Genesis Mission

AMERICAN SCIENCE CLOUD

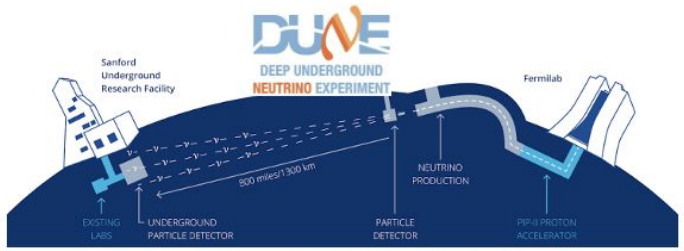
Shared Multi-Facility Inference Platform Across Multiple Experiments

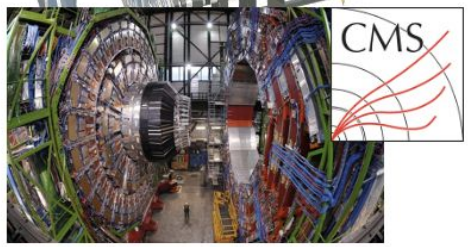
American Science Cloud BES/HEP/NP Scientific User Facility Infrastructure Partnership Project

Meghna Bhattacharya[1], Paolo Calafiura[2], Pengfei Ding[2], Julien Esseiva[2], Xiangyang Ju[2], Dmitry Kondratyev[3], Nhan Tran[1], Steven Timm[1], Mike Wang[1], Nicholas Schwarz[4]

1. Fermi National Accelerator Laboratory, 2. Lawrence Berkeley National Laboratory, 3. Purdue University, 4. Argonne National Laboratory

Science Motivation





Challenge: Turn exabytes of data into scientific discovery!

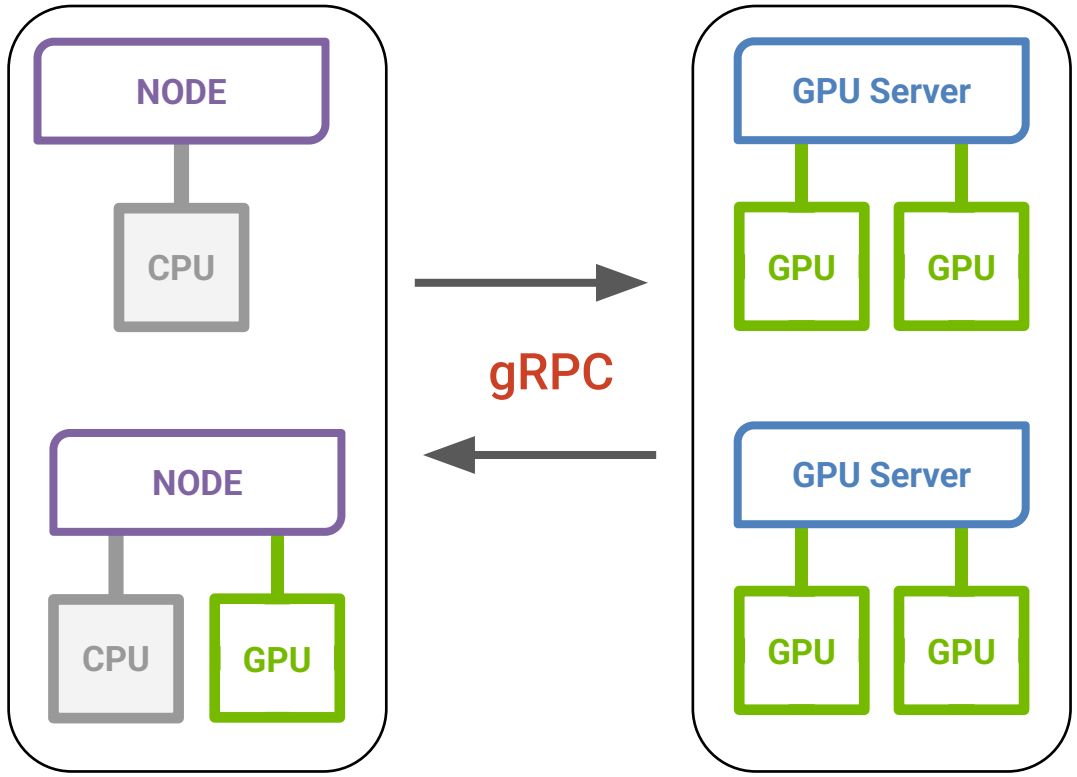
- HEP & NP experiments deploy AI-driven scientific workflows for **fast turnaround of enormous data into discovery**
  - Many ML models have been deployed into production, and more are in development

Need of the hour:

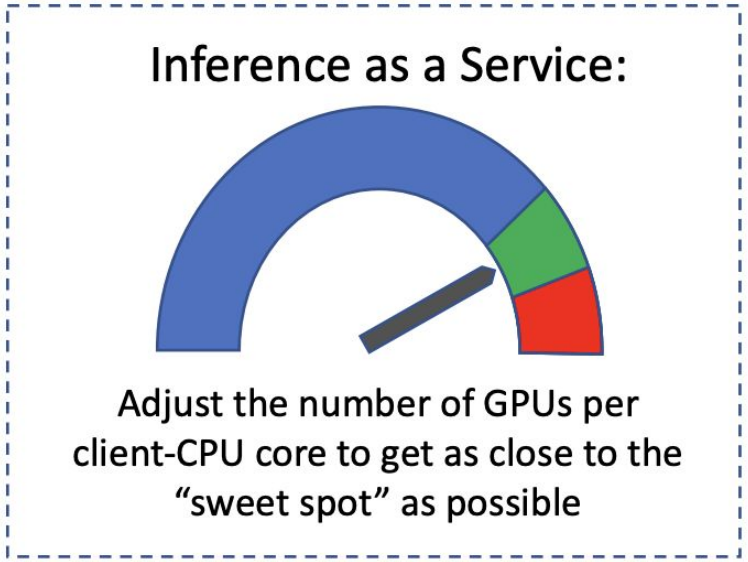
A scalable, maintainable way to accelerate **AI model inference** in production and ensure reproducibility

- **Scalable**: efficiently use heterogeneous resources
- **Maintainable**: validating, versioning, reproducibility, benchmarking

Inference as a Service



Inference as a Service (IaaS) provides a ***flexible*** and **scalable** deployment scheme.

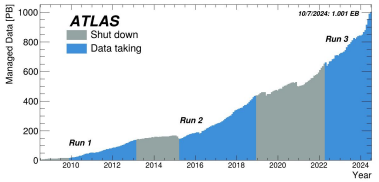


Or a fraction of a GPU for a model

Image credit: P. McCormack.

From Data to Insight: The AmSC-SUF Partnership

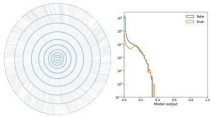
The goal is to establish **Genesis** as a **shared platform** to host and distribute AI **models** and scientific **data** for the research communities of multiple accelerator-based **BES, HEP, and NP Scientific User Facilities (SUFs)**, along with their industry and research partners.



Exabytes of Data

High-quality scientific datasets

arrow_downward



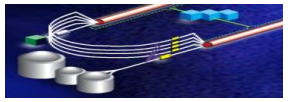
AI Models

Hundreds of production workflows

Initial Representative Demonstrators

Agentic Accelerator Controls

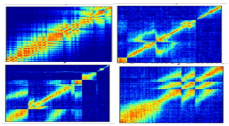
Cross-Facility Data Ingestion



Inference for HEP/NP Data

This talk

Light & Neutron Sources



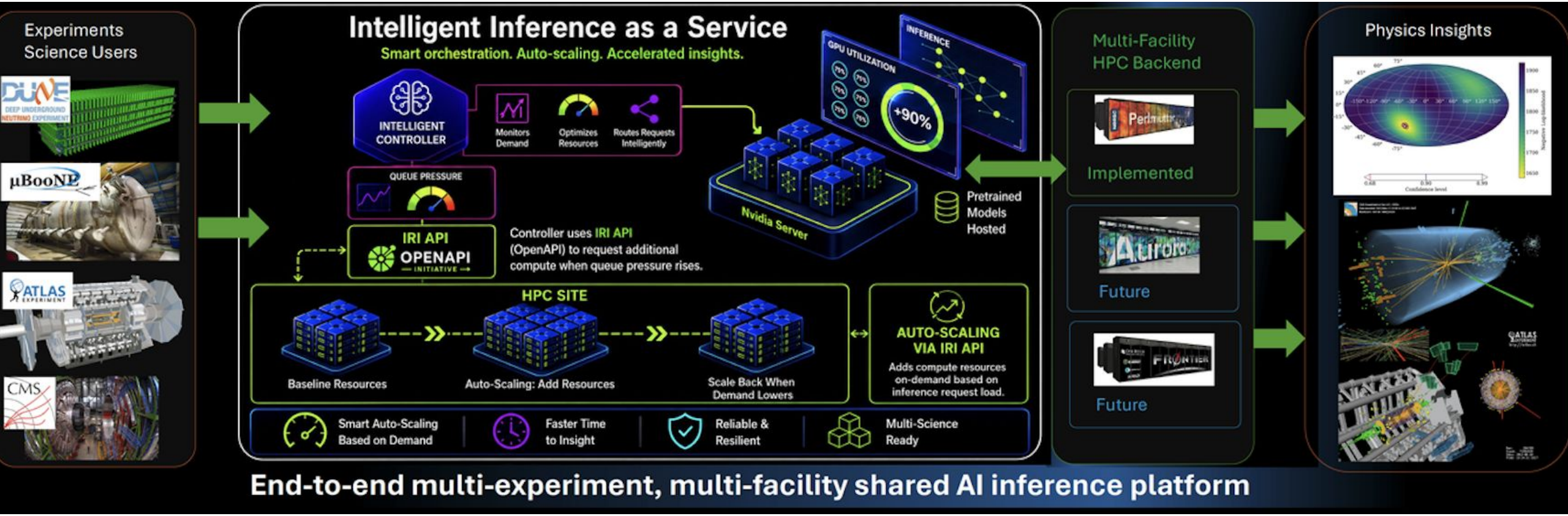
OpenCosmo in AmSC

These demonstrator datasets and models provide the **real-life requirements** and **benchmark solutions** needed to design and validate **AmSC Science Services** with the **Genesis Platform**.

Argonne NATIONAL LABORATORY · BERKELEY LAB · Brookhaven National Laboratory · Fermilab · Jefferson Lab · OAK RIDGE National Laboratory · SLAC

IaaS in a nutshell

Input: Detector data → **AI Model Inference:** Triton Server → **Output:** Physics information



Near-term deliverables

From demonstration to operations: Integrate AI-driven scientific workflows for the scientific user facilities and the user communities

- Expand IaaS to ALCF, OLCF
  - Early setup tests ongoing at ALCF
- Implement users feedback
  - Discussions on-going with DUNE Production team
- Dario Gill's vision of internet for science discovery
  - Dario's suggestion to this work: Add NSF resources and use-cases as well

We welcome collaboration!

American Science Cloud SUFs IP Inference as a Service Team


Pengfei Ding


Xiangyang Ju


Paolo Calafiura


Julien Esseiva

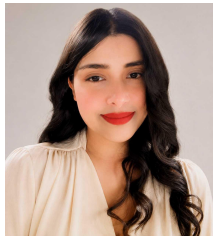

Michael Wang


Nhan Tran


Steven Timm


Meghna Bhattacharya


Giuseppe Cerati


Sparshita Dey


Dmitry Kondratyev

If you want to test this AmSC capability for your experiment please reach out to any of us

Scientific workflows

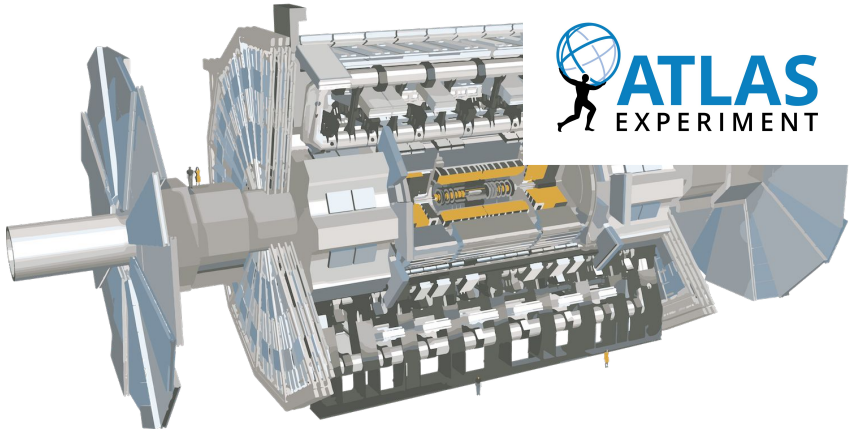
MicroBooNE Nugraph (300MB)
https://amsc.fnal.gov:2880/amsc/neutrino/uboone/

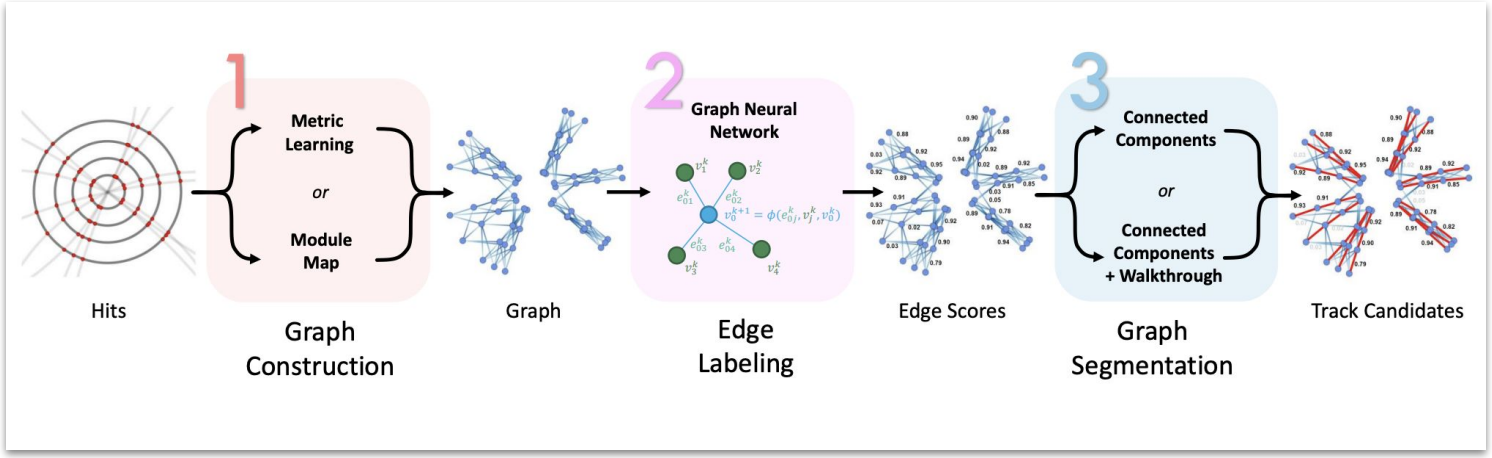
DUNE Supernova (30TB)
https://amsc.fnal.gov:2880/amsc/dune/

(Backup)

Demonstrated four scientific workflows across energy and intensity frontier experiments

Scientific workflows: ATLAS





- **GNN4ITK**: ATLAS Graph Neural Network-based Particle Tracking for next-generation ITk at High Luminosity Large Hadron Collider
  - Requiring high-end GPUs like NVIDIA A100 but not available in Worldwide LHC Computing Grid (WLCG)
  - ***IaaS allows ATLAS to use GPUs from remote sites such as AmSC***

- **DAOD Production**: ATLAS Derived Analysis Object Dataset (DAOD) Production
  - O(10) ML models deployed and more powerful ML models are being developed
  - Production tasks are already limited by memory
  - ***IaaS allows to offload ML algorithms to the Server, saving client memory and enabling large models***

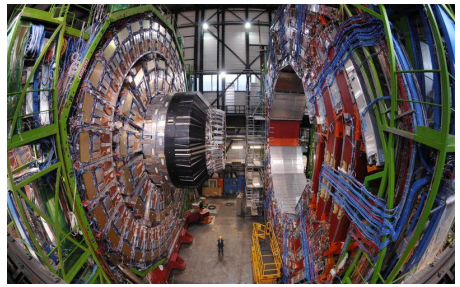
Scientific workflows: CMS

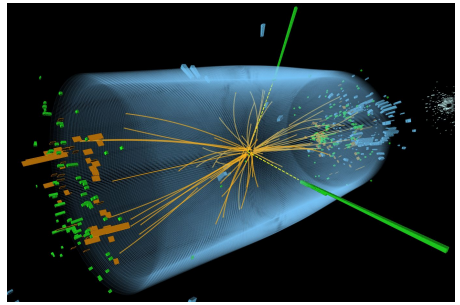
Realistic reconstruction workflow from CMS Run3 campaign

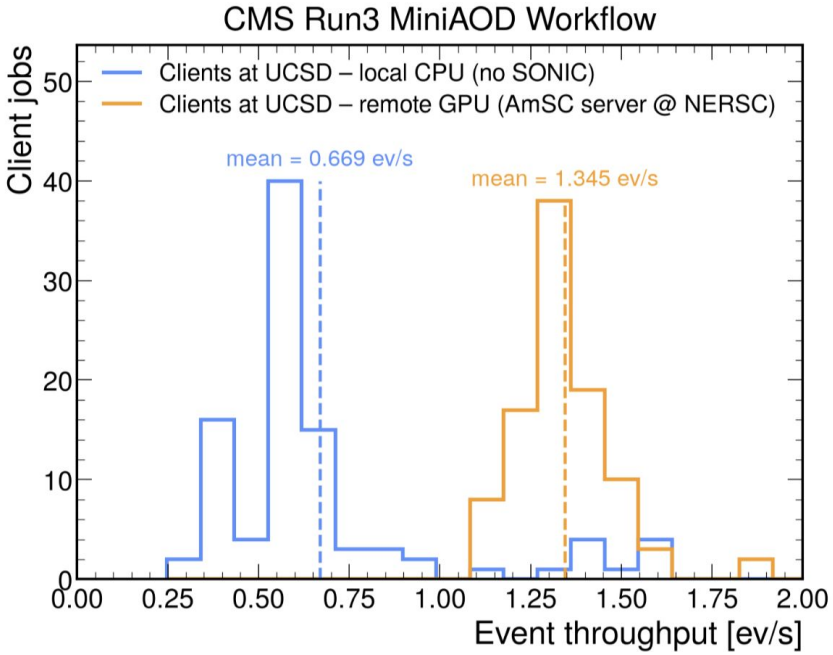
Asynchronous ML inference off-loading is implemented in CMS Software (CMSSW) via **SONIC.**

~10 ML models evaluated (ParticleNET, ParT, DeepTau, etc.)

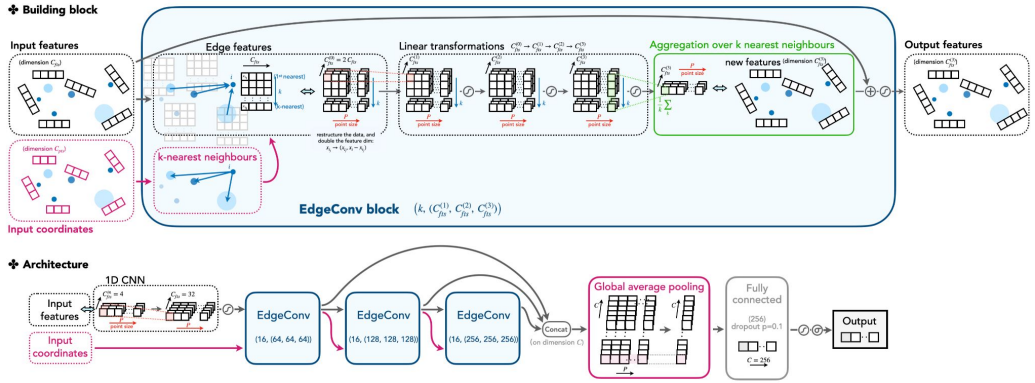
2x throughput improvement from GPU inference at AmSC



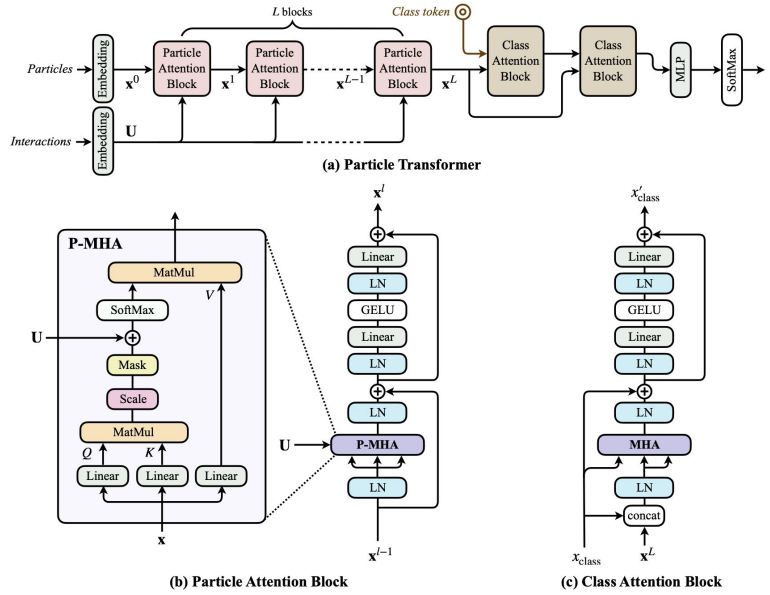




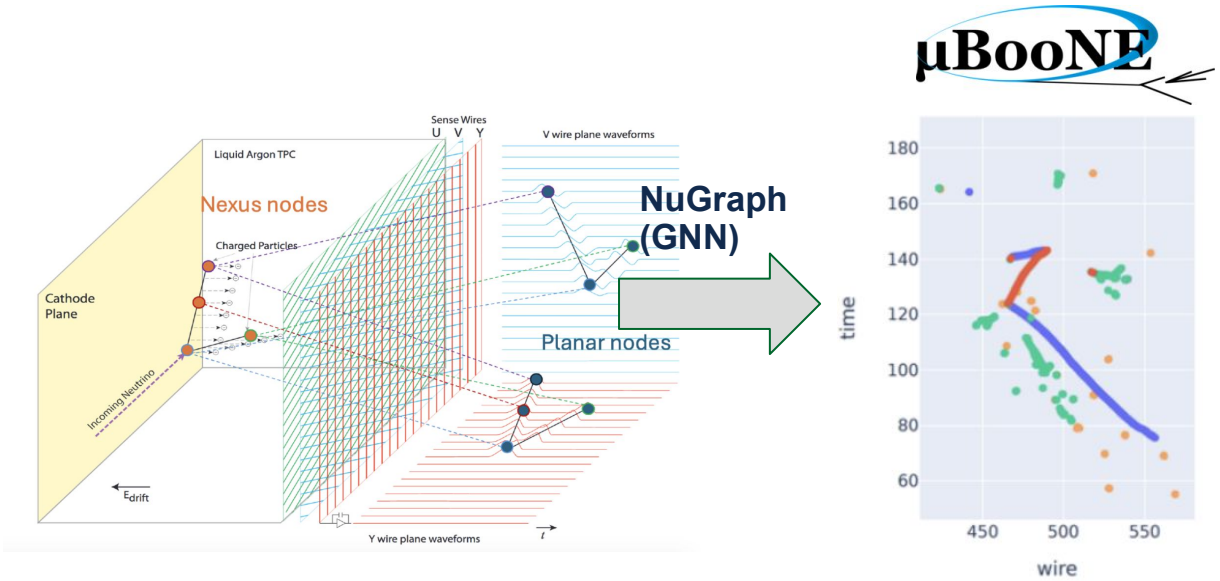
- **ParticleNET:** GNN jet tagging

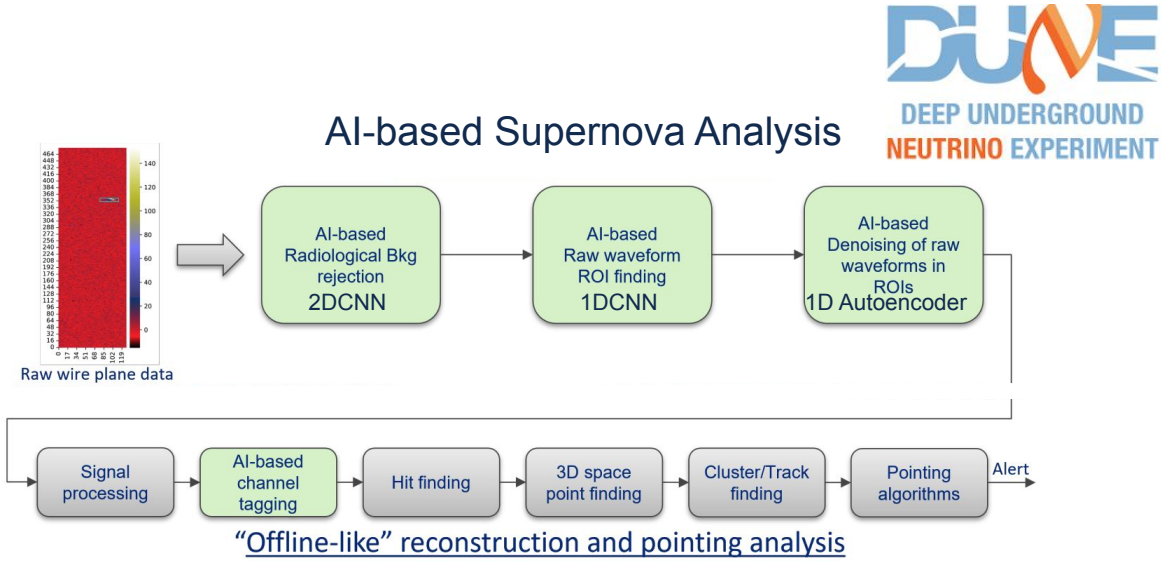


- **Particle Transformer (ParT)**



Scientific workflows: Neutrinos

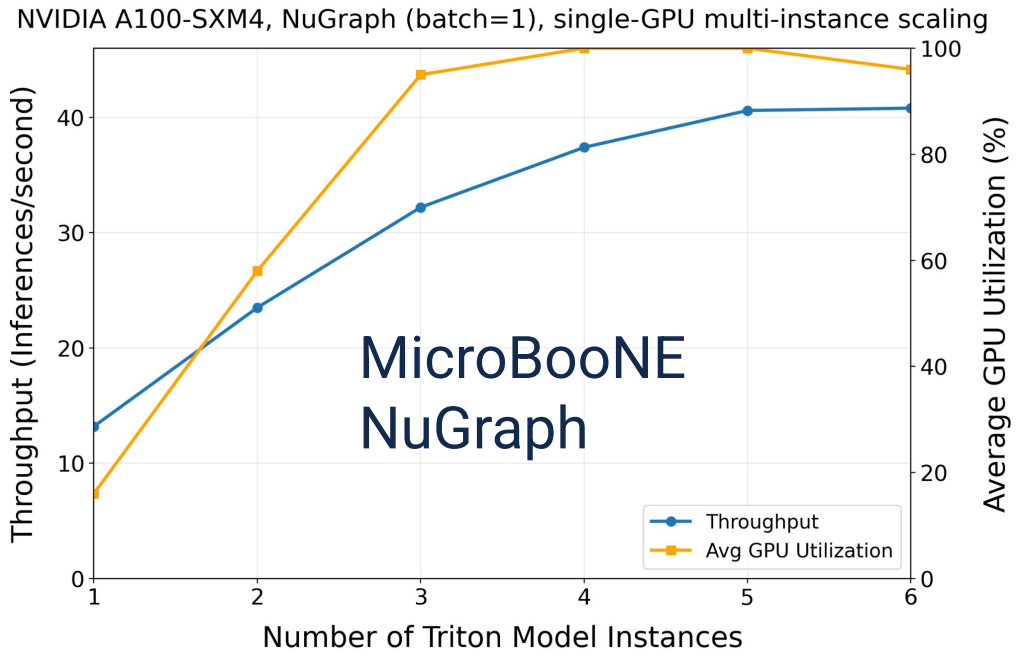


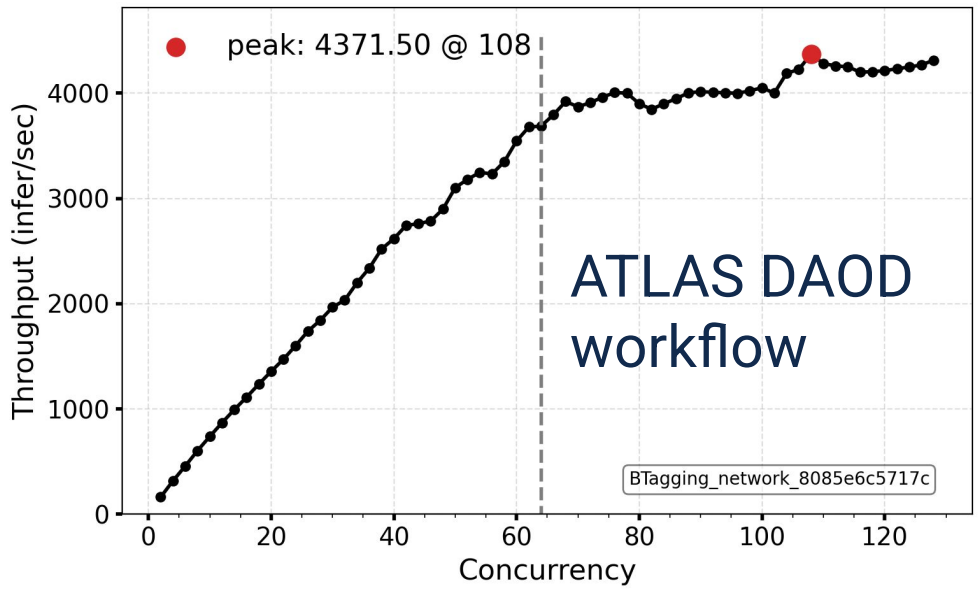


- **NuGraph (GNN-based reconstruction)**:
  - ***IaaS advantage***: *easier managing training and inference environments, avoids duplication of python libraries in C++, makes code maintenance easier*

- **DUNE ML-based Supernova analysis**:
  - Need to process 140 X 4 TB of raw data within minutes for time critical multi-messenger follow-up observations

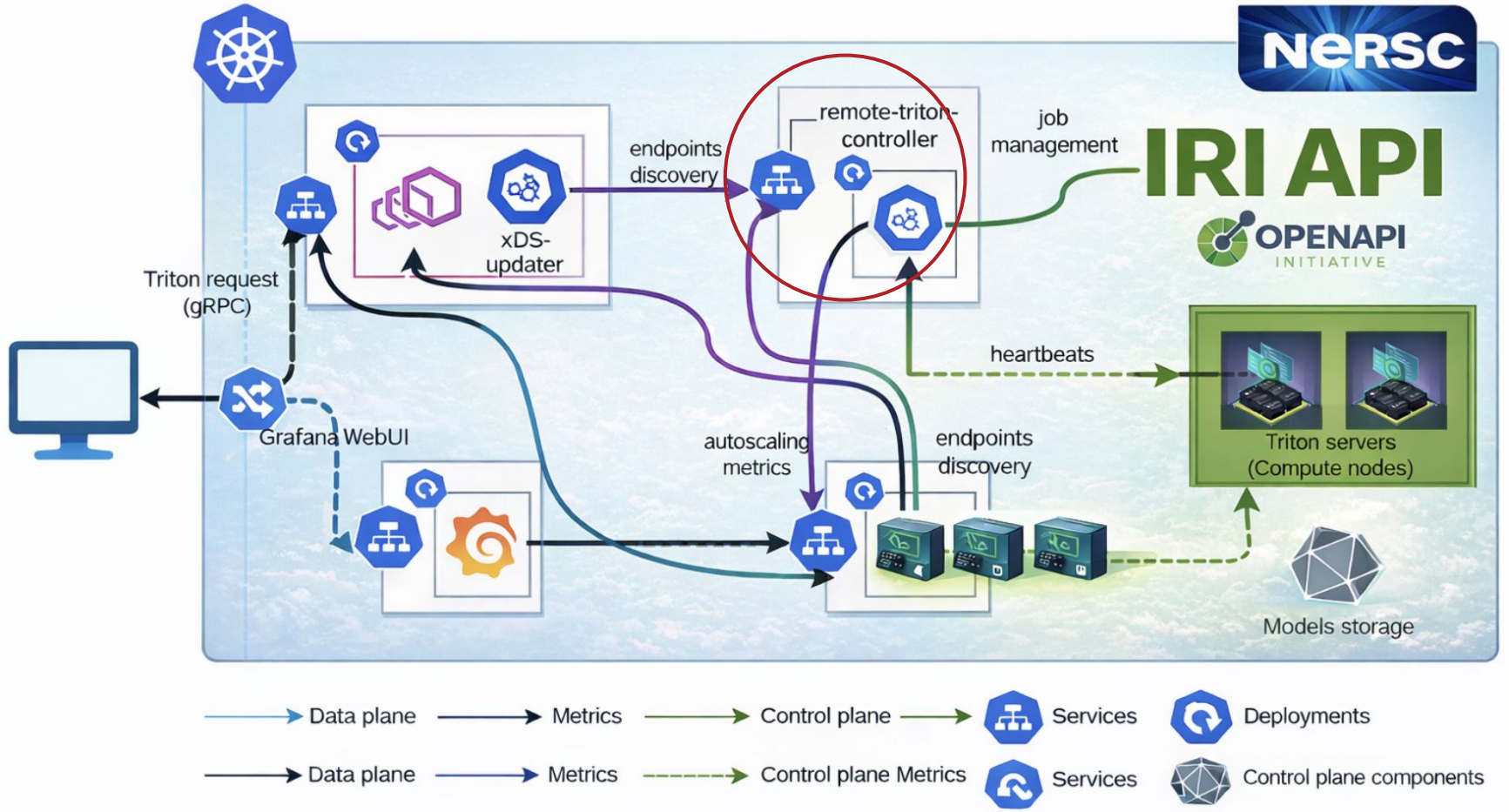
Benchmarking workflows at NERSC





- We benchmarked hosted ML models by varying the IaaS server configurations.
- The IaaS allows us to use GPUs at 100%.

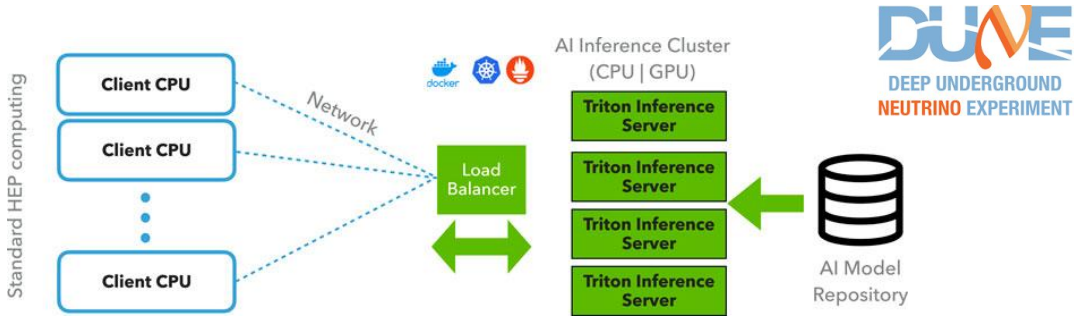
IaaS deployment on AmSC/Genesis Platform



Jointly developed an *intelligent* component: "Remote Triton Server Controller":

- Request computing resources via Integrated Research Infrastructure (IRI) API
- Automatic **scaling** computing resources based on number of requests in the queue
- Monitoring Dash and intelligent status report via Grafana, and more

Already demonstrated efficiency advantages

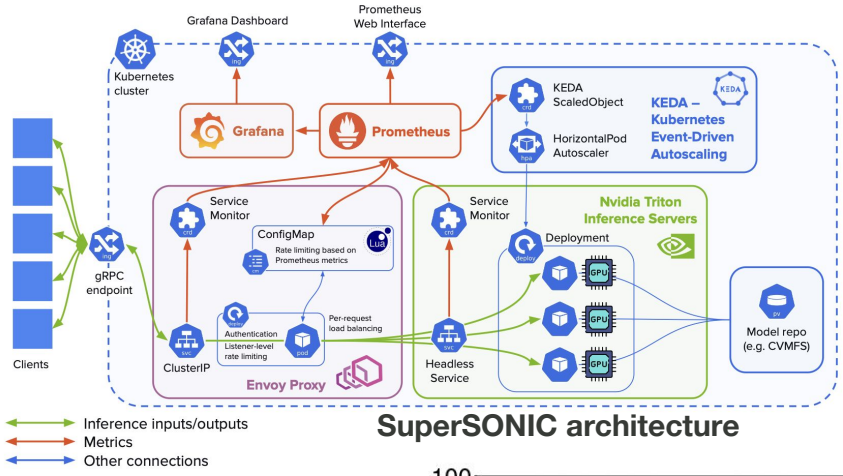


[M.Wang et al., Front. Big Data 3 (2020)]

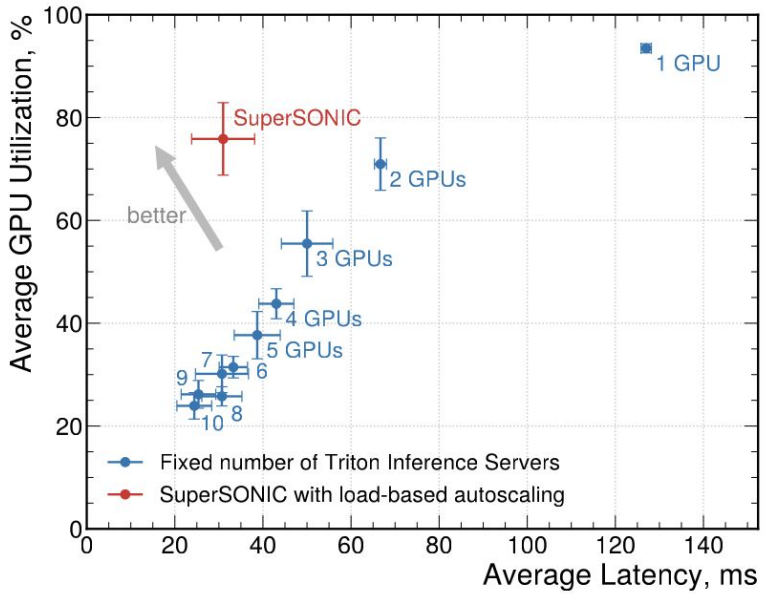
- With GPUaaS: Overall event processing time **2.4x faster,** a ratio of **68:1** simultaneous grid jobs per GPU

- ProtoDUNE beam data in 2021 with cloud-based GPU server

[T. Cai et al., Comput. Soft. Big Sci. 7, 11 (2023)]



- GPU resources provisioned in this way are more effectively used:
better utilization, better latency



[Kondratyev et al, https://arxiv.org/pdf/2506.20657]

First IaaS for HEP paper in 2019: [https://arxiv.org/abs/1904.08986]