HPCNeuroNet: A Neuromorphic Approach Merging SNN Temporal Dynamics with Transformer Attention for FPGA-based Particle Physics

TEMSA R&D Center

Author List: *Murat Isik, Hiruna Vishwamit, Jonathan Naoukin, I. Can Dikmen*

Conference on Computing in High Energy Nuclear Physics (CHEP) October 19-25 Krakow, Poland

Presentation Layout

ABSTRACT

INTRODUCTION

BACKGROUND

- Neural Networks in Particle Physics
- Expressivity of Hybrid Neural Network Models
- Integrating HLS4ML and FPGA

PROPOSED DESIGN METHODOLOGY

- Mathematical models of Datasets
- HPCNeuroNet
- Experimental Setup
- Additional Considerations:

EVALUATION

- Hardware Accelerator MAC Components
- Throughput Analysis

CONCLUSIONS

Introduction

Particle physics constantly deals with complex detector challenges.

Machine learning has improved particle detection, but challenges remain.

Neuromorphic computing offers promising solutions for these challenges.

Introduction

Detector System Challenges	• The expansive realm of Particle Physics continually grapples with intricate challenges, prominently in detector systems.
Particle Identification Issues	• Particle Identification faces concerns like Timing Resolution, Quartz Polishing, and Mechanical systems.
Neuromorphic Computing Potential	• Neuromorphic Computing combines biological systems and artificial intelligence.
HLS4ML Package	• FPGA pros

Background

Neuromorphic computing integrates biological systems with AI for highly efficient computing.

Particle Physics experiments generate vast datasets requiring advanced computational techniques.

Hybrid neural networks enhance particle identification accuracy and efficiency.

FPGAs are crucial for their reconfigurability and parallel processing capabilities in particle physics

The synergy between SNNs and Transformers improves the interpretation of complex detector data.

Background

Machine Learning (ML) techniques have significantly advanced Particle Physics.

The limitations of classical computer architectures lead to exploring new solutions like neuromorphic computing

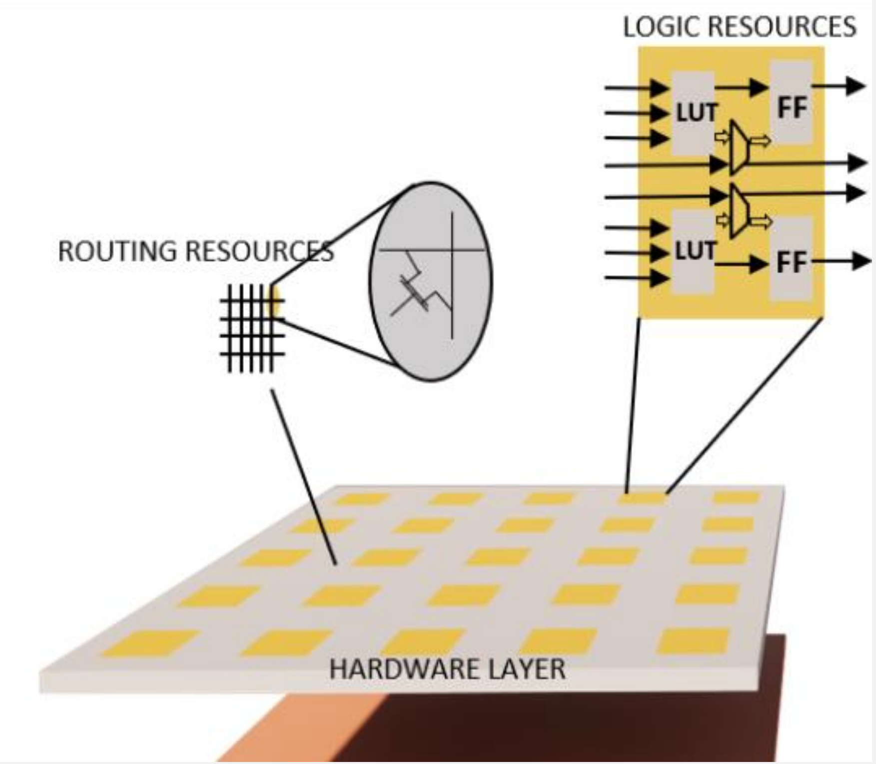
Neuromorphic systems offer rapid parallel processing and energy efficiency.

Hybrid models combining SNNs and Transformer attention mechanisms handle large data streams effectively.

The HLS4ML package simplifies adapting complex ML models to FPGA hardware.

The Fundamental Architecture of FPGA

HPCNeuroNet offers a high-performance computing methodology for particle physics, integrating SNN and Transformer technology on FPGA.



The Fundamental FPGA Architecture

PROPOSED DESIGN METHODOLOGY

Mathematical models of Datasets

- The dataset includes 100k dielectron and 20k proton collision events, showcasing Gaussian and exponential momentum distributions.

HPCNeuroNet

- HPCNeuroNet processes time-series data efficiently using HLS4ML, Transformer layers, Self-Attention, and SNN dynamics to produce refined outputs for advanced analysis.

Experimental Setup

- Implemented using Python on NVIDIA RTX 3060 GPU and Intel i9 12900H CPU.

-Deployed HPCNeuroNet on FPGA using HLS4ML and PYNQ, converting models from TensorFlow/Keras to HLS code, synthesizing with Vivado HLS, and programming the FPGA.

Additional Considerations:

-Quantize and prune HPCNeuroNet to improve speed and save FPGA resources.

-Ensure software compatibility and use debugging tools for deployment issues.

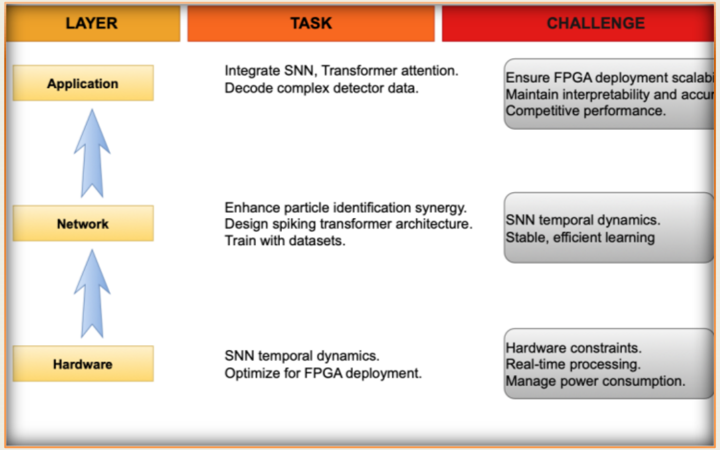
Benefits of Neural Networks in Particle Physics.

Feature	Description
Pattern Recognition	Identifies nuanced features in particle trajectories.
Generalizability	Predicts outcomes on unfamiliar datasets.
Parallel Processing	Evaluates diverse data entities concurrently.
Hardware Synergy	Enables streamlined neural network implementations.

Machine Learning, especially neural networks, has revolutionized particle physics with advanced pattern recognition, generalizability, and efficient data handling, promising further breakthroughs with neuromorphic computing.

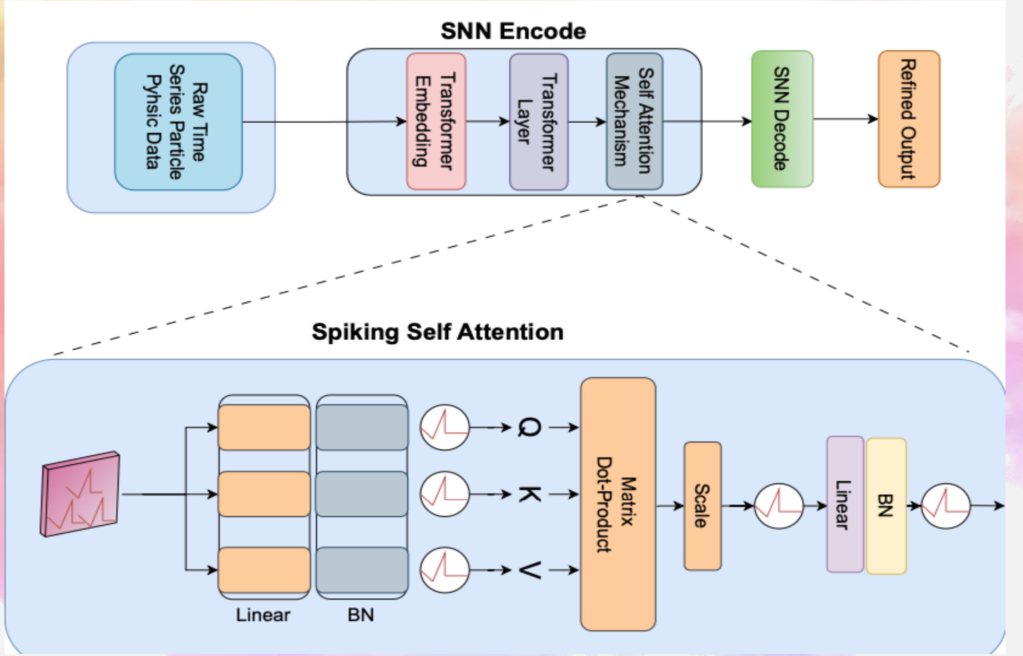
Hybrid Models and FPGA Integration for Efficient Real-Time Processing

- Hybrid models enhance computational capabilities, requiring platforms like HLS4ML to transition ML models into firmware for hardware deployment. Integrating HLS4ML with FPGAs enables efficient, scalable, and precise real-time particle identification, marking a significant technological advancement.



HPCNeuroNet framework depicted across three dimensions: layers, tasks, and challenges of the design.

HPCNeuroNet: Transforming Particle Physics Data with Advanced Neural Networks



The HPCNeuroNet Model processes raw time-series data through a series of sophisticated layers, including Transformer Embedding, Self-Attention Mechanism, Spiking Self Attention, and SNN Encode and Decode stages, resulting in a refined output for further analysis and application.

HPCNeuroNet model processes time series data using Transformer Embedding and Self-Attention, integrates it with Spiking Neural Network dynamics, refines the data, and prepares it for advanced analysis and applications.

Performance and Resource Utilization Analysis of Collision Models on Zynq®-7000 SoC

Zynq®-7000 SoC			
Resource	**Utilization**	**Available**	**% Utilization**
LUT	17579	74000	23.75
FF	20060	106400	18.85
BRAM	1374	3300	41.64
IO	36	150	24
DSP	85	160	53.13

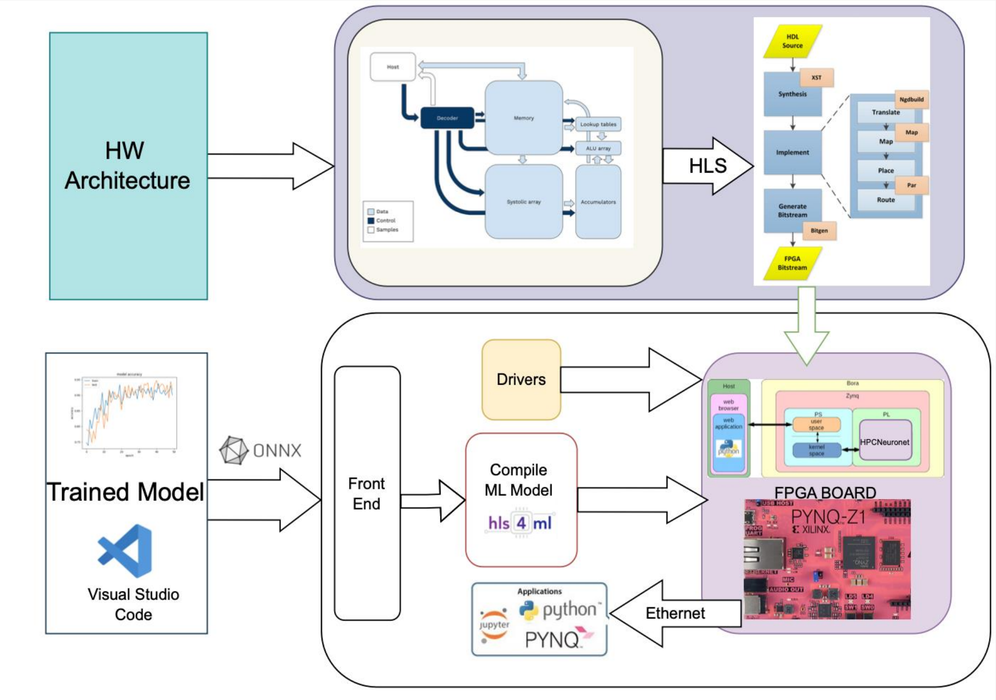
Resource utilization summary

CMS-Electron Collision model exhibits the highest accuracy (94.48%) and the lowest latency (11.5 ms) with minimal computational demands (0.49 GOPs)

In contrast, Geant4-Particle Identification and CMS-Proton Collision models have lower accuracy and higher computational requirements, but all models show minimal latency differences, important for real-time applications.

Performance Evaluation of CERN Electron Collision Models and Particle Identification on HPCNeuronet using FPGA-Based Hardware Architecture

CMS-Electron Collision model outperforms others in accuracy, latency, and computational efficiency



Geant4 and CMS-Proton Collision models have lower accuracy and higher computational demands.

Evaluation results for CERN Electron Collison, LArTPC waveforms, and Particle Identification from Detector Responses on HPCNeuronet.

Comparison of our HPCNeuroNet framework with other models.

Models	DNN	GNN	1-D CNN	**HPCNeuronet (Ours)**
Accuracy (%)	85.1	78.3	85.8	**88.73**
MAC (GOP)	1.32	2.12	0.77	**1.29**
Latency [ms]	24.2	32.9	18.3	**11.5**
Power Efficiency	10.4	3.5	12.2	**22.7**

*Power Efficiency unit is GOP/s/W.

*CMS-Electron Collison dataset was implemented on these models.

Evaluation

HPCNeuroNet integrates SNN temporal dynamics with Transformer attention mechanisms.

The model is optimized using the HLS4ML framework for FPGA deployment.

FPGA's reconfigurability enhances computational speed and accuracy.

Performance benchmarks show HPCNeuroNet's superiority over traditional models.

Future research will explore broader applications and further optimizations.

These advancements highlight the transformative potential of neuromorphic computing in particle physics.

Conclusion

HPCNeuroNet, combining Spiking SNNs and Transformers, excels in particle identification tasks from detector responses.

Using HLS4ML for FPGA deployment, HPCNeuroNet enhances computational speed and accuracy in particle physics.

Despite implementation and compatibility challenges, HPCNeuroNet's architecture shows significant potential in high-energy physics.

HPCNeuroNet's integration of SNNs, Transformers, and FPGA-based computing sets the stage for future research across various scientific domains.

Thank you!

Contact:

can.dikmen@temsa.com

mrtisik@stanford.edu

TEMSA