REACT: A Framework for Rapid Exploration of Approximate Computing Techniques

Mark Wyse, André Baixo, Thierry Moreau, Bill Zorn
James Bornholt, Adrian Sampson, Luis Ceze, Mark Oskin

University of Washington

Motivation

Understand current research

Investigate new techniques

Evaluate impact of existing techniques

Overview

Taxonomy

Dimensions

Conclusions

Framework

Details

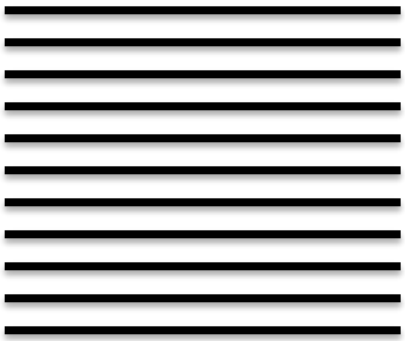
Early Results

Taxonomy

Determinism

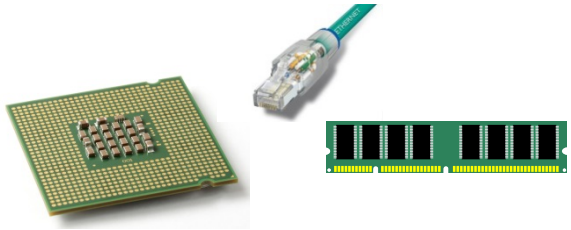
$$|P(x) - A(x)| \leq \varepsilon \; \forall \; x$$
$$\Pr(|P(x) - A(x)| > \varepsilon) < P \; \forall \; x$$

Granularity



Hardware/Software

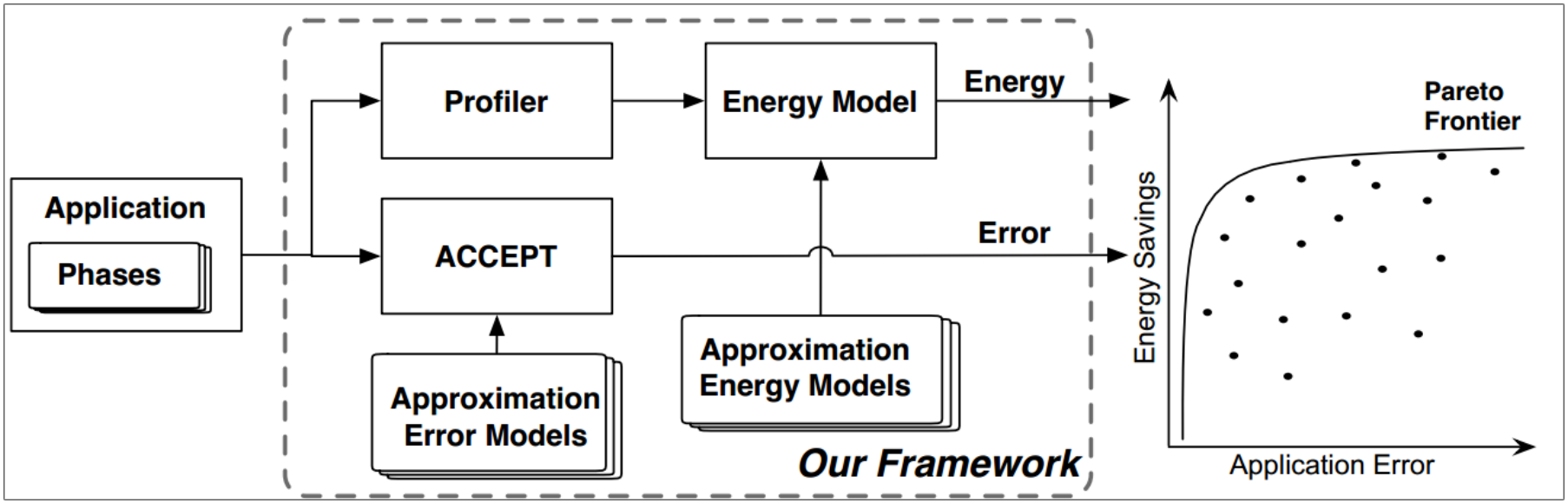
Computational Resource(s)



	Nondeterministic	Deterministic
Fine Grained	DRAM Refresh Rate SRAM Soft Error Exposure Approximate Storage (PCM) Soft Fault Tolerance Synchronization Elision Voltage Overscaling	Bit-Width Reduction Float-to-Fixed Conversion Fuzzy Memoization Hierarchical FPU Load Value Approximation Lossy Compression and Data Packing Precision Scaling ALU Reduced-Precision FPU Underdesigned Multiplier
Coarse Grained	Error F… Neural Acce… …g) 	Algorithm Selection Code Perforation Interpolated Memoization Neural Acceleration (ASIC, FPGA, GPU) Parallel Pattern Replacement Parameter Adjustment

REACT

A Framework for **R**apid **E**xploration of **A**pproximate **C**omputing **T**echniques



Application Profiler & Energy Model

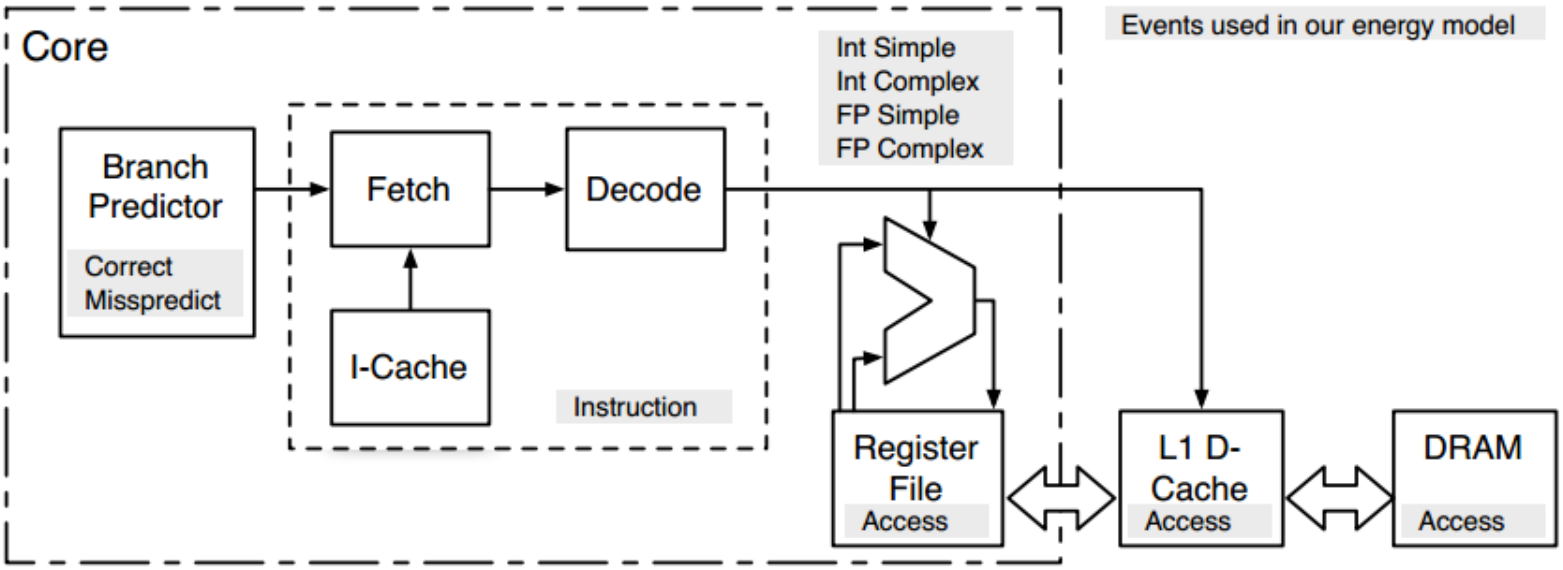
Intel Pin tool

Insn Count + Arch Events

Custom, linear model

Simple, understandable

Validated against McPAT



Error Injection

ACCEPT

Runtime error injection

Simple API

Arbitrary error models

```
int i, p;
APPROX int a;
APPROX int data[N];
a = data[i] * p;
```

Approximation Models

Load Value Approximation

Neural Acceleration

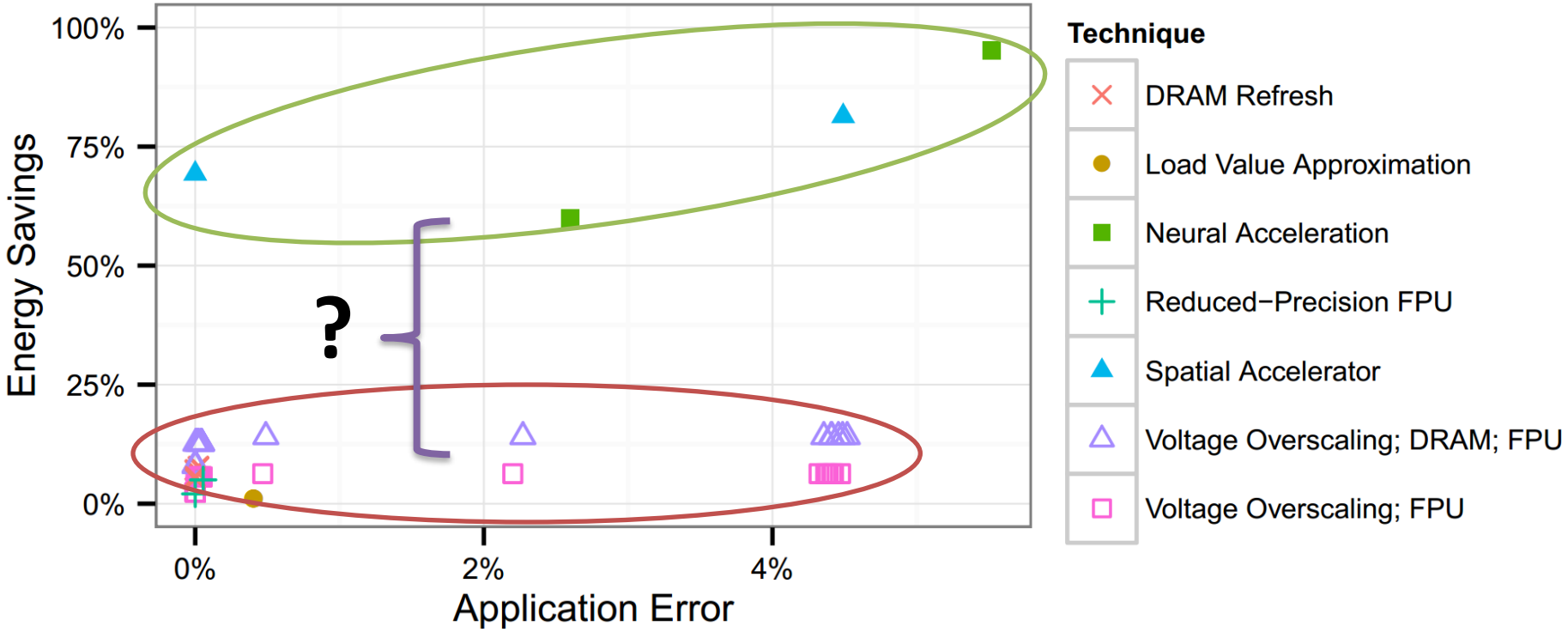
Drowsy SRAM

Reduced Precision FPU

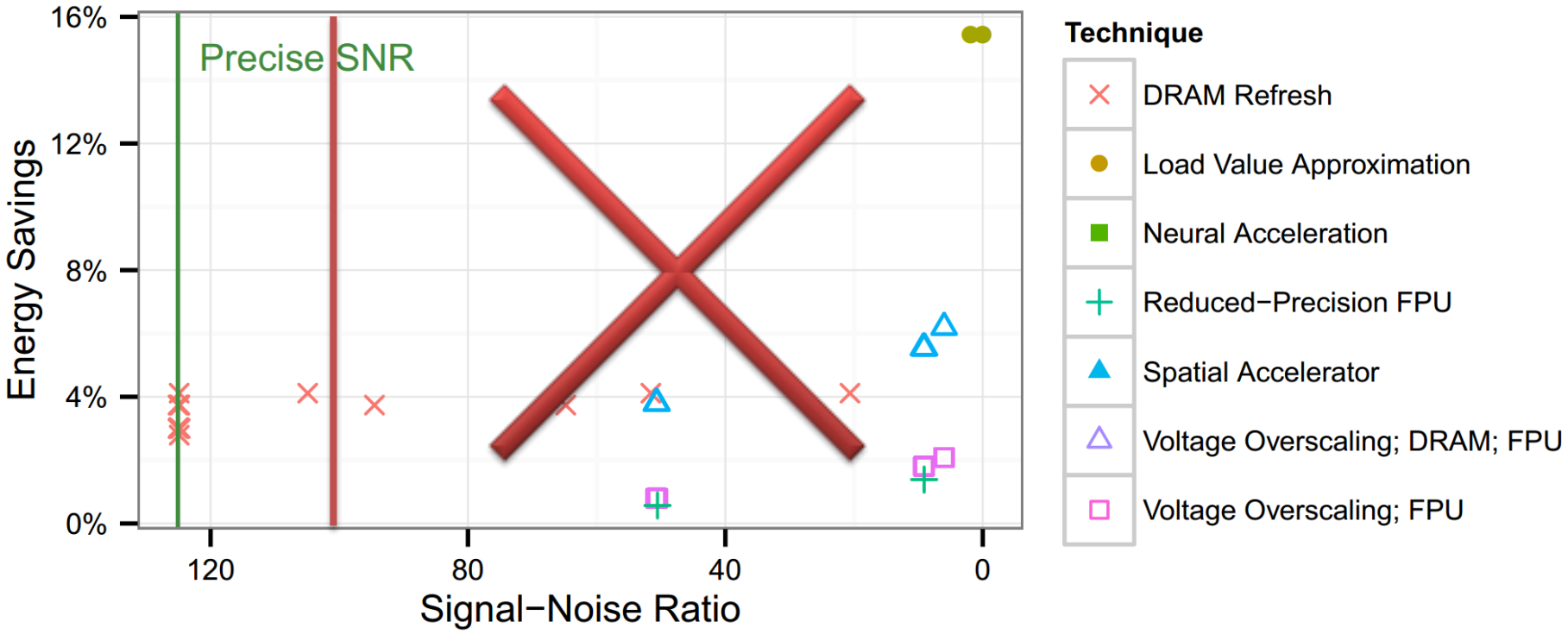
Low refresh rate DRAM

Voltage Overscaled ALU

Early Results - Sobel



Early Results – FFT1D



Conclusions

Coarse-grained superior to fine-grained

Coarse-grained, Nondeterministic!

Thank you!

Questions?