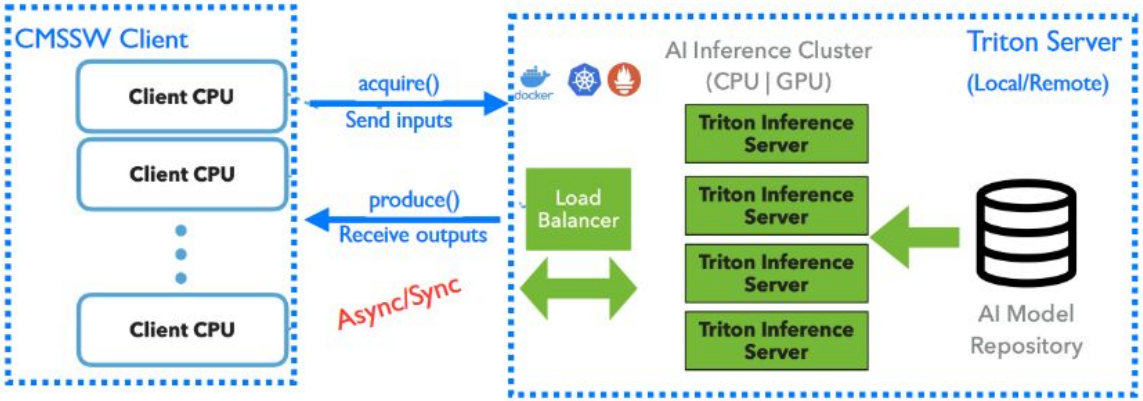
Triton - server

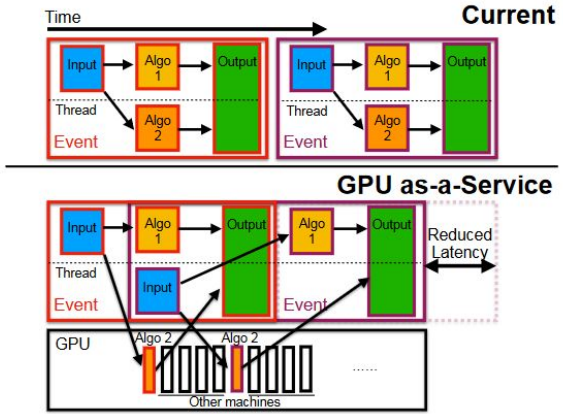
NVIDIA
TRITON INFERENCE SERVER

Kubernetes - load balancing

Service for Optimized Network Inference on Coprocessors (SONIC)

- SONIC aims to **outsource portions of a CMS job to hardware** running NVIDIA's Triton Inference Server.
  - Handles server and client interactions.
  - Offloads parts of processing to GPU for reduction in job time.
  - Highly portable via containerization tools (Docker, Singularity)
  - Still in very early development– more automation work necessary; more models to port.



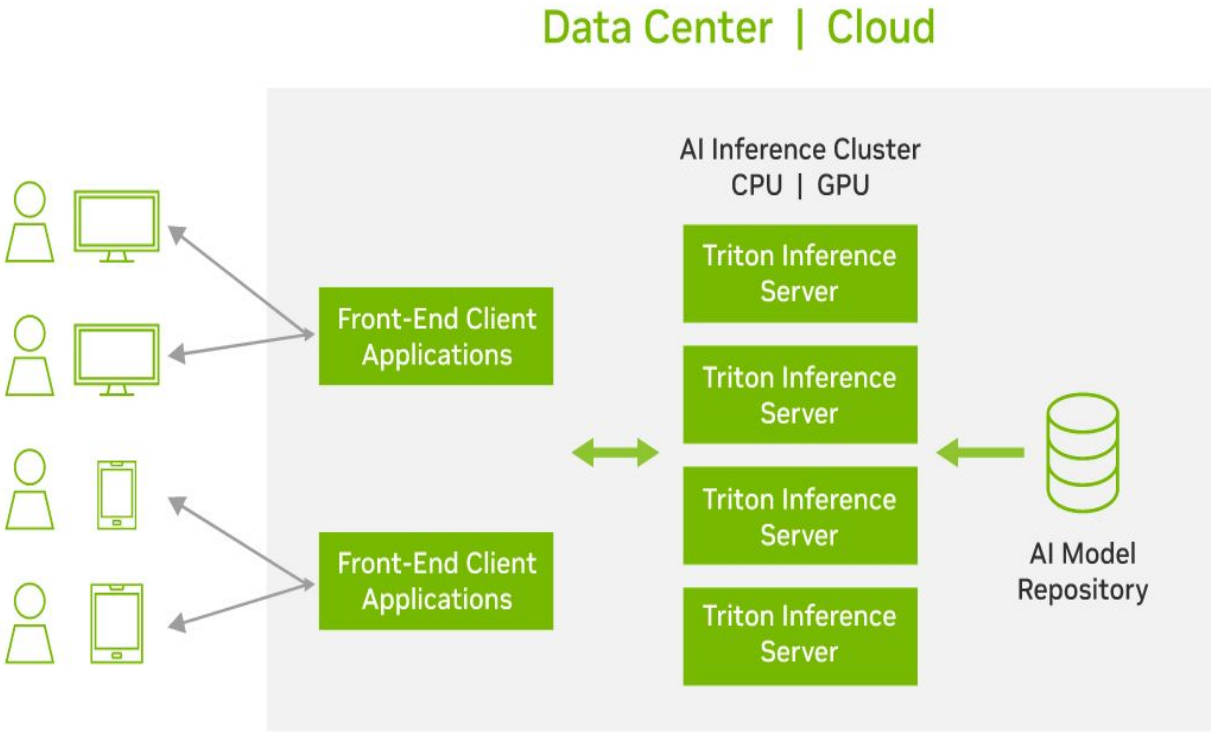


Portable Acceleration of CMS Production Workflow with Coprocessors as a Service (2023), unpublished

Triton Inference Server

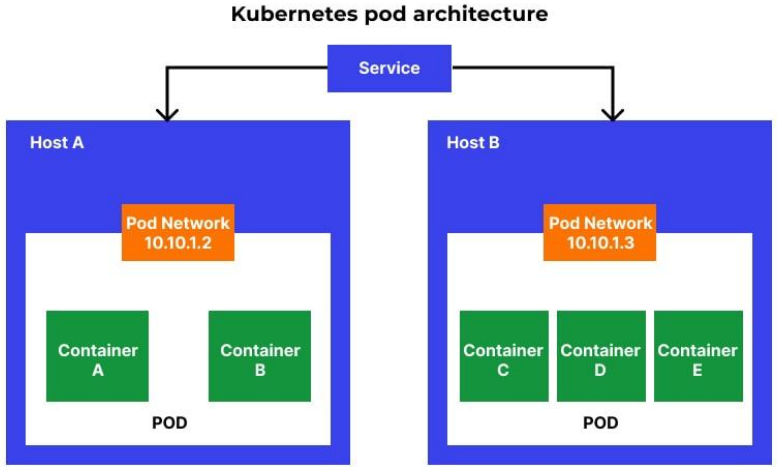
- **Allows for remote ML inference, standardizes it** by allowing models to run on any GPU or CPU
- Can **run multiple models in parallel** on one or many GPUs
- Built-in compatibility with:
  - NVIDIA load balancers or software such Kubernetes to improve inference.
  - Monitoring systems (Grafana, Prometheus)
- **Detailed documentation and resources** to help learn and troubleshoot.

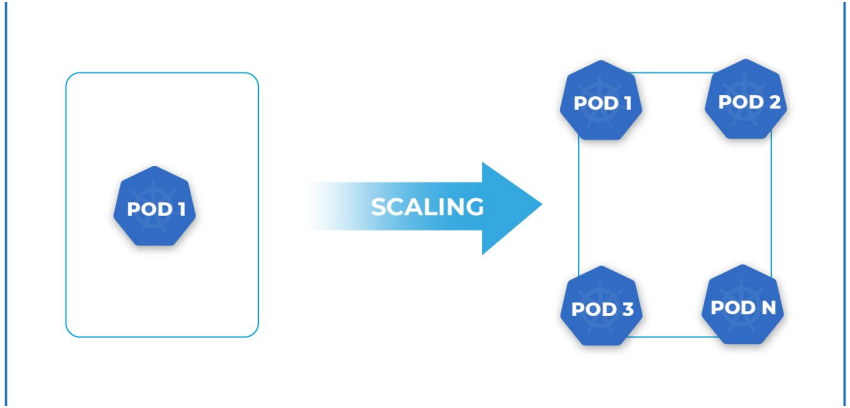
Triton Inference Server



Kubernetes

- Container orchestration (management) platform.
  - Object oriented.
  - One or more applications are containerized in **Pods**.
  - Pods can be cloned using a **Deployment,** which can be used for *auto-scaling*.
  - Pods interact with one another using **Services.**





https://www.wallarm.com/what/what-is-a-kubernetes-pod

Purdue Computing

- Purdue has six hardware clusters
  - The **Hammer cluster** is *completely reserved for CMS processing*.
  - The **Geddes cluster** is the only *K8s enabled* computing cluster at Purdue.
- Hammer has:
  - Series of hardware nodes (a-f) with different specifications.
  - On single A node: 20 Intel Xenon E5-2660 CPUs
  - On single F node: 40 AMD EPYC CPUs, 1 NVIDIA T4 GPU
  - Other nodes were not used in this study.
- Geddes has:
  - User access via interactive JupyterLab interface (Purdue Analysis Facility).
  - On single node:
    - 128 AMD EPYC 7662 CPUs.
    - 2 NVIDIA A100 GPUs (can be partitioned into up to 7 instances).

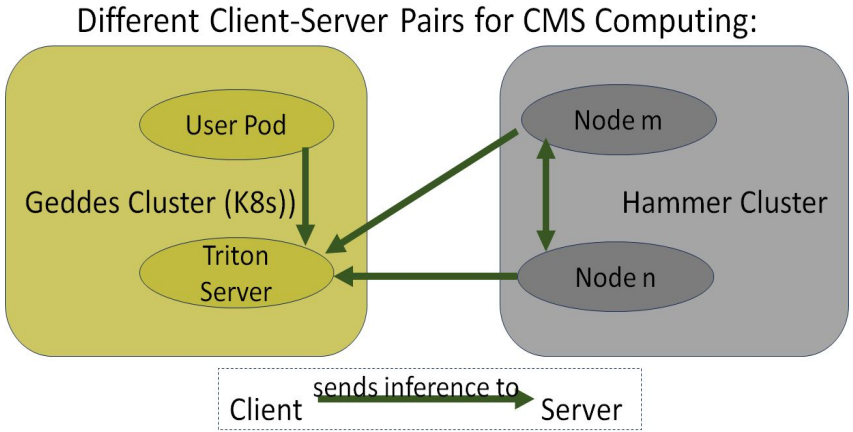
Servers and Clients

Client Options:

- Hammer GPU Node
- Hammer CPU Node
- Geddes user pod w/CPUs (Purdue Analysis Facility)

Server Options:

- Hammer server (no Kubernetes)
- Kubernetes Server via Geddes
- GPU vs CPU server



Current Measurements

1. MiniAOD Workflow

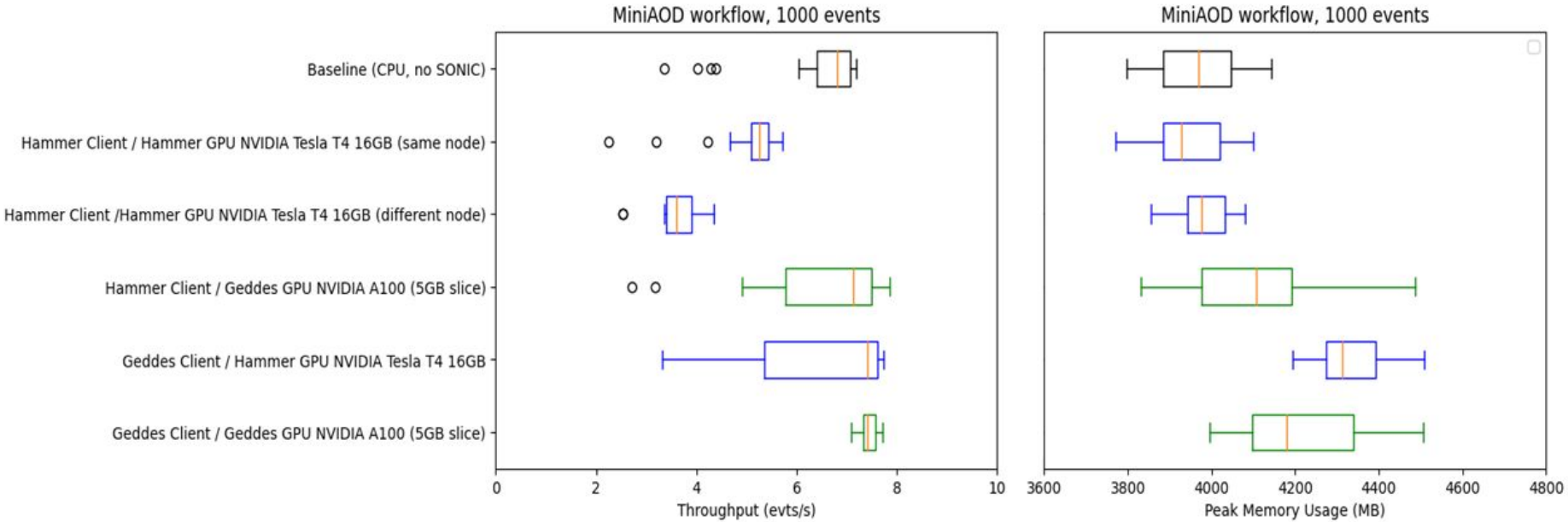
- Contains several data processing steps, including inference for multiple ML models
- In this study, we **offload a portion of the inference** for a model (DeepMET) onto a GPU using SONIC.
- Ran on a ROOT file full of collision data using two SONIC scripts.
- **Goal: Check for possible issues/overheads** for any form of client/server connection, be it from SONIC, Kubernetes, or some other factor.

2. Triton Performance Analyzer

- Array of zeros as input
- **Only includes NN inference**, which can be entirely offloaded to server - unrealistic, but can help us find existing overheads.
- Allows us to compare CMS model CPU performance with GPU
- **Goal: Find bottlenecks** (or lack thereof) in server/client connections and inference requests and **benchmark improvements given by using GPU**.

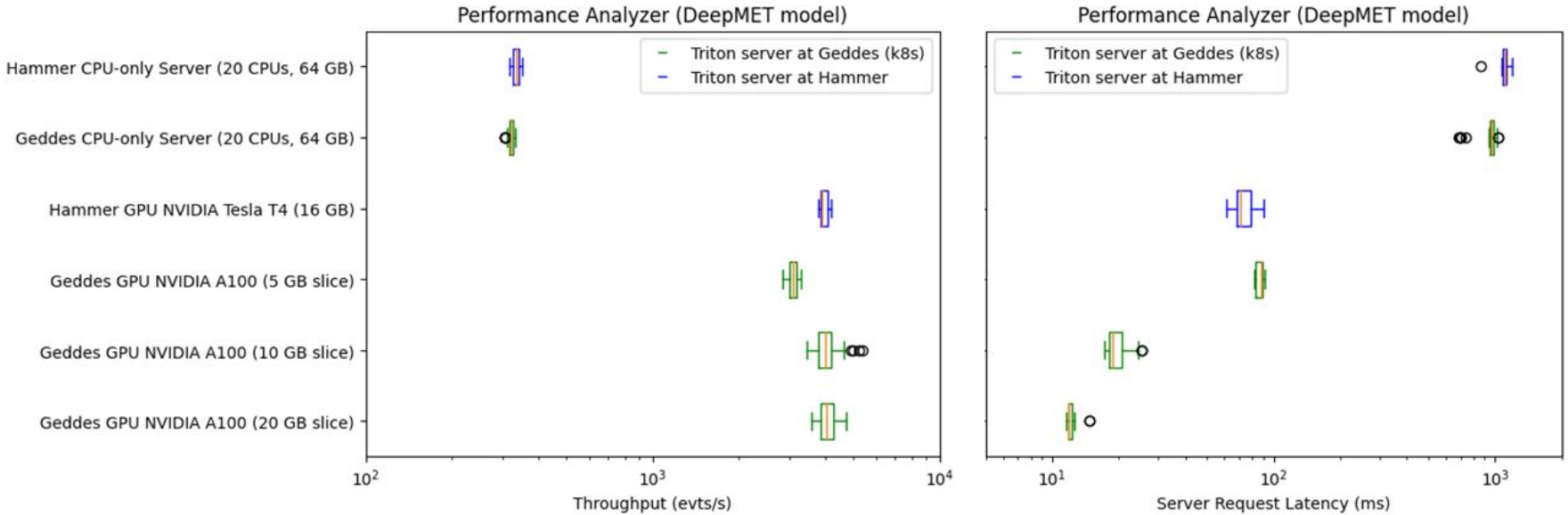
MiniAOD Results

- No obvious bottlenecks for any client/server pair
- No obvious difference with or w/o SONIC.



Triton Performance Analyzer Results

- Large latency decrease when using GPUs as compared to CPUs
- Minimal difference between Hammer/Geddes server for latency + throughput.



Future Plans

- **My Plans:**
  - Will prepare performance analyzer data scraping code for general use for SONIC development team.
  - Run different batch sizes on the performance analyzer to fully understand the extent of Geddes' GPU capabilities.
- **SONIC Plans:**
  - Implementation of load balancing using K8s based auto-scaling.
  - Automation of model transfer from CMSSW to format compatible with Triton Server
  - Gain better understanding of what's required for connection of different workflows to SONIC.
  - Implementation of SONIC at other CMS computing centers for better understanding of setting up SONIC workflows on different clusters/systems.
  - Implement SONIC for non-ML tasks.

References

1. M. Wang et al., GPU-Accelerated Machine Learning Inference as a Service for Computing in Neutrino Experiments, 2021, Front. Big Data 3:604083.

2. CMS Collaboration, “Portable Acceleration of CMS Production Workflow with Coprocessors as a Service”, 2023 (to be published)