

# ArchE-Q: A DSP-Free Dataflow Accelerator for Quantized Neural Networks in Sensor-Aided Millimeter-Wave Edge Connectivity

Arish Sateesan<sup>\*†</sup>, Ljiljana Simić<sup>\*</sup>, and Marina Petrova<sup>‡\*</sup>

<sup>\*</sup>Institute of Networked Systems, RWTH Aachen University, Aachen, Germany

<sup>†</sup>Department of Electronic Systems, Aalborg University, Copenhagen, Denmark

<sup>‡</sup>Mobile Communications and Computing, RWTH Aachen University, Aachen, Germany  
arishs@es.aau.dk, lsi@inets.rwth-aachen.de, petrova@mcc.rwth-aachen.de

**Abstract**—Sensor-aided wireless edge applications, such as LiDAR-based beam prediction for millimeter-wave communications, demand intelligent on-device processing of high-volume sensor data. However, the computational cost of machine learning models often exceeds the tight power and resource constraints of edge hardware. While quantized neural networks (QNNs) reduce resource requirements, typical FPGA accelerators still rely on power-hungry digital signal processing (DSP) slices and incur avoidable data-movement overheads. To bridge this gap, we propose ArchE-Q, a dataflow accelerator for QNNs combined with efficient data preprocessing. Our design is fundamentally multiplier-less, utilizing: (1) an application-specific first-layer kernel that exploits binarized sensor inputs to remove multipliers; (2) the eXtended Vector Activation Unit (XVAU), fusing convolution, activation, and pooling to reduce buffering and data transfers; and (3) memory-centric buffering for efficient data reuse. Implemented on a Xilinx ZCU104 FPGA, ArchE-Q achieves 13.3% lower latency and up to 28% lower dynamic power than the FINN-R baseline, while eliminating DSP usage.

## I. INTRODUCTION

Emerging 6G and millimeter-wave (mm-wave) wireless systems increasingly rely on environment-aware intelligence to meet stringent latency and reliability requirements in highly dynamic scenarios such as vehicle-to-everything (V2X) communication [1], [2]. While mm-wave links offer ultra-high data rates, their susceptibility to blockage and rapid channel variation demands precise directional beamforming. Conventional exhaustive or hierarchical beam sweeping incurs prohibitive latency and signaling overhead [3], motivating sensing-aided beam prediction. In particular, LiDAR provides high-resolution spatial awareness of the environment, enabling machine learning (ML) models to infer optimal beam directions without explicit sweeping [4]. However, processing high-volume sensor data under strict resource and power constraints creates a substantial computational bottleneck to the wireless edge, and sub-millisecond latency requirements render cloud offloading infeasible [5]. General-purpose GPUs often fail to meet the thermal and energy constraints at the edge, demanding specialized hardware acceleration for ML [6].

Quantized neural networks (QNNs) mitigate this challenge, as they drastically reduce computation and memory footprint while retaining competitive accuracy [7]. Field-programmable

gate arrays (FPGAs) are especially well-suited for QNN acceleration due to their ability to implement deeply pipelined, application-specific datapaths [8], [9]. This synergy has driven the adoption of streaming dataflow architectures, which map neural network layers into hardware pipelines that process data as it arrives, maximizing throughput while keeping intermediate feature maps on-chip. FINN-R [9] is a strong baseline for extreme quantization, providing end-to-end automation for generating dataflow accelerators. However, its general-purpose design relies heavily on DSP-based multipliers and uses separate convolution and pooling stages, leading to avoidable data duplication, unnecessary buffering, and energy overhead, which are critical for always-on, sensor-driven workloads.

To address these limitations, we introduce ArchE-Q (Architecture for Efficient Quantized AI), a dataflow accelerator specifically co-designed for LiDAR-aided mm-wave edge connectivity. ArchE-Q achieves superior hardware efficiency through: (1) the eXtended Vector Activation Unit (XVAU), a fused-layer module integrating convolution, batch normalization (realized as thresholding), activation, and pooling; (2) application-specific compute kernels that exploit binarized LiDAR occupancy inputs to replace multipliers with lightweight logic; and (3) an `im2col`-less block ram (BRAM)-based buffering strategy that eliminates redundant data replication and improves data reuse in fully-connected (FC) layers.

## II. PROPOSED ARCH-E-Q ARCHITECTURE

### A. System Model & Data-Aware Preprocessing

We target a LiDAR-aided mm-wave beam prediction system where a user equipment (UE) predicts the optimal base station (BS) beam index  $b^*$  from a codebook  $\mathcal{F}$ , as represented in Figure 1. The predicted beam index is calculated as  $\hat{b} = \mathcal{M}_\theta(B)$ , where  $\mathcal{M}_\theta$  is a QNN and  $B$  is a preprocessed LiDAR input. As standard convolutional neural networks (CNN) struggle with sparse LiDAR point clouds  $\mathcal{P}$ , we implement a data-aware preprocessing pipeline. Raw polar scans are converted to cartesian coordinates and projected onto a fixed-size  $H \times W$  ( $H=W=20$ ) Bird’s-Eye-View (BEV) grid based on a global region of interest. We then binarize this grid into a single-channel occupancy map  $B \in \{0,1\}^{H \times W}$ . This 1-bit representation is a deliberate co-design choice that enables multiplier-less computation in hardware.

This work has received funding by the German BMFTI in the course of the 6GEM research hub, grant number 16KISK036.

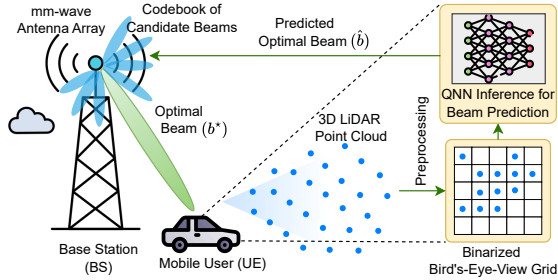


Fig. 1. System model for LiDAR-aided mm-wave beam prediction. A UE-mounted LiDAR captures a 3D point cloud, which is preprocessed into a binarized BEV grid and used by a QNN accelerator at the UE to select the optimal beam at the BS from a codebook of  $\mathcal{F} = 64$  candidate beams.

### B. DSP-Free Dataflow Accelerator

ArchE-Q is a fully pipelined streaming accelerator optimized for 2-bit QNNs. The design follows a sequential pipeline of XVAU instances followed by FC layers, where shallow FIFOs act as elastic buffers between stages to decouple processing and maintain high throughput. The architectural representation of ArchE-Q is shown in Figure 2. It employs *neuron folding* [9] to time-multiplex a limited set of compute units, sharing arithmetic resources and weight memories to balance resource usage and throughput.

*eXtended Vector Activation Unit:* XVAU fuses convolution, thresholding (batch-norm folded into thresholds), activation, and max-pooling into a single streaming module, reducing intermediate buffering and latency. Furthermore, the XVAU implements an `im2col-less` max-pooling, where convolution outputs are written to on-chip BRAM, and a controller performs pooled window reads, eliminating redundant sliding-window restructuring and reducing on-chip data traffic.

*DSP-Free Compute Kernels:* To eliminate power-hungry DSP slices, ArchE-Q uses application-specific processing elements (PE) that exploit the low-precision data format. A vector arithmetic and threshold unit (VATU) within the XVAU includes the application-specific PEs and their associated weight and threshold memories. In the first convolution layer, binarized inputs allow select-accumulate (SAC) units to replace DSP-based multiply-accumulate arrays by selecting (or zeroing) the corresponding ternary weight via lightweight control logic. For subsequent layers with 2-bit activations, conditional-add-accumulate (CAA) units replace low-bit multiplications using conditional addition logic built from XOR gates and adders, reducing lookup table (LUT) cost relative to generic multipliers while preserving throughput.

*Efficient Buffering Strategy:* FC layers replace standard streaming FIFOs with BRAM-based buffering to store flattened vectors and enable data reuse across computation cycles without re-streaming from preceding layers, addressing the memory-bound nature of FC computations.

## III. RESULTS & CONCLUSIONS

We evaluate ArchE-Q using the DeepSense 6G dataset (scenario 8) [10] and a lightweight 2-bit quantized CNN tailored for LiDAR occupancy grids. The model, having a size of 5.2 KB, is trained offline using Brevitas [11] and deployed

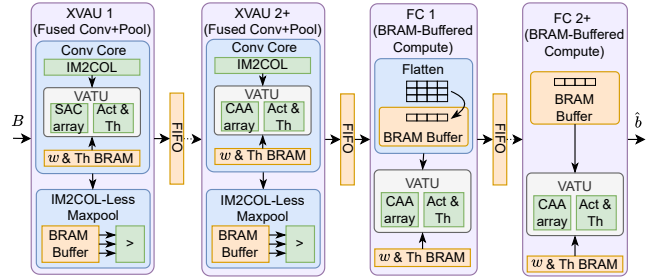


Fig. 2. Architectural representation of ArchE-Q layers. The design features the fused XVAU for convolution/pooling and custom DSP-free computation kernels. Key notations:  $w$  - weights,  $Th$  - thresholds,  $Act$  - activations.

on a Xilinx ZCU104 FPGA using Vivado 2022.2. We compare the design against a FINN-R baseline generated from the same model. Both designs target 100 MHz for a fair comparison.

*Hardware Efficiency:* As summarized in Table I, ArchE-Q shows superior efficiency across key metrics. The notation FINN-R- $X$  represents the FINN-R model with an MVAU width of  $X$ , and ArchE-Q- $X$  represents the ArchE-Q model with  $X$  number of PEs in the first two XVAUs layers, with accuracy remaining unchanged. The PEs in the XVAU-3, FC-1 and FC-2 layers are fixed as 2, 32, and 16. ArchE-Q eliminates DSP usage entirely, which drives the reduction in dynamic power consumption: 13% lower for ArchE-Q-8 and 28% lower for ArchE-Q-4 compared to FINN-R-4. Beyond power, the architectural fusion in the XVAU reduces logic, achieving a 33% reduction in LUT for the ArchE-Q-4 compared to FINN-R-2. Furthermore, ArchE-Q-8 achieves 13% lower end-to-end latency with 11% lower LUT usage than FINN-R-4.

*Accuracy & System Implications:* The 2-bit QNN achieves a top-1 prediction accuracy of 48.5% and top-5 accuracy of 92.0%. While there is a noticeable drop in accuracy compared to a full-precision model (top-1: 57.5%, top-5: 95.6%) [4], the quantized model retains strong predictive performance with a 96.6% reduction in the memory footprint. More importantly, the system-level performance of ArchE-Q compensates for this quantization loss. With a throughput of  $\approx 7300$  fps, ArchE-Q exceeds typical LiDAR data rates (30 fps) by  $\approx 240\times$ , allowing the UE to have ample time to perform a rapid, targeted beam sweep of the Top-3 or top-5 predicted candidates to mitigate quantization loss without violating latency constraints.

*Conclusions:* This work shows that application-specific, DSP-free dataflow architectures can outperform generic QNN accelerators and address the computational bottleneck of sensor-driven mm-wave edge connectivity. By co-optimizing preprocessing, quantization, and datapath design, ArchE-Q reduces latency, power, and resource usage, supporting efficient edge intelligence for next-generation wireless systems.

TABLE I  
HARDWARE EVALUATION RESULTS

Model	Latency (ms)	LUT	DSP	BRAM	TP (fps)	Power (mW)
FINN-R-4	0.158	3531	39	4	6329	54
FINN-R-2	0.299	3528	20	4	3344	49
ArchE-Q-8	0.137	3155	0	5.5	7299	47
ArchE-Q-4	0.179	2349	0	5.5	5586	39

## REFERENCES

- [1] Y. Mao, C. You, J. Zhang, K. Huang, and K. B. Letaief, "A survey on mobile edge computing: The communication perspective," *IEEE Commun. Surveys Tuts.*, vol. 19, no. 4, pp. 2322–2358, 2017.
- [2] R. Meneguette, R. De Grande, J. Ueyama, G. P. R. Filho, and E. Madeira, "Vehicular edge computing: Architecture, resource management, security, and challenges," *ACM Comput. Surv.*, vol. 55, no. 1, pp. 1–46, 2021.
- [3] Q. Xue, C. Ji, S. Ma, J. Guo, Y. Xu, Q. Chen, and W. Zhang, "A survey of beam management for mmWave and THz communications towards 6G," *IEEE Commun. Surveys Tuts.*, vol. 26, no. 3, pp. 1520–1559, 2024.
- [4] S. Jiang, G. Charan, and A. Alkhateeb, "LiDAR aided future beam prediction in real-world millimeter wave V2I communications," *IEEE Wireless Commun. Lett.*, vol. 12, no. 2, pp. 212–216, 2022.
- [5] Z. Zhou, X. Chen, E. Li, L. Zeng, K. Luo, and J. Zhang, "Edge intelligence: Paving the last mile of artificial intelligence with edge computing," *Proc. IEEE*, vol. 107, no. 8, pp. 1738–1762, 2019.
- [6] J. Chen and X. Ran, "Deep learning with edge computing: A review," *Proc. IEEE*, vol. 107, no. 8, pp. 1655–1674, 2019.
- [7] A. Sateesan and L. Simić, "FPGA-powered environment awareness via quantized neural networks for LiDAR-aided mm-wave beam prediction," in *Proc. IEEE Int. Conf. Mach. Learn. Commun. Netw. (ICMLCN)*, 2025, pp. 1–6.
- [8] W. Chen, H. Qiu, J. Zhuang, C. Zhang, Y. Hu, Q. Lu, T. Wang, Y. Shi, M. Huang, and X. Xu, "Quantization of deep neural networks for accurate edge computing," *ACM J. Emerg. Technol. Comput. Syst.*, vol. 17, no. 4, pp. 1–11, 2021.
- [9] M. Blott, T. B. Preußer, N. J. Fraser, G. Gambardella, K. O'Brien, Y. Umuroglu, M. Leeser, and K. Vissers, "FINN-R: An end-to-end deep-learning framework for fast exploration of quantized neural networks," *ACM Trans. Reconfigurable Technol. Syst.*, vol. 11, no. 3, pp. 1–23, 2018.
- [10] A. Alkhateeb, G. Charan, T. Osman, A. Hredzak, J. Morais, U. Demirhan, and N. Srinivas, "DeepSense 6G: A large-scale real-world multi-modal sensing and communication dataset," *IEEE Commun. Mag.*, vol. 61, no. 9, pp. 122–128, 2023.
- [11] A. Pappalardo, "Xilinx/brevitas," [Online]. Available: <https://doi.org/10.5281/zenodo.3333552>. Accessed: Jan. 4, 2026, 2023.