“AI/ML Computing Across Scientific Disciplines”

Invited Presentation

Sixth National Research Platform Conference (6NRP)

UC San Diego

January 30, 2025

Dr. Larry Smarr

Founding Director Emeritus, California Institute for Telecommunications and Information Technology;

Distinguished Professor Emeritus, Dept. of Computer Science and Engineering

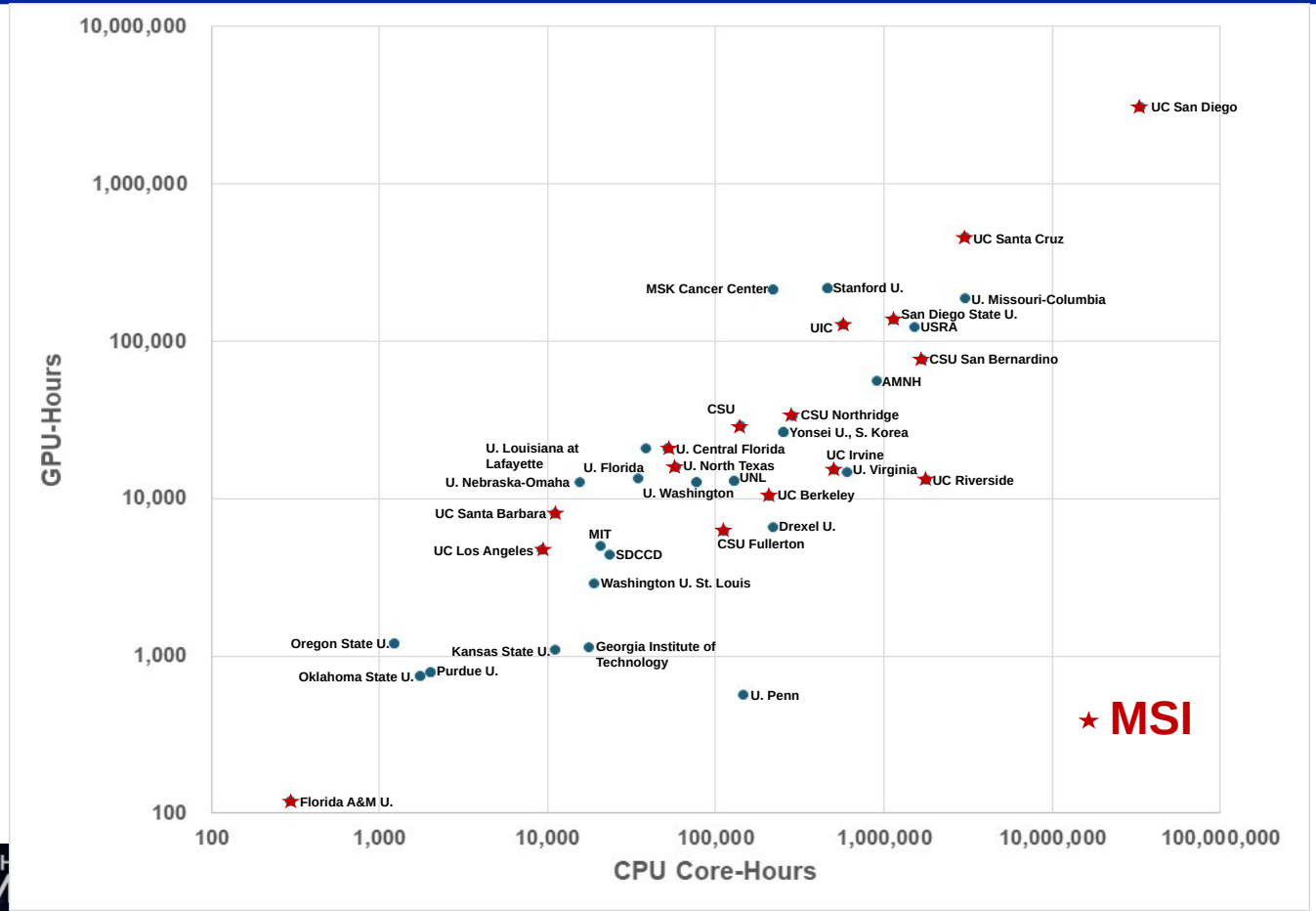
Jacobs School of Engineering, UCSD

http://lsmarr.calit2.net

NRP NATIONAL RESEARCH PLATFORM

NRP Nautilus

Nearly 40 Campuses Used Hundreds to Millions of NRP GPU-Hrs & CPU Core-Hrs in Calendar 2024, Including 16 MSIs and 4 EPSCoR States



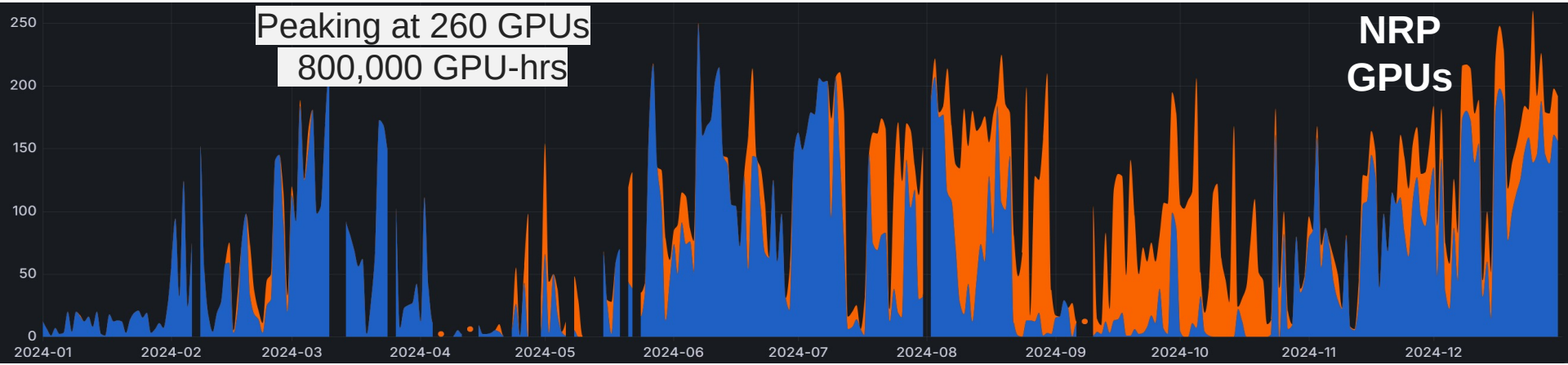
NRP NATIONAL RESEARCH PLATFORM

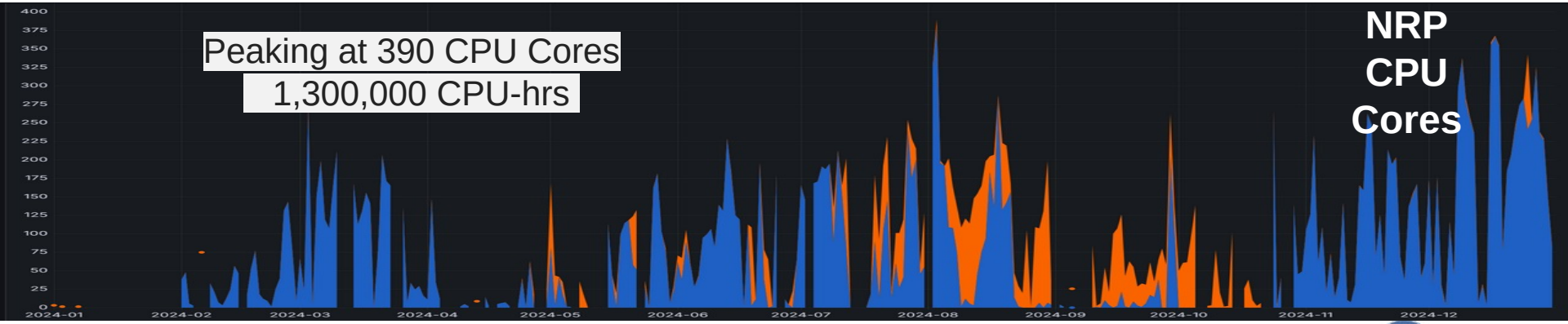
NRP Nautilus

We are Seeing an Increasing Number of Scientific Disciplines Using NRP ML/AI/DS Computations on Large Datasets

- **Particle Physics**
- **Astronomical Observations**
- **Biomedical Data**

IceCube Neutrino & LIGO Gravitational Wave Detectors Are Major NRP Users - GPU/CPU Core Usage Per Day, Calendar 2024





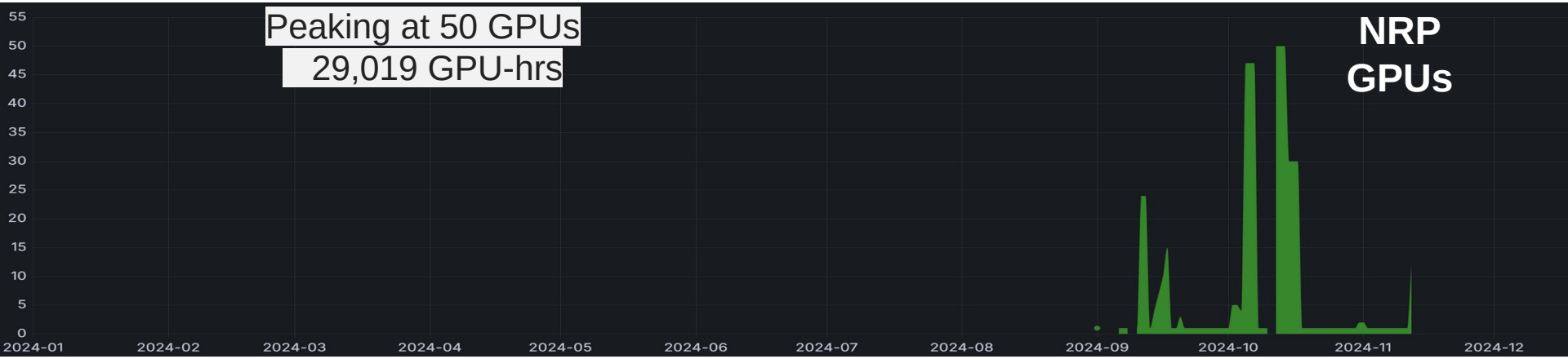
NRP NATIONAL RESEARCH PLATFORM

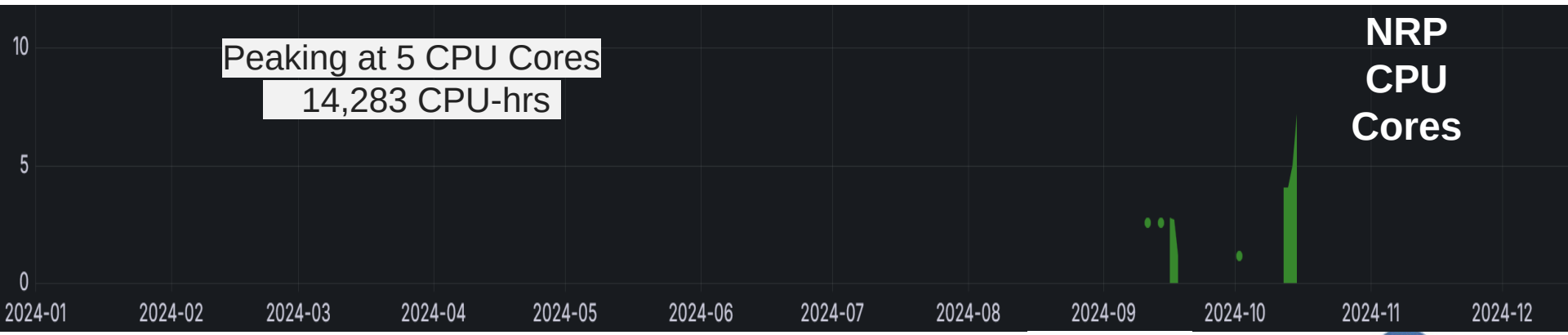
Namespaces osg-icecube & osg-ligo

Grafana

NRP Nautilus

IceCube Neutrino Detector Is Just Beginning to use ML
GPU/CPU Core Usage Per Day, Calendar 2024





NRP NATIONAL RESEARCH PLATFORM

Namespaces icecube-ml

Grafana

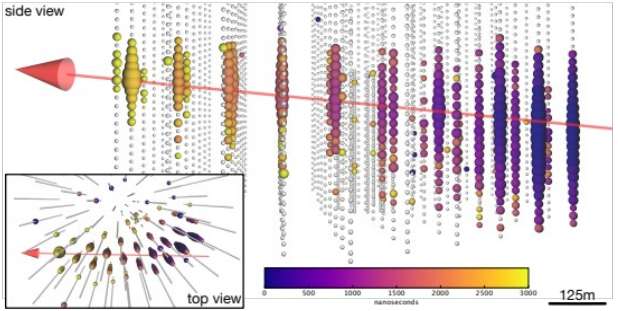
NRP Nautilus

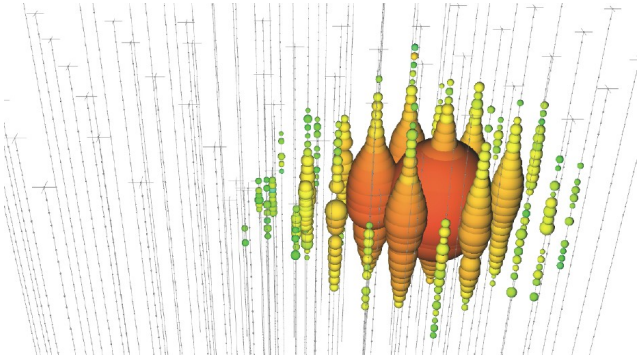
IceCube ML – Differentiate Event Topology

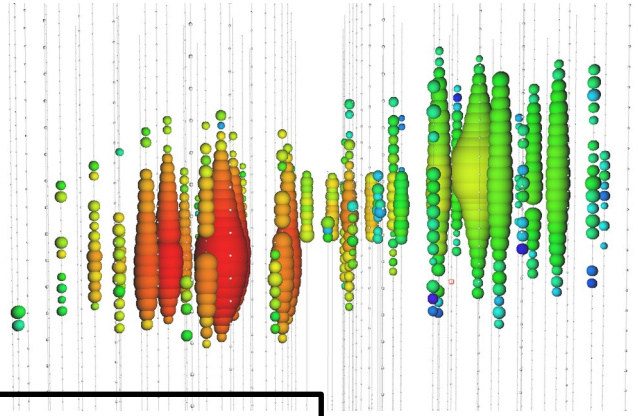
Track – Line

Cascade - Ball

Double Bang – Dumb Bell







Using a CNN to differentiate topologies in a single algorithm for the first time

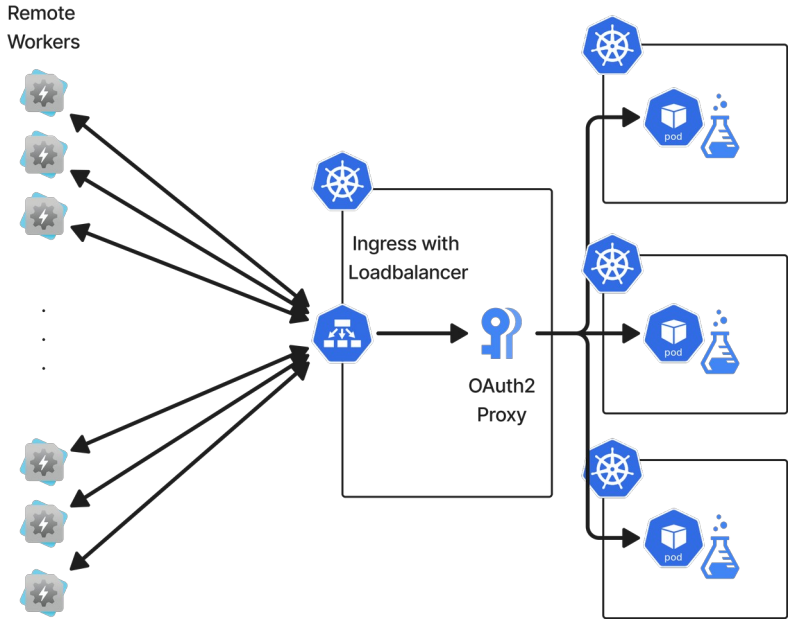
Source: Benedikt Riedel
IceCube Neutrino Observatory
Accelerated AI Algorithms for Data-Driven Discovery
University of Wisconsin-Madison

NRP NATIONAL RESEARCH PLATFORM

NRP Nautilus

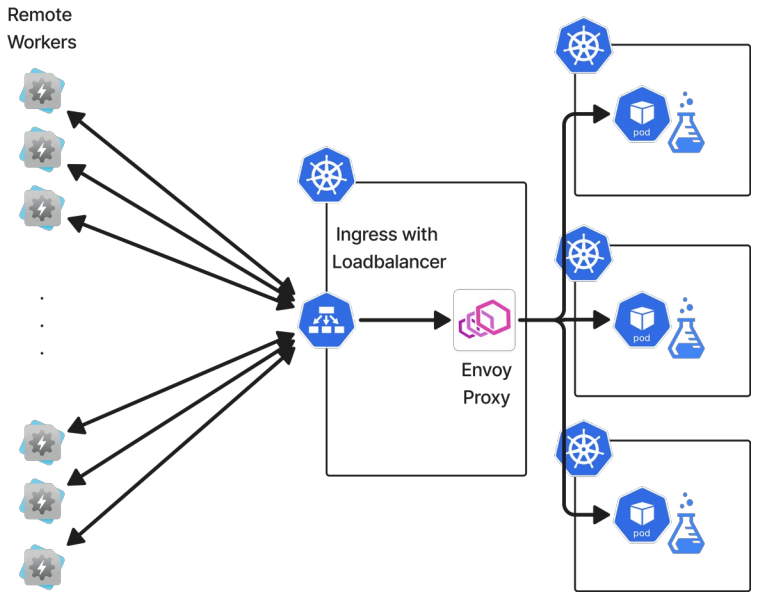
ML Inference-as-a-Service – Taking it to Production

- National resources have GPUs, but not Kubernetes interface – Except NRP and JetStream2
- Deploying on NRP has some challenges
  - Not in control of the cluster
  - Barebone setup in terms of Kubernetes extensions
    - Horizontal Pod Scalers are not readily supported
    - Custom Metrics take admin intervention
  - Distributed infrastructure
    - Issues with connection between ingress, service, and pods – Service not found in time
    - Overloading regional ingresses – DNS look-up issues

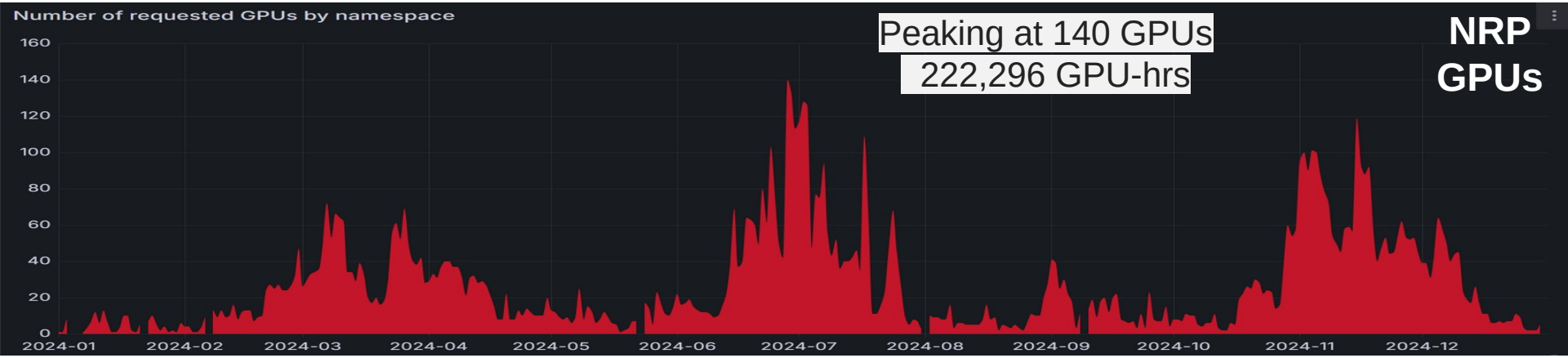


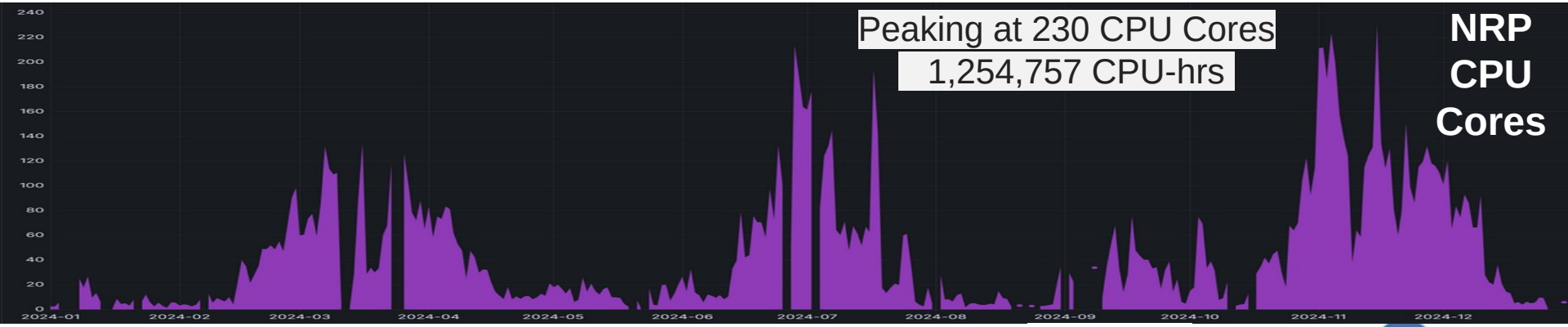
ML Inference-as-a-Service – Taking it to Production

- **CMS, LIGO, ATLAS and IceCube are essentially doing/wanting to do the same thing – Lets work together!**
- **SuperSONIC**
  - **Common Helm chart to deploy SONIC on NRP (or other Kubernetes cluster)**
  - **Enough config parameters to make it work for all experiments!**
  - **Biggest Change is using envoy proxy and KEDA for scaling pods**



GPU/CPU Core Usage Per Day, Calendar 2024





NRP NATIONAL RESEARCH PLATFORM

Namespace cms-ml

Grafana

NRP Nautilus

Self-Supervised Learning (SSL) for Jet Tagging

Using NRP to train and evaluate foundation models for jet tagging, and (in the future) other jet related tasks

Evaluating the dataset scalability of foundation models

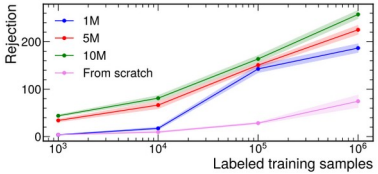


Figure 3: Background rejection at 50% signal efficiency of different training strategies, with legend indicating the number of unlabeled training samples used in pretraining

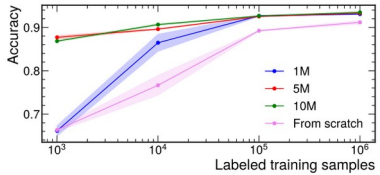


Figure 4: Accuracy of different training strategies, with legend indicating the number of unlabeled training samples used in pretraining

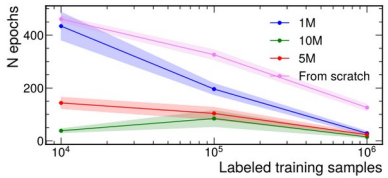
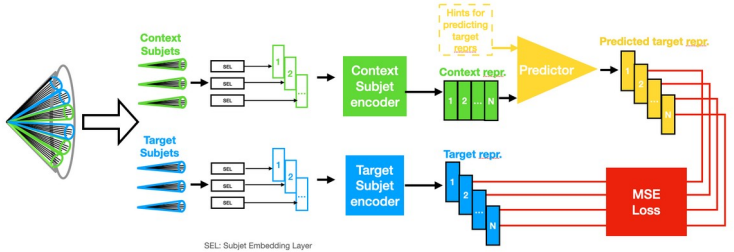


Figure 5: Number of epochs required to reach within 1% of the final accuracy of different training strategies, with legend indicating the number of unlabeled training samples used in pretraining

https://arxiv.org/abs/2408.09343

Developing new SSL strategies for training foundation models



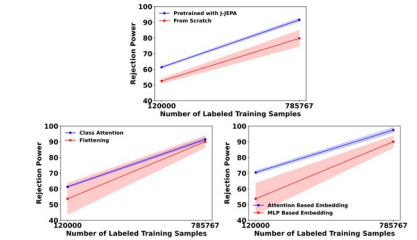


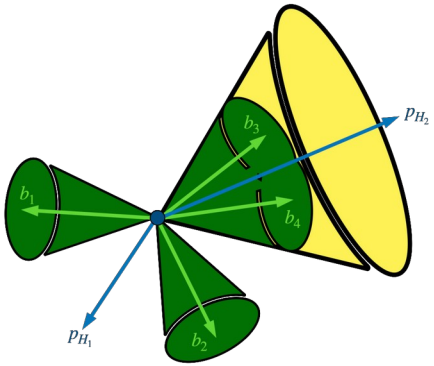
Figure 2: Comparison of the background rejection metric $1/\varepsilon_B(\varepsilon_S = 0.5)$ as a function of the number of labeled training samples used for finetuning for a pretrained model versus the one trained from scratch (top), for a pretrained model using Cls Attn versus Flatten for aggregation (bottom left), and for a pretrained model using an attention-based empbedding versus an MLP (bottom right). The shaded bands represent standard deviations calculated from 5 identical trials with random initialization.

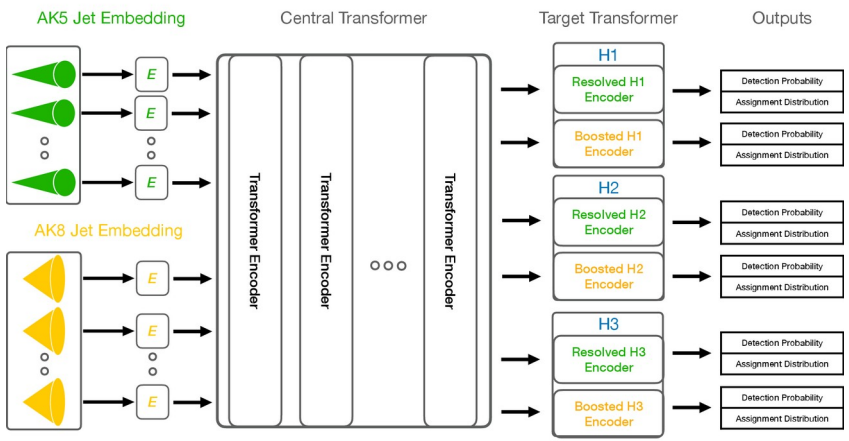
https://arxiv.org/abs/2412.05333

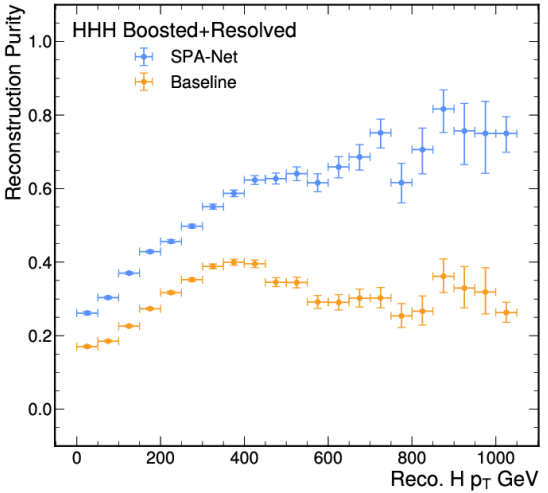
Zihan Zhao, Billy Li

Reconstructing Higgs Bosons with Symmetry-Preserving Transformers

Symmetry-preserving transformers applied to simultaneously reconstruct multiple Higgs bosons from low to high energies at the CERN LHC







https://arxiv.org/abs/2412.03819

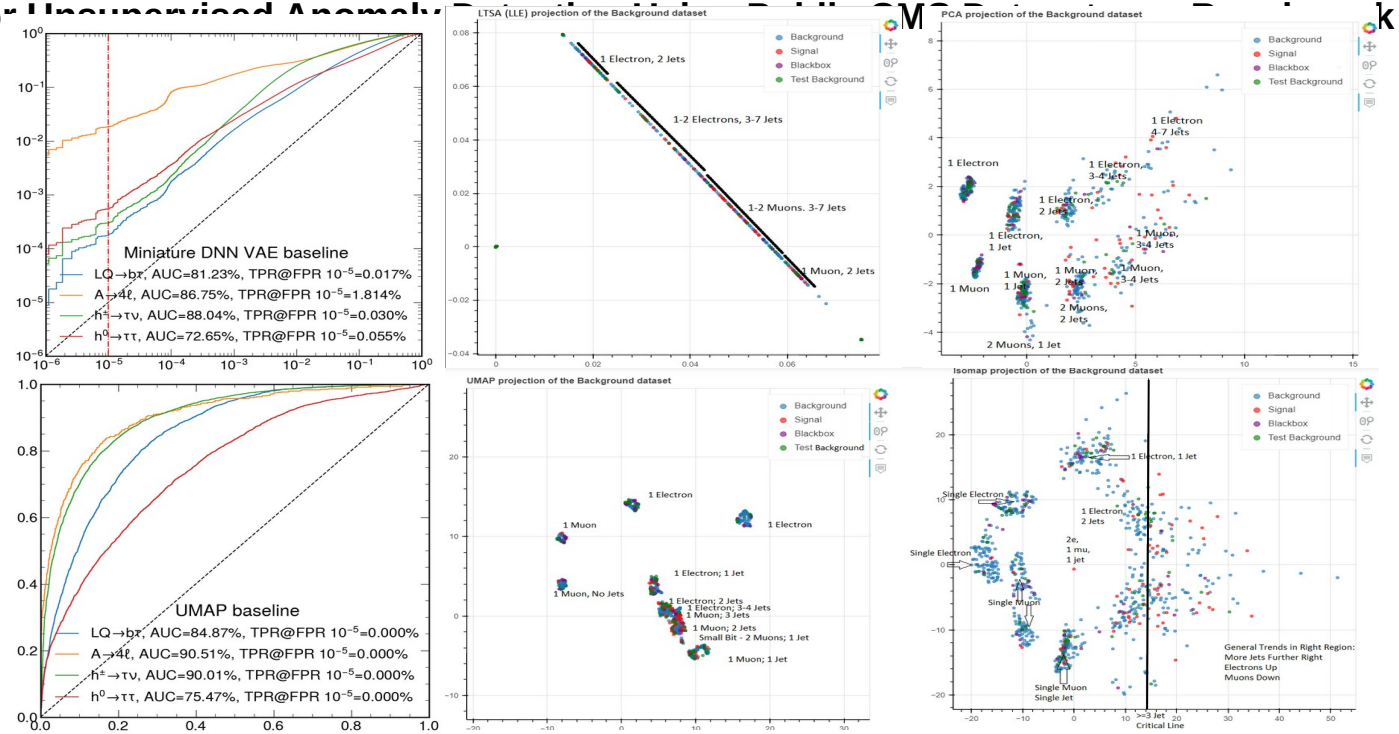
Billy Li

NRP NATIONAL RESEARCH PLATFORM

NRP Nautilus

Anomaly Detection for LHC Trigger

Exploring Alternative Dimension Reduction Based ML Architectures & Algorithms



Benchmark variational autoencoder vs. UMAP, LDLE, t-SNE …

Rohan Sachdeva, Melissa Quinnan

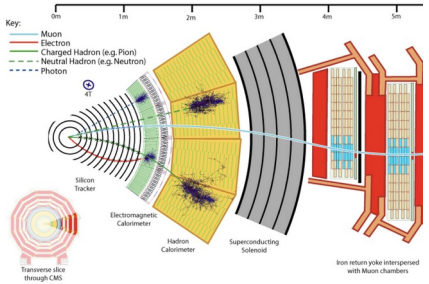
NRP NATIONAL RESEARCH PLATFORM

NRP Nautilus

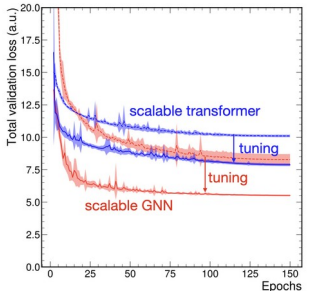
Machine-Learned Particle-Flow Reconstruction

We study *Scalable neural networks* for event reconstruction at current and future colliders.

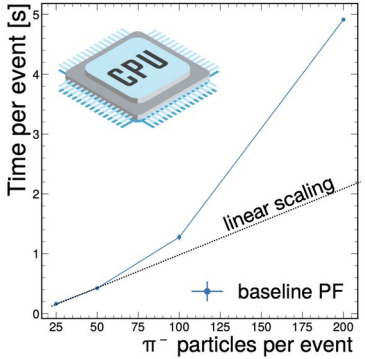
Current and future multilayered detectors need complex data reconstruction → particle flow algorithm



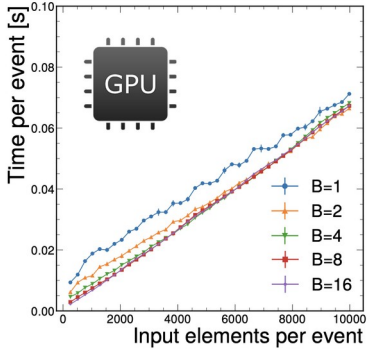
Model hypertuning achieves better than SOTA performance!



Baseline (untuned) algo runs **only on CPU**, scales ~quadratically, **runtime per event is in seconds**.



ML model **scales linearly**, runs in **milliseconds per event** on a consumer 8GB GPU





Farouk Mokhtar

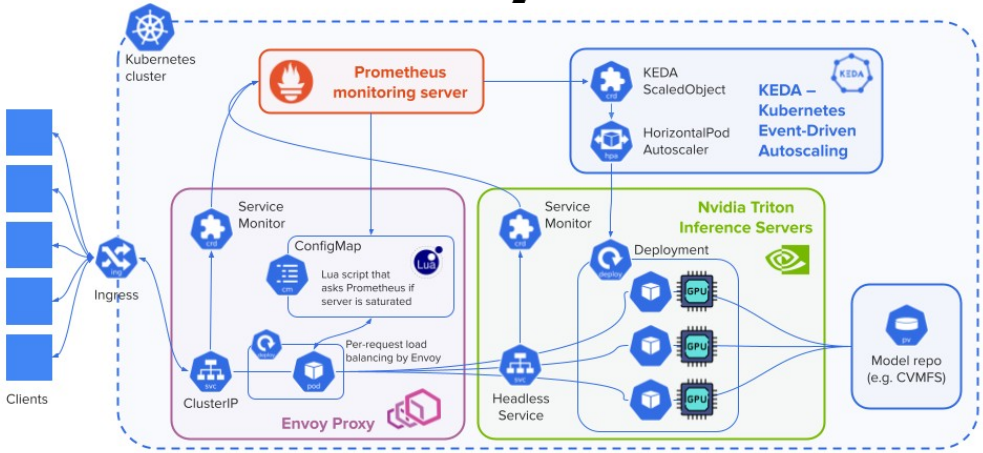
NRP NATIONAL RESEARCH PLATFORM

NRP Nautilus

SuperSONIC:
GPU Inference As-a-Service for Science

- **SONIC** is a computing paradigm for integration of **coprocessors as-a-service** in large physics experiments
- CMS, ATLAS, IceCube, LIGO are working on integrating SONIC in their workflows

- **SuperSONIC**: portable Kubernetes-based server infrastructure for GPU inference-as-a-service
  - Nvidia Triton inference servers
  - Envoy Proxy: load balancing, rate limiting
  - KEDA: load-based autoscaling
- First version of SuperSONIC Helm chart just released
- Active testing at Nautilus cluster
  - CMS: sonic-server namespace
  - ATLAS: triton-atlas namespace
  - IceCube: icecube-ml namespace



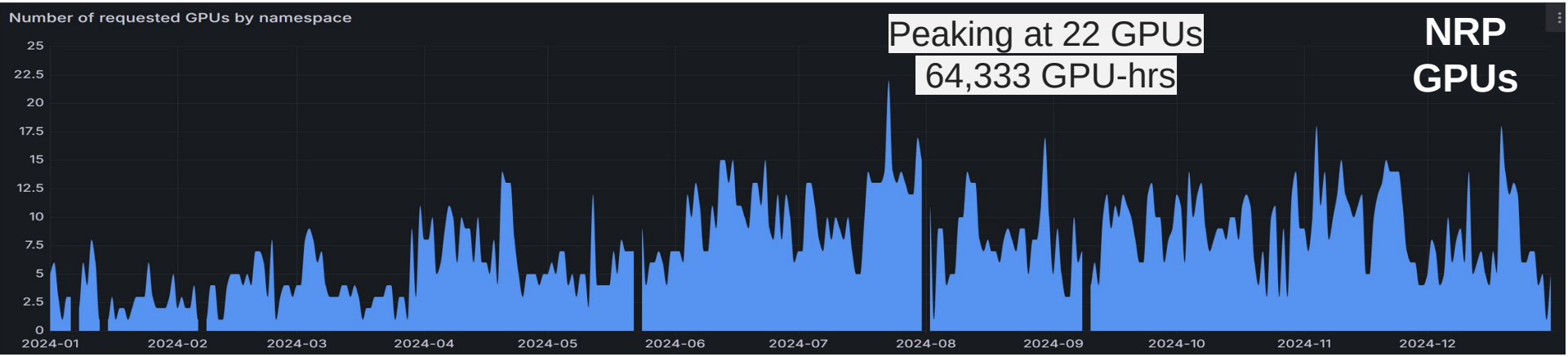
Dmitry Kondratyev

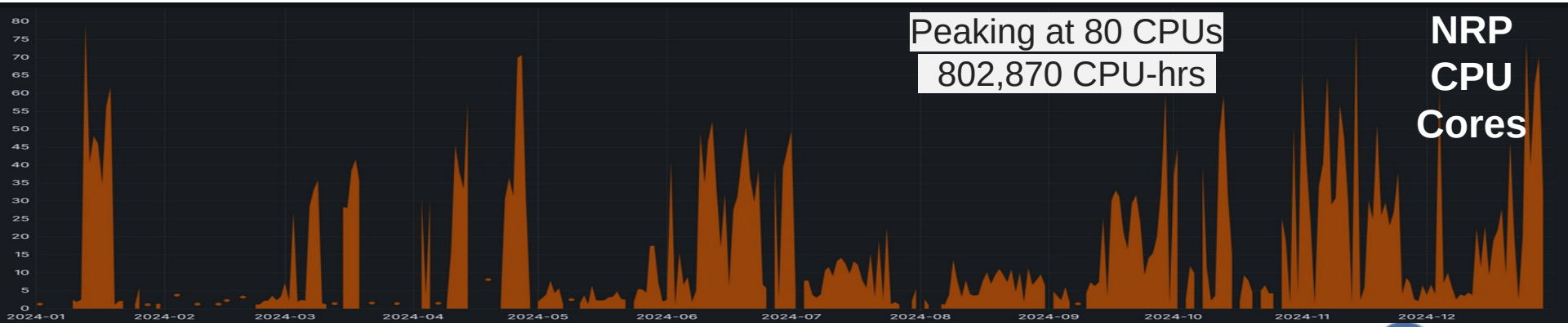
jupyterhub is the
Multi-User Version of the Jupyter Notebook



A multi-user version of the notebook designed for companies, classrooms and research labs

GPU/CPU Core Usage Per Day, Calendar 2024





NRP NATIONAL RESEARCH PLATFORM

Namespace jupyterlab

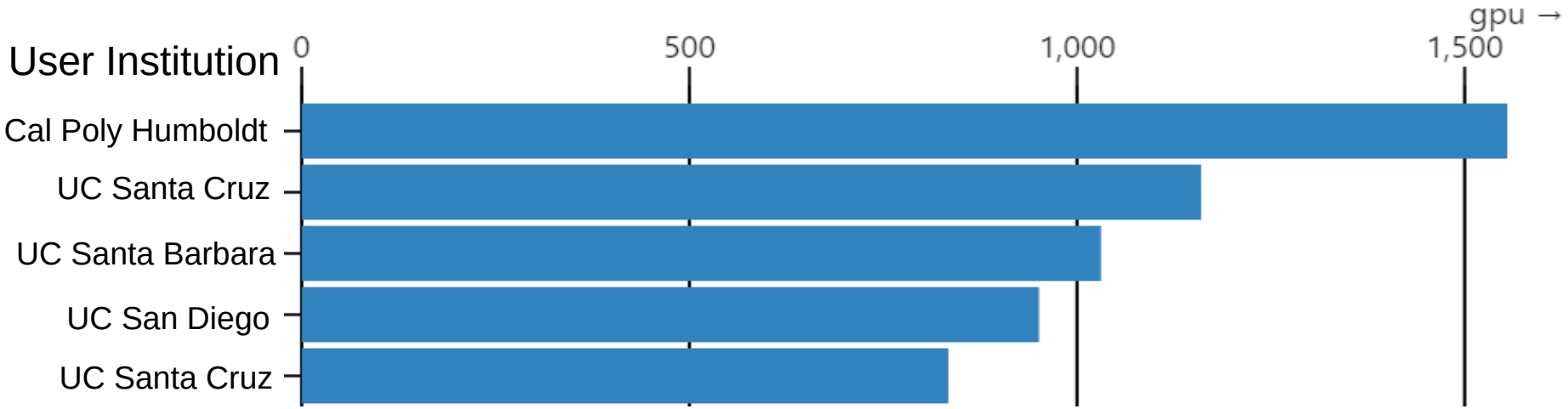
Grafana

NRP Nautilus

Nautilus Namespace *jupyterlab* Has an Active JupyterHub: *jupyterlab*'s Top 5 GPU-Hour Users Over Last 4 Months

July 2024

GPU Hours per User



Users Are Identified Only by Email Addresses

https://observablehq.com/

NRP NATIONAL RESEARCH PLATFORM

NRP Nautilus

Tom DeFanti Emailed the Largest GPU User of Namespace Jupyterlab, Asking "Who Are You and What is Your Research Project?"

We Received This Reply:

- **"Wow! Who'd have thought a botanist would be using the most GPU resources 😂"**
  - **Alana Chin**
  - **Assistant Professor of Plant Physiology**
  - **Cal Poly Humboldt**



"I am running some big Fully Convolutional Network (FCN) models to segment some of the highest resolution CT scans ever made on cores of wood collected from redwood trunks across their full height, up to 100 m above ground.

We are segmenting out tissue types for 3D measurements and brightness histograms to understand wood density and hydraulic parameters."

Source: Prof. Alana Chin
Cal Poly Humboldt

The Future of Forestry is Carbon Sequestration
The NRP is Helping Us Get There!

Forest carbon storage is tied to 2 factors –
1) biomass accumulation associated with wood growth and
2) density and the longevity and decay resistance of trees.

Dr. Alana Chin's Tree Function Lab at Cal Poly Humboldt focuses on questions related to the physiology of carbon storage and tree climatic tolerance aimed at predicting tree survival and improving management for extending and optimizing natural carbon storage.

Their recent work on Nautilus asks:
1) What are the anatomical components of wood density? and
2) Can wood anatomical structure predict future carbon storage potential ?

Prof. Alana Chin
Cal Poly Humboldt

How the Heck is a Tiny Group of Tree-Climbing Botanists Burning So Much GPU Time!?

“The Tree Function Lab works with ultra high-resolution synchrotron-based X-ray tomographic microscopy. Their final wood images are often >14 billion voxels and 50+ GB in size.

They run fully convolutional models for semantic segmentation of images into their anatomical components with up to 98% accuracy.

This allows for 3D anatomical measurements and the acquisition of cell-level wood density information based on brightness.

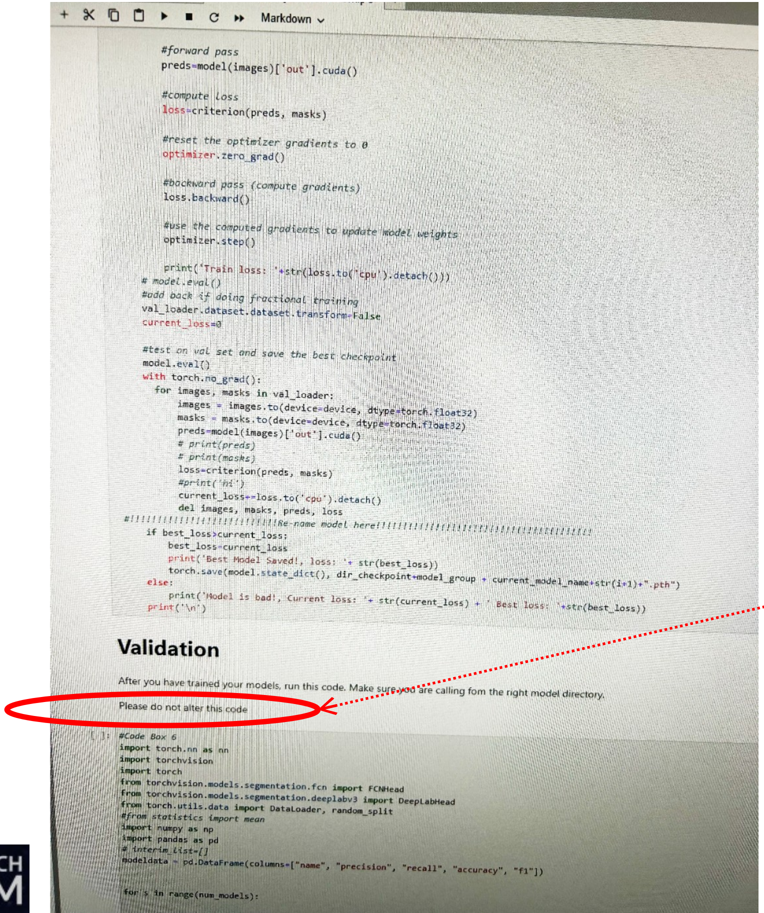
No one has ever done this before!”



NRP NATIONAL RESEARCH PLATFORM

Prof. Alana Chin
Cal Poly Humboldt

Why jupyter Notebooks?



The most honest answer is that *botanists don't bash*, their training is most definitely **not** in computer science, and Jupyter makes it **easy** to share and run Python (and sometimes R) code on a range of platforms.

It is great for handing students and less techy collaborators too!

Hooray for accessibility!

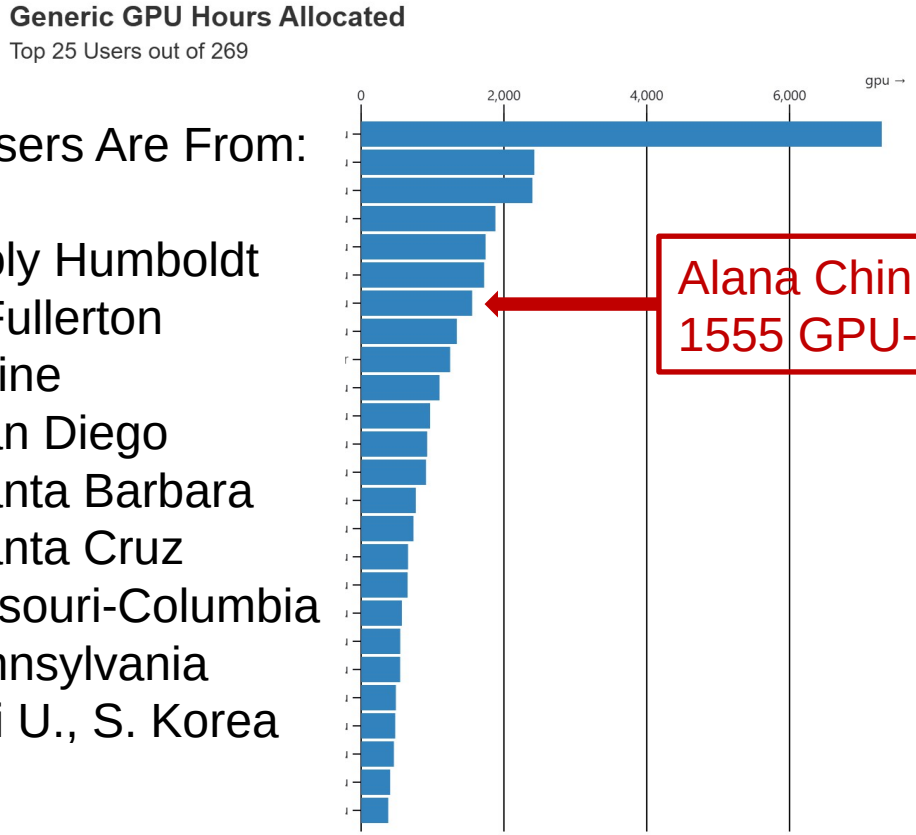
Prof. Alana Chin
Cal Poly Humboldt

NRP NATIONAL RESEARCH PLATFORM

Cal Poly Humboldt.

NRP Nautilus

Namespace jupyterlab, PI Larry Smarr
GPU-Hours Per User, Calendar 2024



Top 25 Users Are From:

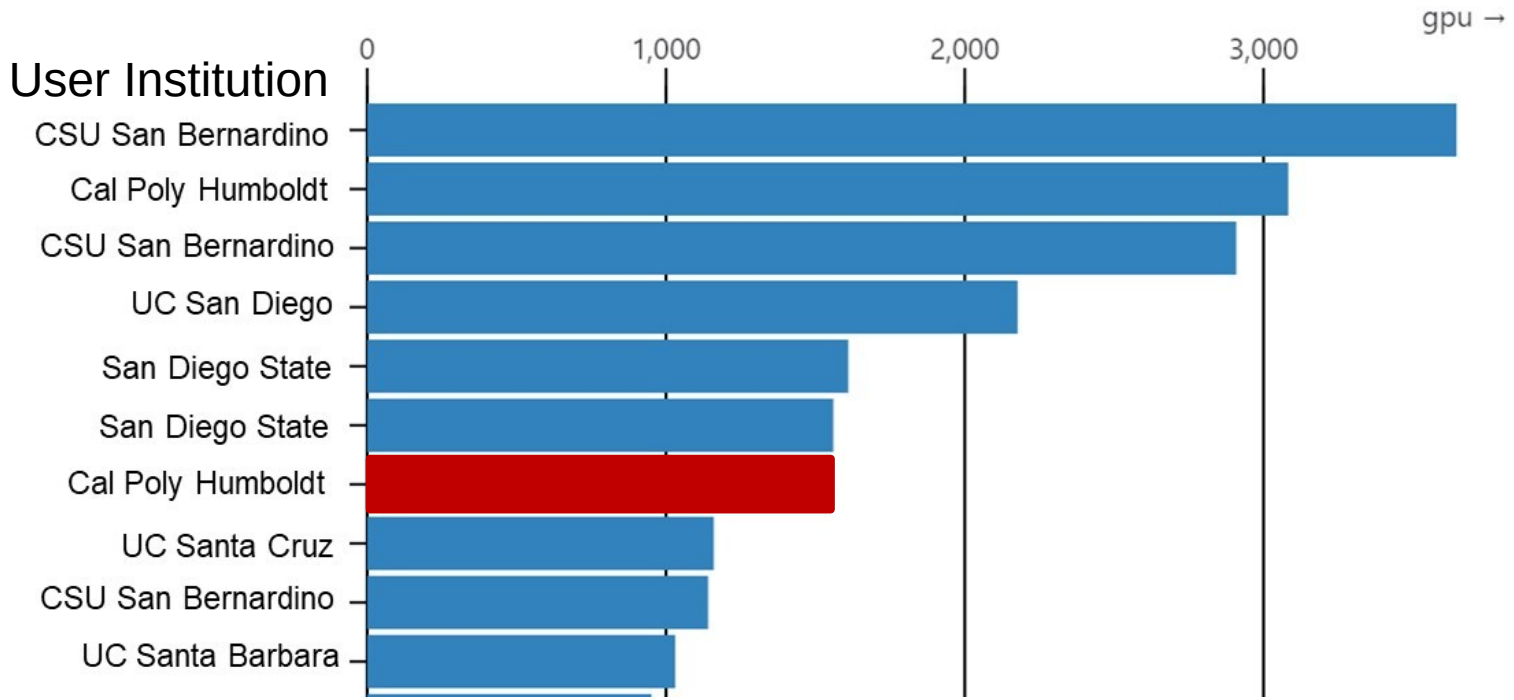
- Cal Poly Humboldt
- CSU Fullerton
- UC Irvine
- UC San Diego
- UC Santa Barbara
- UC Santa Cruz
- U. Missouri-Columbia
- U. Pennsylvania
- Yonsei U., S. Korea

Alana Chin, CPH
1555 GPU-Hours

https://observablehq.com/

There Were 92 Namespaces
That Had an Active JupyterHub in the Last 4 Months

Top 10 GPU Hours per User



Namespace ucsd-haosulab

NRP NATIONAL RESEARCH PLATFORM

NRP Nautilus