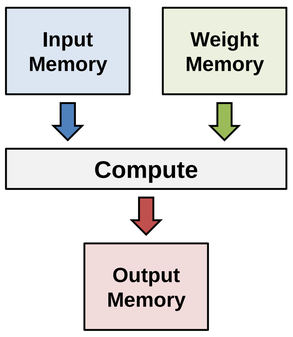
RAELLA: Reforming the Arithmetic for Efficient, Low-Resolution, and Low-Loss Analog PIM: No Retraining Required!

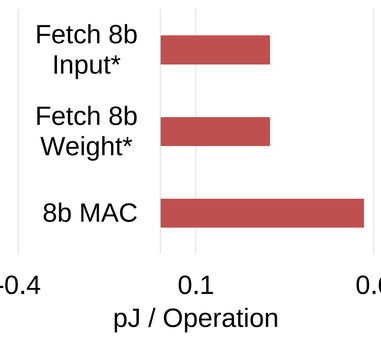
Tanner Andrulis, Joel S. Emer, Vivienne Sze

Motivation

Conventional Deep Neural Network (DNN) Accelerator

High energy for data movement & compute

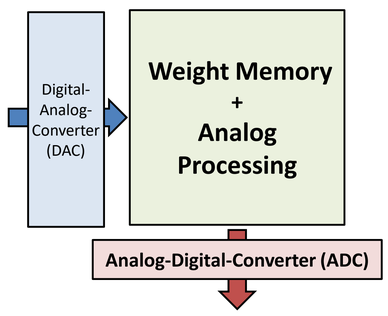


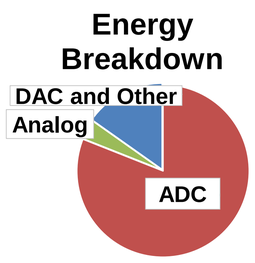


Processing in Memory (PIM) DNN Accelerator

Less data movement, low-power compute

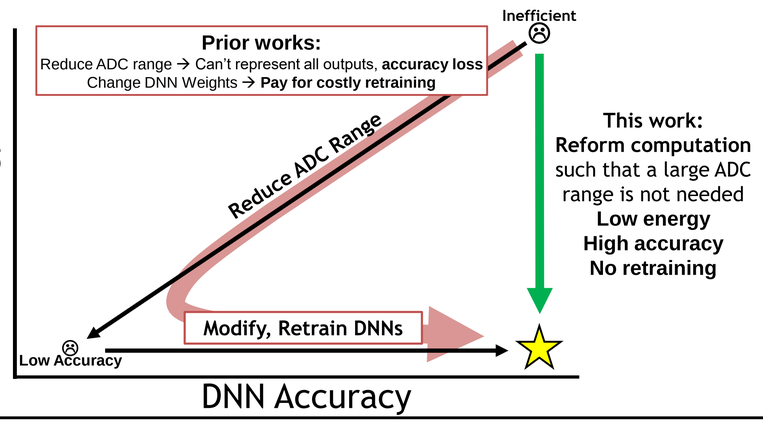
High energy for analog-digital conversion (ADC)



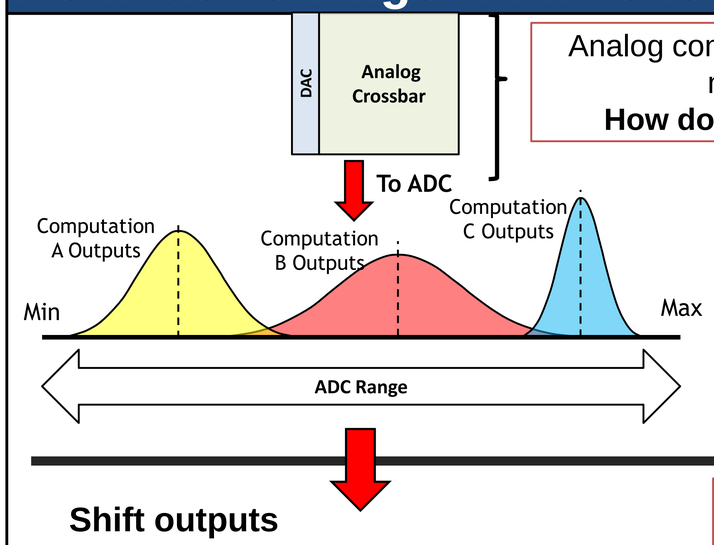


ADC Energy Increases with ADC Range

Reduce ADC range → Save energy

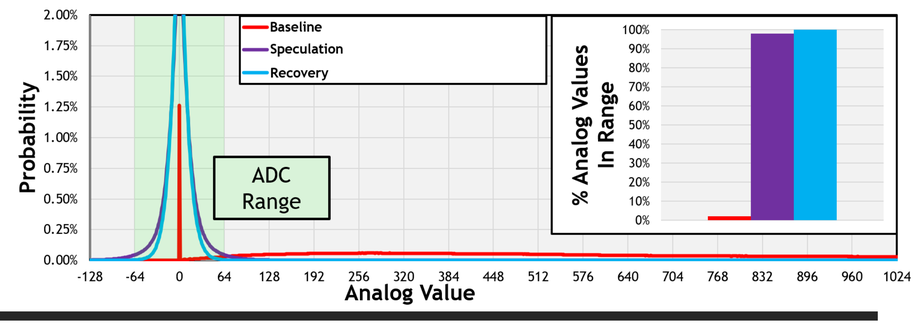


Methods: Change Arithmetic to Reduce ADC Range

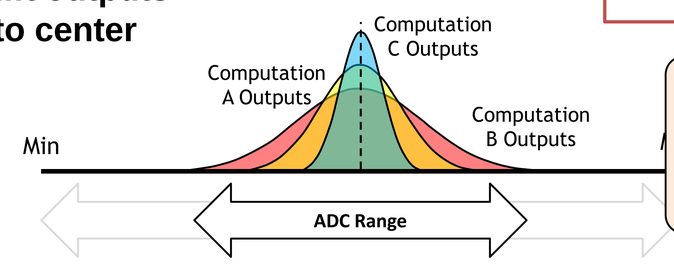


Analog computations produce distributions of outputs. ADC range must include all outputs to maintain accuracy

How do we change computations to reduce ADC range?

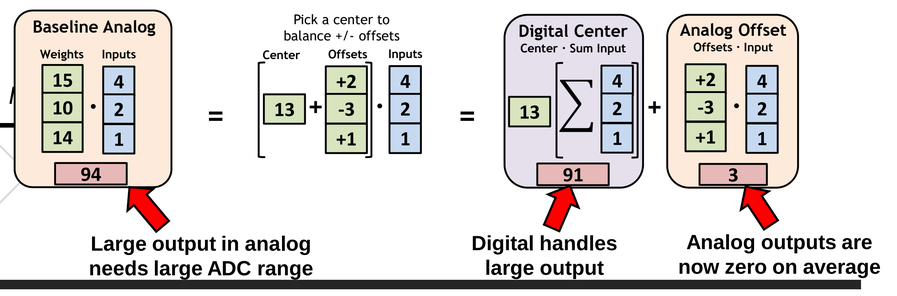


Shift outputs to center

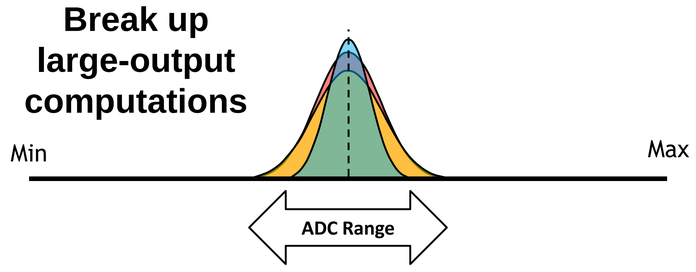


Center + Offset Encoding

Shift DNN weights to make outputs zero-average

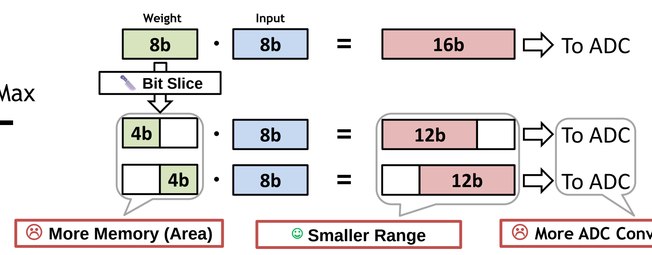


Break up large-output computations



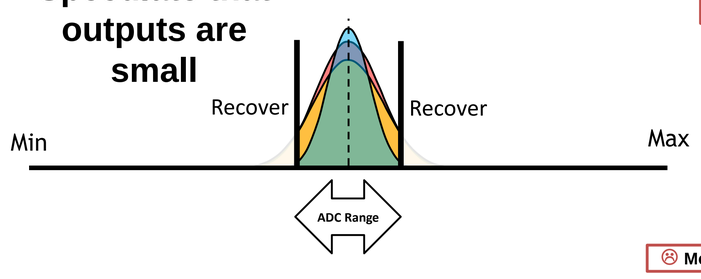
Adaptive Weight Slicing

Break up large-range computations into smaller pieces



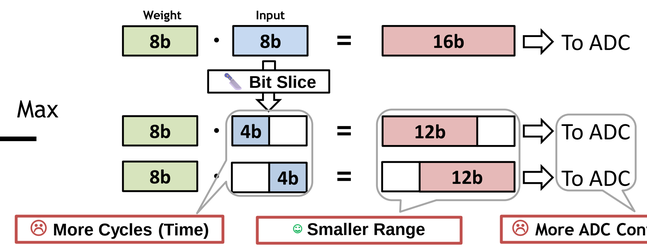
DNN weights known ahead of time → Slice weights in advance, **adaptive**

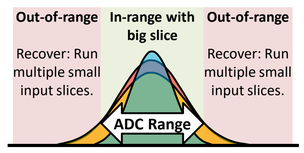
Speculate that outputs are small



Dynamic Input Slicing

Speculate that outputs will be small, recover large outputs by breaking up computations





DNN inputs not known ahead of time → Speculate **dynamically**

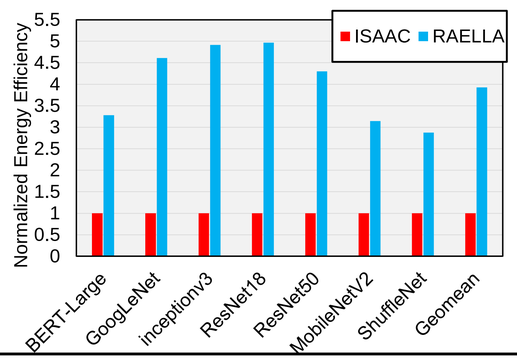
Results: Impacts of a lower-range ADC

Compare to popular PIM DNN accelerator ISAAC

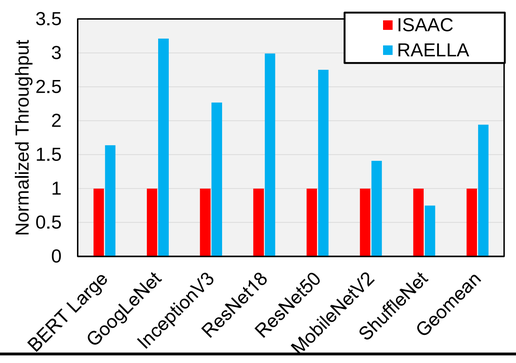
Less ADC energy → **Up to 5x higher energy efficiency,**
Less ADC area, spend more chip area on compute → **3x higher throughput**

Partition analog/digital compute + adaptive & dynamic strategies → **Protect from noise-induced accuracy loss**

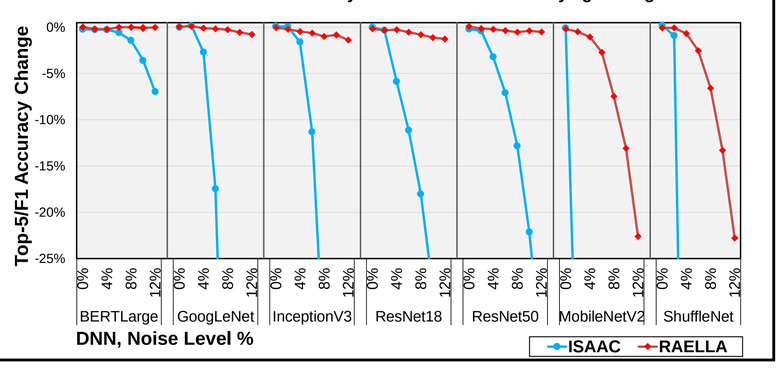
ISAAC Versus RAELLA Energy Efficiency on Various DNNs



ISAAC Versus RAELLA Throughput on Various DNNs



ISAAC Versus RAELLA Accuracy on Various DNNs for Varying Analog Noise



Published at ISCA 2023

Article

https://doi.org/10.1145/3579371.3589062

Artifact

https://github.com/mit-emze/raella

This work was funded in part by Ericsson, TSMC, the MIT AI Hardware Program, and MIT Quest.