AdaptivFloat: A Floating-Point Based Data Type for Resilient Deep Learning Inference

Thierry Tambe[1] En-Yu Yang[1] **Zishen Wan[1]** Yuntian Deng[1]

Vijay Janapa Reddi[1] Alexander Rush[2] David Brooks[1] Gu-Yeon Wei[1]

1. Harvard University, Cambridge, MA, USA
2. Cornell Tech, New York, NY, USA

HARVARD
John A. Paulson
School of Engineering
and Applied Sciences

Outline

- Motivation
- AdaptivFloat Methodology
- Experimental Results
- PE Architecture
- Hardware Evaluation
- Summary

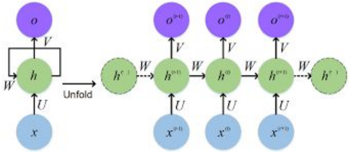
Motivation

- Deep neural networks (DNNs) are deployed at all computing scales:

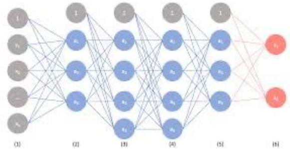
   

Resource-constrained IoT edge devices ⟹ massive datacenter farms

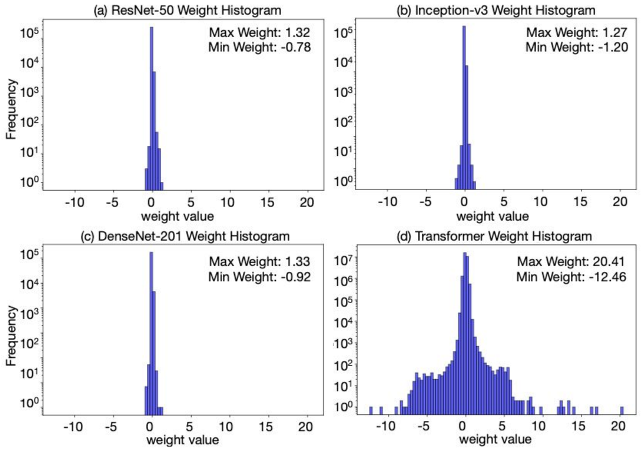
- Recurrent Neural Networks (RNNs):



- Convolutional neural networks (CNNs):



Motivation



Weight distribution for :

Top-Left: ResNet50 Top-Right: Inception-v3
Bot-Left: DenseNet **Bot-Right: Transformer**

Transformer: much **wider** parameter distribution than CNNs

Resource-constrained ⇨ Reduced Precision!

Quantization:

- Previous techniques focus on shallow models or with narrow parameter distribution
- Binary, ternary, quaternary
- Larger weight ⇨ higher impact

Outlier-Aware Quantization

✔ reserve large magnitude weights

❌ complicate hardware implementation

Motivation

Hardware-Friendly Encodings

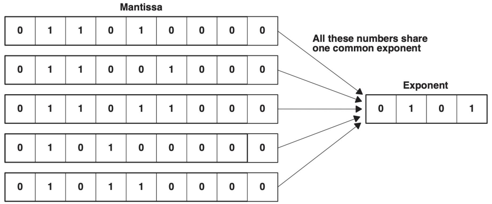
- Linear fixed-point or uniform integer quantization:

  ✓ CNN ✗ Transformer

- Block floating-point:

  ✓ Achieve floating-point-like dynamic range and fixed-point like HW cost

  ✗ Elements with small magnitudes are more prone to data loss

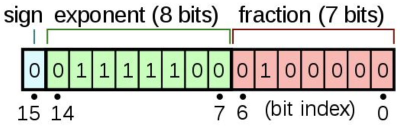


Number Formats with Higher Dynamic Range

- Bfloat16 (2nd TPU, Intel FPGA):

  ✓ Preserves the dynamic range of 32-bit float

  ✗ Incurs reduces precision with 7 fractional bits

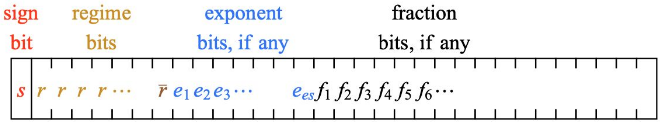


- Posit:

  ✓ higher accuracy on small values

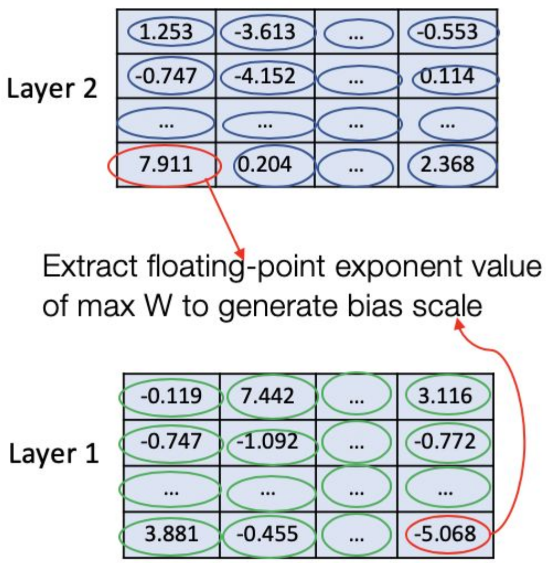
  ✗ lower accuracy on large values

  ✗ worse energy-delay product in hardware implementation



AdaptivFloat: **Adaptive Floati**ng-Point

$$Q = sign * 2^{\exp + bias} * mantissa$$



- A Floating-point based data encoding for deep learning
- Dynamic maximizes and optimally clips its dynamic range
- Achieve higher inference accuracy on a diverse models (CNN, RNN, Transformer, etc)
- At neural network layer granularity
- Hardware friendly: higher energy efficiency and low power consumption

Outline

- Motivation
- AdaptivFloat Methodology
- Experimental Results
- PE Architecture
- Hardware Evaluation
- Summary

Methodology

Floating points w/o denormal

+0.25	-0.25
+0.375	-0.375
+0.5	-0.5
+0.75	-0.75
+1	-1
+1.5	-1.5
+2	-2
+3	-3

Floating points w/o denormal, but sacrifice ±min for ±0

+0	-0
+0.375	-0.375
+0.5	-0.5
+0.75	-0.75
+1	-1
+1.5	-1.5
+2	-2
+3	-3

- No denormal values, sacrificing pos/neg min values for zero representation
- Introduce exp_{bias}, dynamically shift the range of exponent values.

Algorithm 1 AdaptivFloat Bit Vector to Value

Input: bit vector x, number of bits n, number of exponent bits e and exp_{bias}

```
// Get Mantissa bits
```
$m := n - e - 1$

```
// Extract sign, exponent, mantissa
```
$sign := 1$ if $x[n-1] = 0$, otherwise -1
$exp := x[n-2:m] + exp_{bias}$
$mant := 1 + x[m-1:0]/2^m$

```
// Map to 0 if exp, mant bits are zeros
```
if $x[n-2:0] = 0$ **then**
  $val := 0$
else
  $val := sign * 2^{exp} * mant$
end if
return val

Methodology (Quantization)

Algorithm 2 AdaptivFloat Quantization

Input: Matrix W_{fp}, number of bits n and number of exponent bits e

`// Get Mantissa bits`

$m := n - e - 1$

`// Obtain sign and abs matrices`

$W_{sign} := sign(W_{fp})$

$W_{abs} := abs(W_{fp})$

`// Determine` exp_{bias} `and range`

Find normalized exp_{max} for $max(W_{abs})$ such that

$$2^{exp_{max}} \leq max(W_{abs}) < 2^{exp_{max}+1}$$

$exp_{bias} := exp_{max} - (2^e - 1)$

$value_{min} := 2^{exp_{bias}} * (1 + 2^{-m})$

$value_{max} := 2^{exp_{max}} * (2 - 2^{-m})$

`// Handle unrepresentable values`

Round $value < value_{min}$ in W_{abs} to 0 or $value_{min}$

Clamp $value > value_{max}$ in W_{abs} to $value_{max}$

`// Quantize` W_{fp}

Find normalized W_{exp} and W_{mant} such that

$$W_{abs} = 2^{W_{exp}} * W_{mant}, \text{ and } 1 \leq W_{mant} < 2$$

$W_q :=$ quantize and round W_{mant} by $scale = 2^{-m}$

`// Reconstruct output matrix`

$W_{adptiv} := W_{sign} * 2^{W_{exp}} * W_q$

return W_{adptiv}

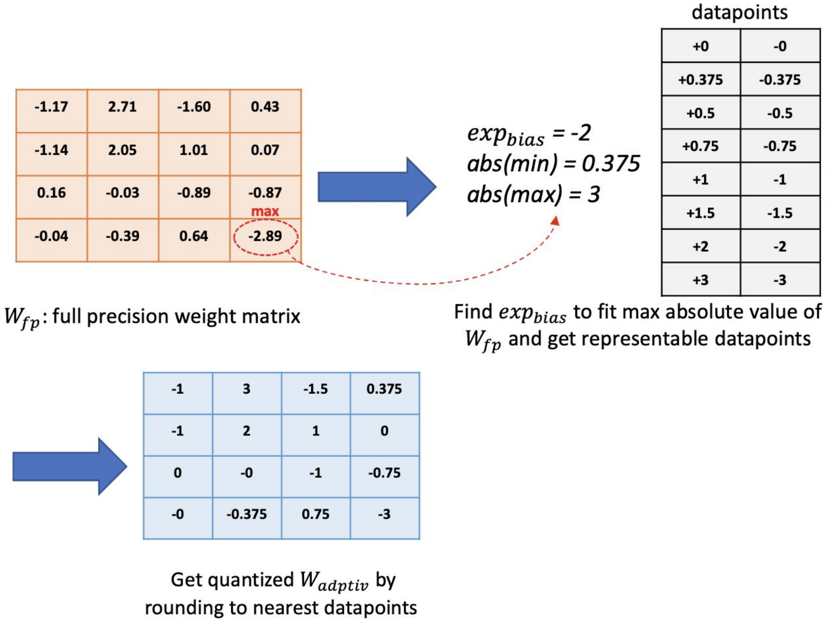
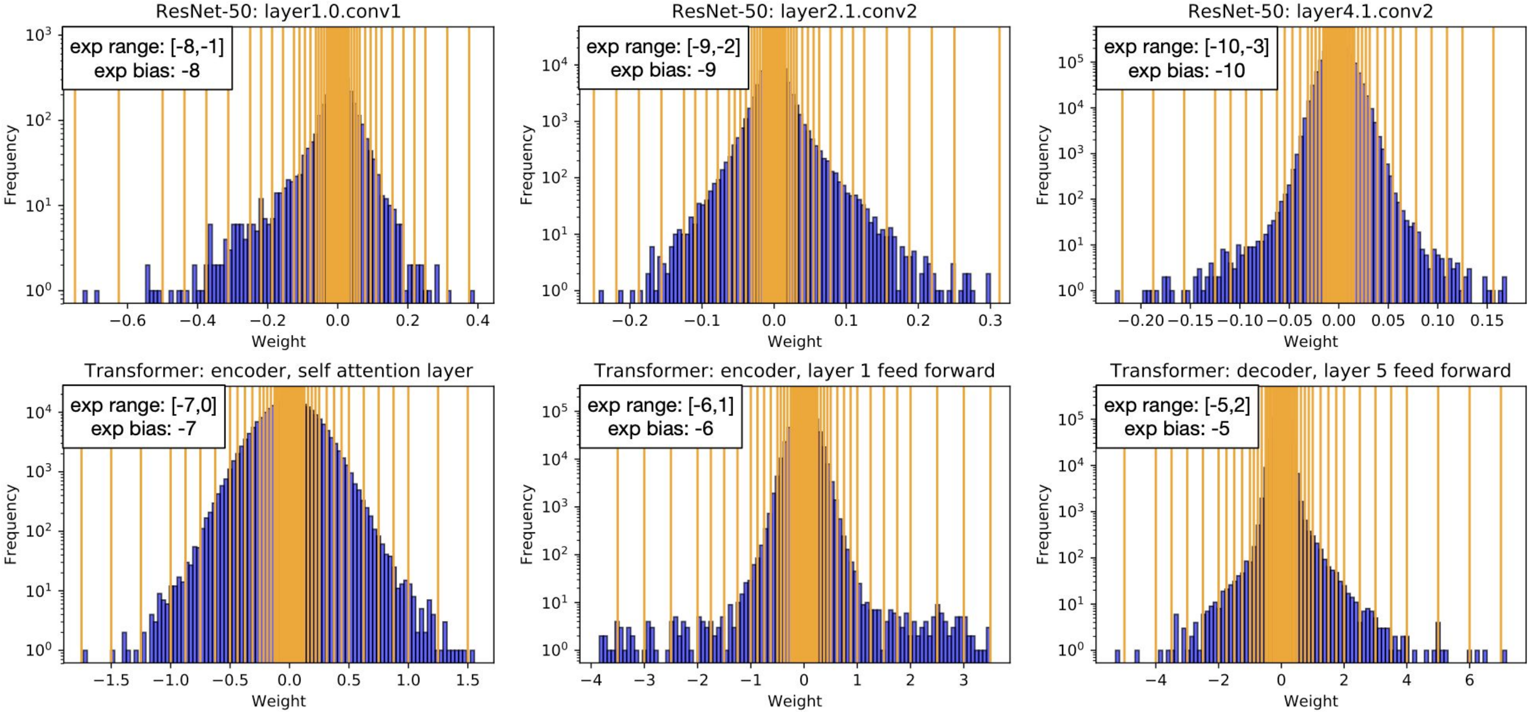


Illustration of AdaptivFloat < 4, 2 > quantization from a full precision weight matrix



AdaptivFloat < 6, 3 > quantization on ResNet50 (top) and Transformer(bottom)

Narrower weight distribution → Smaller max value in weight tensor → More negative exp_{bias}

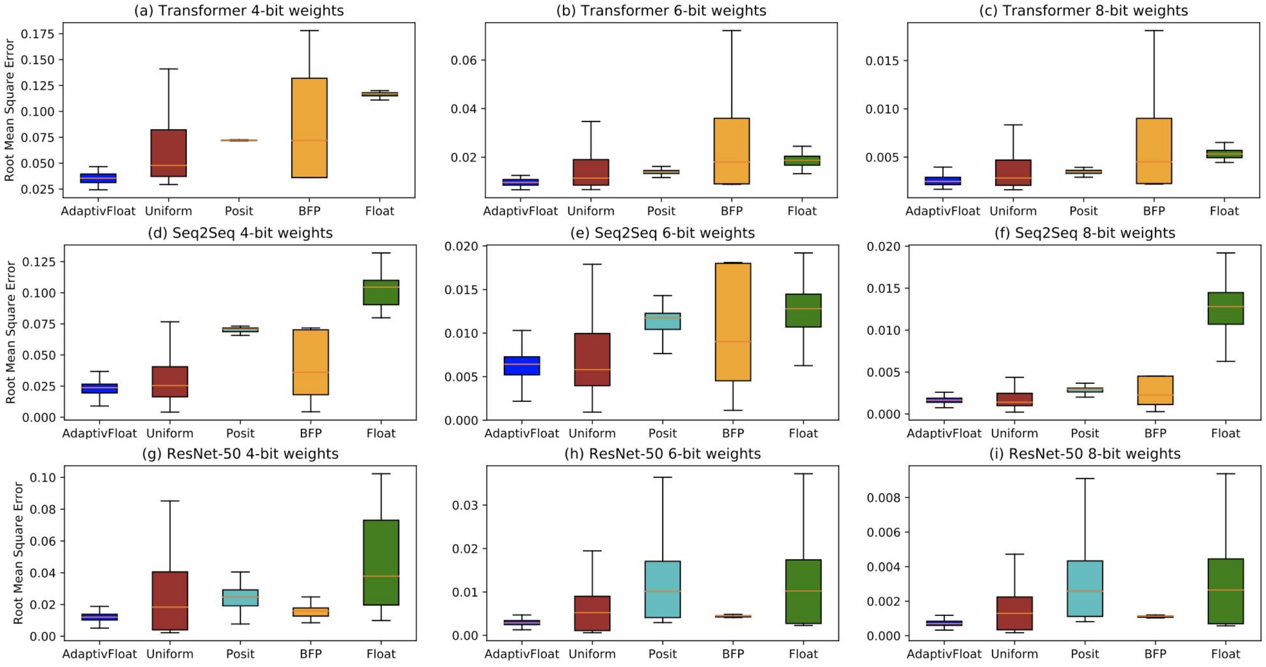
Outline

- Motivation
- AdaptivFloat Methodology
- Experimental Results
- PE Architecture
- Hardware Evaluation
- Summary

Model Evaluation

Model	Application	Dataset	Structure	Number of Parameters	Range of Weights	Fp32 Performance
Transformer	Machine Translation	WMT'17 En-to-De	Attention, FC layers	93M	[-12.46, 20.41]	BLEU: 27.40
Seq2Seq	Speech-to-Text	LibriSpeech	Attention, LSTM, FC layers	20M	[-2.21, 2.39]	WER: 13.34
Resnet-50	Image Classification	ImageNet	CNN, FC layers	25M	[-0.78, 1.32]	Top1 Acc: 76.2

Root Mean Squared Error



- AdaptivFloat produces lower average quantization error
- AdaptivFloat exhibits the tightest error spread

Inference Performance Analysis

- Transformer: BLEU Score on WMT'17 (Fp 32 Baseline: 27.4)

Bit Width	Float	BFP	Uniform	Posit	AdaptivFloat
16	27.4 / 27.4	27.4 / 27.4	27.4 / 27.4	27.4 / 27.5	27.4 / 27.6
8	27.2 / 27.5	26.3 / 27.3	27.3 / 27.4	27.3 / 27.5	27.3 / 27.7
7	27.1 / 27.5	16.9 / 26.8	26.0 / 27.2	27.3 / 27.4	27.3 / 27.7
6	26.5 / 27.1	0.16 / 8.4	0.9 / 23.5	26.7 / 27.2	27.2 / 27.6
5	24.2 / 25.6	0.0 / 0.0	0.0 / 0.0	25.8 / 26.6	26.4 / 27.3
4	0.0 / 0.0	0.0 / 0.0	0.0 / 0.0	0.0 / 0.0	16.3 / 25.5

- Seq2Seq: Word Error Rate on LibriSpeech (Fp32 Baseline: 13.34)

Bit Width	Float	BFP	Uniform	Posit	AdaptivFloat
16	13.40 / 13.07	13.30 / 13.14	13.27 / 12.82	13.29 / 13.05	13.27 / 12.93
8	14.06 / 12.74	13.23 / 13.01	13.28 / 12.89	13.24 / 12.88	13.11 / 12.59
7	13.95 / 12.84	13.54 / 13.27	13.45 / 13.37	13.36 / 12.74	13.19 / 12.80
6	15.53 / 13.48	14.72 / 14.74	14.05 / 13.90	15.13 / 13.88	13.19 / 12.93
5	20.86 / 19.63	21.28 / 21.18	16.53 / 16.25	19.65 / 19.13	15.027 / 12.78
4	INF / INF	76.05 / 75.65	44.55 / 45.99	INF / INF	19.82 / 15.84

a/b:

a is result from post training quantization

B is result from quantization aware retraining

- ResNet-50: Top1 Accuracy on ImageNet (Fp32 Baseline: 76.2)

Bit Width	Float	BFP	Uniform	Posit	AdaptivFloat
16	76.1 / 76.3	76.2 / 76.3	76.1 / 76.3	76.1 / 76.3	76.2 / 76.3
8	75.4 / 75.9	75.7 / 76.0	75.9 / 76.1	75.4 / 76.0	75.7 / 76.3
7	73.8 / 75.6	74.6 / 75.9	75.3 / 75.9	74.1 / 75.8	75.6 / 76.1
6	65.7 / 74.8	66.9 / 74.9	72.9 / 75.2	68.8 / 75.0	73.9 / 75.9
5	16.1 / 73.6	13.2 / 73.4	15.1 / 74.0	33.0 / 73.9	67.2 / 75.6
4	0.5 / 66.3	0.5 / 66.1	2.6 / 67.4	0.7 / 66.7	29.0 / 75.1

- AdaptivFloat demonstrates much greater resiliency at very low precision (≤ 6-bit)
- For resilient performance at low word size, it is critical to have quantization scheme that can adjust its available dynamic range to represent network's weight

Effect of both Weight and Activation Quantization

- Transformer: BLEU Score on WMT'17 (Fp 32 Baseline: 27.4)

Bit Width	Float	BFP	Uniform	Posit	**AdaptivFloat**
W8/A8	27.4	27.4	10.1	26.9	27.5
W6/A6	25.9	0.0	5.7	25.7	27.1
W4/A4	0.0	0.0	0.0	0.0	0.3

- ResNet-50: Top1 Accuracy on ImageNet (Fp32 Baseline: 76.2)

Bit Width	Float	BFP	Uniform	Posit	**AdaptivFloat**
W8/A8	75.7	75.7	75.9	75.8	76.0
W6/A6	73.5	73.4	74.1	73.6	75.0
W4/A4	63.3	63.0	64.3	63.0	72.4

- Seq2Seq: Word Error Rate on LibriSpeech (Fp32 Baseline: 13.34)

Bit Width	Float	BFP	Uniform	Posit	**AdaptivFloat**
W8/A8	12.77	12.86	12.86	12.96	12.59
W6/A6	14.58	14.68	14.04	14.50	12.79
W4/A4	INF	78.68	48.86	INF	21.94

Wn/An: n-bit weight, n-bit activation

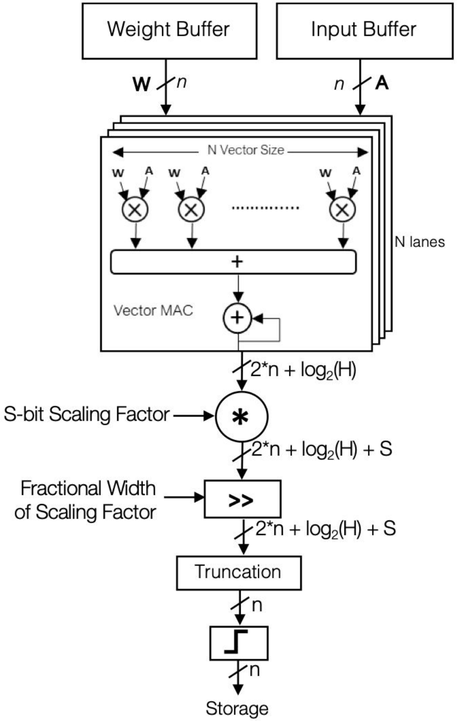
- AdaptivFloat demonstrates greater resiliency compared to other datatypes.
- Performance degradation is steeper on sequence models than on ResNet-50.

Outline

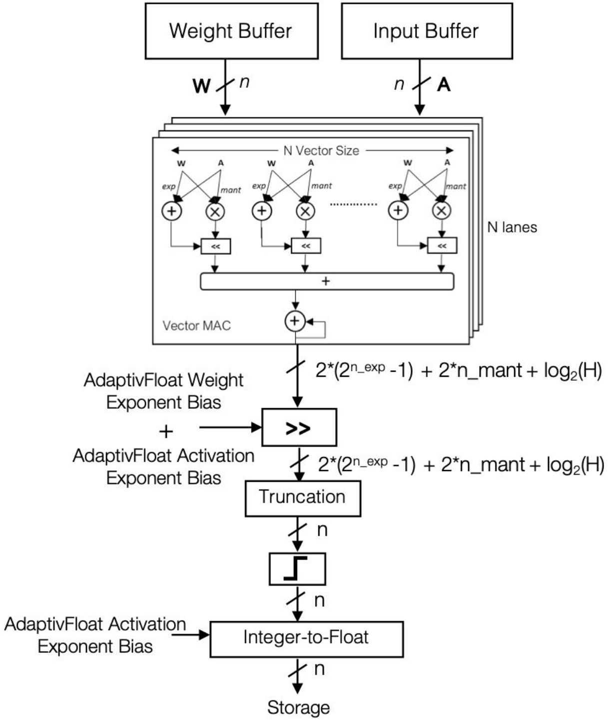
- Motivation
- AdaptivFloat Methodology
- Experimental Results
- PE Architecture
- Hardware Evaluation
- Summary

PE Architecture

- Conventional *n*-bit Integer-based PE



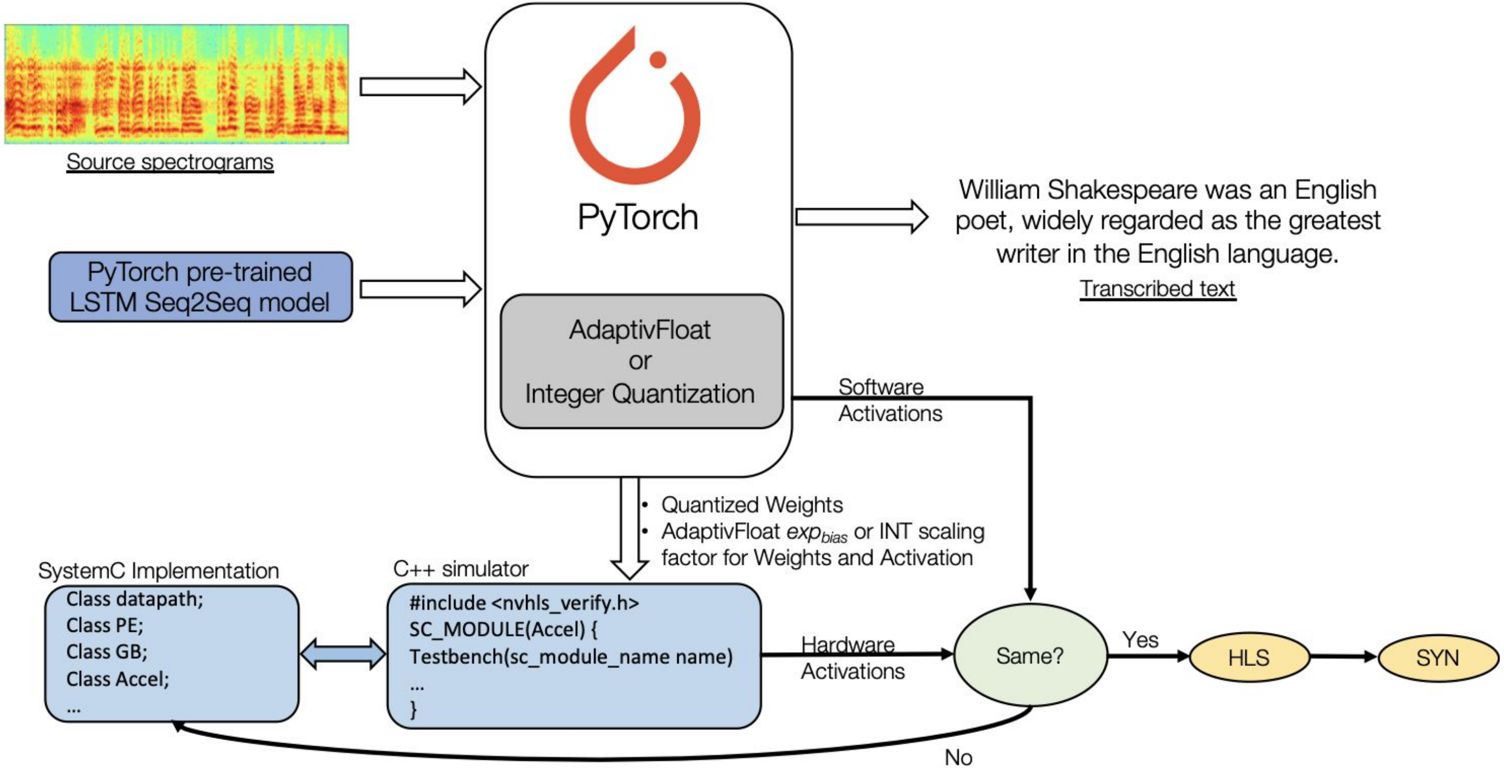
- *n*-bit Hybrid Float-Integer PE



Outline

- Motivation
- AdaptivFloat Methodology
- Experimental Results
- PE Architecture
- Hardware Evaluation
- Summary

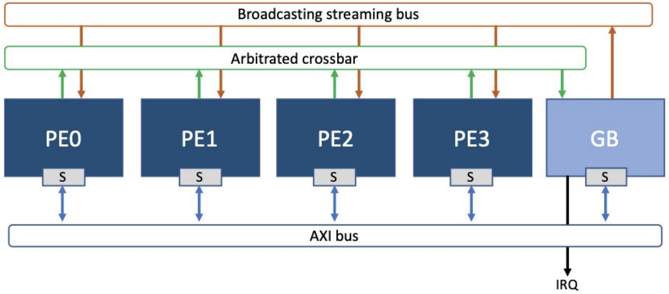
Algorithm-Hardware Co-design Methodology



HARVARD
John A. Paulson
School of Engineering
and Applied Sciences

Algorithm-Hardware Co-design Methodology

- INT and HFINT Accelerator system with 4PEs and a global buffer (GB) targeting sequence-to-sequence networks

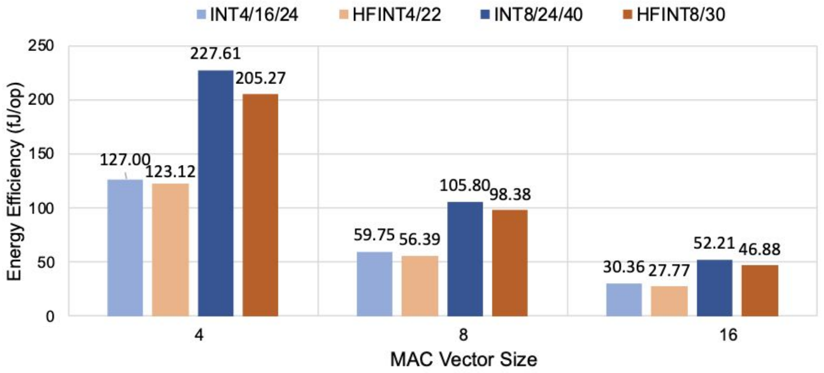


Same evaluation methodology

- Simulation workload: 100 LSTM time steps, 256 hidden units, weight stationary
- Energy and Performance: post-HLS Verilog netlists, Catapult tool @ 1GHz, 16nm FinFET
- Area: Synopsys Design Compiler, after placement and timing-aware logic synthesis

Energy, Performance and Area Analysis

- Energy Efficiency (energy per operation)



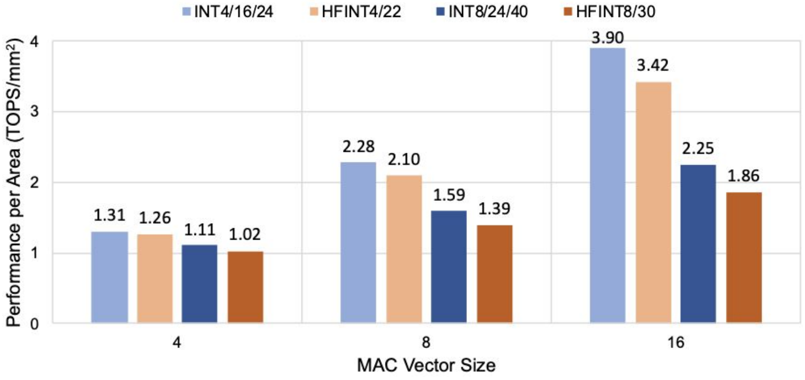
HFINT / INT:

4-bit op, 4-bit Vector Size: 0.97x

4-bit op, 16-bit Vector Size: 0.91x

8-bit op, 4-bit Vector: 0.90x

- Performance per Area



HFINT / INT:

4-bit op, 4-bit Vector Size: 0.96x

4-bit op, 16-bit Vector Size: 0.88x

8-bit op, 4-bit Vector: 0.92x

Energy, Performance and Area Analysis

- PPA results of 8-bit INT and 8-bit HFINT accelerators:

	Power (mW)	Area (mm^2)	Computational Time for 100 LSTM Timesteps (μs)
Int Accelerator with 4 INT8/24/40 PEs	61.38	6.9	81.2
HFInt Accelerator with 4 HFINT8/30 PEs	56.22	7.9	81.2

HFINT accelerator reports 0.92x the power and 1.14x the area of integer-based adaptation.

Outline

- Motivation
- AdaptivFloat Methodology
- Experimental Results
- PE Architecture
- Hardware Evaluation
- Summary

Summary

- AdaptivFloat: a resilient floating-point based encoding solution that dynamically maximizes and optimally clips its available dynamic range, at a layer granularity

- AdaptivFloat demonstrates marked robustness at very low precision (≤6-bit) on Transformer, LSTM-based seq2seq and ResNet-50 networks.

- Proposed AdaptivFloat algorithm-hardware co-design framework

- Proposed Hybrid Float-Integer PE that leverages AdaptivFloat mechanism, demonstrated 0.90x and 0.97x per-operation energy compared to integer-based adaptations.