Quality Programmable Vector Processors for Approximate Computing

Swagath Venkataramani[1], Vinay Chippa[1], Srimat Chakradhar[2], Kaushik Roy[1], Anand Raghunathan[1]

[1]Integrated Systems Laboratory, School of ECE, Purdue University

[2]Systems Architecture Department, NEC Laboratories America

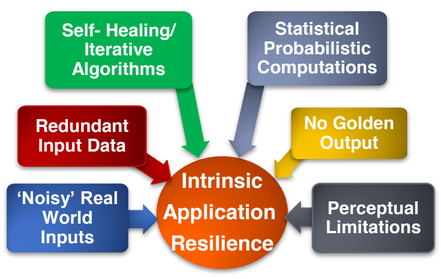
Approximate Computing - Motivation

Intrinsic Application Resilience

Ability of applications to produce outputs of acceptable quality despite underlying computations executed *imprecisely*

Recognition | Mining | Synthesis | Search | Vision | Video

Sources of Resilience



- Notion of correctness is relaxed

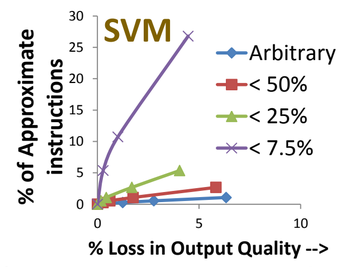
Good enough answers !!!

Need programmable platforms for approximate computing!

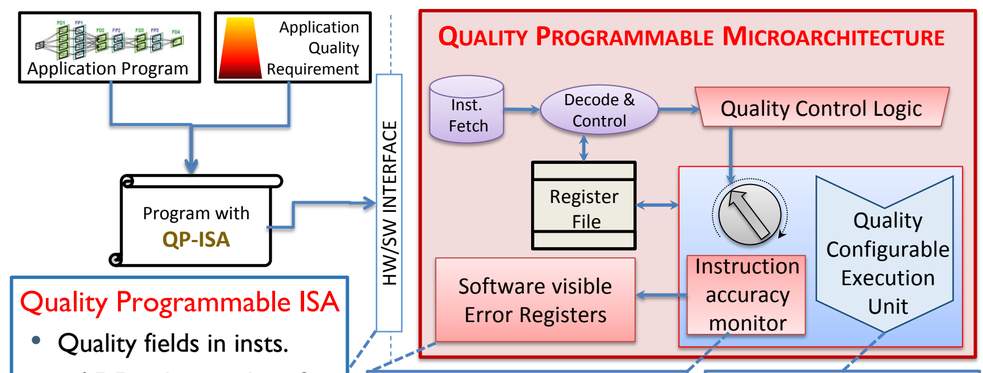
Quality Programmable Processors

An abstract model for programmable approximate processors

- **Quality Programmability**: Ability to specify the desired accuracy requirement to HW
- Notion of quality explicitly built into the instruction set



- Constraining errors enables 25-100X more approximate instructions compared to allowing arbitrary errors



Quality Programmable ISA

- Quality fields in insts.

qpADD dest, op1, op2, MAG, 1%

Quality-programmable add — Error magnitude < 1%

- Purely based on instruction semantics

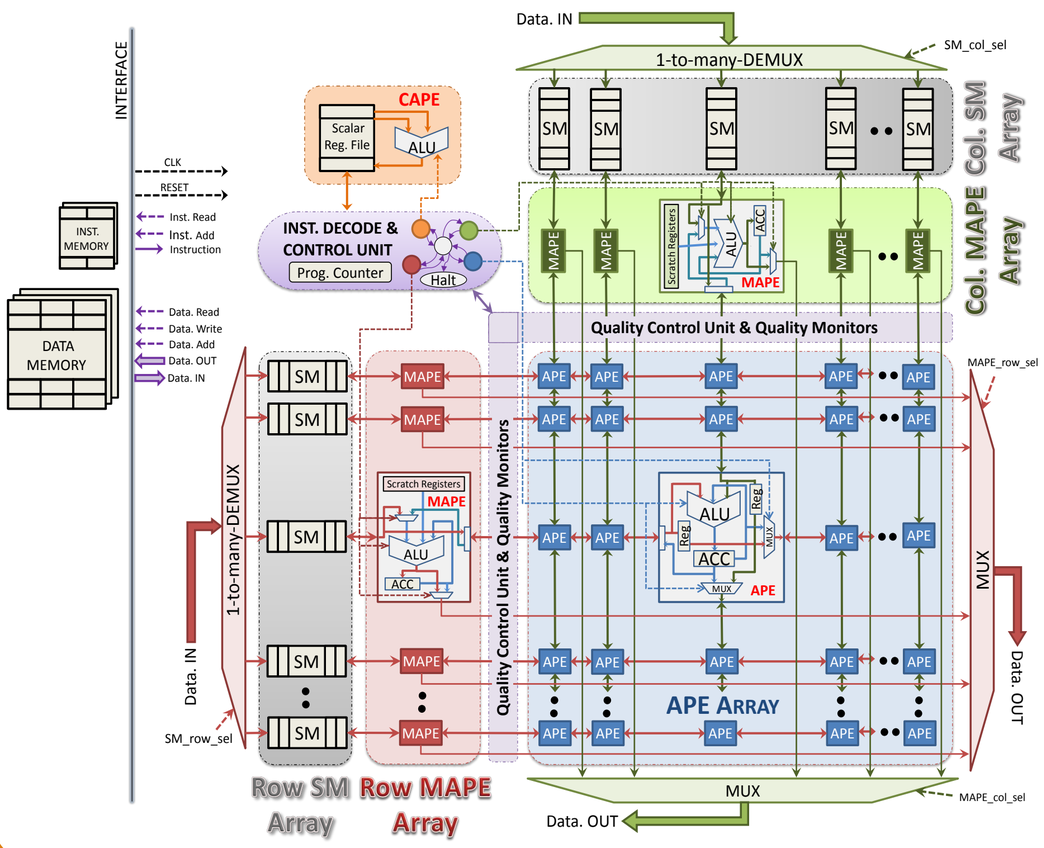
Feedback about actual error - used by software to determine quality levels of future instructions

Capable of executing instructions with different quality levels

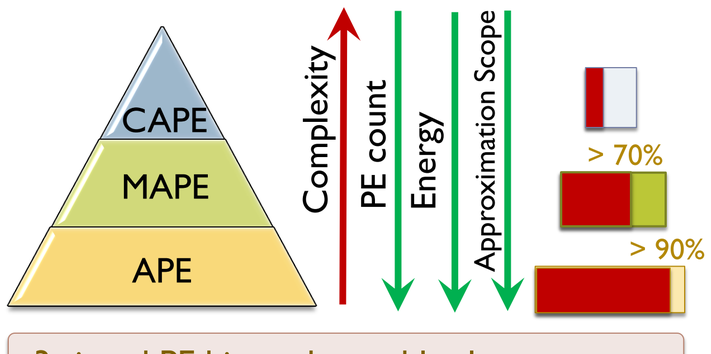
- Micro-architecture ***guarantees*** instruction level quality bounds

QUORA: Quality Programmable 1D/2D Vector Processor

Block Diagram of QUORA Micro-architecture



Processing Element Hierarchy



3-tiered PE hierarchy enables larger energy benefits from approximate computing

Quality Programmable Instruction Set

- 47 Instructions – 9 APE, 22 MAPE, 13 CAPE, 3 SM
- Instructions extended with 2 quality fields

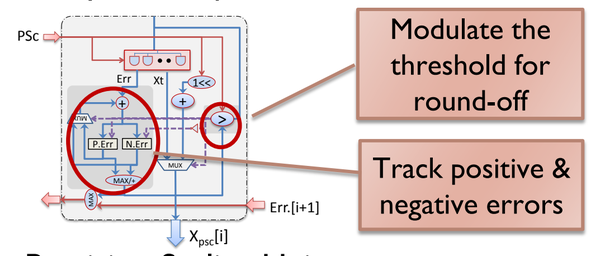
e.g., **qpMAC** R_length, R_row_enb, R_col_enb, ***R_q_type, R_q_amt***

Type of error – 3 quality metrics e.g., $MaxErr = |O_{orig} - O_{approx}|$

Amount of error

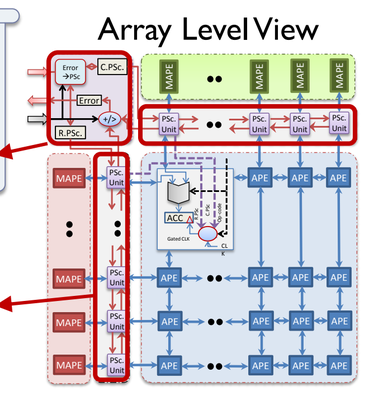
Quality Configurable Execution

- Quality specified @ vector inst. outputs
- Scale the precision of input operands
- Key idea: Compensate errors across many scalar operations



Precision Scaling Unit

Quality Control Unit
- Set precision values



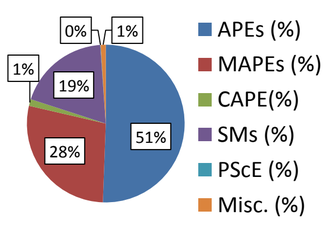
PSc. units shared across row/col <1% energy

Experimental Methodology and Results

Micro-architectural Parameters

Micro-architectural Parameters	Value
Array Dimensions	16 X 16
No. of PEs (APEs + MAPEs+ CAPE)	289 (256 + 32 + 1)
No. of SM elements	32
Depth of SM elements	64
Operating Frequency	250 MHz

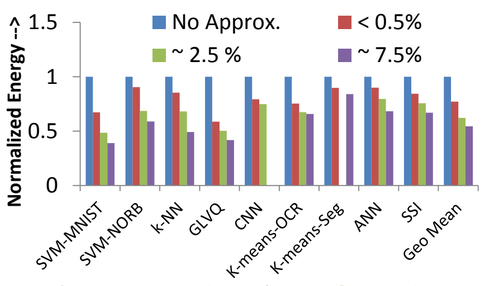
Metric	Value
Feature Size	45nm
Area	2.6 mm^2
Power	367.8 mW
Gate Count	502042



Benchmarks

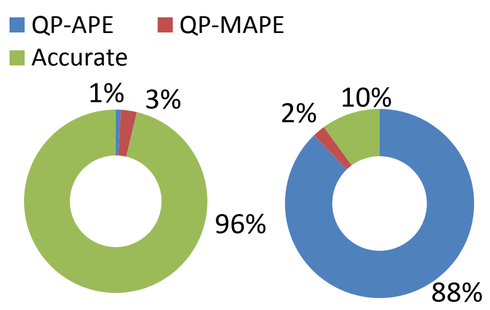
Applications	Dataset	Quality Metric
Handwritten Digit Recognition (SVM-MNIST)	MNIST	Percentage classification accuracy
Object Recognition (SVM-NORB)	NORB	
Digit Classification (CNN)	MNIST	
Eye Detection (GLVQ)	NEC labs.	
Optical Character Recognition (k-NN)	OCR digits	
Census Data Analysis (ANN)	Adult	
Document Search (SSI)	Subset of Wikipedia	No. correct in top 25 results
Image Segmentation (K-Means-Seg)	Berkeley dataset	Mean distance of clustered points from respective centroids
Optical Character Clustering (K-Means-OCR)	OCR digits	

Energy Benefits



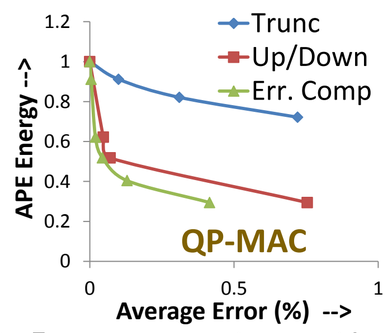
- **1.05-1.7X** savings for **NO** quality loss
- **1.18-2.1X** savings for **< 2.5%** quality loss
- **> 2.5X** savings for **< 7.5%** quality loss

QP-Instructions



- **90%** of energy in QP instructions

Precision Scaling Mechanisms



- Error compensation provides **superior** Q-E trade-off

Summary

- Intrinsic application resilience: A new dimension to optimize HW and SW
- Objective: Energy-efficient & programmable processor for approximate computing
- Quality programmable processors: Quality codified as part of the instruction set
- QUORA: Quality programmable 1D/2D vector processor
  - Quality programmable ISA and microarchitecture

Acknowledgement

National Science Foundation

NEC laboratories America