Artificial Intelligence Accelerated Discoveries
At the Large Hadron Collider

Mia Liu
Purdue University

Oct 13, 2021
University of Birmingham
The Standard Model of Fundamental Particles

HISTORY OF THE UNIVERSE

Particle era

Accelerators

RHIC & LHC heavy ions

High-energy cosmic rays

Inflation

Big Bang

ASSIME DARK MATTER RELICS

Nucleosynthesis

Nuclei

Cosmic Microwave Background radiation is visible

Structure formation

Dark energy accelerated expansion

Size of visible universe

Today

Time (seconds, years)

Energy of photons (units GeV, 1.6 x 10^-10 joules)

Key:

quark
neutrino
electron
ion
star
The Standard Model of Fundamental Particles
Last Missing Piece
2012: Higgs Discovered at the LHC!
Remaining puzzles

Fine tuning?
Dynamical origin?

Experimental:
Dark matter/dark energy
Not in SM!

Neutrinos in SM, masses?…

Anomalies: Muon g-2, LHCb lepton flavour universality?
Extensive searches performed at the LHC
Can we do better at the LHC?

Unconventional signatures: Long lived
Loop contributions: e.g. Lepton flavor violating
The Fast and Furious

Extreme data volume & rate from LHC collisions.

Multiple pp collisions in the same beam crossing:
LHC: 20-50. HL-LHC: 140-200
How to be prepared for LHC Run-3/4?

Now: The LHC

Higher pileup, fine granularity detectors.
Advanced algorithms to maintain/improve the acceptance of (un)conventional signatures

LHC Run-4
Measure what is measurable and make measurable what is not so with Artificial Intelligence?
From Collisions to Discoveries

- LHC L1 Trigger (pipelined)
- LHC High Level Trigger
- LHC/DUNE Offline processing

1 µs - 1 ms - 1 s

- ~7 PB/s
- 40 MHz
- 100 kHz
- ~7 PB/day
- 1 kHz
- 1 MB/evt
- Offline
Real-time ML... @ Level-1

FPGAs/ASICs - high bandwidth low latency specialized compute hardware
NN on FPGAs

\[ x_n = g_n(W_{n,n-1}x_{n-1} + b_n) \]

Activation functions
Precomputed, and stored in BRAMs
Multiplications
Digital Signal Processing
DSPs

Addition
Logic cells

FPGA diagram

DSP slice
RAM
Natural fit for FPGAs… limited resources

\[ x_n = g_n(W_{n,n-1}x_{n-1} + b_n) \]

Activation functions
Precomputed, and stored in BRAMs
Multiplications
DPSs
Addition
Logic cells

\[ N_{\text{multiplications}} = \sum_{n=2}^{N} L_{n-1} \times L_n \]

Small network: thousands of connections

Limitation: Number of DSPs

Virtex Ultrascale+ VU9P
6800 DSPs
1M LUTs
2M FFs
75 Mb BRAM
Fit NN on FPGA: Quantization & Reuse

3-layer pruned, Kintex Ultrascale

- Reuse Factor = 1
- Reuse Factor = 2
- Reuse Factor = 3
- Reuse Factor = 4
- Reuse Factor = 5
- Reuse Factor = 6

Max DSP

6 parallel DSPs: 75 ns

1 DSP serial: 175 ns

Severe performance drop below 12 bits

FPGA AUC / Expected AUC

- g tagger
- q tagger
- w tagger
- z tagger
- t tagger

Fixed-point precision

integer fractional width
Neural Network compression is a widespread technique to reduce the size, energy consumption, and overtraining of deep neural networks.

Several approaches in literature:

arxiv.1510.00149, arxiv.1712.01312, arxiv.1405.3866, arxiv.1602.07576,
doi:10.1145/1150402.1150464
High Level Synthesis 4 Machine Learning

PYTORCH

Keras
TensorFlow
PyTorch
...

model

compressed model

HLS conversion

HLS project

Vivado™ HLS

Co-processing kernel

Custom firmware design

https://fastmachinelearning.org/hls4ml/

Usual ML software workflow

+ TensorFlow

ONNX
Recent developments

First paper demonstrated a fully connected NN in 100 ns.

HLS4ML in CMS
Run 3: muon momentum regression in CMS
More models demonstrated for Phase-2 trigger upgrade TDR

Advanced models:
binary/ternary, CNNs, RNNs, auto-encoders. Support for Graph Neural Network Models

Advanced Pruning/quantization:
Quantization-aware training with QKeras/Quantization-aware pruning

On ASICs and Low power devices.

For latest status: please check hls4ml website, CPAD 2021 talk,
Try it out: hls4ml tutorials
Including Graph Neural Networks

Application Algorithms drive HLS4ML developments.

Graph Neural Networks (GNNs):
Represent data as nodes and edges
appropriate representation of particle physics data: irregular, structural, relational

GNN developments driven by social media.

Active development and applications in science domains: how to adapt to domain knowledge & applications

Many variations of GNNs. Problem formulation is important: Node/edge classification, graph classification, identifying subgraphs
Will show two GNN studies at Purdue.
τ→3μ : motivation and current trigger

Very rare in the SM, Neutrino oscillations:
BR ~O(10^{-14})

R-parity violating SUSY

CMS Phase-2 Simulation

- Collimated, low pT and very forward muons
- Current CMS trigger for HL-LHC uses track trigger tracks, 25% efficiency.
GNN \( \tau \rightarrow 3\mu \) classifier

**Graph Inputs:** Muons hits (Coordinates and bending angle) from L1 primitives, represented as nodes in graph.

**Attention mechanism** for information aggregation between local nodes and global node.

**Training setting:**

\( \tau \rightarrow 3\mu \) signals mixed with PU backgrounds vs pure PU sample

> 90% efficiency for 30 kHZ trigger bandwidth

Fully connected network: \(~26\%

**On-going:** Model interpretations, Data augmentation. Regression of muon kinematics, Model adaption (to other signals & other experiments), implementation with HLS4ML.
The Fast and Furious

Extreme data volume & rate from LHC collisions.

Multiple pp collisions in the same beam crossing:
LHC: 20-50. HL-LHC: 140-200
Semi-supervised Pileup mitigation with GNN

Improve Per-Particle Pileup Mitigations With better

- Trained on charged particles and applied to neutral ones —> can learn from data.
- Outperforms Puppi, comparable to fully supervised method.
- Next: Apply to CMS simulation & data. Neutral particle vertex association in for the forward region.
Speeding up HLT & Offline

LHC L1 Trigger (pipelined)

LHC High Level Trigger

LHC Offline processing

Traditionally CPU; Moving towards heterogeneous computing
HL-LHC: Big Data Challenge

Event complexity

Data Volume

>5x

>10x

Current: ~5 minutes per HL-LHC event

Moore’s Law continues
…but Dennard Scaling fails.
No faster computers for free!
#Trending in Industry: Heterogeneous Computing

Advances driven by big data explosion & machine learning
Machine Learning Inference as-a-service

Services for Optimized Network Inference on Co-processors (SONIC paper)

Speed of light → 10 ms

External processing
CMSSW module

callback
acquire

FPGA, GPU,
Latency: 10 ms (60 ms) for local (remote cloud) server, (10/100) faster than CPU-only

Max data throughput: 600-700 images/sec
SONIC: recent explorations

GPU-as-a-service

FPGA-as-a-service Toolkit
Open source tools: flexibility

Hardware platforms

Algorithm complexity

More benchmarks driven by use cases to test scaling for HLT/offline: 2k-10M parameters

GPU-as-a-service for DUNE

NVIDIA Triton
SONIC in CMS miniAOD

- SONIC miniAOD workflow has been developed, machine learning algorithms offloaded to GPU servers.
- Testing at Purdue T2 with local GPUs/GPUs in google cloud.
- Infrastructure development in plan such as server management, authentication etc
- A HLT workflow has also been developed for non-ML algorithms (patatrk, tracking on GPUs)

<table>
<thead>
<tr>
<th>Latency (ms)</th>
<th>Fraction (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td>718.3</td>
</tr>
<tr>
<td>ParticleNet</td>
<td>90.3</td>
</tr>
<tr>
<td>DeepTau</td>
<td>24.6</td>
</tr>
<tr>
<td>DeepMET</td>
<td>12.8</td>
</tr>
</tbody>
</table>

Please see slides and interactive pie chart
Big Data Era

LHC Science data
~200 PB

Facebook uploads
180 PB

SKA Phase 1 –
2023
~300 PB/year
science data

LSST 2021

Google searches
98 PB

Yearly data volumes

DUNE 2026

SKA Phase 2 – mid-2020’s
~1 EB science data

HL-LHC – 2026
~600 PB Raw data

HL-LHC – 2026
~1 EB Physics data

LHC – 2016
50 PB raw data

Google Internet archive
~15 EB
Community built upon hls4ml & sonic effort: monthly general meetings, alternating hls4ml & co-processor meetings.

Held workshops at Fermilab/SMU (virtual)

Fruitful discussions on common challenges across science domains & interesting intersections with industry and other fields

HEP, neutrino, astrophysics, plasma physics (fusion control), material science, Xilinx, Nvidia, Neuromorphic compute.

White papers: 2019, 2020 submitted to frontier in big data.

Next (mini) Fast ML workshop in Spring 2022
Harnessing the data revolution Institute grant awarded by National Science Foundation (NSF) Accelerated Artificial Intelligence For Data-Driven Discoveries.

The challenge for domain scientists is that a broad range of expertise is required to arrive at full ML device implementations.

15 M grant, 9 Institutions. HEP, astrophysics, neural science, AI algorithm, Hardware acceleration. Interface with frontier algorithm & engineering research (machine learning compiler)
Summary

Machine learning methods offer opportunities to significantly boost the discovery potential at the LHC (e.g. \( \tau \rightarrow 3 \mu \)).

Accelerated machine learning inference in online & offline processing.

User-friendly prototype tools for domain experts.

Multidisciplinary teams to realize optimal ML on targeted (e.g. CMS L1 trigger) or heterogeneous systems (e.g. CMS HLT & offline).

Look forward to the visions unfold in the next few years!
Adaptive algorithms and tools

Dark sector searches at SeaQuest @ Fermilab
CNNs, Graphs, RNNs, auto-encoders, binary/ternary
e.g. Lepton flavor violation: $\tau \rightarrow 3\mu$

Measuring muon EDM with frozen spin techniques.
How many CPU can this GPU serve?

CPU-to-GPU ratio:

t\_total\_cpu - t\_othercpu = t\_ml

t\_total\_sonic - t\_othercpu - t\_sonic\_cpu = t\_transfer + t\_scheduling + t\_sonic\_gpu

t\_ml/ (t\_sonic\_cpu + t\_transfer + t\_scheduling + t\_sonic\_gpu) =

t\_sonic\_cpu\_part = t\_sonic - t\_transfer - t\_scheduling

ratio = t\_sonic\_cpu\_part/t\_gpu

Passes this ratio, GPU saturates and average processing time increase.
Semi-supervised Graph Puppi

- **Charged LV particles**
- **Charged PU particles**
- **Neutral particles**

**Common feature domain**

**Charged-specific feature domain**

**Neutral-specific feature domain**

$p_T$, charge, PUPPI weight

$\Delta \eta, \Delta \phi, \Delta R$
Semi-supervised Graph Puppi

(b). Randomly select charged LV/PU particles, and mask charged-specific feature domain for training

- Charged LV particles
- Charged PU particles
- Neutral particles

- Common feature domain
- Charged-specific feature domain
- Neutral-specific feature domain
(b). Randomly select charged LV/PU particles, and mask charged-specific feature domain for training

- Algorithm outperforms Puppi, comparable to fully supervised method. Can be adapted to different pile up conditions. No need for tuning as the particle itself is represented as a node.


- Next: Apply to CMS simulation & data. Neutral particle vertex association in for the forward region.
Summary and outlooks

LHC data let us probe the new physics scale at the LHC

SUSY searches and examine SM’s description of triboson processes

Accelerated discovery potential with ML

Fast machine learning inference for CMS data processing

Crucial in maximizing the HL-LHC physics potential

Look forward to continue with this exciting journey at UCR!
HLS - High Level Synthesis - compiler for C, C++, SystemC into FPGA IP cores

HLS 4 Machine learning: Prototype ML algorithms for FPGA WITHOUT Verilog/VHDL: firmware in a few hours
• SUSY partners of the SM electroweak sector:
  • U(1) -> Bino, SU(2)->Winos
  • Higgs -> Higgsinos
  • Leptons -> sleptons
• Could be light and accessible at the LHC

Important to search for them at the LHC!
Typical mass spectrums of chargino-neutralinos

- Depending on the mass scales of Bino/Winos/Higgsinos:
  lightest chargino/neutralinos form different mass spectrums
- Two main mass spectrums explored at the LHC: **Wino-like, Higgsino-like**
- My focus: search for Wino-like production with WH events:
  - Larger cross section.
  - Loosely constrained in 8 TeV searches compared to WZ

\[ \tilde{\chi}_1^+ \rightarrow W \tilde{\chi}_1 \]
\[ \tilde{\chi}_2^0 \rightarrow \tilde{\nu} \tilde{\nu} \]

\[ \Delta m \sim \text{hundreds of MeV to tens of GeV} \]

Wino-like: 45 fb*  
Higgsino-like: 11 fb*

*Cross-sections for 500 GeV sparticles at 13 TeV (\( \tilde{\chi}_2^0, \tilde{\chi}_1^\pm \) only)
Search for electroweakinos with WH events

- 1 lepton (e/μ)+bb + large missing transverse energy:
  - leptons: trigger and handle against backgrounds
- Higgs→bb (60%). Mass peak in the SM kinematic tails. ttbar→2L background directly controlled in the mass sideband.
- First result to probe chargino mass to 500 GeV in WH decay
WH(lvbb) + MET: pushing Wino limits

- Probes chargino mass up to 500 GeV in the WH topology
- 300 GeV improvement wrt 8 TeV reach
- Dominates the sensitivity in the bulk.
**Table 1:** Event selection criteria for the SS category, which contains events with two same-sign leptons and at least two hadronic jets

<table>
<thead>
<tr>
<th>Variable</th>
<th>$\text{e}^{\pm}\text{e}^{\pm}$</th>
<th>$\text{e}^{\pm}\mu^{\pm}$</th>
<th>$\mu^{\pm}\mu^{\pm}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Signal leptons</td>
<td>exactly 2 tight equally-charged leptons with $p_T &gt; 25$ GeV</td>
<td>no additional rejection lepton</td>
<td></td>
</tr>
<tr>
<td>Additional leptons</td>
<td>no (additional) isolated tracks</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Isolated tracks</td>
<td>$\geq 2$ jets with $p_T &gt; 30$ GeV, $</td>
<td>\eta</td>
<td>&lt; 2.5$</td>
</tr>
<tr>
<td>Jets</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>b-tagged jets</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dijet mass (closest $\Delta R$)</td>
<td>$65 &lt; M_{jj} &lt; 95$ GeV ($M_{jj}$-in) OR $</td>
<td>M_{jj} - 80$ GeV$</td>
<td>\geq 15$ GeV ($M_{jj}$-out)</td>
</tr>
<tr>
<td>Dijet mass (leading jets)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\Delta \eta$ of two leading jets</td>
<td>$&lt; 400$ GeV</td>
<td>$&lt; 1.5$</td>
<td></td>
</tr>
<tr>
<td>$p_T^{\text{miss}}$</td>
<td>$&gt; 60$ GeV</td>
<td>$&gt; 60$ GeV if $M_{jj}$-out</td>
<td></td>
</tr>
<tr>
<td>$M_{\ell\ell}$</td>
<td>$&gt; 40$ GeV</td>
<td>$&gt; 30$ GeV</td>
<td>$&gt; 40$ GeV</td>
</tr>
<tr>
<td>$M_{\ell\ell}^{\text{max}}$</td>
<td>$</td>
<td>M_{\ell\ell} - M_Z</td>
<td>&gt; 10$ GeV</td>
</tr>
<tr>
<td>$M_T$</td>
<td>—</td>
<td>$&gt; 90$ GeV</td>
<td>—</td>
</tr>
</tbody>
</table>
Three leptons

Table 2: Event selection criteria for the 3$\ell$ category, which contains events with exactly three leptons

<table>
<thead>
<tr>
<th>Variable</th>
<th>0 SFOS</th>
<th>1 SFOS</th>
<th>2 SFOS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Signal leptons</td>
<td>exactly 3 tight charged leptons with $p_T &gt; 25/20/20$ GeV and charge sum = $\pm 1e$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Additional leptons</td>
<td></td>
<td>no additional rejection lepton</td>
<td></td>
</tr>
<tr>
<td>Jets</td>
<td>$\leq 1$ jets with $p_T &gt; 30$ GeV, $</td>
<td>\eta</td>
<td>&lt; 5$</td>
</tr>
<tr>
<td>b-tagged jets</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$p_T(\ell\ell\ell)$</td>
<td></td>
<td></td>
<td>$&gt; 60$ GeV</td>
</tr>
<tr>
<td>$\Delta\phi(\vec{p}_T(\ell\ell\ell), \vec{p}_T^{miss})$</td>
<td></td>
<td></td>
<td>$&gt; 2.5$</td>
</tr>
<tr>
<td>$p_T^{miss}$</td>
<td>$&gt; 30$ GeV</td>
<td>$&gt; 45$ GeV</td>
<td>$&gt; 55$ GeV</td>
</tr>
<tr>
<td>$M_T^{max}$</td>
<td>$&gt; 90$ GeV</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$M_T^{3rd}$</td>
<td></td>
<td>$&gt; 90$ GeV</td>
<td></td>
</tr>
<tr>
<td>SF lepton mass</td>
<td>$&gt; 20$ GeV</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Di-electron mass</td>
<td>$</td>
<td>M_{ee} - M_Z</td>
<td>&gt; 15$ GeV</td>
</tr>
<tr>
<td>$M_{SFOS}$</td>
<td></td>
<td>$</td>
<td>M_{SFOS} - M_Z</td>
</tr>
<tr>
<td>$M_{\ell\ell\ell}$</td>
<td></td>
<td>$</td>
<td>M_{\ell\ell\ell} - M_Z</td>
</tr>
</tbody>
</table>
Populations of objects show dark matter, dark energy

Region-based CNNs on heterogeneous compute devices
- LSST: 20 Tb / night
- 1 Billion transient alerts / night

Long: competition between faint galaxies, transient objects
Short: Weather, annual modulation of sky positions
Smart telescopes: reinforcement learning for optimal scheduling and control
Improved isolation definition

- **Smaller cone-size:** 0.4 → 0.3
- **Add lepton candidates to Isolation:** improves rejection against heavy flavor decay (B→D→2 leptons + X), one of the leptons is selected as good lepton.

3.5 X background rejection for muons @ 70% efficiency
Towards Run-2 Result

- Improve cut-based analysis:
  - e.g. OSFOS (e⁺/e⁻/μ⁺/μ⁻, m⁺/m⁻/e⁺/e⁻)
  - Large fraction of fake leptons in 2016 analysis:
    - Dilepton ttbar
  - Low event yield:
    - susceptible to statistical fluctuations.

- New selection features:
  - Customized IDs for e⁺/e⁻/μ⁺/μ⁻, m⁺/m⁻/e⁺/e⁻
  - Soft b jet veto: 30% fake rejection, no signal loss
  - Lifted kinematic selections
  - Overall improvement >50%.

- In parallel, exploring MVA based analysis.
Covering all VVV processes

<table>
<thead>
<tr>
<th>Quantities</th>
<th>WWZ</th>
<th>WZZ</th>
<th>ZZZ</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\sigma_{\text{total}} \times B_{VVV \rightarrow 4\ell}$ (fb)</td>
<td>4.12</td>
<td>0.74</td>
<td>1.19</td>
</tr>
<tr>
<td>$\sigma_{\text{total}} \times B_{VVV \rightarrow 5\ell}$ (fb)</td>
<td>-</td>
<td>0.36</td>
<td>-</td>
</tr>
<tr>
<td>$\sigma_{\text{total}} \times B_{VVV \rightarrow 6\ell}$ (fb)</td>
<td>-</td>
<td>-</td>
<td>0.05</td>
</tr>
<tr>
<td>$\sigma_{\text{total}} \times B_{VVV \rightarrow 4\ell} \times 137 \text{ fb}^{-1}$ (N evt.)</td>
<td>564</td>
<td>101</td>
<td>163</td>
</tr>
<tr>
<td>$\sigma_{\text{total}} \times B_{VVV \rightarrow 5\ell} \times 137 \text{ fb}^{-1}$ (N evt.)</td>
<td>-</td>
<td>49.3</td>
<td>-</td>
</tr>
<tr>
<td>$\sigma_{\text{total}} \times B_{VVV \rightarrow 6\ell} \times 137 \text{ fb}^{-1}$ (N evt.)</td>
<td>-</td>
<td>-</td>
<td>6.85</td>
</tr>
</tbody>
</table>

- Handful of other VVV events produced with full Run-2 data (Leptonic channels).
- WWZ(4l) has the best expected sensitivity: 3.5 $\sigma$
  - Events categorized with W leptons (ee/mm vs em).

Expect the first evidences of WWW and WWZ production with full Run-2 dataset.
Some topics need HL-LHC dataset

- 3000 fb-1 data expected at the HL-LHC
- e.g. Higgsinos: Low cross section, challenging signatures
  - $\Delta m \sim$ tens of GeV: Soft decay products
  - $\Delta m \sim$ hundreds of MeV: Long-lived signatures

$\tilde{\chi}_1^0 \rightarrow W^+ \tilde{\chi}_2^0 \rightarrow W^+ Z/h \rightarrow W^+ W^+ W^+ W^+$

$\Delta m \sim$ hundreds of MeV to tens of GeV

Wino-like: 45 fb*

Higgsino-like: 11 fb*

*Cross-sections for 500 GeV sparticles @ 13 TeV ( $\tilde{\chi}_2^0 \tilde{\chi}_1^{$\pm$} $ only)
Microsoft Brainwave

- **Mature service** at scale (more than just a single co-processor)
- Multi-FPGA/CPU fabric accelerates *both computing* and *network*
- Models supported:
  - ResNet50, ResNet152, DenseNet121, VGGNet16…
  - Partially fixed neural network architecture. weights can be retuned.
**WWZ: smaller rate but clean**

- **Z → ττ** peak
- **4ℓ**: Tag Z→ℓℓ, WW→ (eμ/ee/μμ)
  - eμ shown as example: kinematic selections against ZZ

40% smaller than WWW, leptonic decays

\[ m_{T2} = \min_{\tilde{p}_T^{(1)}, \tilde{p}_T^{(2)}, \tilde{p}_T^{(3)}, \tilde{p}_T^{(4)}} \left[ \max \left( m_T^{(1)}(\tilde{p}_T^{(1)}, \tilde{p}_T^{e}), m_T^{(2)}(\tilde{p}_T^{(2)}, \tilde{p}_T^\mu) \right) \right] \]
Edge data box at Feynmann computing center

- Gain experience in deploying co-processors in local clusters with cloud native tools: docker image, kubernetes
- Benchmark latency and scaling performance, compare with previous studies
- Can be used for neutrino and cosmology experiments as of ~today—> next slide
Brainwave cloud service

Same setup as last page
Single Service
Parallel CPU jobs:
5000 images/job

Comparable max data throughput: 600-700 images/sec
Neutrinos oscillate:
Lepton number not conserved

At the LHC
What about in charged leptons? $\tau \to 3\mu$
Improved lepton isolation definition

- Smaller cone improves rejection of $B \rightarrow J/\psi \gamma$ decay
- Add lepton isolation improves signal to background ratio

3.5 X background rejection for muons

$|\cos \theta_j|$
Need to cope with more challenging LHC environment in Run 2 & Run 3 (300 fb⁻¹) until HL-LHC upgrade (2023).

Module designed to reduce dynamic inefficiency
Digital readout chip (ROC). Faster readout.

Geometry design: ensure tracking and vertex quality
Added layers, channels doubled

Services: reduce material budget
CO2 cooling, DCDC powering, Service electronics out of tracker volume.
**Half cylinders and service electronics**

- **Portcard:**
  - Distributes power and bias voltages, clock, trigger and calibration signals to modules. Programs Modules (TBM and ROCs)
- **Electric/optical Converters mounted**
  - Digital opto-hybrid (DOH): Optical→Electrical
  - Pixel opto-hybrid (POH): Electrical→Optical
- **CCU:** Communication & Control Unit
- **uTCA crate hosting front-end controller/drivers.**
FPIX assembly at Fermilab

All four half-cylinders tested with full DAQ readout chain at Fermilab
challenge: build the detectors!

- CMS Phase 0 detector designed for LHC nominal luminosity
  - Tracking efficiency drops to 80% at PU=40
- Phase 1 pixel: designed for LHC Run 2 & Run 3 data-taking (300 fb⁻¹) until HL-LHC upgrade (2024).
  - Improved module design, geometry, material budget.
Higgs boson detection in CMS

Higgs boson decays

- ZZ 3%
- cc 3%
- γγ 0.2%
- ττ 6%
- gg 9%
- WW 22%
- bb 58%

Construct invariant mass from decay products

\[ E_{cm} = \left[ (E_1 + E_2)^2 - (p_1 + p_2)^2 \right]^{1/2} \]
Signal vs background

Higgs boson discovery: decay modes of lower backgrounds ($WW/ZZ/\gamma\gamma$).
Re-train Res-Net 50 to tag top jets

Quantized model Brainwave’s implementation of ResNet50 on FPGA

State-of-art performance achieved with quantized ResNet50 on BrainWave service

Floating point: acc = 90.1%, AUC = 98.0%, \( \frac{1}{\varepsilon_B} = 671 \)
Quant.: acc = 84.1%, AUC = 97.5%, \( \frac{1}{\varepsilon_B} = 415 \)
Quant, f.t.: acc = 98.2%, AUC = 93.0%, \( \frac{1}{\varepsilon_B} = 971 \)
Brainwave: acc = 92.6%, AUC = 98.22%, \( \frac{1}{\varepsilon_B} = 935 \)
Brainwave, f.t.: acc = 93.5%, AUC = 98.3%, \( \frac{1}{\varepsilon_B} = 1000 \)

30% signal eff.

Better

Emulation

Background efficiency

Signal efficiency