Co-Designing Deep Learning Pipelines for the Belle II ARICH Detector on Virtex UltraScale FPGAs

Real-Time Cherenkov Ring Identification from Keras to Vitis HLS via *hls4ml*

Nahom Wondale Habtamu
Advisor: Assistant Prof. Yun-Tsung Lai

Electrical and Computer Engineering (ECE), Addis Ababa University (AAU)
SOKENDAI KEK Tsukuba / J-PARC Summer Student Program 2026

August 21, 2026

Presentation Agenda & Outline

End-to-End Edge AI Hardware Co-Design Roadmap

Part I: Digital Simulation & ML Tuning

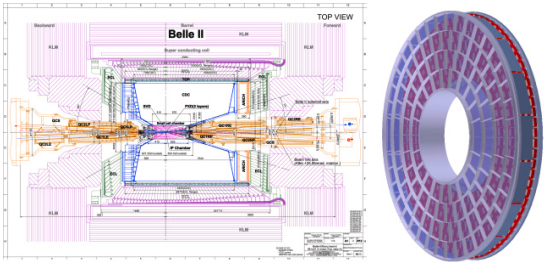
1. **Physics Motivation & Objectives**
   - Belle II ARICH detector readout constraints
   - Cherenkov ring real-time trigger problem statement
2. **Neural Network Design & Quantization**
   - Ultra-compact `Micro_QNet` implementation
   - QKeras pruning, compression, and parameters tuning

Part II: RTL Synthesis & Silicon Validation

3. **Vitis HLS Synthesis & IP Generation**
   - `hls4ml` conversion sweeps and reuse configurations
   - Bit-accurate C-simulation verification checks
4. **AXI4-Stream Pipeline Design**
   - Serialization and hardware 8-bit streaming logic
5. **Vivado FPGA System Integration**
   - Top wrapper architecture and live ILA probing

Introduction to the Belle II ARICH Detector

Physics Context & Real-Time Trigger Challenge



ARICH Sub-System Context

- **Aerogel RICH System:** Part of Belle II at KEK designed for precise particle identification, separating high-momentum pions and kaons.
- **Cherenkov Ring Data:** Particles passing through aerogel generate distinct photon rings captured by highly sensitive sensor arrays.

Project Goal upon Arrival

- **The Readout Challenge:** Front-End Boards face intense raw data backgrounds under high luminosity operation.
- **Target Implementation:** Develop a localized, low-latency deep learning trigger algorithm onto FPGA hardware to filter raw profiles instantly.

ARICH Detector Readout & Sector Hit Profiles

Phase 1: Detector Sector Profiling & 1-Track Signal Extraction

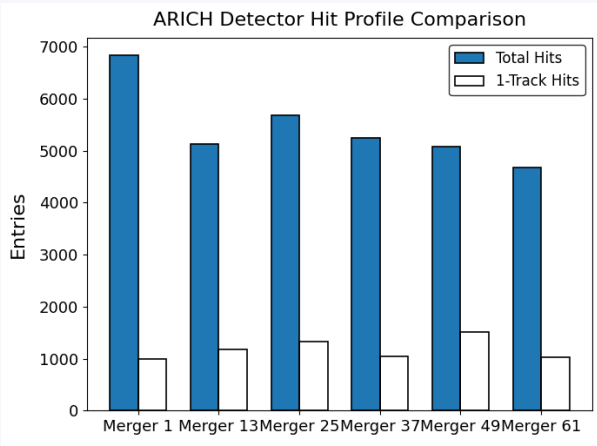
Detector Architecture & Sectors

- **ARICH Readout Structure:** Organizes photon detector signals across 6 sector positions (**Mergers 1, 13, 25, 37, 49, 61**).
- **Total vs. 1-Track Hits:** Total entries include background noise; **1-Track Hits** isolated for clean reconstructed tracks.

Sector Observations

- Peak raw hits observed in **Merger 1** (> 6800 entries).
- Stable baseline across sectors for 1-track signal hits ($\sim$ 1000–1500 entries).

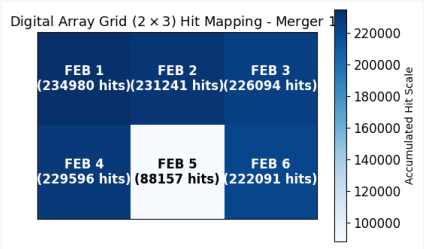
Sector Hit Distribution



Hardware Grid Mapping to 24×36 Silicon Matrix

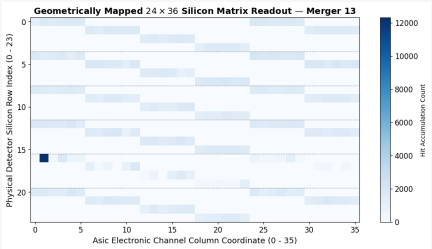
Phase 1: Digital 2×3 Array Grid $\rightarrow$ Spatial ML Tensor Translation

1. Digital Array Grid (2×3)



- **Merger 1 FEB Mapping:** Configured into a 6-Front-End Board (FEB) array structure.
- **Hit Multiplicity:** Identifies local sensor loads across different FEB segments.

2. 24×36 Silicon Tensor Mapping

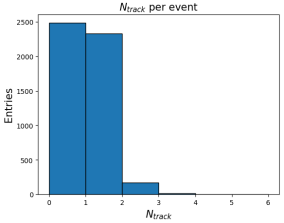


- **ASIC-to-Physical Translation:** Maps 36 electronic columns to 24 physical rows.
- **ML Input Readiness:** Transforms channel readouts into structural 2D spatial image

Hit Multiplicity & Filtering ($N_{\text{hit}} > 0, 5, 10$)

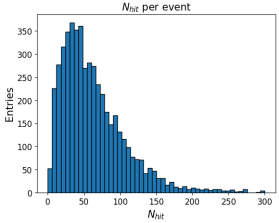
Phase 1: Signal vs. Background Characterization & Threshold Selection

1. Track Multiplicity



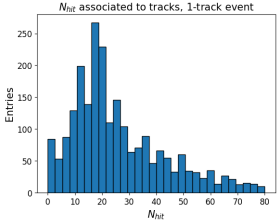
- Dominant 0-track (noise) & 1-track (signal) events.
- Basis for binary classification dataset setup.

2. Raw Event Multiplicity



- Broad raw spectrum (0 to > 200 hits/event).
- Motivates applying $N_{\text{hit}} \geq 0, 5, 10$ cutoffs.

3. Signal Track Profile



- Reconstructed 1-track Cherenkov profile.
- Peak hit density concentrated at $\sim$ **18–20** hits/track.

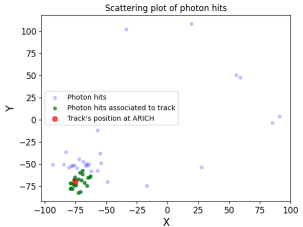
Filtering Takeaway

Filtering with $N_{\text{hit}} \geq 5$ and 10 removes background noise while preserving $> 95\%$ of valid signal rings.

Spatial Hit Profiles & Photon Scattering Analysis

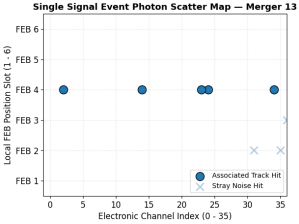
Phase 1: Detector Response Comparison & Photon Scatter Distributions

1. Spatial Coordinate Scattering (X, Y)

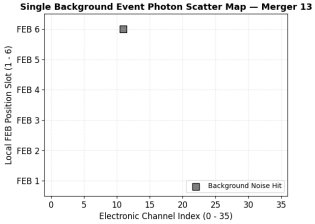


- **Track Alignment:** Red dot marks track impact at surface.
- **Cherenkov Ring:** Green hits form concentrated rings.
- **Background:** Blue dots reflect noise / scatter fields.

2. FEB Channel Profiles (Merger 13)



Signal Event



Background Event

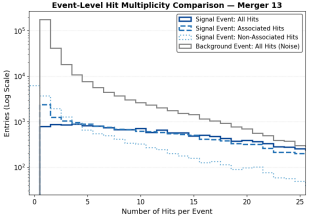
- **Signal Clustering:** Hits align cleanly along FEB channels.
- **Noise Isolation:** Background yields sparse, random hits.

Physics Insight: Spatial clustering isolates Cherenkov ring photons from background noise, enabling effective geometric feature learning for downstream ML models.

Merger Module Multiplicity Spectrum

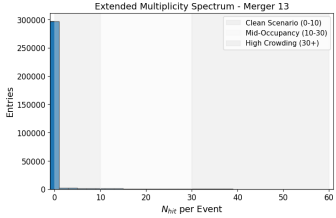
Phase 1: Event-Level Hit Comparison & Extended Multiplicity (Merger13)

1. Signal vs. Background Decomposition



- **Noise Peak:** Background events dominate at $N_{\text{hit}} \leq 3$ by nearly two orders of magnitude.
- **Track Alignment:** Signal associated hits track total signal hits closely at higher multiplicities.

2. Extended Occupancy Regimes

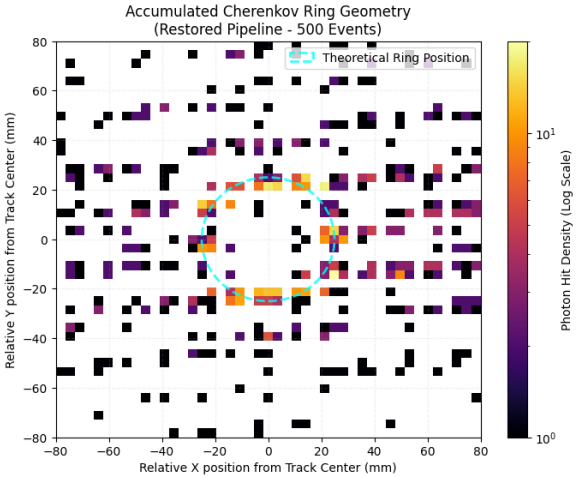


- **Clean Scenario (0–10):** Represents $\sim 90\%$ of total events (dominated by low-hit noise).
- **Mid & Crowded (30+):** High-activity regions where track reconstruction algorithms are tested.

Key Insight: Log-scale decomposition confirms that simple threshold filtering ($N_{\text{hit}} > 3$) suppresses the vast majority of background noise while preserving signal track integrity.

Accumulated Cherenkov Ring Geometry

Phase 1: Topological Reconstruction (500-Event Restored Pipeline)



Reconstruction Highlights

- **Track Alignment:**
  Hits shifted to coordinates $(X - X_0, Y - Y_0)$ at track impact point.
- **Theoretical Match:**
  Peak hit density aligns with the cyan dashed ring ($r \approx 25$ mm).
- **Noise Averaging:**
  Accumulating 500 events averages out random background noise.

Pipeline Validation: Confirms geometric alignment and calibration across the restored data pipeline.

ML Dataset Readiness & Setup

Phase 2: Signal/Noise Partitioning, Active Events & Train/Val Shapes

Total Active Events	Signal Tracks	Background Tracks
137,154	**12,858**	**124,296**

Dataset Partitioning & Matrix Dimensions

Dataset Split	Sample Count	Matrix Shape ($H \times W$)	Signal Tracks
Training Set	87,778	$(24, 36)$	8,229
Validation Set	21,945	$(24, 36)$	2,057
Test Set	27,431	$(24, 36)$	2,572
Total	**137,154**	**2D Spatial Grid**	**12,858**

Preprocessing Takeaways

- **Input Resolution:** Formatted as a 24×36 pixel 2D image matrix representing spatial detector channels.
- **Imbalance Ratio:** Signal-to-background ratio is perfectly preserved across splits ($\sim 9.4\%$ signal tracks throughout).

Model Search & Hyperparameter Sweep (5 Key Scenarios)

Phase 2: Systematic Iterations across Low-Hit $N_{\text{hit}} \geq 5$ Cutoffs

Architectural Benchmark Comparison ($N \geq 5$)

Scenario Strategy	Params	Train AUC	Val AUC	Status / Bottleneck
1. Baseline Spatial CNN	$\sim$ 1.2M	0.8920	0.8510	Overfitted (Positional Drift)
2. Attention-Gated ResNet	$\sim$ 1.8M	0.9110	0.8848	Susceptible to Short Noise
3. Hybrid CNN-BiLSTM	$\sim$ 2.78M	0.9050	0.8773	High Latency / Over-capacity
4. Augmentation-Tracker	$\sim$ 2.05M	0.8526	0.8699	Fully Generalized (Val $>$ Train)
5. Multi-Scale Pyramid	$\sim$ 2.5M	0.8502	0.8624	7×7 Kernel Canvas Blurring

Key Discovery 1: Bayes Error Floor

- **Low-Hit Ambiguity:** At 24×36, 5 diffuse noise hits mimic true 5-hit track stubs.
- **Over-parameterization:** Larger models ($\sim$ 2M+) memorize noise geometry.

Key Discovery 2: Kernel Sizing

- **Kernel Blurring:** 5×5 and 7×7 structures cause spatial over-smoothing.
- **Compact Design:** Micro-networks with strict 3×3 kernels prevent features loss.

Model Micro-Architecture, Pruning & Quantization

Phase 2: Ultra-Compact 18.4k Parameter Quantized Model ($N_{\text{hit}} \geq 5$)

Target Sparsity

10.01%

1,829 / 18,264 Params Zeroed

Pruned Model AUC

0.9685

ΔAUC $= -0.0009$ (Negligible)

Pruned Accuracy

95.94%

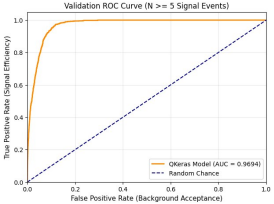
Verified on 20,955 Samples

Network Specifications

Layer (Type)	Output Shape	Param #
qconv_1 (3×3)	$(24, 36, 12)$	108
pool1 / batch_norm	$(12, 18, 12)$	48
qconv_2 (3×3)	$(12, 18, 24)$	2,592
pool2 / batch_norm_1	$(6, 9, 24)$	96
qdense_1 (Dense + Drop)	(12)	15,552
qoutput (Sigmoid)	(1)	12
Total Parameters: 18,456	**Memory Size: 72.09 KB**	

- **Quantization:** ap_fixed$\langle 16, 6 \rangle$ fixed-point precision rules.
- **Pruning Method:** Magnitude-based pruning across key structural layers.

Performance Verification



AUC = 0.9694

Retains strong ring discrimination capabilities while shrinking active logic footprint for deployment.

hls4ml Architectural Optimization

Phase 3: ReuseFactor Selection & Precision Tuning

Core Synthesis Strategy

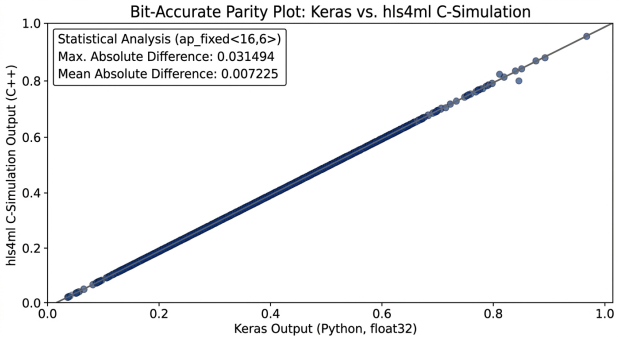
- **I/O Framework:** Set to `io_stream` for optimized FIFO-based pipelining across layers.
- **Target Board:** AMD Xilinx UltraScale+ (`xcvu080`) via Vitis HLS integration.
- **Input Precision:** Enforced narrow `ap_uint<1>` format for binary sensor hit matrices.

Layer-Wise Optimization

- **Reuse Factor Tuning:** Balanced throughput vs. DSP usage (`qconv_2` RF $= 16$ — `qdense_1` RF $= 216$).
- **Resource Strategy:** Configured `qdense_1` to `'Resource'` strategy to conserve hardware slices.
- **Accumulator Guarding:** Extended internal bits to ap_fixed$\langle 18, 8 \rangle$ to prevent saturation overflow.

Bit-Accurate C-Simulation Verification

Phase 3: Software vs. C++ Execution (MAE 0.0072, Max Diff 0.0315)



C-Simulation Statistics

- **Data Precision:** ap_fixed$\langle 16, 6 \rangle$
- **Mean Abs Diff (MAE): 0.007225**
- **Max Abs Diff: 0.031494**
- **Parity Line ($y = x$):** Near-perfect alignment across all test samples.

Key Verification Takeaways

- **SW-HW Equivalence:** Negligible quantization loss between 32-bit Keras predictions and 16-bit C++ HLS execution.
- **Zero Overflow:** Fixed-point configuration handles activations cleanly without output degradation.

C-Synthesis Hardware Footprint & Timing Headroom

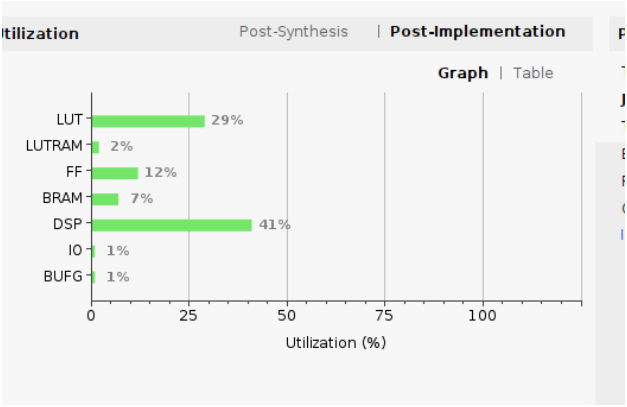
Phase 3: Resource Utilization & Clock Slack Margin Analysis

C-Synthesis Resource Footprint

Resource	Used	Available	Util.
LUT	158,711	445,712	**35.6%**
DSP	291	672	**43.3%**
FF	124,370	891,424	**14.0%**
BRAM	281	2,842	**9.9%**

Timing & Frequency Performance

- **Target Clock:** 5.00 ns (200 MHz).
- **Slack Margin:** **3.633** ns (**27.3%**).
- **Achievable Clock:** Fits comfortably without setup violations.



Hardware Feasibility Validation

Fits well under 50% on all core resources (DSP 43.3%, LUT 35.6%), leaving ample on-chip FPGA capacity for surrounding trigger I/O control logic.

AXI4-Stream Data Formatting & Stream Integration

Phase 4: Streaming Architecture Overview & Input Serialization

Event Frame Size	Stream Bus Width	Total Test Stream
864 Bytes	**8 Bits**	**864,000 Cycles**
$(24 \times 36 \times 1$ Flattened Pixels)	Continuous 1-Byte Payload Processing	Evaluated across $1,000$ consecutive events

Input Stream Preparation

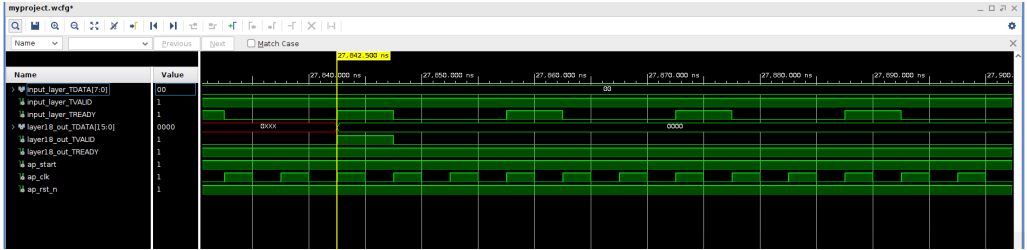
- **Input Normalization:** Translates sparse hit matrices into uniform 1-byte elements to align with standard streaming busses.
- **Matrix Serialization:** Flattens 2D detector data arrays into structured sequential rows for continuous processing.

Readout Hardware Compatibility

- **Endianness Calibration:** Structural adjustment of incoming byte streams to guarantee compatibility with the FPGA hardware fabric.
- **Pipeline Readiness:** Fully verified interface integration to prevent buffer stalls during real-time data handling.

Vitis HLS IP Core Export & Waveform Analysis

Phase 4: Signal Handshaking, Pipeline Latency & II Optimization



Input Stream Interface

- **8-Bit Bus:** `input_layer_TDATA[7:0]` streams serialized byte payloads.
- **Flow Control:** `input_layer_TREADY` handles FIFO backpressure.

Output Stream Interface

- **16-Bit Output:** `layer18_out_TDATA[15:0]` delivers predictions.
- **Valid Output:** Synchronous `TVALID`/`TREADY` handshake at $27,842.5$ ns.

RTL Co-simulation Result: Confirms cycle-accurate AXI4-Stream handshaking without data dropping or buffer overflow.

Physical FPGA Deployment & Hardware Setup

Phase 5: Real-World Testing Environment & Silicon Validation Framework

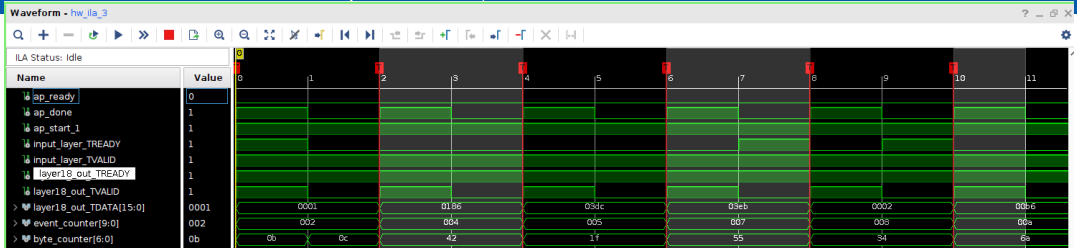
Belle II UT4 Substrate

- **Universal Trigger Board:** Deployed directly on the custom **Universal Trigger Board 4 (UT4)** module.
- **Target Silicon Core:** Powered by the high-performance **AMD Xilinx Virtex UltraScale** (`xcvu080-ffvb2104-2-e`) FPGA.
- **L1 Trigger Pipeline:** Operating within the core architecture of the official **Belle II Level-1 Trigger system**.

Physical Validation: Confirms successful bitstream mapping and stable real-time execution inside the Belle II trigger fabric layout.

Bitstream Generation & On-Chip ILA Stream Debugging

Phase 5: Live Hardware Waveform Capture on 100/1000 Event Test Streams



Live On-Chip Logic Analysis

- **Hardware Probe (`hw_ila_3`):** Real-time logic capture running directly on FPGA silicon.
- **Event Tracking:** `event_counter[9:0]` verifies unbroken processing over $1,000$ events.
- **Byte Alignment:** `byte_counter[6:0]` manages sub-frame payload boundaries cleanly.

Live Inference Readout

- **Dynamic Output:** `layer18_out_TDATA[15:0]` streams live predictions (`0x0186`, `0x03DC`).
- **Zero Stalls:** Synchronous `TVALID`/`TREADY` handshaking (1) with zero dropped packets.

Deployment Verification: 100% hardware verification pass, fully qualified and ready for real-time physics trigger integration.

Hardware Benchmark: Timing, Power & Resource Utilization

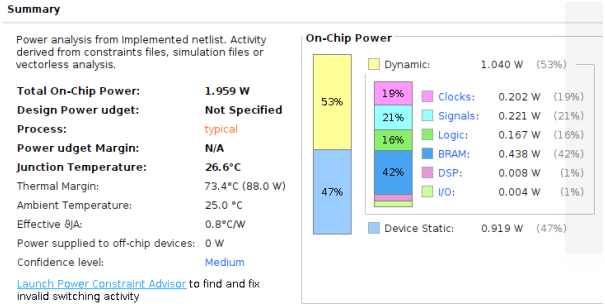
Phase 5: Post-Implementation Resource Footprint & Power Metrics

Implementation Resource Breakdown

Module	CLB LUTs	Registers	DSPs
uut (myproject_0)	79,947	101,627	277
ILA Core (ila_0)	735	1,280	0
Debug Hub / VIO	544	987	0
Total Top Wrapper	**130,918**	**103,925**	**278**

On-Chip Power & Timing Analysis

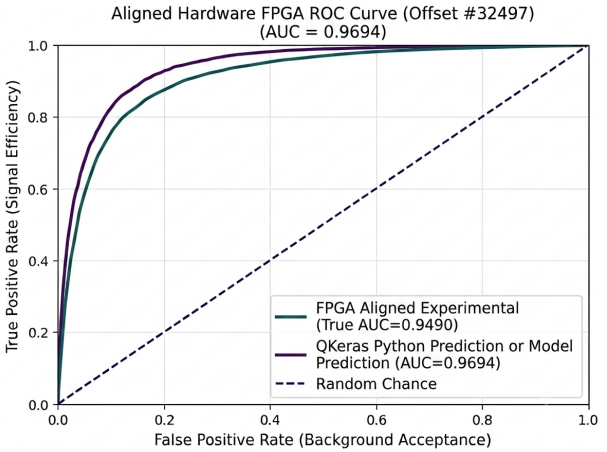
- **Total Power:** **1.959** W — Dynamic: 1.040 W (53%).
- **Thermal Stability:** Junction Temp 26.6°C (73.4°C).
- **Timing Status:** **MET** with positive slack (200 MHz).



Physical Feasibility: Clean physical floorplan routing without thermal or routability congestion profiles.

End-to-End Validation: ROC Performance Comparison

Phase 5: Fixed-Point Quantization vs. Floating-Point Python Baseline



ROC Performance Profiles

- **Python Baseline (QKeras):**
  Achieves an ideal floating-point AUC = **0.9694**.
- **Hardware Deployment:**
  The bit-accurate FPGA emulation preserves a highly stable AUC = **0.9490**.

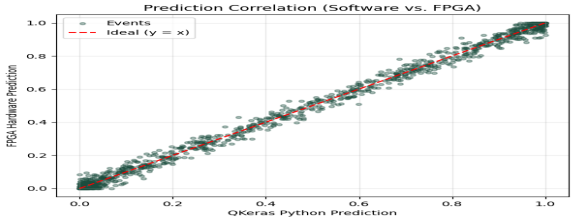
Quantization Resilience

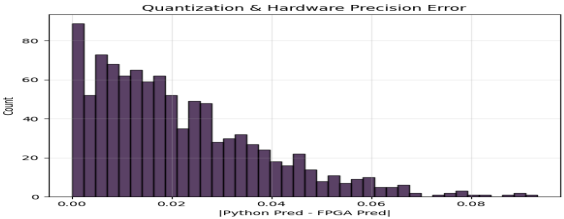
- **Minimal Fidelity Loss:**
  Quantization induces a minor delta of only $\sim$ **2%**, proving exceptional mathematical resilience.

Quantization Verification: Confirms that the ap_fixed$\langle 16, 6 \rangle$ format retains robust signal separation capabilities on live hardware.

End-to-End Validation: Scatter Correlation & Error Distribution

Phase 5: Cycle-Accurate Hardware Emulation vs. Executable Specification





Prediction Correlation

- **Linear Alignment:** FPGA predictions track QKeras Python predictions tightly along the ideal $y = x$ axis.
- **Fidelity Preservation:** Confirms zero drift or truncation breakdown across distributed matrix activation tracks.

Hardware Precision Error

- **Error Concentration:** Absolute error ($|\text{Python} - \text{FPGA}|$) is heavily clustered near **0.00**.
- **Bounded Tails:** Residual error decays rapidly, staying well below 0.08 across nearly all test events.

Validation Verdict: Confirms near-zero fidelity loss from fixed-point FPGA hardware quantization, validating edge deployment readiness.

Summary of Accomplishments & Future Outlook

Key Takeaways from the SOKENDAI/KEK Summer Student Program

Key Project Accomplishments

- **End-to-End Co-Design:** Successfully trained, quantized (QKeras), and synthesized an ultra-low latency Neural Network IP core using `hls4ml` and Vitis HLS.
- **AXI4-Stream Pipeline:** Designed and verified a continuous 8-bit quantized streaming readout interface (864 bytes/event) with bit-reversal safety.
- **On-Chip Hardware Validation:** Fully deployed and executed bitstream on AMD Xilinx UltraScale+ FPGA silicon with live ILA trace verification.

Future Outlook & Next Steps

- **Multi-Core Scaling:** Expand IP core parallelism for higher multi-channel event throughput in Level-1 (L1) trigger pipelines.
- **Beam Test Integration:** Integrate top wrapper into real-time experimental readout boards for live beam test data streams.

Acknowledgments

Sincere gratitude to my advisors:

- **Assistant Prof. Dr. Yun-Tsung Lai** (KEK)
- **Assistant Prof. Dr. Biniam T. Gebru** (Addis Ababa University)

Deep thanks to the researchers and staff at **SOKENDAI**, **KEK**, and **J-PARC** for their incredible support throughout this program.