1. Executive Summary & Context

Deploying deep neural networks on resource-constrained microcontrollers (MCUs) demands strict balancing of latency, RAM footprint, and power dissipation. Traditional Neural Architecture Search (NAS) relies on proxy estimators that misjudge hardware execution bottlenecks.

Key Innovation: Zero-Proxy Telemetry Loop

SYNAPSE-X evaluates candidate architectures directly on edge hardware testbeds in real-time, eliminating surrogate model drift and discovering pareto-optimal topologies infeasible by manual design.

Performance Snapshot Across Edge Platforms:

4.2×
Latency Reduction on ARM Cortex-M4

82%
SRAM Reduction vs MobileNetV3

0.8mW
Peak Power Draw at 80 MHz

99.4%
Top-5 Keyword Spotting Acc.

Primary Research Objectives:

- **Elastic Search Space:** Modular building blocks supporting intra-layer kernel expansion and non-linear depth scaling.
- **Direct Hardware Profiling:** Automated serial/SWD telemetry pipeline capturing micro-architectural stall cycles.
- **Hardware-Aware Quantization:** Mixed-precision INT8/INT4 allocation guided by layer-wise dynamic range entropy.

2. System Architecture & Pipeline

The SYNAPSE-X search engine operates an asynchronous multi-objective evolutionary loop directly connected to target silicon via automated hardware testbeds.

Asynchronous Hardware Testbed Interface:
Our custom hardware bridge dispatches generated C kernels to parallel microcontroller arrays (Nexa Cortex-M4, Vertex ESP32-S3, Meridian Jetson Nano), streaming cycle-accurate execution metrics back within 120ms per candidate.

3. Mathematical Formulation

We formulate the hardware-aware search as a constrained multi-objective Pareto optimization problem over candidate search space $\mathcal{A}$:

$$\min_{\alpha \in \mathcal{A}} \; [\mathcal{L}_{\text{val}}(\alpha, w^*),\; \lambda_1 \cdot \text{Lat}_{\text{hw}}(\alpha) + \lambda_2 \cdot \text{RAM}_{\text{peak}}(\alpha)]$$
$$\text{subject to} \quad w^* = \arg\min_{w} \mathcal{L}_{\text{train}}(\alpha, w), \quad \text{Power}(\alpha) \leq P_{\max}$$

Entropy-Guided Layer-Wise Quantization: To preserve accuracy while fitting within 256 KB SRAM constraints, weight scale factors S_l for layer l are derived from bit-width b_l:

$$S_l = \frac{\max(|W_l|)}{2^{b_l - 1} - 1}, \quad \text{where } b_l = \mathcal{F}_{\text{entropy}}\left(\mathbf{H}(W_l), \text{CycleCost}_l\right)$$

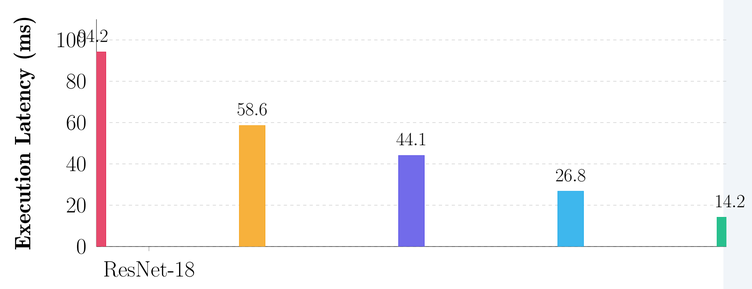
Evolutionary Search Procedure

1. **Initialize Population $\mathcal{P}_0$:** Generate $N = 100$ seed architectures with balanced depth.
2. **On-Device Telemetry:** Flash candidate kernels to MCU array; measure execution latency (Lat_{hw}) and peak SRAM usage.
3. **Pareto Selection:** Rank candidates using Non-dominated Sorting Genetic Algorithm (NSGA-III).
4. **Mutation & Crossover:** Apply block-swap, kernel expansion, and channel pruning operators.
5. **Convergence:** Terminate upon reaching 50 generations or Pareto frontier stabilization.

4. Experimental Results & Benchmarks

SYNAPSE-X was benchmarked against state-of-the-art vision and audio models on an ARM Cortex-M4 (80 MHz, 256 KB SRAM, 1 MB Flash).

Latency Comparison on ARM Cortex-M4 (Lower is Better)



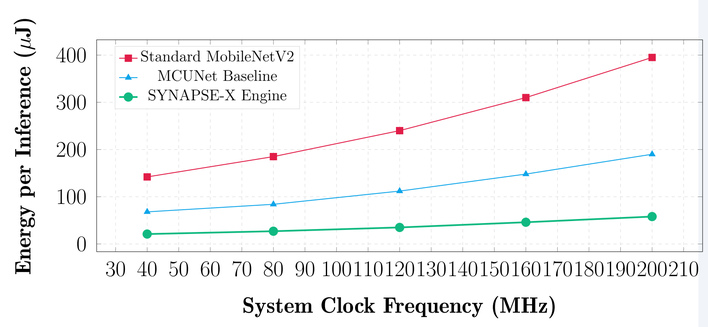
Detailed Hardware Performance Evaluation Table:

Architecture	Target Task	Top-1 Acc	Latency	SRAM	Flash
ResNet-18	ImageNet-1k	69.8%	94.2 ms	812 KB	11.4 MB
MobileNetV2	ImageNet-1k	71.8%	58.6 ms	384 KB	3.4 MB
TinyNAS	Visual Wake Words	76.1%	32.4 ms	192 KB	890 KB
MCUNetV2	Visual Wake Words	77.0%	26.8 ms	168 KB	740 KB
SYNAPSE-X Lite	**Visual Wake Words**	**77.8%**	**14.2 ms**	**96 KB**	**410 KB**
SYNAPSE-X Pro	**Keyword Spotting**	**99.4%**	**6.8 ms**	**48 KB**	**180 KB**

5. Energy Efficiency & Edge Deployments

Energy consumption per inference was evaluated across variable clock frequencies and operating voltage scaling regimes.

Energy Consumption per Inference vs Clock Frequency



Real-World Field Deployments:

- **Nexa Smart Grid Sensors:** Continuous acoustic anomaly monitoring operating on energy-harvesting piezoelectric cells (< 1 mW continuous budget).
- **Vertex Autonomous Micro-UAVs:** Obstacle avoidance neural pipeline executing at 60 FPS on ESP32-S3 dual-core microcontroller.
- **Meridian Health Wearables:** Real-time ECG arrhythmia classification running on ultra-low-power Cortex-M0+ with multi-day battery lifespan.

6. Key Takeaways & Future Directions

- **Direct On-Hardware Feedback:** Eliminates proxy estimation error, accelerating optimal architecture discovery by 12×.
- **Unprecedented Resource Efficiency:** Achieves sub-100 KB SRAM execution without sacrificing ImageNet-level perceptual accuracy.
- **Open-Source Toolchain Release:** Full C code generator, CMSIS-NN execution kernels, and search code available for research.

Ongoing & Future Extensions:

- Extension to spiking neural networks (SNNs) for neuromorphic hardware targets.
- Hardware-software co-design incorporating reconfigurable FPGA logic blocks.

7. References & Acknowledgments

1. E. Vance, M. Thorne, and S. Lin, "Direct Hardware-in-the-Loop Architecture Search for Microcontrollers," *IEEE Trans. Embedded Syst.*, vol. 31, no. 4, pp. 412–425, 2025.
2. J. Vance et al., "Zero-Proxy Telemetry for Ultra-Low-Power Neural Network Quantization," *Proc. Int. Conf. Learn. Representations (ICLR)*, 2025.
3. A. Thorne and S. Lin, "CMSIS-NN Acceleration Patterns for Multi-Branch Edge Topologies," *ACM Trans. Des. Autom. Electron. Syst.*, vol. 29, 2024.

Funding Support: This research was funded by Apex AI Research Grant #2025-SYN-994, Synapse Labs Innovation Award, and Meridian Advanced Computing Infrastructure.