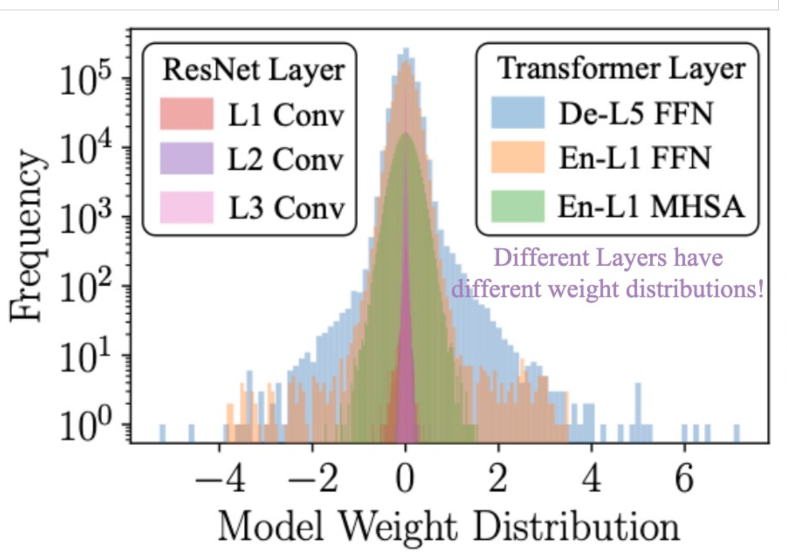
ALGORITHM-HARDWARE CO-DESIGN OF DISTRIBUTION-AWARE LOGARITHMIC-POSIT ENCODINGS FOR EFFICIENT DNN INFERENCE

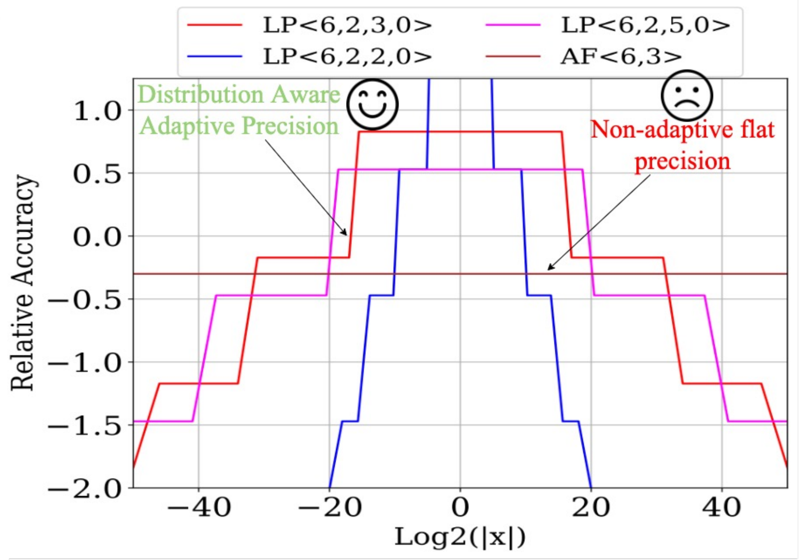
Akshat Ramachandran[1], Zishen Wan[1], Geonhwa Jeong[1], John Gustafson[2], Tushar Krishna[1]

[1]*Georgia Institute of Technology*, [2]*Arizona State University*

DESIGN AUTOMATION CONFERENCE 60 — FROM CHIPS TO SYSTEMS — LEARN TODAY, CREATE TOMORROW

① Motivation





- **Uniform Quantization:** Substantial **distributional variance** and orders of **magnitude difference** in DNN parameters causing significant quantization errors.
- **Floating-Point Techniques:** Fail to adapt to the **tapered distribution** of DNN parameters and use **flat accuracy, have increased hardware complexity.**
- **Why Posits?:** Posit-based representations outperform floats in DNN inference, offering **improved dynamic range**, **higher accuracy**, **simpler exception handling** and **tapered accuracy**. But still have **complex hardware**.
- **Logarithmic Posits:** A composite data type that blends the adaptability of posits with the hardware efficiency of LNS.

② Logarithmic Posits (LP)

$$x\langle n, es, rs, sf\rangle = (-1)^{sign} \times 2^{2^{es}\times k - sf} \times 2^{\mathbf{ulfx}}$$

❖ Parameterizations for incorporating distribution-aware properties:

✓ **Bits (n):** Identify optimal precision for a DNN layer.

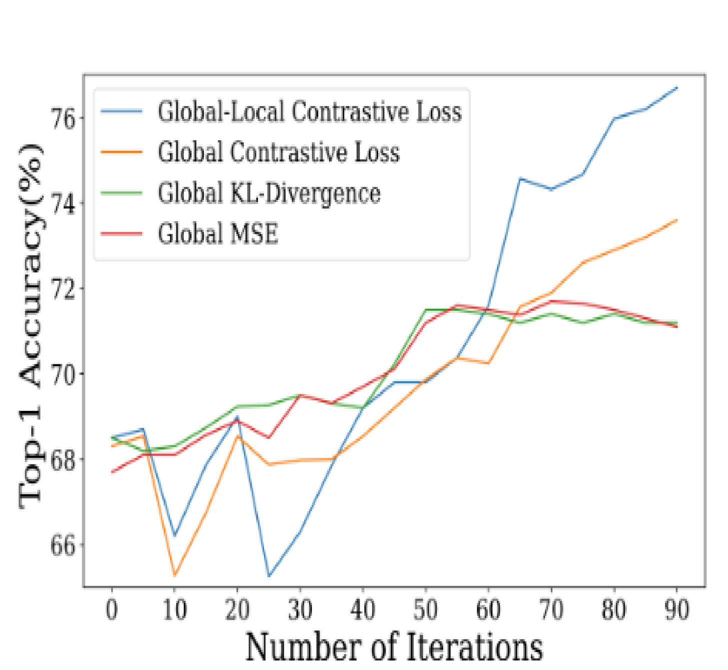
✓ **Exponent Size (es):** Controls dynamic range.

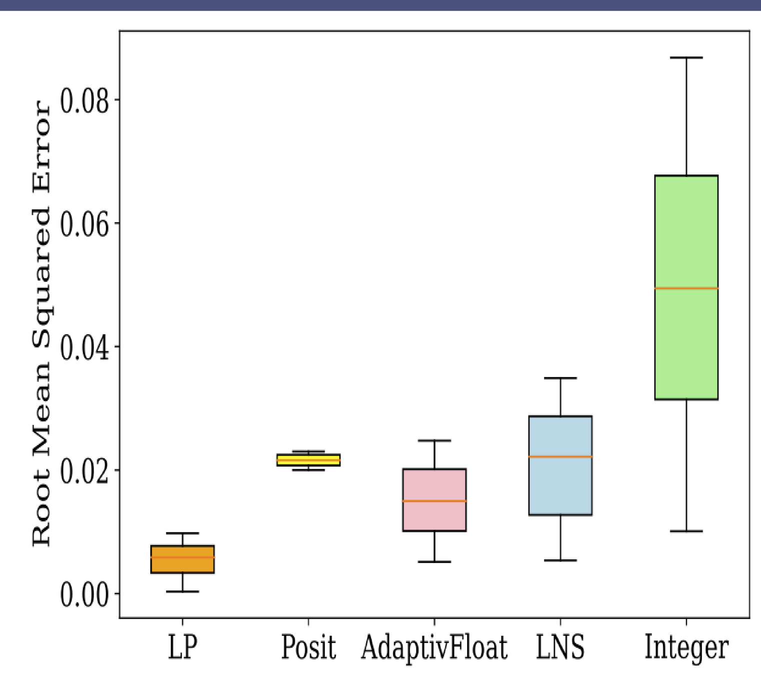
✓ **Regime Size (rs):** Controls distribution shape.

✓ **Scale Factor (sf):** Adjusts distribution position.

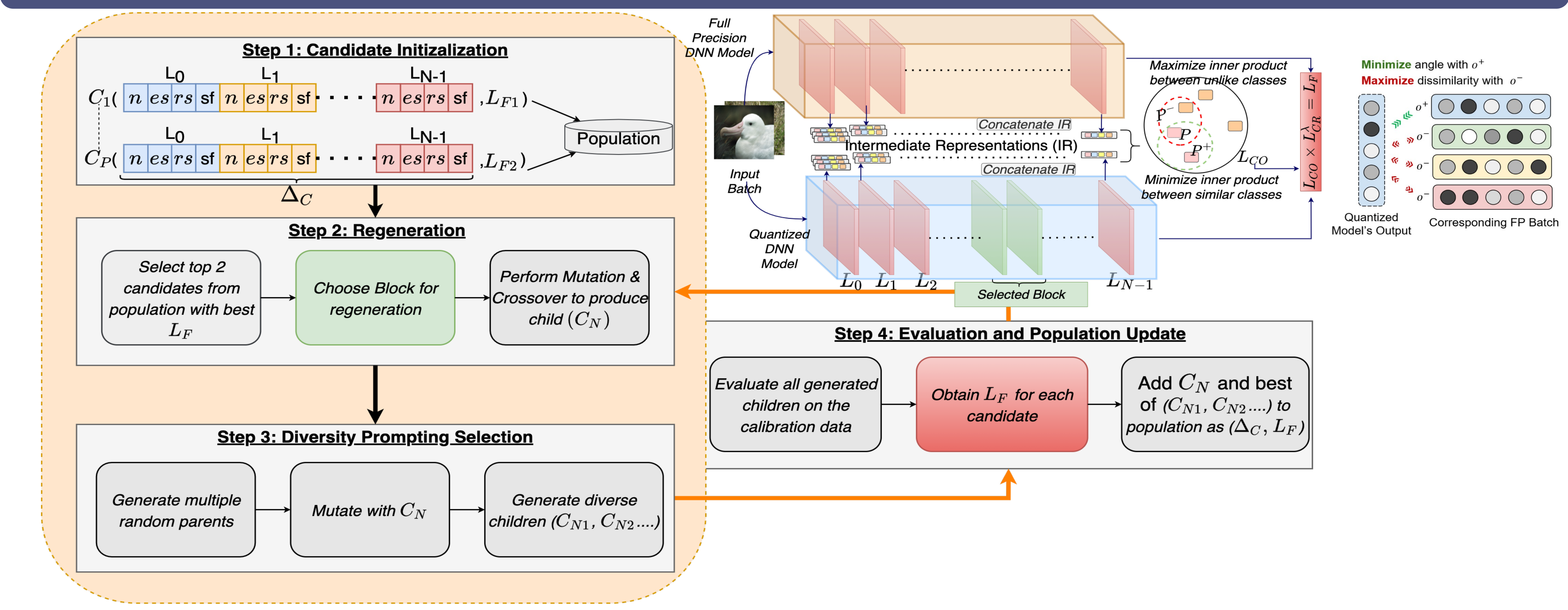
❖ Express standard fraction and exponent in the logarithmic domain as a unified fixed-point exponent of the power of two as 2^{ulfx}, where $ulfx=e+f$.

⑤ Algorithm Component Effectiveness





③ Algorithm: Genetic-Algorithm Based LP Quantization (LPQ)



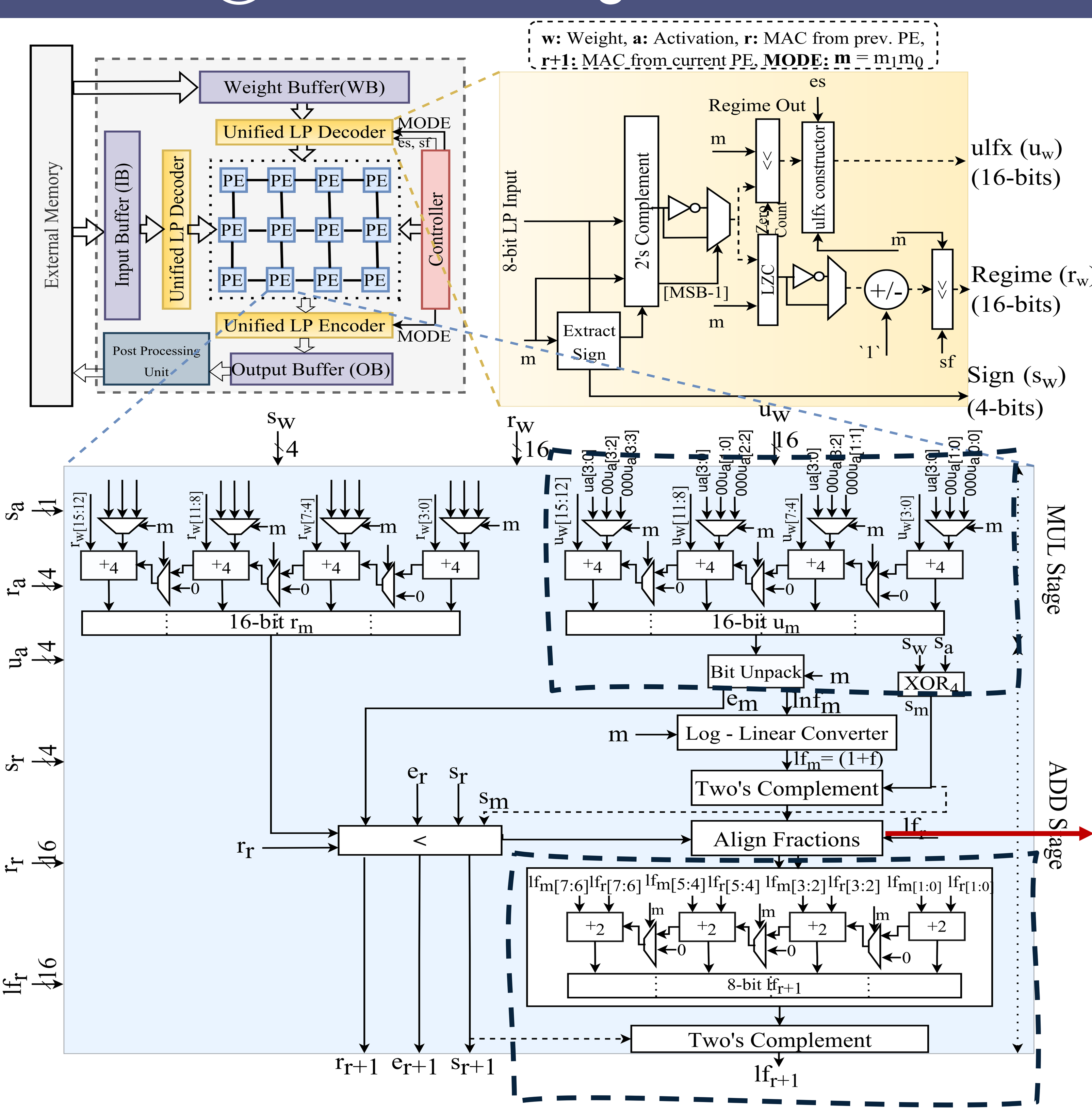
➤ **Fitness Function:**

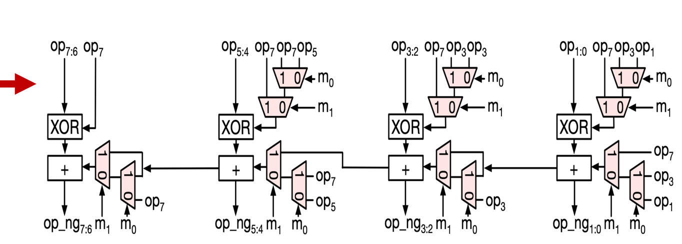
✓ A novel global-local contrastive loss, combats overfitting to calibration data and prevents premature convergence by minimizing representational divergence of intermediate layers.

✓ Also includes a compression loss that drives the optimization to identify lower bit widths.

✓ This combination of fitness function drives the genetic algorithm for layer-wise quantization.

④ Hardware: Logarithmic-Posit Accelerator (LPA)





Mixed-precision LP MAC unit made entirely of integer adder building blocks.

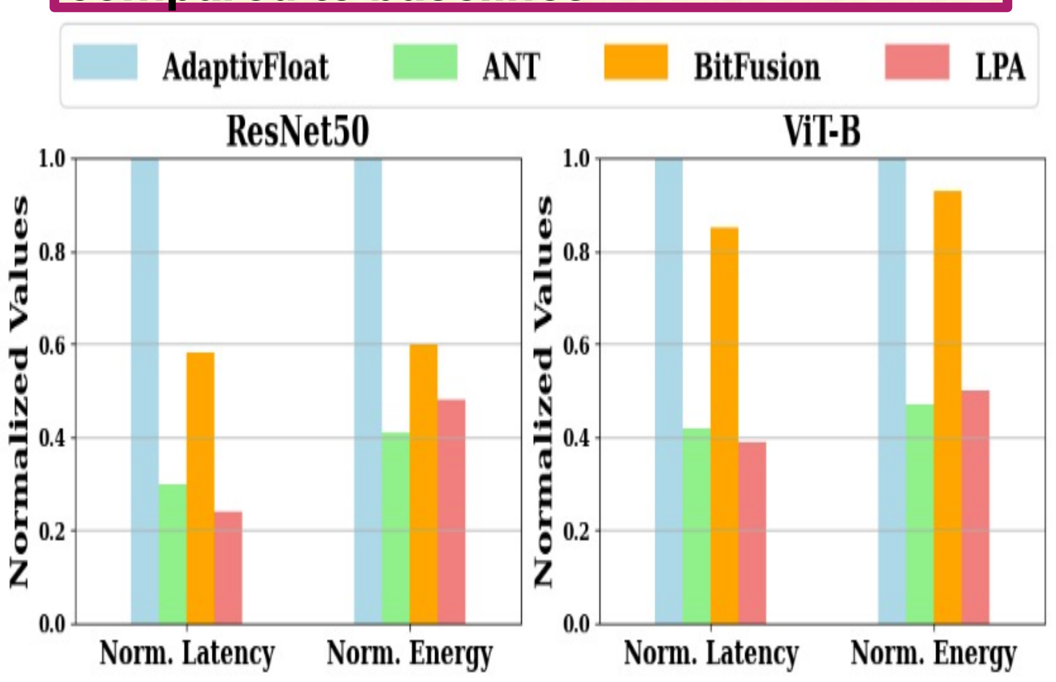
Multiplication in Log Domain expressed as 4-bit addition.

Log-Linear Converter implemented as gates.

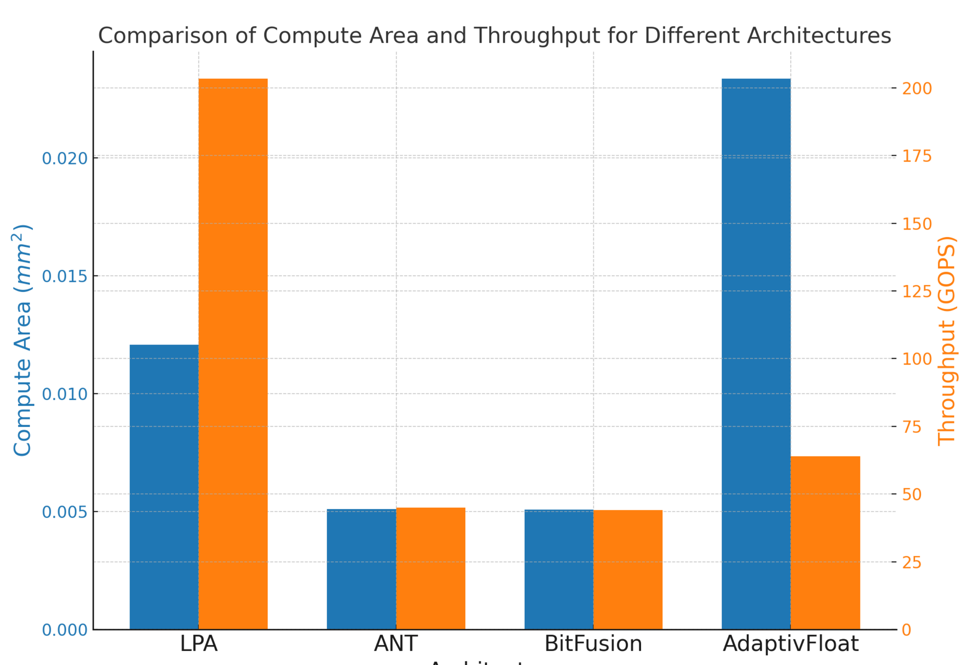
Linear Domain Addition using multiple 2-bit adders.

⑥ Co-Design Results

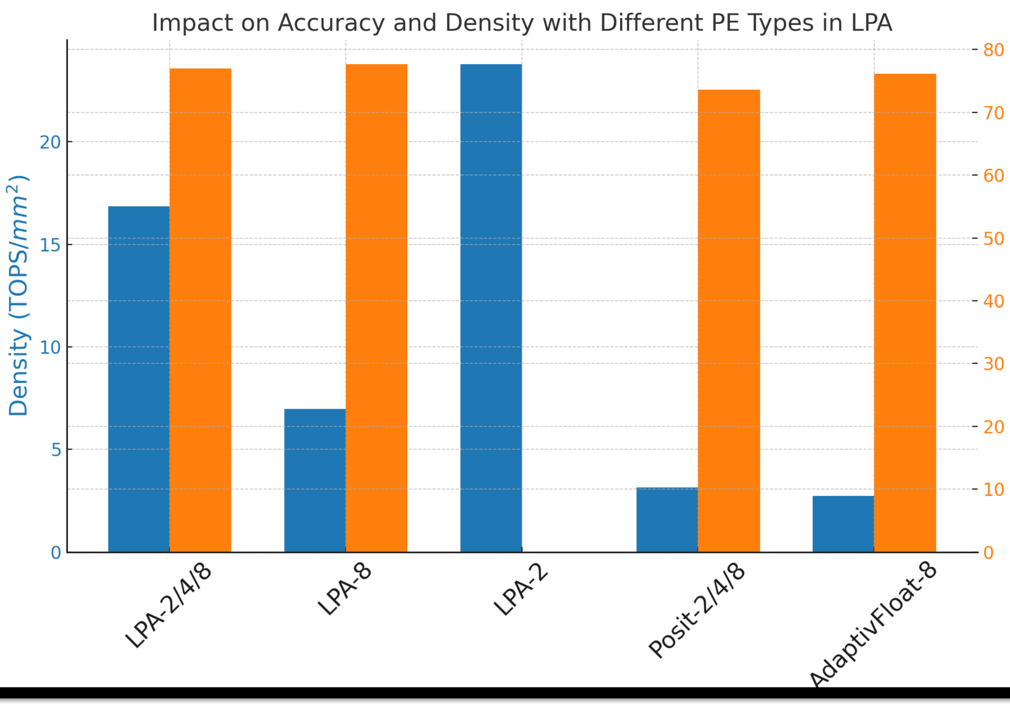
40% Lower Latency on average compared to baselines



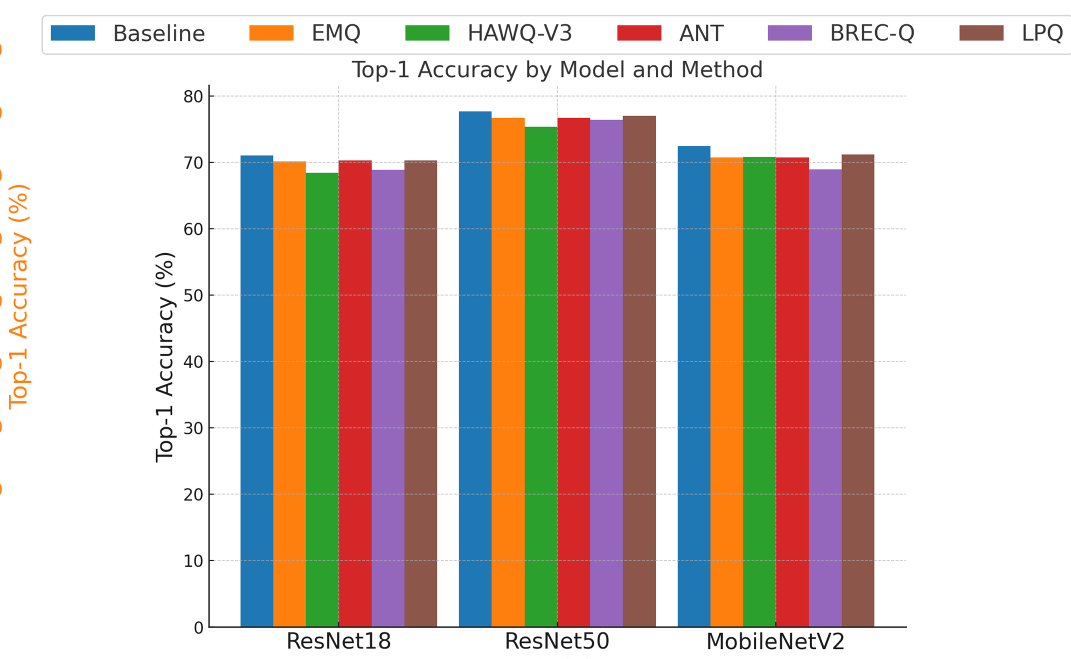
5x higher throughput with only modest area increase due to MP-support



Mixed-precision LP PEs provides highest TOPS/area with best accuracy



<1 % Accuracy Degradation after Mixed-Precision Quantization



Acknowledgements: This work was supported in part by CoCoSys, one of seven centers in JUMP 2.0, a Semiconductor Research Corporation (SRC) program sponsored by DARPA.

Semiconductor Research Corporation · Georgia Tech · SYNERGY

Paper