# SECO: A Scalable Accuracy Approximate Exponential Function via Cross-Layer Optimization

ISLPED 2019
Lausanne, Switzerland | July 31

**Di Wu**, **Tianen Chen**, **Chien-Fu Chen**, **Oghenefego Ahia**, **Joshua San Miguel**, **Mikko Lipasti**, and **Younghyun Kim**

WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON

## I Problem Statement

- Power-hungry and area-demanding **multiplication** and **division** operations
- In **signal processing** and **spiking neural networks**, exponentiation is a key operation
- Typically, **exponentiation** has no hardware support, but implemented instead in a math software library

## II Proposed Solution

- In this experiment, we exploit the Taylor Series approximation of exponents to provide a **fast, energy efficient exponential functional unit (EFU)**
- Replace **multiplication** and **division** within the exponent operation with the **shift** operation

## III Taylor Series Expansion

$$\exp(x) = \sum_{n=0}^{\infty} \frac{x^n}{n!} = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots \backsimeq \sum_{n=0}^{N} s_n \cdot \frac{x^{p_n}}{2^{q_n}}$$

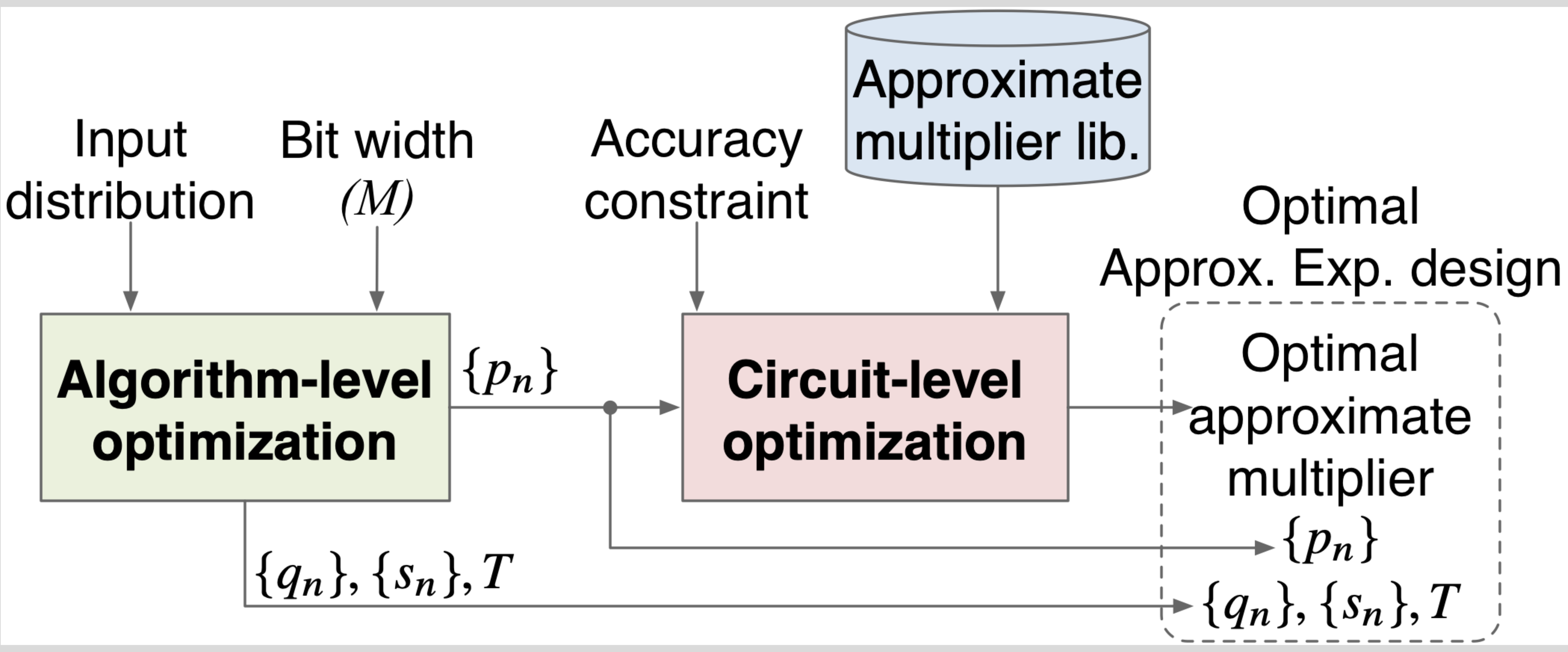
- **Increase** or **decrease** the number of terms within the expansion for the operand x ∈ [0, 1)
- The approximation is **initially centered around x = 0** for x ≈ but then switches to be **centered around x = 1** for x ≈ 1

## IV Cross-Layer Optimization

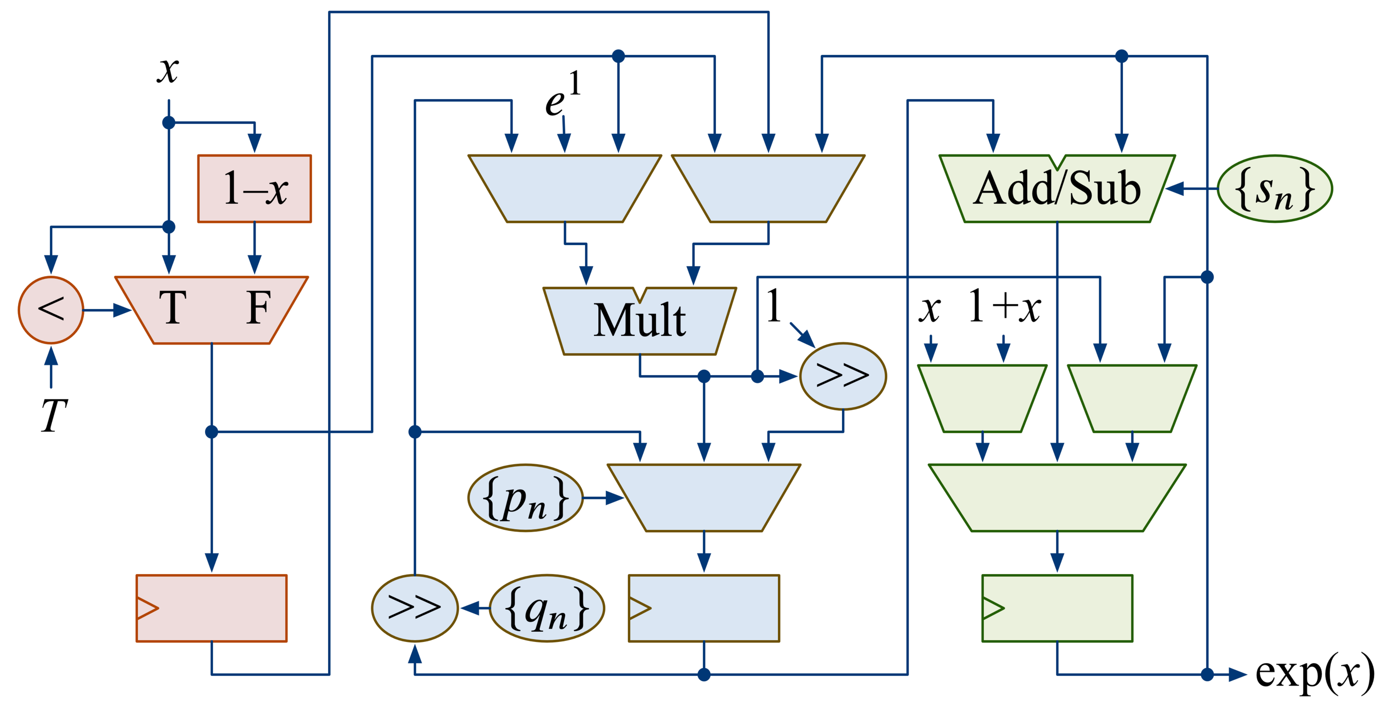
$$s_n \cdot \frac{x^n}{n!} \approx s_n \cdot \frac{x^{p_n}}{2^{q_n}}$$

- **Algorithm-level optimization** by optimizing the four design parameters, $p_n$, $q_n$, $s_n$, T
- **Circuit-level optimization** by finding the best approximate multiplier (EvoApproxLib)
- Parameters optimized by minimizing **weighted mean relative error** (WMRE)

## V Design Flow

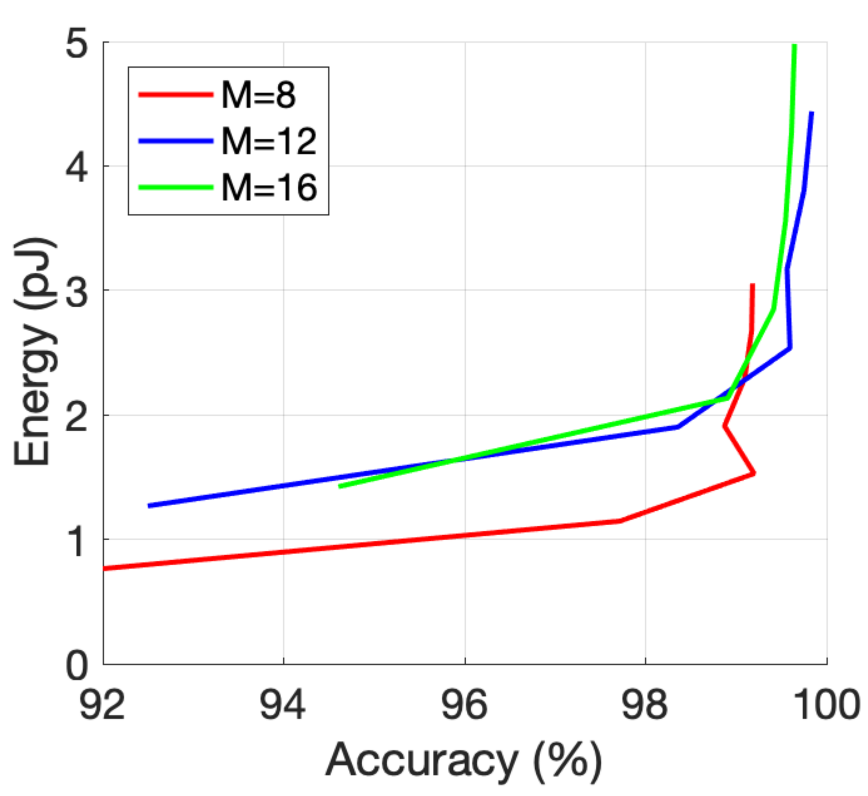



## VI Hardware Implementation

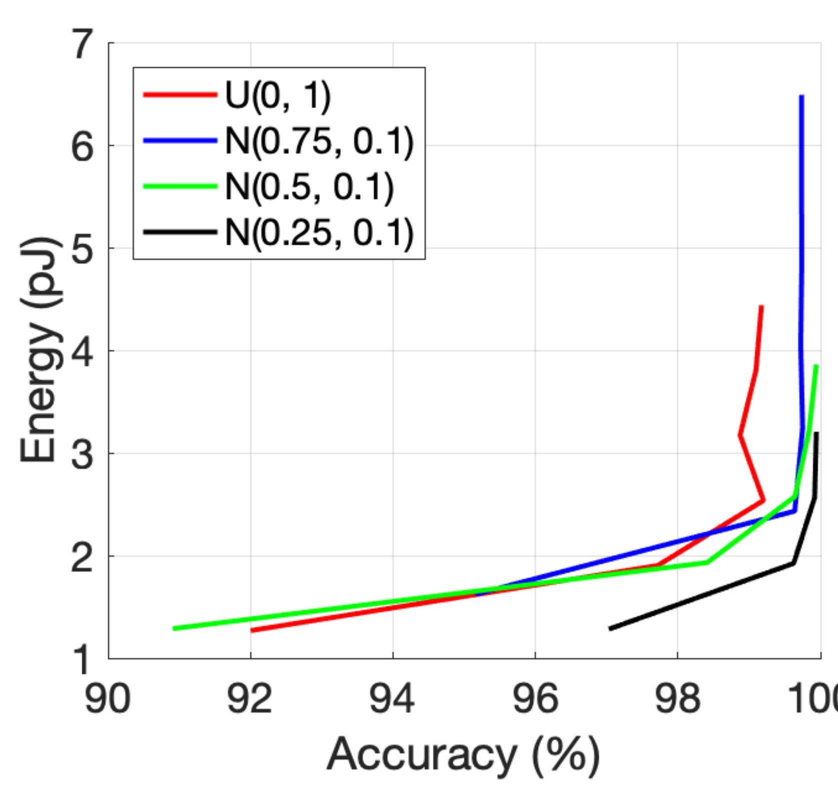



## VII Energy-Accuracy Evaluation

| Input Distribution | Optimal Parameters | | |
|---|---|---|---|
| U(0,1) | $\{p_n\}$ | 0, 1, 2, 3, 4, 5, 5 | |
| | $\{s_n\}$, $\{q_n\}$ | 0, 0, 1, 2, -3, 4, 5 | |
| | T | 0.875 | Multiplier mul12u_2QN |
| N(0.75, 0.1) | $\{p_n\}$ | 0, 1, 2, 3, 4, 4, 4 | |
| | $\{s_n\}$, $\{q_n\}$ | 0, 0, 1, 3, -4, 5, -6, 7 | |
| | T | 0.375 | Multiplier mul12u_2PM |

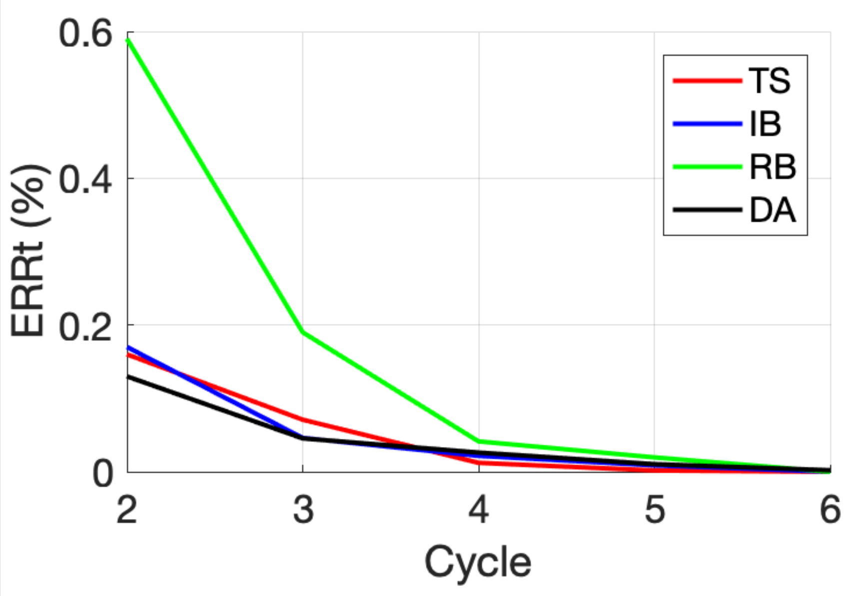



Varying bit-width

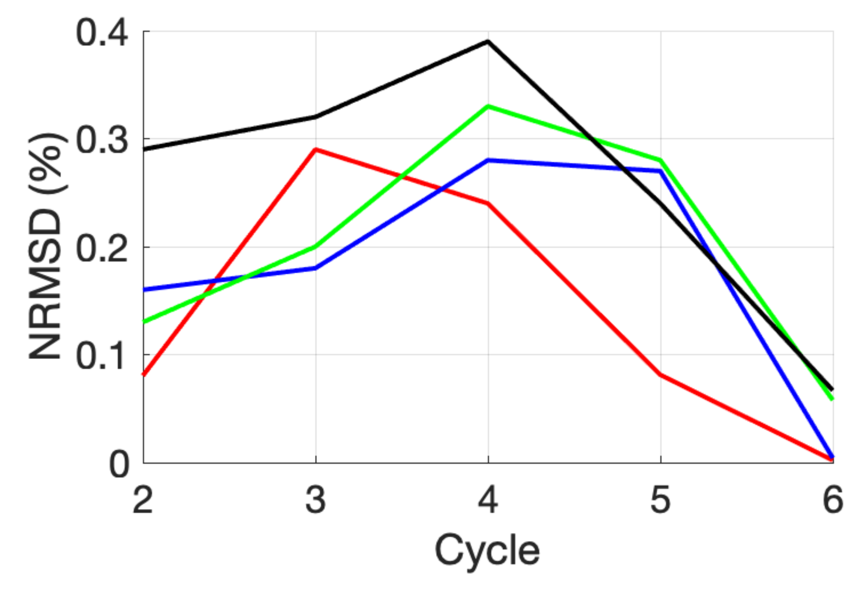



Varying input distribution

## VIII Adaptive Exponential Neuron Case Study




Time error




Value error

## IX Conclusion

- **Negligible accuracy loss** with a significant drop in **power, area,** and **latency**
- Accuracy drop from 99.997% (baseline design) to 99.7% while saving 96% energy, 94.5% area, and 82.5% latency
- **Cross-layer optimization framework** for SECO generalizable to other designs
- Evaluated the algorithm and design's efficacy on **Adaptive Exponential Neuron**