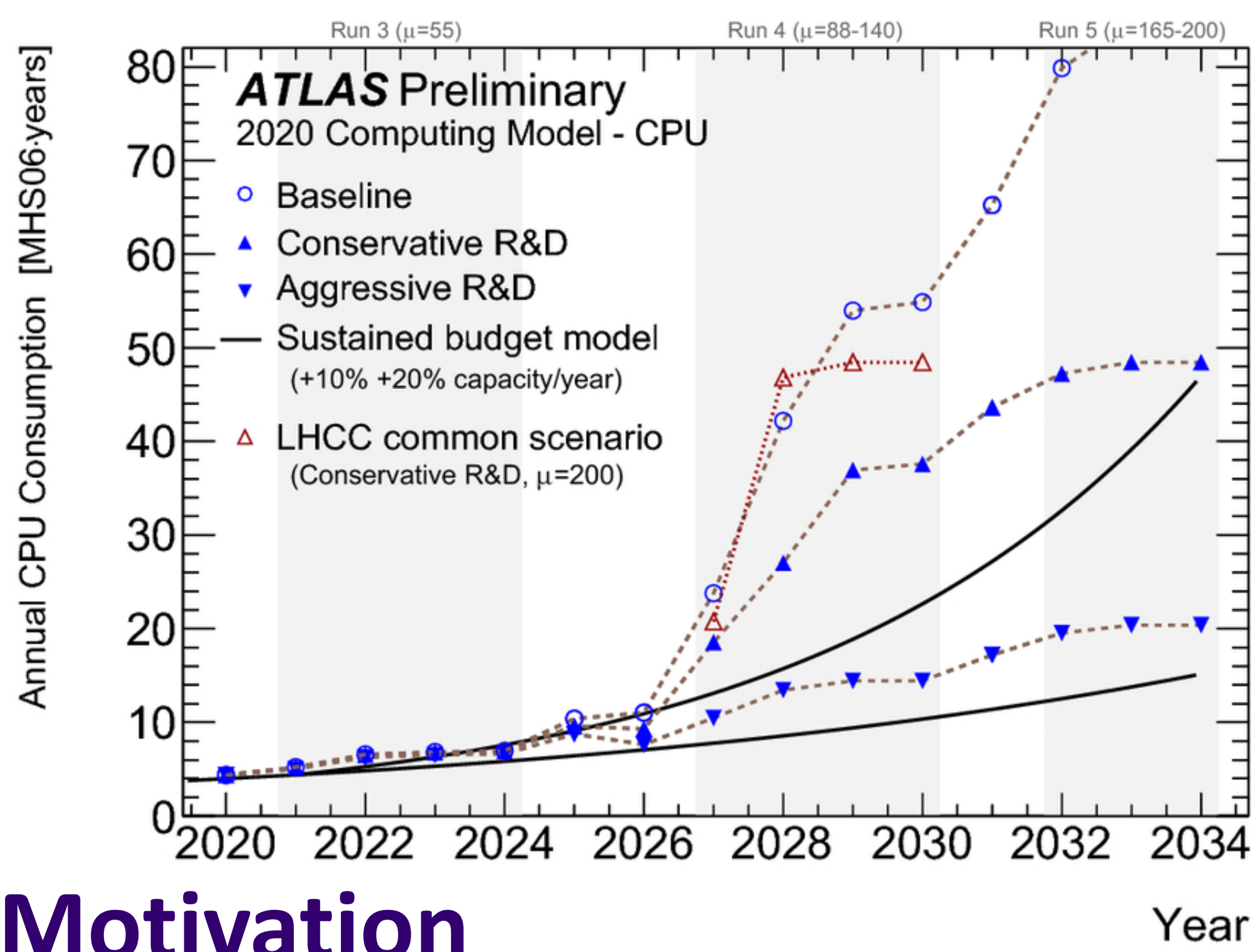# Accelerating CNNs on FPGAs for Particle Energy Reconstruction

Chijui Chen, Yanlun Huang, Lingchi Yang, Ziang Yin, Bo-Cheng Lai, Phil Harris, Scott Hauck, Shih-Chieh Hsu, Dylan Rankin, **Alex Schuy**

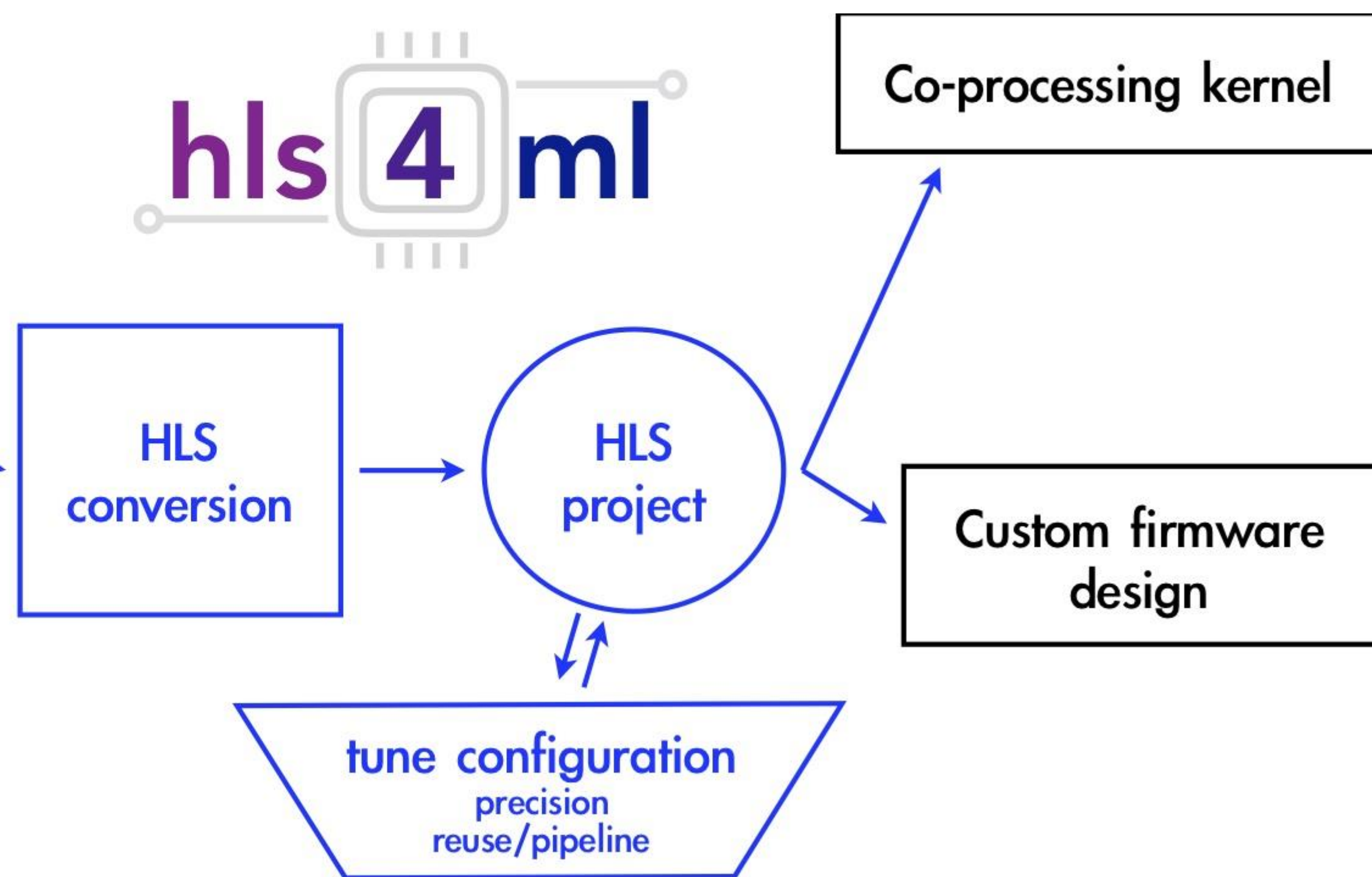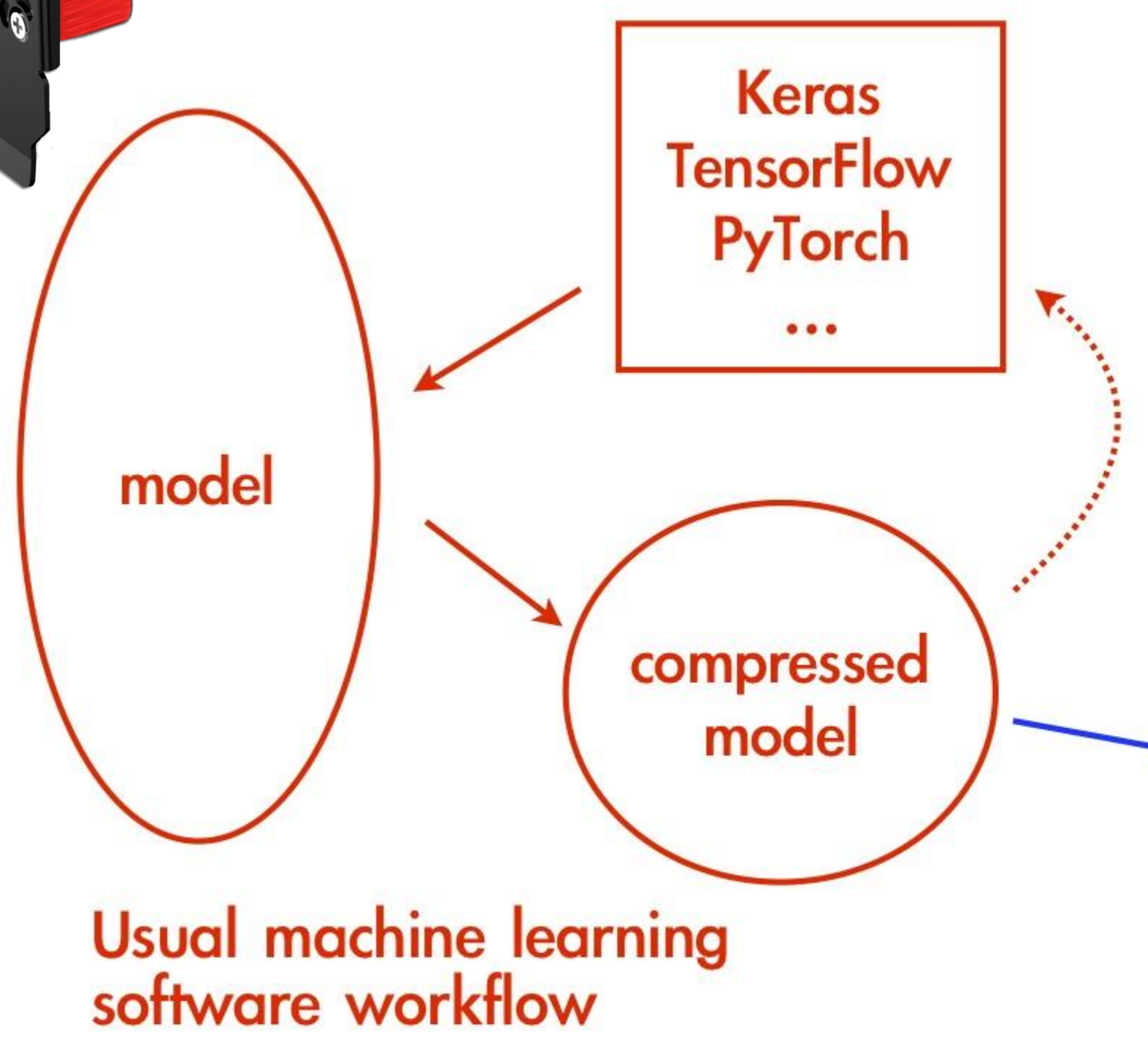Penn · 陽明交大 NYCU · MIT · W · a3d3

Motivation

- **Search for new physics** → particle colliders at higher energy & luminosity!
  - **exceed computing budget** ☹
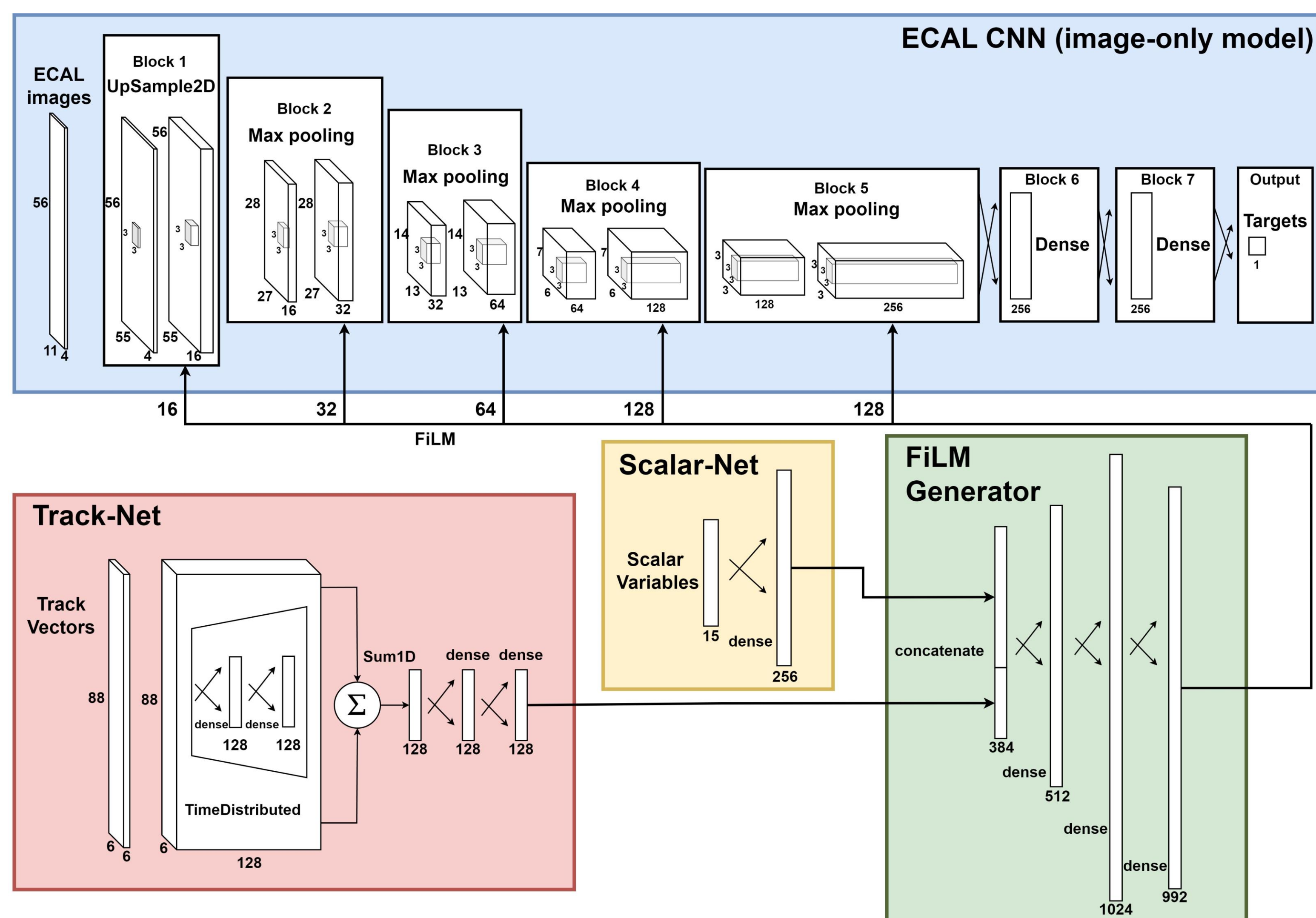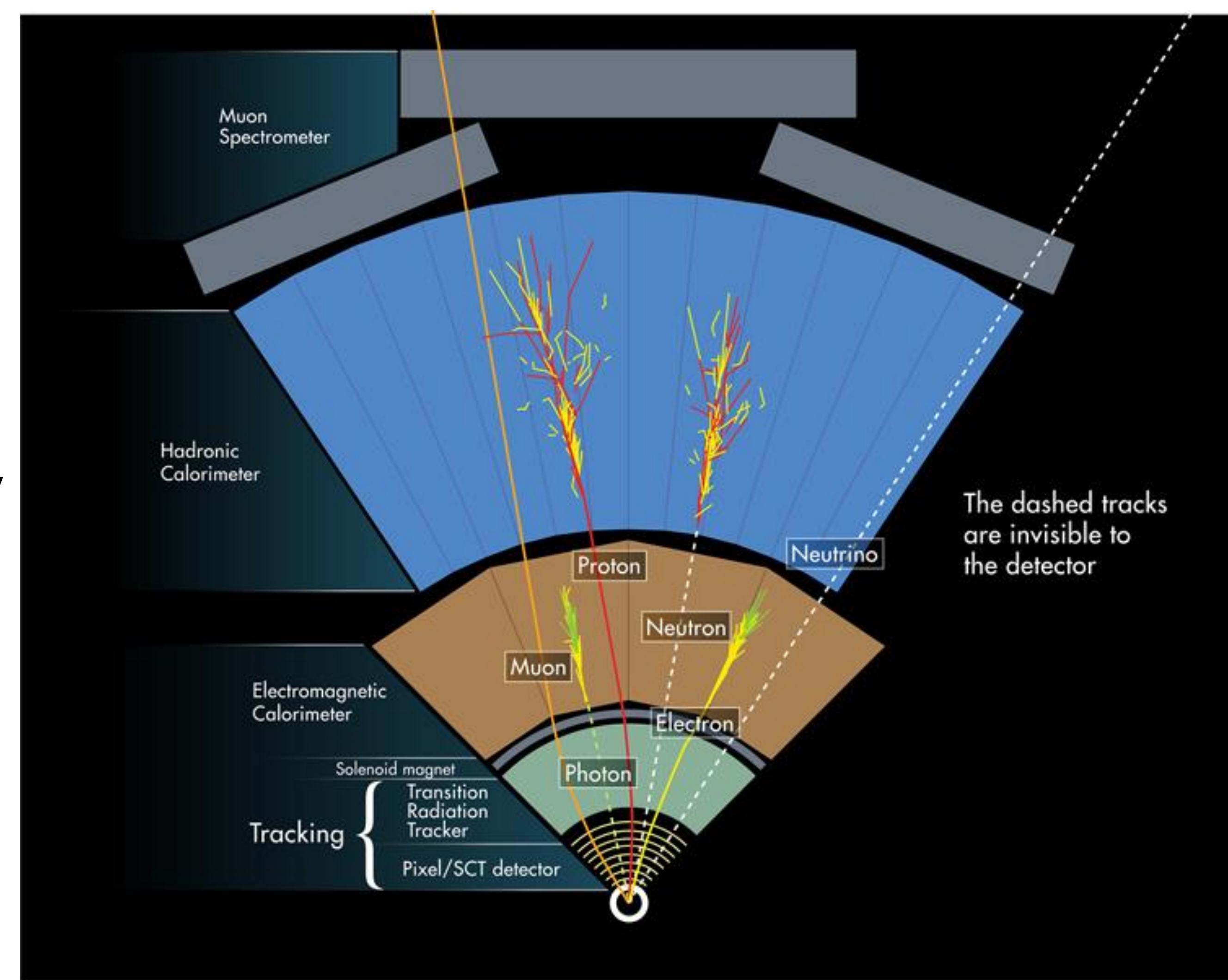- **Deep learning** is prevalent, but can be slow...
  - Need accelerators: GPUs, **FPGAs**

Usual machine learning software workflow

Event Reconstruction

- **Complicated process:**
  - Tracking
  - Vertexing
  - **EM Calorimetry**
  - Hadronic Calorimetry
  - ...
- **Many particles**
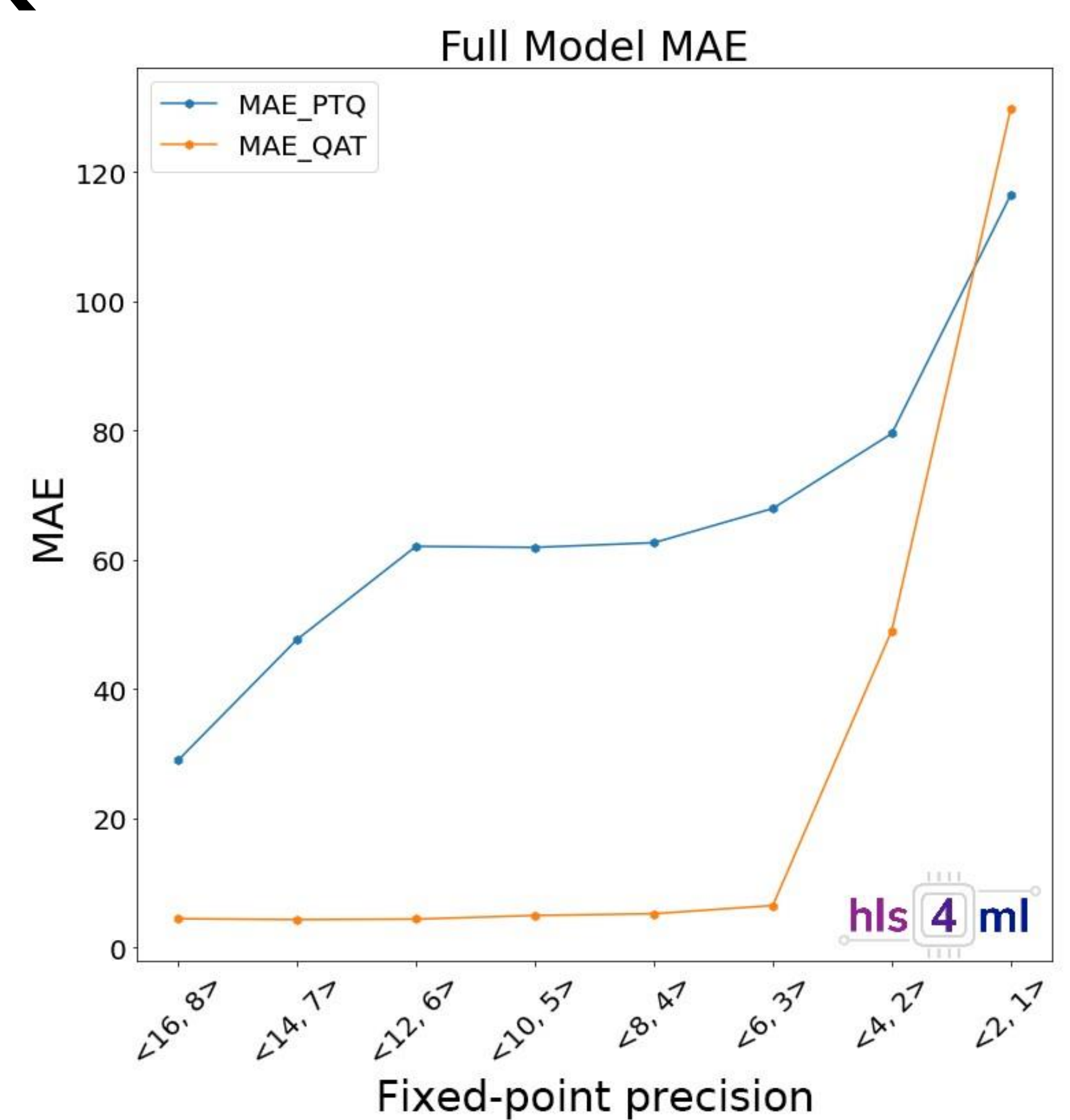 - **Electrons, photons**, muons, protons, neutrons, pions, ...

The dashed tracks are invisible to the detector

DeepCalo

- **Energy regression**
- **CNN-based model**
  - **Large (~3.6M par.)**
- **~20% improvement over BDT method**

HLS4ML

- **High-level → FPGA conversion**
- **Popular in particle physics**
- **Not well-optimized for large models**

Timing Studies

- **FPGA outperforms CPU/GPU at batch size = 1 & 5 on latency, speedup, power & energy**

Quantization

- **Extend DeepCalo w/ QKeras**
- **Quantization-Aware Training**

Full Model MAE

Coprocessor	CPU			GPU			FPGA	
Type	Ryzen 7 3700X	Ryzen 5 5600H	AMD EPYC 7262	RTX 2070 Super	Tesla V100	RTX 2080 Ti	single-stream	mixed-type
Batch=1								
Latency	7.52ms	8.75ms	5.865ms	8.47ms	4.8ms	8.2ms	1.106ms	0.898ms
Speedup	1.164×	**1×**	1.492×	1.033×	1.823×	1.067×	7.911×	**9.744×**
Power	53.73W	29.13W	42.65W	49.77W	60.11W	64.54W	19.76W	20.75W
Energy	404.05mJ	254.888mJ	250.142mJ	421.552mJ	288.528mJ	529.228mJ	21.855mJ	18.634mJ
Batch=5								
Latency	11.5ms	13.45ms	10.545ms	9.75ms	5.1ms	7ms	2.695ms	1.485ms
Speedup	1.17×	**1×**	1.275×	1.379×	2.637×	1.921×	4.991×	9.057×
Power	62.44W	37.67W	48.94W	51.83W	61.73W	84.18W	21W	23.775W
Energy	718.06mJ	506.66mJ	516.07mJ	505.345mJ	314.825mJ	589.26mJ	56.595mJ	35.305mJ