

# Yash Akhauri

Ph.D. Student at Cornell University

☎ (+91) 7891512802  
✉ ya255@cornell.edu  
📁 akhauriyash.github.io

## Education

- 2022 : **Ph.D. in Electrical & Computer Engineering**, Cornell University, New York.
- 2016–2020 : **B.E. in Electronics & Instrumentation**, Birla Institute of Science & Technology, Rajasthan.  
Thesis: Exposing Hardware Building Blocks to Machine Learning Frameworks [\[arXiv\]](#)

## Research Portfolio

- 2023 Encodings for Prediction-based Neural Architecture Search,  
**Yash Akhauri**, Mohamed S Abdelfattah, Under Review at ICML'24.
- 2023 On Latency Predictors for Neural Architecture Search,  
**Yash Akhauri**, Mohamed S Abdelfattah, Under Review at MLSys'24.
- 2023 Multi-Predict: Few Shot Predictors For Efficient Neural Architecture Search,  
**Yash Akhauri**, Mohamed S Abdelfattah, **AutoML'23**.
- 2022 EZNAS: Evolving Zero Cost Proxies For Neural Architecture Scoring,  
**Yash Akhauri**, J. Pablo Muñoz, Nilesh Jain, Ravi Iyer, **NeurIPS'22**.
- 2021 Enabling One-Shot NAS With Automatic Super-Network Generation,  
J. Pablo Muñoz, Nikolay Lyalyushkin, **Yash Akhauri**, Anastasia Senina, Alexander Kozlov, Nilesh Jain, **Practical-DL AAAI'22**.
- 2021 A Genetic Programming Approach To Zero-Shot Neural Architecture Ranking,  
**Yash Akhauri**, J. Pablo Muñoz, Nilesh Jain, Ravi Iyer, **AIPLANS NeurIPS'21**.
- 2020 LogicNets: co-designed neural networks and circuits for extreme-throughput applications,  
**Yash Akhauri\***, Yaman Umuroglu\*, Nicholas James Fraser, Michaela Blott, **FPL'20**.  
Stamatis Vassiliadis (best paper) Award & DATE Special Session Presentation
- 2020 High-throughput dnn inference with logicnets,  
Yaman Umuroglu, **Yash Akhauri**, Nicholas James Fraser, Michaela Blott, **FCCM'20**.
- 2019 Hadanets: Flexible quantization strategies for neural networks,  
**Yash Akhauri**, **UAVision CVPR'19 Oral**.
- 2021 RHNAS: Realizable Hardware and Neural Architecture Search,  
**Yash Akhauri\***, Adithya Niranjana\*, J Pablo Muñoz, Suvadeep Banerjee, Abhijit Davare, Pasquale Cocchini, Anton A Sorokin, Ravi Iyer, Nilesh Jain, .
- 2021 BootstrapNAS: Automated Generation Of Super-Networks From Pre-Trained Models For Neural Architecture Search,  
J. Pablo Muñoz, Nikolay Lyalyushkin, Daniel Cummings, Anastasia Senina, Chaunté W Lacewell,  
**Yash Akhauri**, Alexander Kozlov, Nilesh Jain, Anthony Sarah, .

## Patents

- 2023 Apparatuses, methods and systems for instructions for structured-sparse tile matrix FMA.
- 2022 System for universal hardware-neural network architecture search (co-design).
- 2022 Two-stage decompression pipeline for non-uniform quantized neural network inference on reconfigurable hardware.
- 2022 Apparatus, articles of manufacture, and methods for composable machine learning compute nodes.
- 2022 Methods and apparatus to perform weight and activation compression and decompression.

---

## Experience

- Aug'23 – **Student Researcher**, *Google Research* *New York, USA.*  
Present ○ **Large Language Model Parallelization:** Building out an analytical simulation tool for deploying transformer model training/inference on n-dimensional TPU topologies. Currently working on improving the collective insertion strategies to handle arbitrarily partitioned tensors as well as pipelining operations and collectives to improve simulation accuracy.
- May'23 – **Research Intern**, *Google Research* *California, USA.*  
Aug'23 ○ **Large Language Model Parallelization:** Set up a fully unconstrained, customizable computational graph for transformers, with support for arbitrary partitioning strategies on n-dimensional server topologies. Enabled analytical simulation of computational graph with simple roofline performance modelling techniques, and integrated a hyper-parameter optimizer with the simulation framework demonstrate a potential order of magnitude improvement in latency with novel parallelization strategies.
- May'20 – **Research Scientist**, *Intel Labs* *Bangalore, India.*  
June'22 ○ **Dynamic Inference Optimization:** Worked on a closed-loop framework to dynamically optimize cache, memory bandwidth and core allocation. Investigated dynamic depth classifiers to maximize inference performance at minimal model switching cost.  
○ **DLRM Optimization:** Formulated a NAS strategy for Deep Learning Recommendation Models (DLRM) and studying static cache and memory bandwidth allocation along with model switching with pareto optimal DLRM models.  
○ **Zero Shot NAS:** Proposed a framework to represent Neural Architecture Ranking algorithms as genetic programs. Utilized evolutionary search on genetic programs to discover SoTA Zero Shot Neural Architecture Ranking programs.  
○ **Sparse Acceleration:** Enabled pruning of image classification and transformer models and its software acceleration on next generation Intel Xeon CPUs.  
○ **AutoML:** Enabled efficient and realizable co-design of configurable hardware accelerators with arbitrary neural network search spaces.  
○ **Neural Network Compression:** Proposed a clustering based non-uniform neural network weight quantization scheme to maximize accuracy and amortize memory bandwidth requirement on CPUs and a LUT based multi-stage decompression framework for FPGAs.
- Aug'19 – **Visiting Scholar**, *Xilinx Research* *Dublin, Ireland.*  
May'20 ○ **LogicNets:** Developed the LogicNets library, explored extremely sparse and quantized MLPs and convolutional networks with PyTorch. Developed a Verilog code generator to convert MLPs in PyTorch to Verilog netlist for FPGA synthesis.  
○ **FPGA4HEP-Brevitas:** Developed a library to demonstrate Brevitas quantization library on the Jet Classification and Regression task to CERN.
- Jan'19 – **Research Intern**, *Uranium* *France (Remote).*  
Jul'19 ○ **Semantic Segmentation:** Utilized DeepLabV3 to enable semantic segmentation of facial features.
- Jun'18 – **Undergraduate Researcher**, *Wolfram Summer School* *Massachusetts.*  
Jul'18 ○ **Neural Network Quantization:** Conducted research on neural network quantization and implemented a custom quantized neural network library from scratch in the Wolfram Language.

---

## Talks & Awards

Intel Labs High-5 Patent Award	
Intel Labs Invention Disclosure Award	
Exceptional Reviewer Award	AIPLANS@NeurIPS
Stamatis Vassiliadis Award (Best Paper Award)	FPL'20
Intel Nervana Early Innovators Grant	\$5000
Intel CVPR Travel Grant	\$3000
Wolfram Student Aid (Full Scholarship)	\$2400
KVPY Fellowship by Dept of Science and Tech., Govt. of India	
HadaNets: On Quantization	Intel Demo Booth CVPR'19
On Neural Network Quantization	Champaign, IL