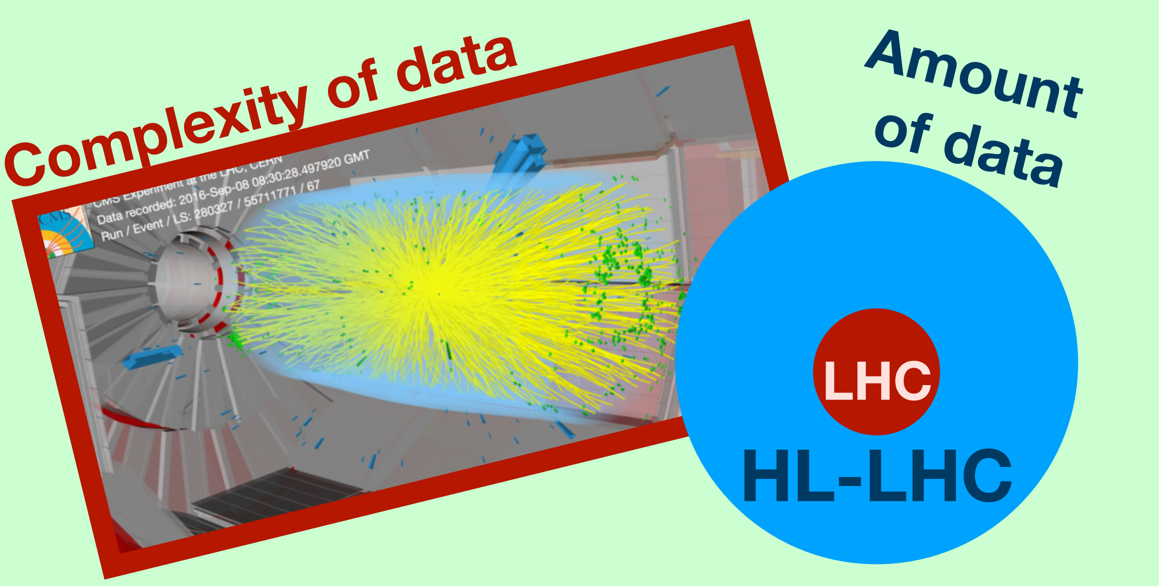
Accelerated Machine Learning as a Service for Particle Physics Computing

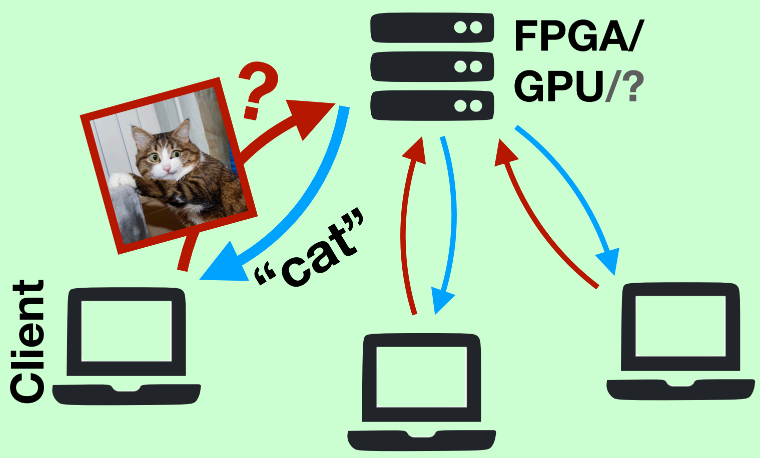
Javier Duarte[1,2], Burt Holzman[1], Sergo Jindariani[1], **Thomas Klijnsma**[1], Benjamin Kreis[1], Mia Liu[1], Kevin Pedro[1], Nhan Tran[1], Aristeidis Tsaris[1], Phil Harris[3], Dylan Rankin[3], Vladimir Loncar[4], Jennifer Ngadiuba[4], Maurizio Pierini[4], Suffian Khan[5], Brian Lee[5], Brandon Perez[5], Ted W. Way[5], Colin Versteeg[5], Scott Hauck[6], Shih-Chieh Hsu[6], Matthew Trahms[6], Dustin Werran[6], Zhenbin Wu[7]

1: Fermi National Accelerator Laboratory, 2: University of California San Diego, 3: Massachusetts Institute of Technology, 4: CERN, 5: Microsoft, 6: University of Washington, 7: University of Illinois Chicago

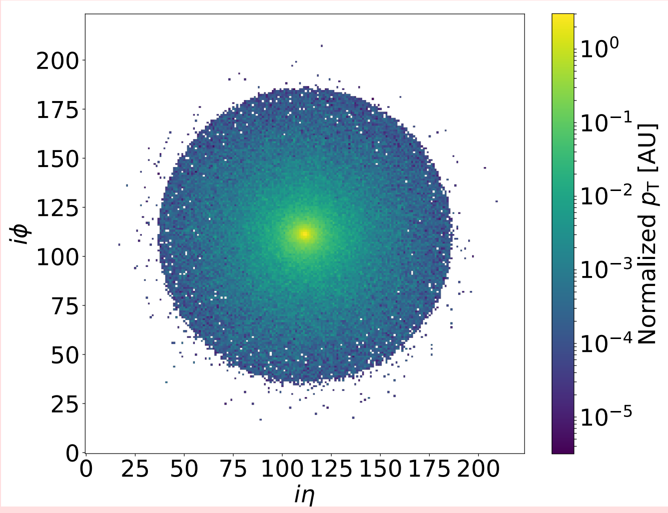
FERMILAB-POSTER-19-143-E



Inference-as-a-service

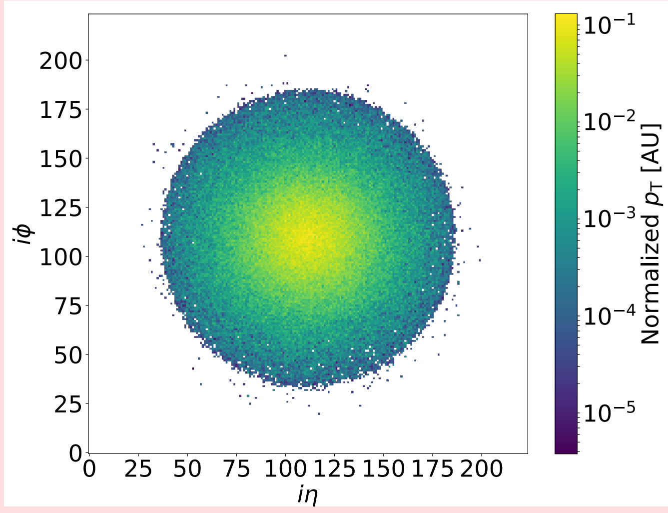


- Amount and complexity of high-energy physics data increases dramatically from 2025 onward
- Traditional algorithms will require too much CPU time
- Machine learning can solve **combinatorially-scaling** problems in **constant time**, but must be fast enough



QCD

(averages of 5000 images)



Top quark

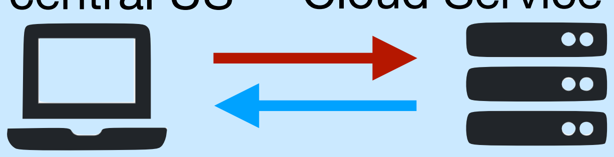
Example challenge

Separate "*top quarks*" (interesting!) from "*QCD jets*" (uninteresting)

- Inputs 224x224 'single-color' images in a ResNet50 architecture
- Pixels are energy collections in the CMS electromagnetic calorimeter crystals

Inference-as-a-service

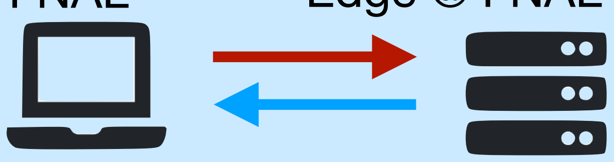
VM in central US — Microsoft Azure Cloud Service



60-80 ms — **660 inf/s**

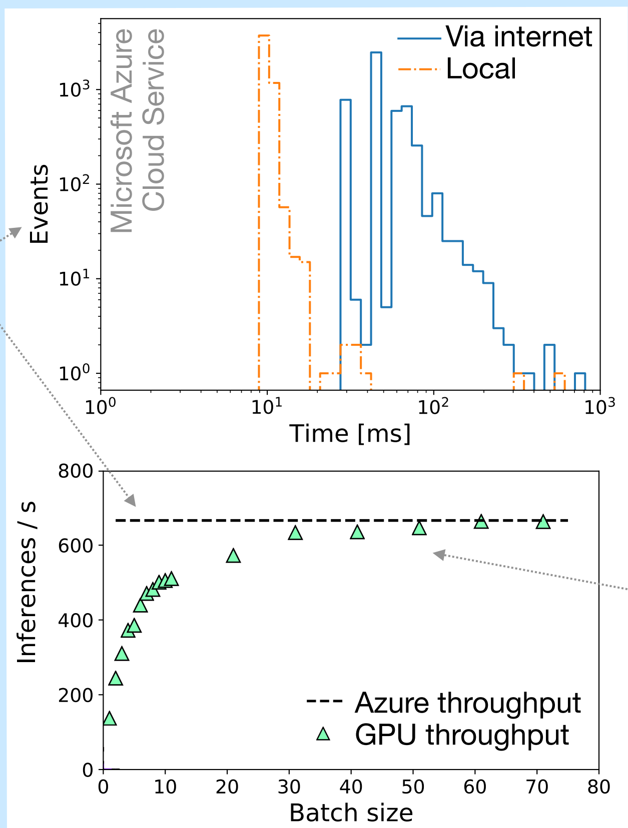
Optimal case: 10 ms

CPU @ FNAL — Azure Data Box Edge @ FNAL



20 ms — **>200 inf/s**

Currently limited by network equipment

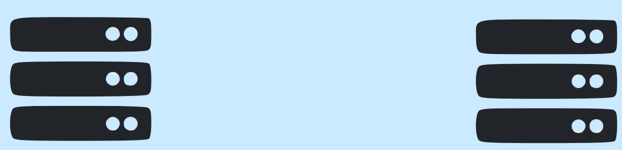


Local

1 CPU — Azure Data Box Edge @ Fermilab

~500 ms, 2 inf/s — **14 ms, 70 inf/s**

GPU batch size 32 — GPU batch size 1



1.5 ms, 667 inf/s — **7 ms, 143 inf/s**

Results

An **FPGA-aaS** reaches the same throughput as a **locally connected GPU**, the former by having many CPUs access it and the latter by setting a high batch size

- What NN architectures are suitable for our physics problems ***and*** IaaS?
- How scalable are these solutions to HL-LHC data volumes?

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