

# SuperSONIC

Cloud-Native Infrastructure for ML Inferencing

**Yuan-Tang Chou<sup>3</sup>,**

Dmitry Kondratyev<sup>1</sup>, Benedikt Riedel<sup>2</sup>, Miles Cochran-Branson<sup>3</sup>,  
Noah Paladino<sup>4</sup>, David Schultz<sup>2</sup>, Mia Liu<sup>1</sup>, Javier Duarte<sup>5</sup>,  
Philip Harris<sup>4</sup>, and Shih-Chieh Hsu<sup>3</sup>

<sup>1</sup> Purdue University, <sup>2</sup> UW-Madison, <sup>3</sup> University of Washington, <sup>4</sup> MIT, <sup>5</sup> UCSD



**fast machine learning  
for science**

# ML and Coprocessors in HEP/MMA

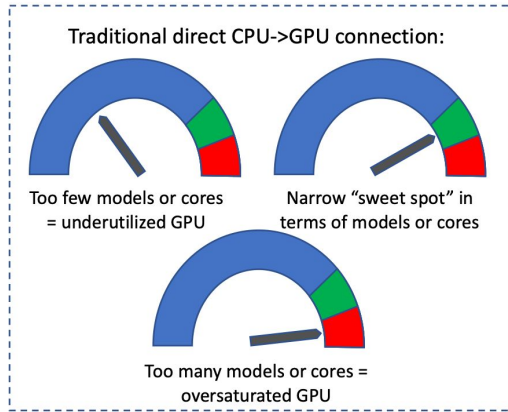
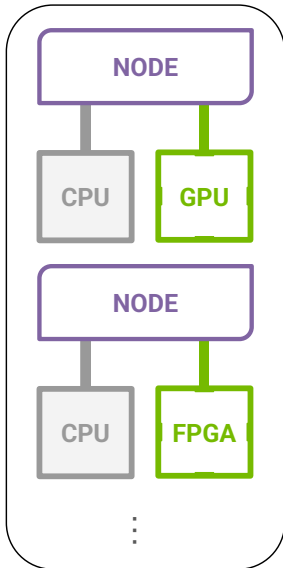
- The role of ML algorithms in HEP/MMA is growing, and so does their computational share in workflows.
- It's typical for trained models to run in production for years, accumulating high computational costs at **inference** stage.

**Use of coprocessors (GPU, FPGA, etc) is a must, but they are scarce and expensive – we need to use them efficiently.**

# Coprocessor Integration

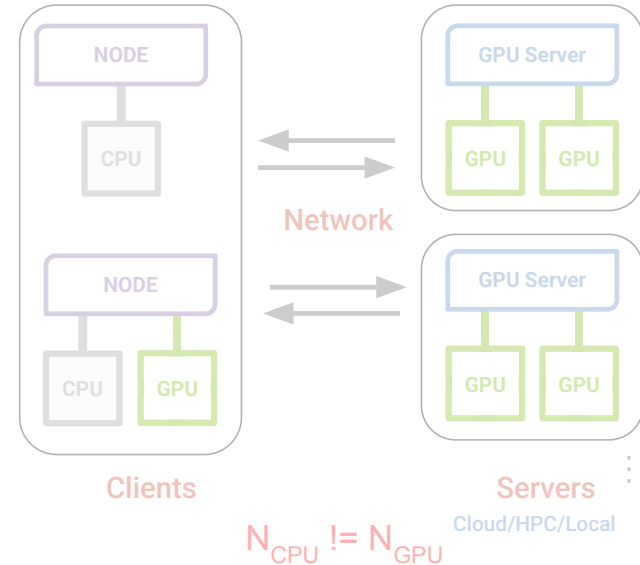
## Direct connection

Efficient only if inference load is known in advance, otherwise wastes resources or hurts throughput



## “As a service”

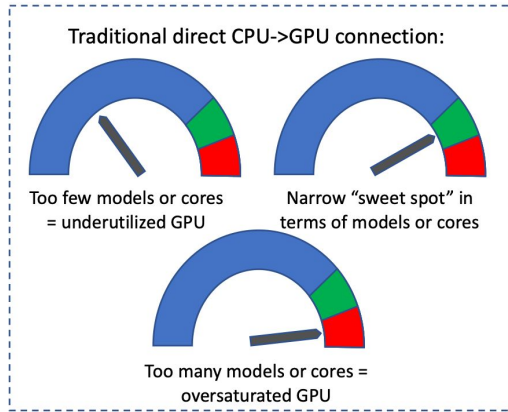
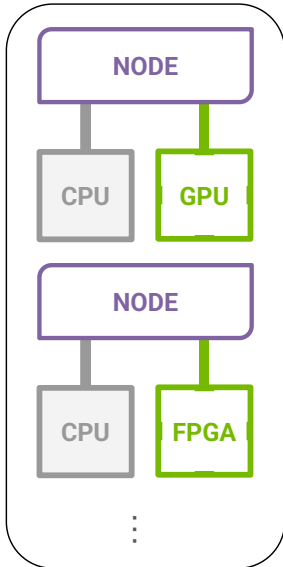
Allows to dynamically optimize resource usage, but adds R&D costs



# Coprocessor Integration

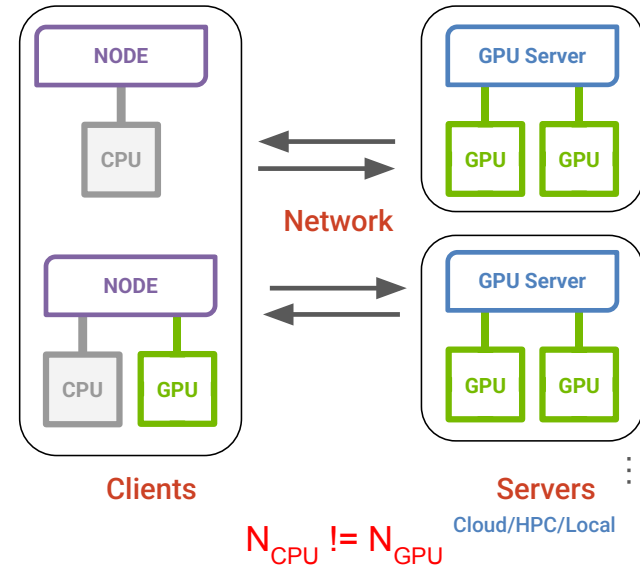
## Direct connection

Efficient only if inference load is known in advance, otherwise wastes resources or hurts throughput



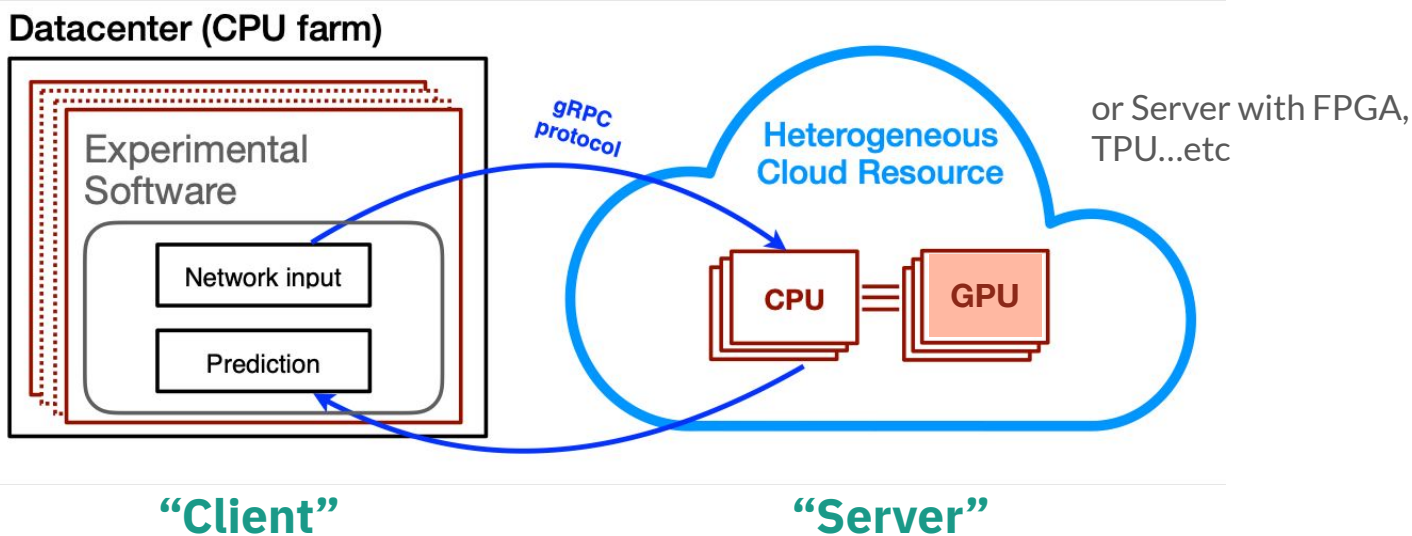
## “As a service”

Allows to dynamically optimize resource usage, but adds R&D costs



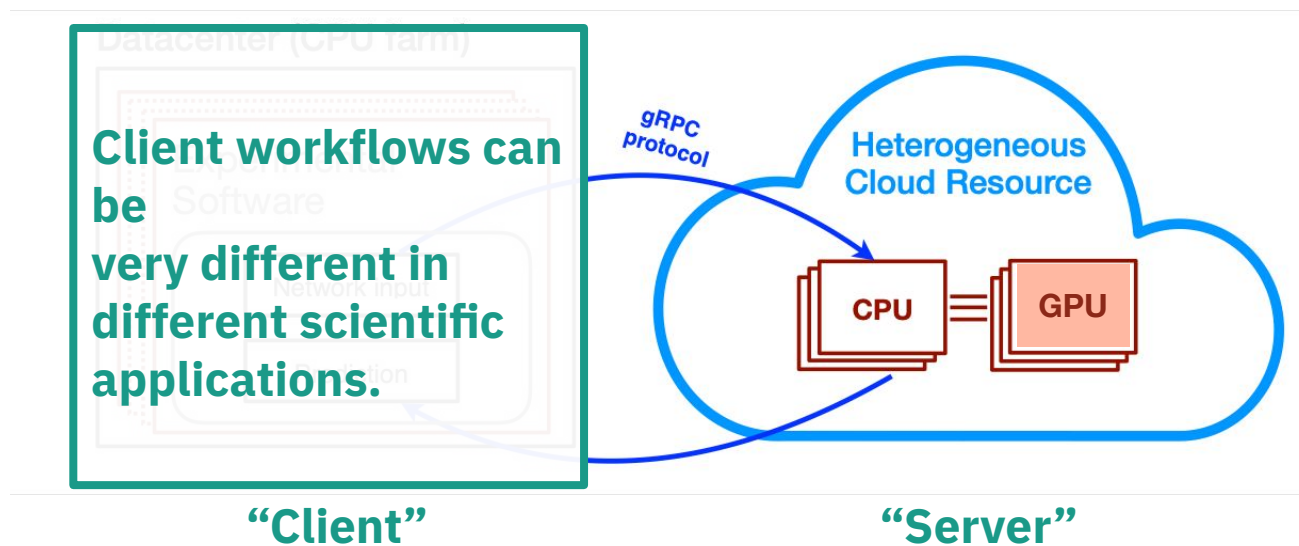
# Inference as a Service

Design principles formulated over past 7 years under the umbrella of **SONIC: Services for Optimized Inference on Coprocessors**



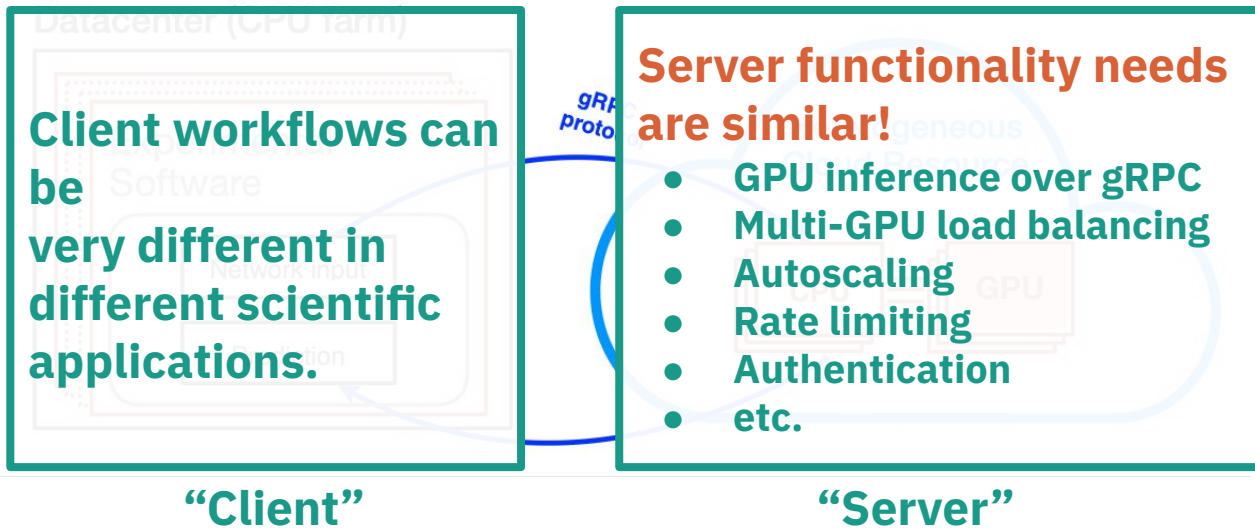
# Inference as a Service

Design principles formulated over past 7 years under the umbrella of **SONIC: Services for Optimized Inference on Coprocessors**



# Inference as a Service

Design principles formulated over past 7 years under the umbrella of **SONIC: Services for Optimized Inference on Coprocessors**



# **SuperSONIC**

This idea sprouted at FastML 2024@Purdue last year!

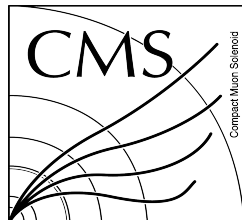
Engineers from multiple HEP / MMA experiments joined efforts to develop **SuperSONIC** – a common server infrastructure for GPU-accelerated inference!

CMS: ParticleNet, mini-AOD productions

ATLAS: GNN4ITK and tracc tracking

ICeCube: DNN/CNNs Evt classifier

LIGO: ML-based matched-filtering pipeline



# Key Features

Designed for Kubernetes deployment

Built around NVIDIA Triton Inference Server

Relies on mature open-source cloud-native tools from Cloud Native Computing Foundation:

**Envoy Proxy:**

- Load balancing
- Rate limiting
- Token-based authentication

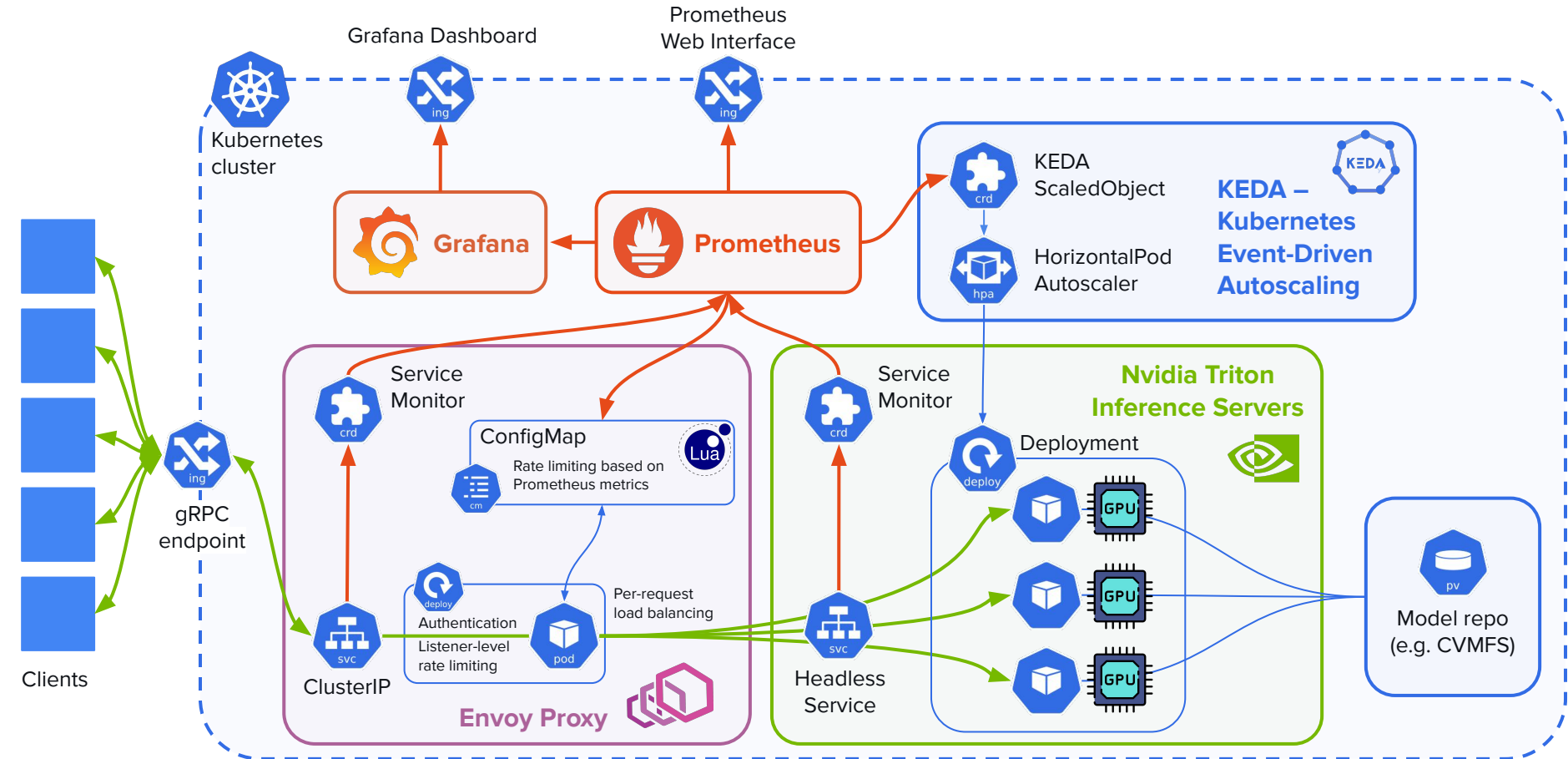
**KEDA:**

- Autoscaling

**Prometheus/Grafana/OpenTelemetry:**

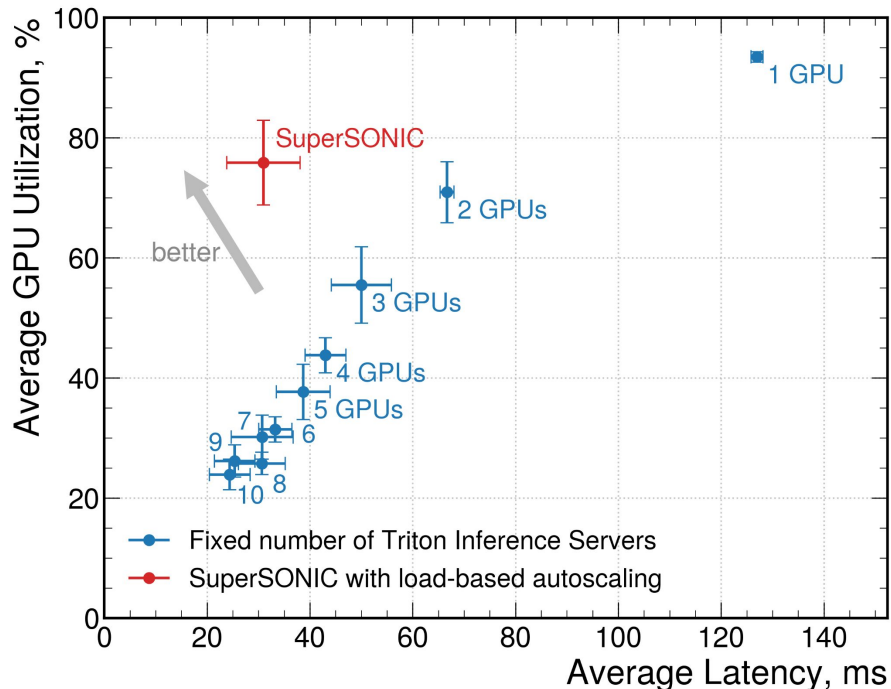
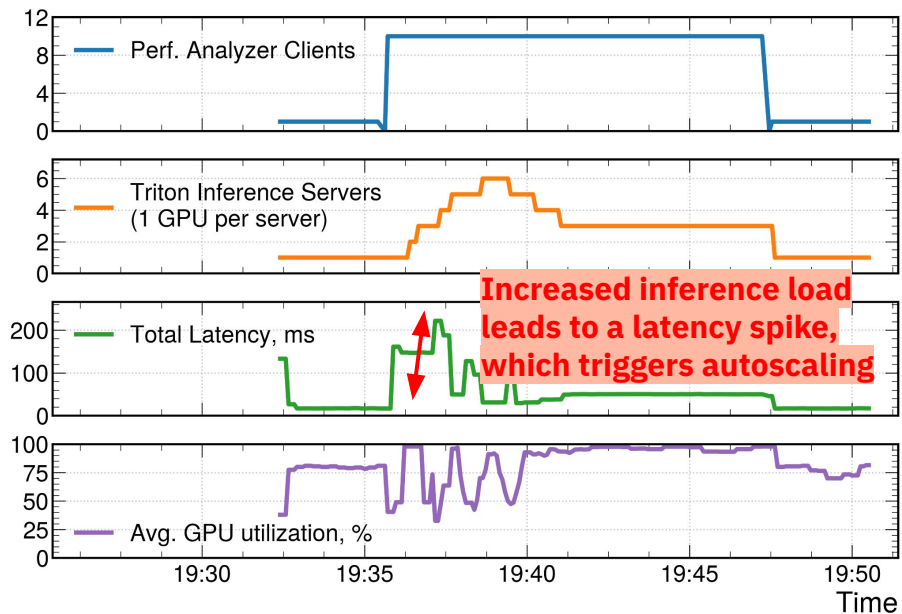
- Monitoring





# Performance

Autoscaling allows us to dynamically adjust number of GPUs based on inference load. This leads to high average GPU utilization, while keeping the inference latency low.



# Ongoing Work

## Extend Kubernetes into batch for GPU provisioning

- Currently, SuperSONIC requires a Kubernetes cluster with GPUs on the nodes.
- We are working on integrating [interLink](#) project, which allows to extend Kubernetes cluster into batch (Slurm, HTCondor, etc.) using Virtual Kubelets.

## Dynamic model loading

- When serving multiple models in a SuperSONIC server, we currently load all models into all GPUs. So far it is not a problem because models are small.
- We are working on a solution to dynamically load / unload models to address potential issues due to limited GPU memory.

# Summary

A cross-experiment solution to deploy SONIC workflow to K8s clusters using Nvidia Tritons

Working toward production ready.

Add your favorite HPCs here!

For more info:

- [github.com/fastmachinelearning/SuperSONIC](https://github.com/fastmachinelearning/SuperSONIC)
- [arXiv:2506.20657](https://arxiv.org/abs/2506.20657)

We thank the NSF HDR A3D3 institute and the US CMS software and computing program to support in this project!

## Status of deployment

	<a href="#">CMS</a>	<a href="#">ATLAS</a>	<a href="#">IceCube</a>
<a href="#">Purdue Geddes</a>	✓	-	-
<a href="#">Purdue Anvil</a>	✓	-	-
<a href="#">NRP Nautilus</a>	✓	✓	✓
<a href="#">UChicago</a>	-	✓	-



Accelerated AI  
Algorithms for  
Data-Driven  
Discovery

