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

Di Wu, Tianen Chen, Chienfu Chen, Oghenefego Ahia, Joshua San Miguel, Mikko Lipasti, and Younghyun Kim
University of Wisconsin-Madison

Department of Electrical and Computer Engineering
UNIVERSITY OF WISCONSIN–MADISON

Executive Summary

- Incremental approximation of exponentiation via Taylor series.
- Cross-layer optimization framework for energy-accuracy tradeoff, including algorithm-level and circuit-level.
- Application of the proposal to adaptive exponential integrate-and-fire neuron.

Department of Electrical and Computer Engineering
UNIVERSITY OF WISCONSIN–MADISON

Outline

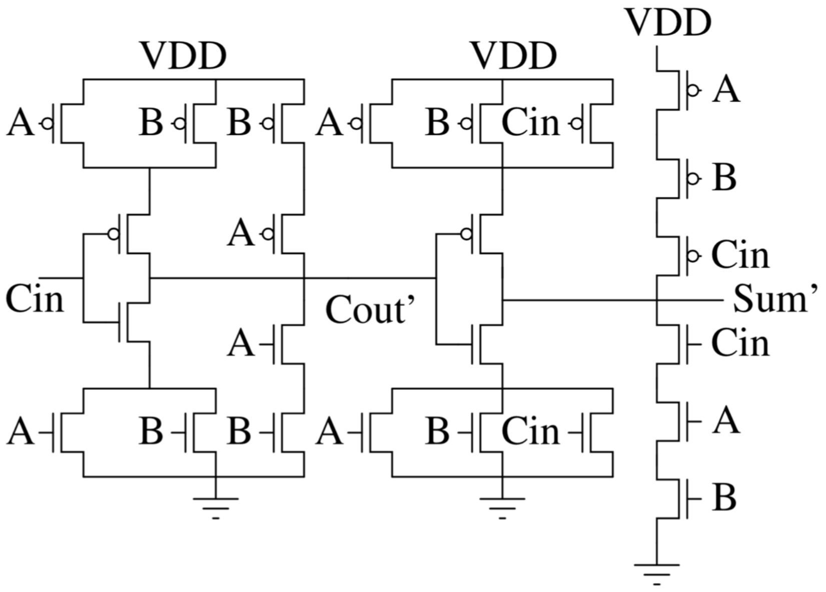
- ❑ **Background**
- ❑ Approximate Taylor series
- ❑ Cross-layer optimization
- ❑ Performance evaluation
- ❑ Case study on AdEx neuron
- ❑ Conclusion
- ❑ Future work

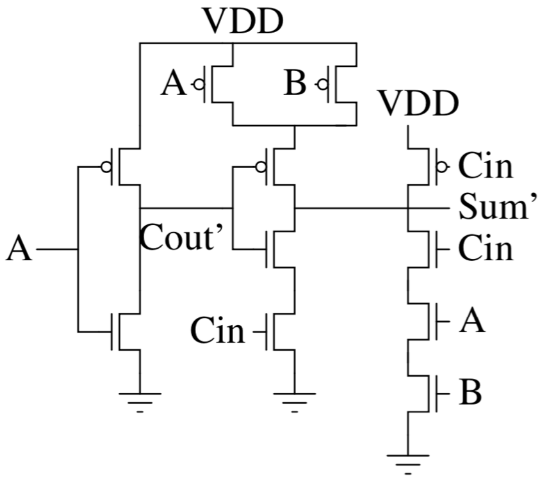
Department of Electrical and Computer Engineering
UNIVERSITY OF WISCONSIN–MADISON

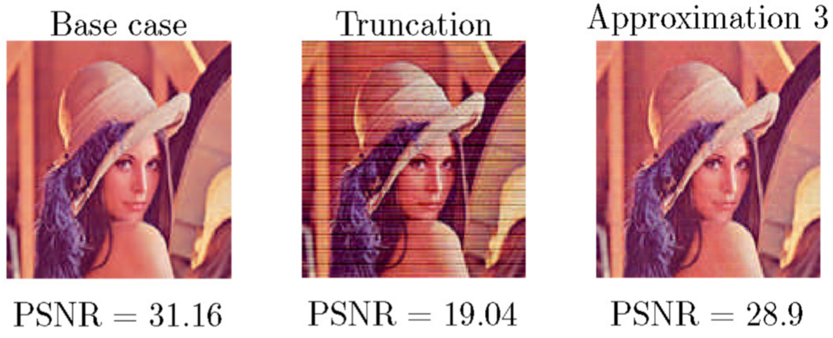
Background – Approximate computing

- Reduced accuracy for high energy efficiency
  - Circuit-level

Mirror adder







V. Gupta, etc., IMPACT: IMPrecise adders for low-power approximate computing, 2011

Background – Approximate computing

- Reduced accuracy for high energy efficiency
  - Circuit-level
  - Algorithm-level

```
for ( int i = 0; i < N; i++ ) {
    // do things
}
```

```
for ( int i = 0; i < N; i++ ) {
    // do things
    i = i + skip_factor;
}
```

Loop perforation

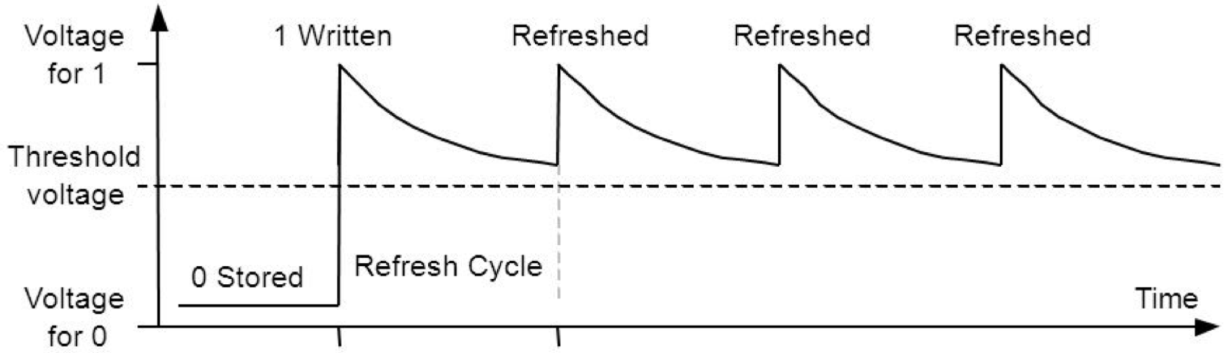
Department of Electrical and Computer Engineering
UNIVERSITY OF WISCONSIN–MADISON

Background – Approximate computing

- Reduced accuracy for high energy efficiency
  - Circuit-level
  - Algorithm-level
  - Storage-level

$$\mathrm{trunc}(x, n) = \frac{\lceil 10^n \cdot x \rceil}{10^n}$$

Truncation



DRAM refresh time tuning

Department of Electrical and Computer Engineering
UNIVERSITY OF WISCONSIN–MADISON

Background – Approximate computing

- Reduced accuracy for high energy efficiency
  - Circuit-level
  - Algorithm-level
  - Storage-level
  - System-level: A combination of all

Department of Electrical and Computer Engineering
UNIVERSITY OF WISCONSIN–MADISON

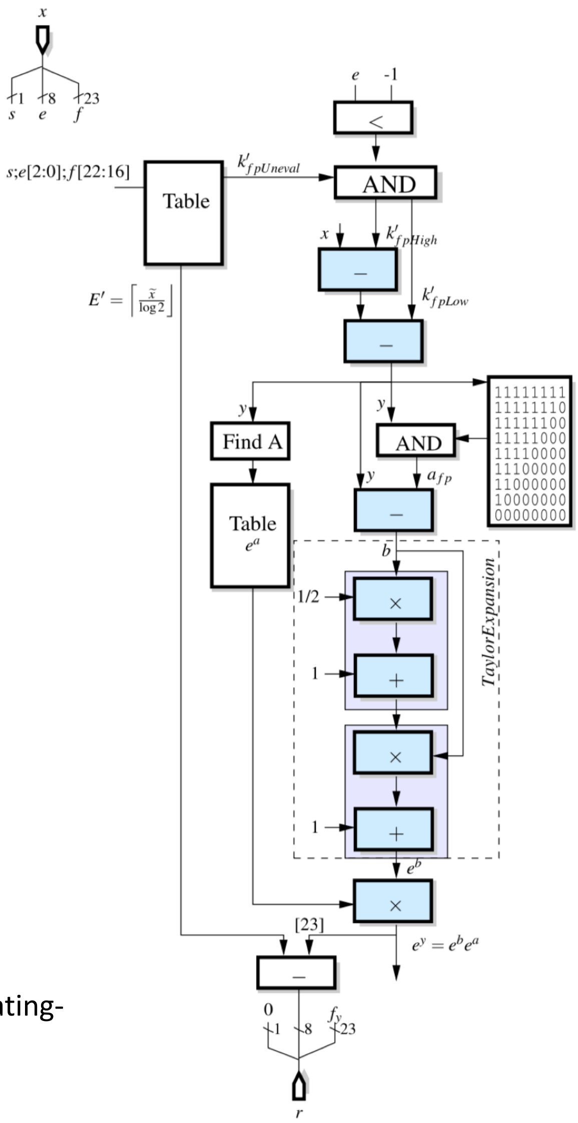
Background – Approximate computing

- Reduced accuracy for high energy efficiency
  - Circuit-level
  - Algorithm-level
  - Storage-level
  - System-level

- Energy-accuracy tradeoff
  - Design-time fixed
  - Run-time tunable
  - Input-aware

Background – Existing Exponentiation Unit

- Floating point
  - High accuracy
  - Long latency
  - High energy

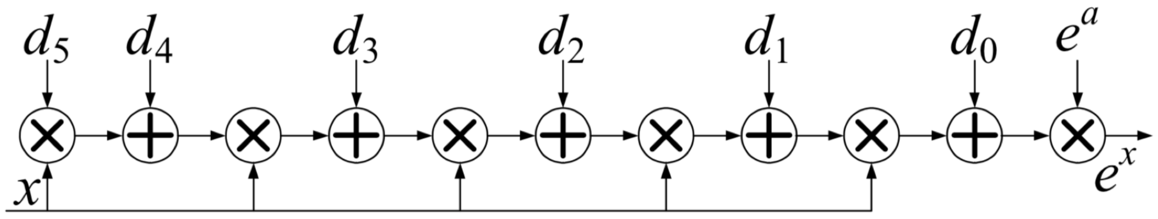


M. Langhammer, Single Precision Logarithm and Exponential Architectures for Hard Floating-Point Enabled FPGAs, 2017

Background – Existing Exponentiation Unit

- Floating point
  - High accuracy
  - Long latency
  - High energy

- Fixed point
  - Taylor series
  - Fixed Taylor terms
  - Fixed precise coefficients



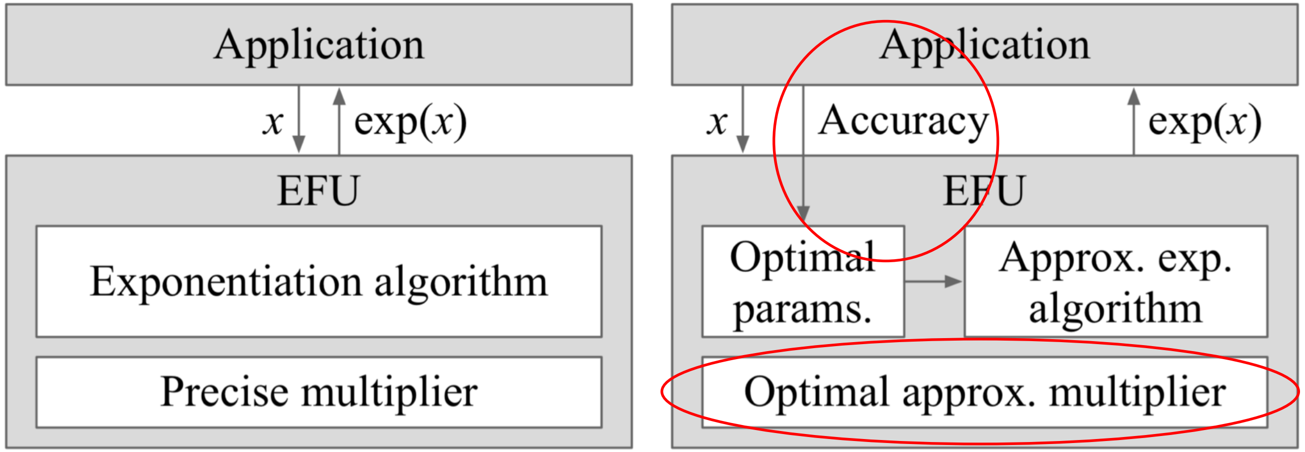
	Coefficient	Numerical value
d_0	$1-a+\frac{a^2}{2}-\frac{a^3}{6}+\frac{a^4}{24}-\frac{a^5}{120}+\frac{a^6}{720}$	0.606532118055556
d_1	$1-\frac{2a}{2}+\frac{3a^2}{6}-\frac{4a^3}{24}+\frac{31a^4}{720}-\frac{6a^5}{720}$	0.606597222222222
d_2	$\frac{1}{2}-\frac{3a}{6}+\frac{6a^2}{24}-\frac{64a^3}{720}+\frac{14a^4}{720}$	0.302604166666667
d_3	$\frac{1}{6}-\frac{4a}{24}+\frac{66a^2}{720}-\frac{16a^3}{720}$	0.103472222222222
d_4	$\frac{1}{24}-\frac{34a}{720}+\frac{9a^2}{720}$	0.021180555555556
d_5	$\frac{7}{720}-\frac{2a}{720}$	0.008333333333333

P. Nilsson, Hardware implementation of the exponential function using Taylor series, 2014

Our goal

- Reduced accuracy for high energy efficiency
  - Circuit-level
  - Algorithm-level
  - Storage-level
  - System-level

- Energy-accuracy tradeoff
  - Design-time fixed
  - Run-time tunable
  - Input-aware



(a) Conventional EFU (b) Proposed EFU

Department of Electrical and Computer Engineering
UNIVERSITY OF WISCONSIN–MADISON

Outline

- ❑ Background
- ❑ **Approximate Taylor series**
- ❑ Cross-layer optimization
- ❑ Performance evaluation
- ❑ Case study on AdEx neuron
- ❑ Conclusion
- ❑ Future work

Department of Electrical and Computer Engineering
UNIVERSITY OF WISCONSIN–MADISON

Approximate Taylor series

- Conventional Taylor series
  - Accurate but not efficient

$$\exp(x) = \sum_{n=0}^{\infty} \frac{x^n}{n!} = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots ,$$

Department of Electrical and Computer Engineering
UNIVERSITY OF WISCONSIN–MADISON

Approximate Taylor series

- Conventional Taylor series
  - Accurate but not efficient

$$\exp(x) = \sum_{n=0}^{\infty} \frac{x^n}{n!} = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots,$$

- Approximate Taylor Series
  - Multiplication skipping

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

where, for $n = 0, 1, \ldots, N$,

$$p_n = \begin{cases} p_{n-1} + 1 & \text{if multiplication is not skipped,} \\ p_{n-1} & \text{if multiplication is skipped,} \end{cases}$$

Department of Electrical and Computer Engineering
UNIVERSITY OF WISCONSIN–MADISON

Approximate Taylor series

- Conventional Taylor series
  - Accurate but not efficient

$$\exp(x) = \sum_{n=0}^{\infty} \frac{x^n}{n!} = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots,$$

- Approximate Taylor Series
  - Multiplication skipping
  - Approximate division

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

where, for $n = 0, 1, \ldots, N$,

$$p_n = \begin{cases} p_{n-1} + 1 & \text{if multiplication is not skipped,} \\ p_{n-1} & \text{if multiplication is skipped,} \end{cases}$$

Q_n is the shifting offset

S_n is the sign

Department of Electrical and Computer Engineering
UNIVERSITY OF WISCONSIN–MADISON

Approximate Taylor series

- Conventional Taylor series
  - Accurate but not efficient

$$\exp(x) = \sum_{n=0}^{\infty} \frac{x^n}{n!} = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots,$$

- Approximate Taylor Series
  - Multiplication skipping
  - Approximate division
  - Double-sided expansion

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

where, for $n = 0, 1, \ldots, N$,

$$p_n = \begin{cases} p_{n-1} + 1 & \text{if multiplication is not skipped,} \\ p_{n-1} & \text{if multiplication is skipped,} \end{cases}$$

Q_n is the shifting offset

S_n is the sign

Department of Electrical and Computer Engineering
UNIVERSITY OF WISCONSIN–MADISON

Approximate Taylor series

- Conventional Taylor series
  - Accurate but not efficient

$$\exp(x) = \sum_{n=0}^{\infty} \frac{x^n}{n!} = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots,$$

- Approximate Taylor Series
  - Multiplication skipping
  - Approximate division
  - Double-sided expansion

$\{p_n\}$	0, 1, 2, 3, 4, 5, 5
$\{s_n \cdot q_n\}$	0, 0, 1, 2, -3, 4, 5

- Example

$$\exp(x) = \frac{x^0}{2^0} + \frac{x^1}{2^0} + \frac{x^2}{2^1} + \frac{x^3}{2^2} - \frac{x^4}{2^3} + \frac{x^5}{2^4} + \frac{x^5}{2^5}$$

Department of Electrical and Computer Engineering
UNIVERSITY OF WISCONSIN–MADISON

Approximate Taylor series

- Conventional Taylor series
  - Accurate but not efficient

$$\exp(x) = \sum_{n=0}^{\infty} \frac{x^n}{n!} = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots ,$$

- Approximate Taylor Series
  - Multiplication skipping
  - Approximate division
  - Double-sided expansion

$\{p_n\}$	0, 1, 2, 3, 4, 5, 5
$\{s_n \cdot q_n\}$	0, 0, 1, 2, -3, 4, 5

- Example

$$\exp(x) = \frac{x^0}{2^0} + \frac{x^1}{2^0} + \frac{x^2}{2^1} + \frac{x^3}{2^2} - \frac{x^4}{2^3} + \frac{x^5}{2^4} + \frac{x^5}{2^5}$$

Department of Electrical and Computer Engineering
UNIVERSITY OF WISCONSIN–MADISON

Approximate Taylor series

- Conventional Taylor series
  - Accurate but not efficient

$$\exp(x) = \sum_{n=0}^{\infty} \frac{x^n}{n!} = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots,$$

- Approximate Taylor Series
  - Multiplication skipping
  - Approximate division
  - Double-sided expansion

$\{p_n\}$	0, 1, 2, 3, 4, 5, 5
$\{s_n \cdot q_n\}$	0, 0, 1, 2, -3, 4, 5

- Example

$$\exp(x) = \frac{x^0}{2^0} + \frac{x^1}{2^0} + \frac{x^2}{2^1} + \frac{x^3}{2^2} - \frac{x^4}{2^3} + \frac{x^5}{2^4} + \frac{x^5}{2^5}$$

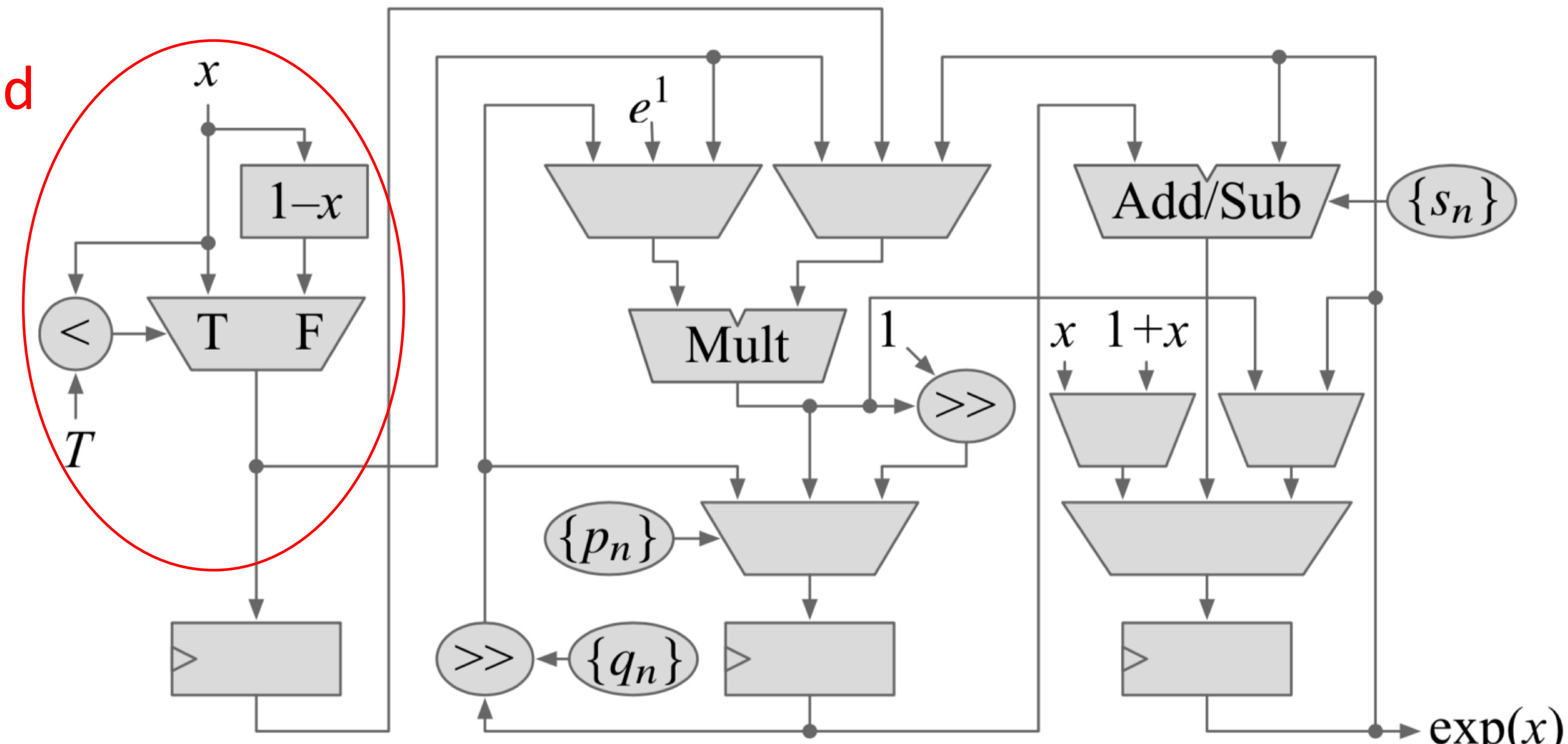
Error compensation

Department of Electrical and Computer Engineering
UNIVERSITY OF WISCONSIN–MADISON

Approximate Taylor series

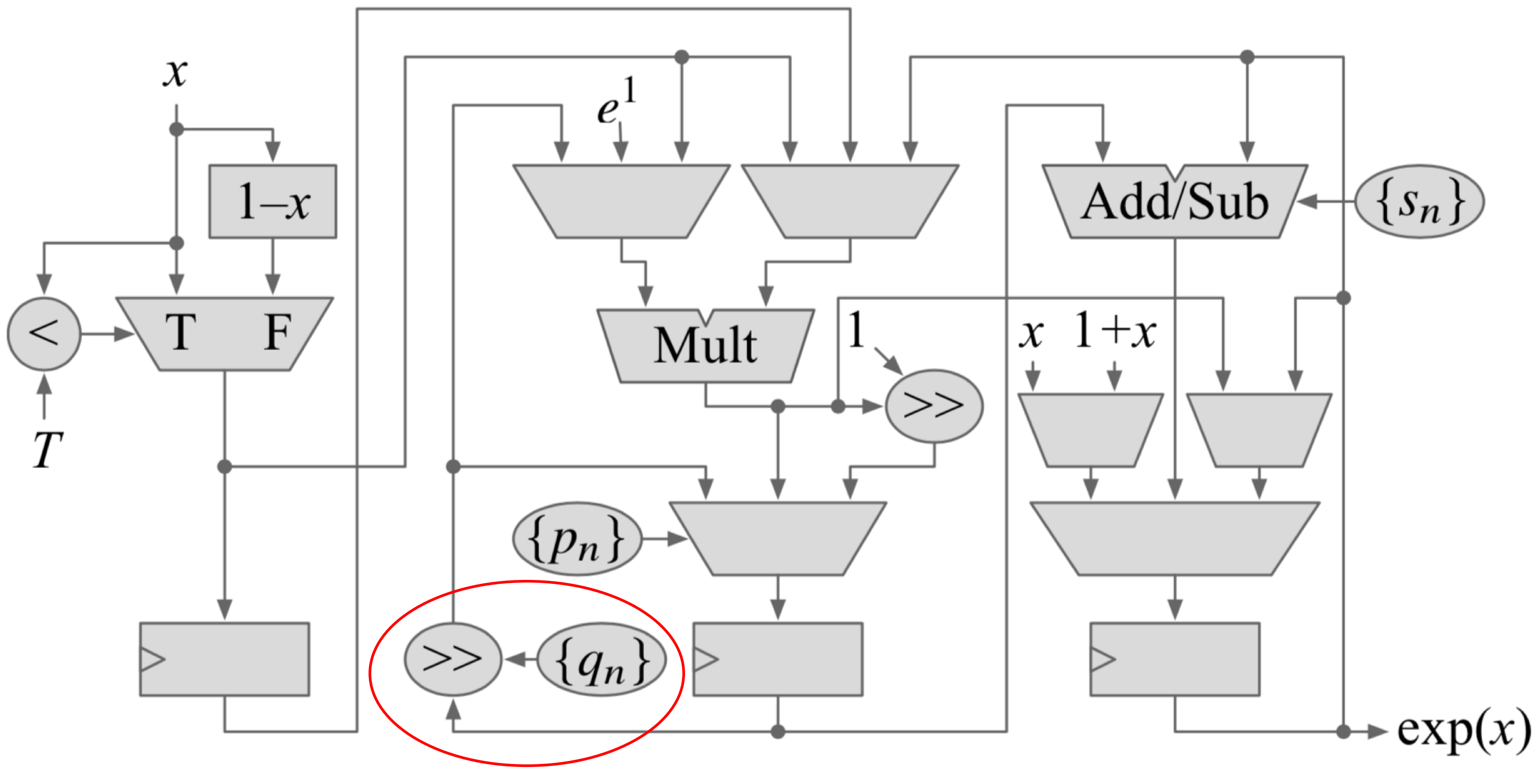
- Resultant architecture

Double-sided expansion



Approximate Taylor series

➤ Resultant architecture

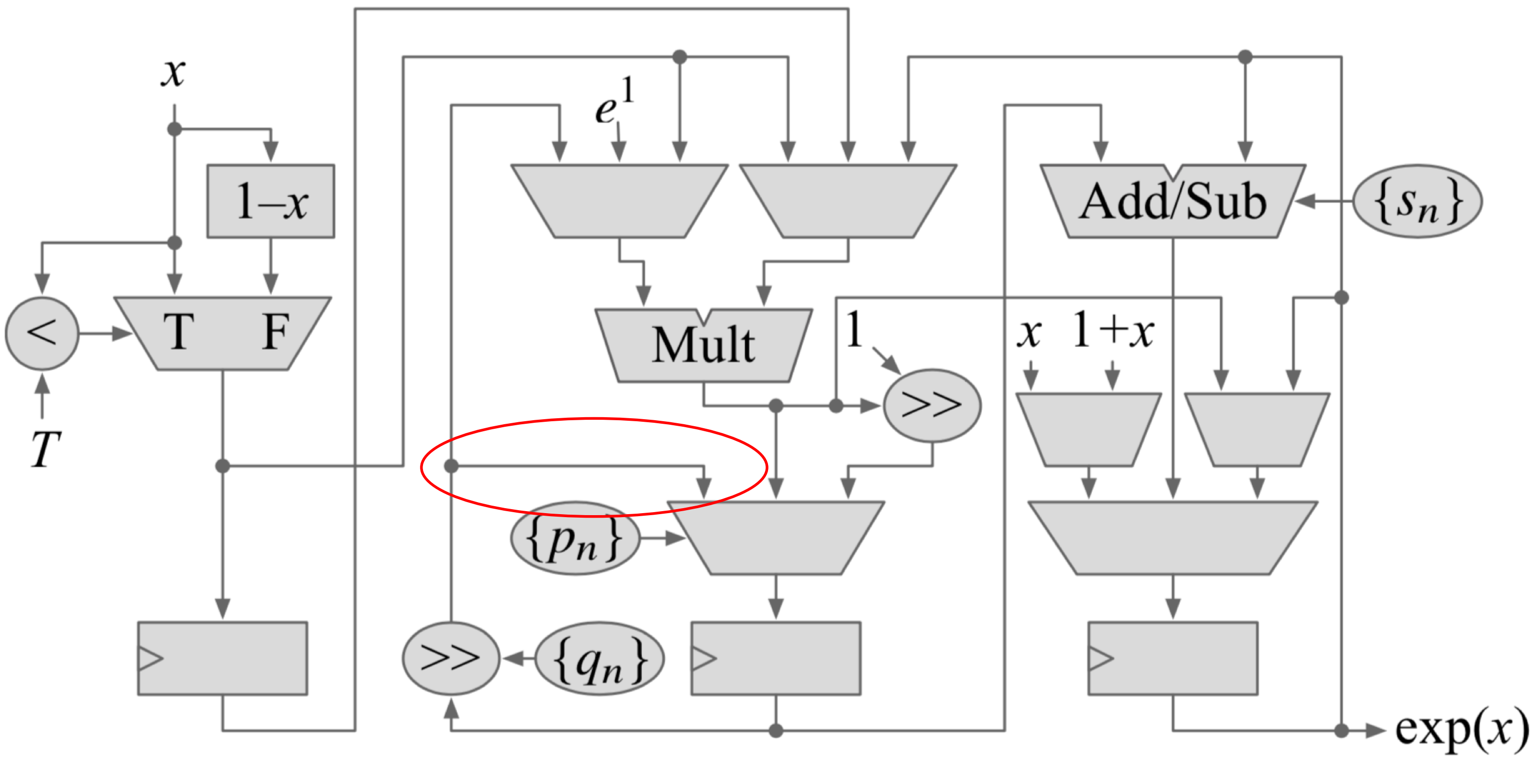


Approximate Division

Department of Electrical and Computer Engineering
UNIVERSITY OF WISCONSIN–MADISON

Approximate Taylor series

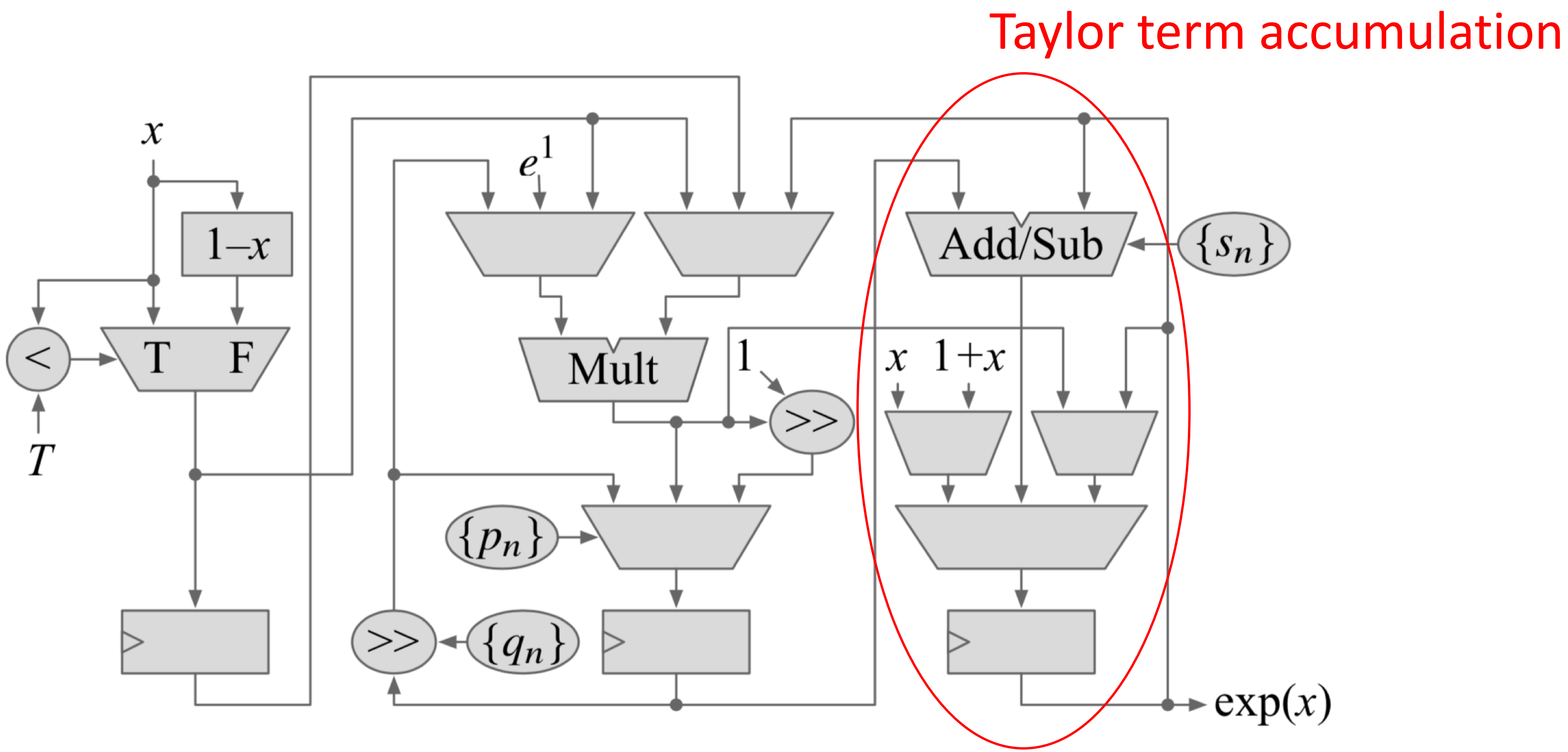
- Resultant architecture



Multiplication skipping

Approximate Taylor series

- Resultant architecture



Outline

- ❑ Background
- ❑ Approximate Taylor series
- ❑ **Cross-layer optimization**
- ❑ Performance evaluation
- ❑ Case study on AdEx neuron
- ❑ Conclusion
- ❑ Future work

Cross-layer optimization – Benefit

- Varying run-time demands
  - Static design-time optimization is not always satisfying

- Achieving the best application-level approximation
  - Isolated circuit design is usually for uniform input
  - Unknown input distribution leads to uncontrollable output quality
  - To the limits of approximation

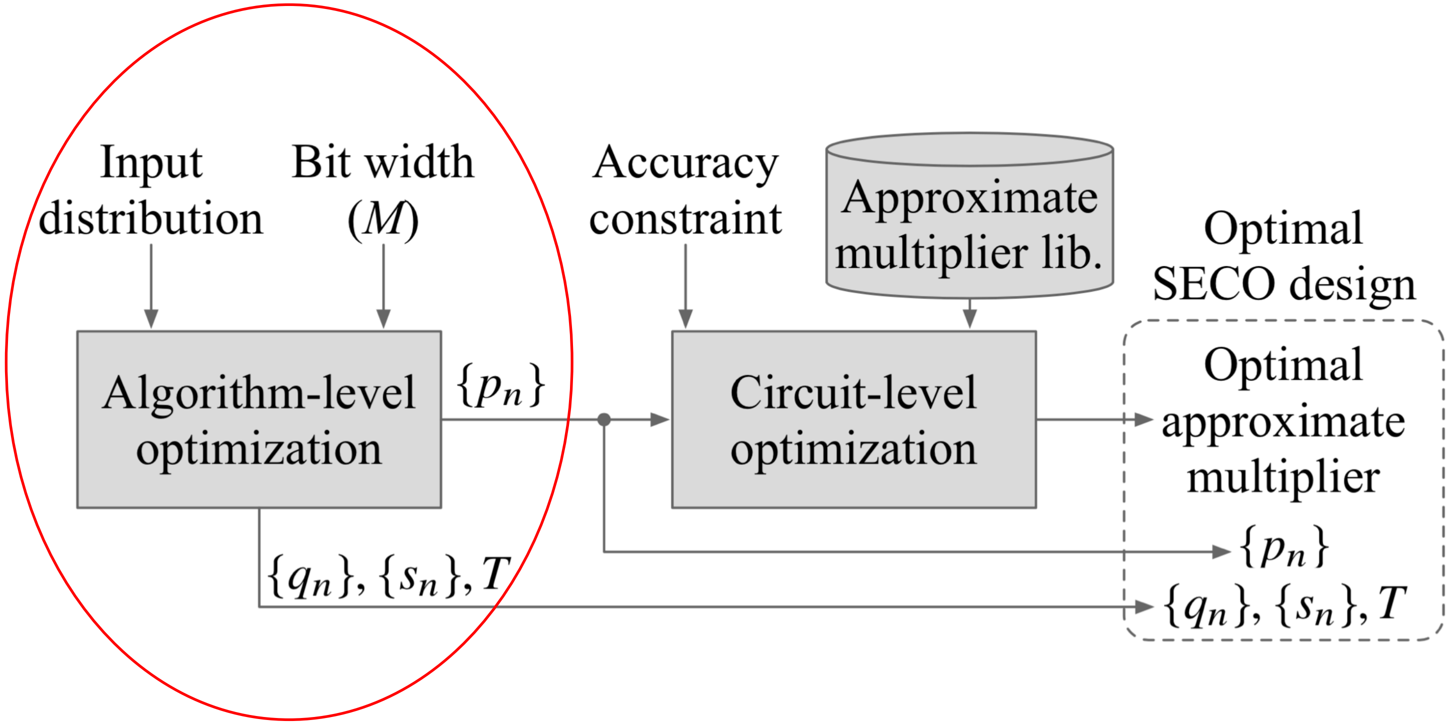
- Cross-layer optimization bridges approximate circuits and real applications

Department of Electrical and Computer Engineering
UNIVERSITY OF WISCONSIN–MADISON

Cross-layer optimization – Flow

- Algorithm level
  - Find the best parameters

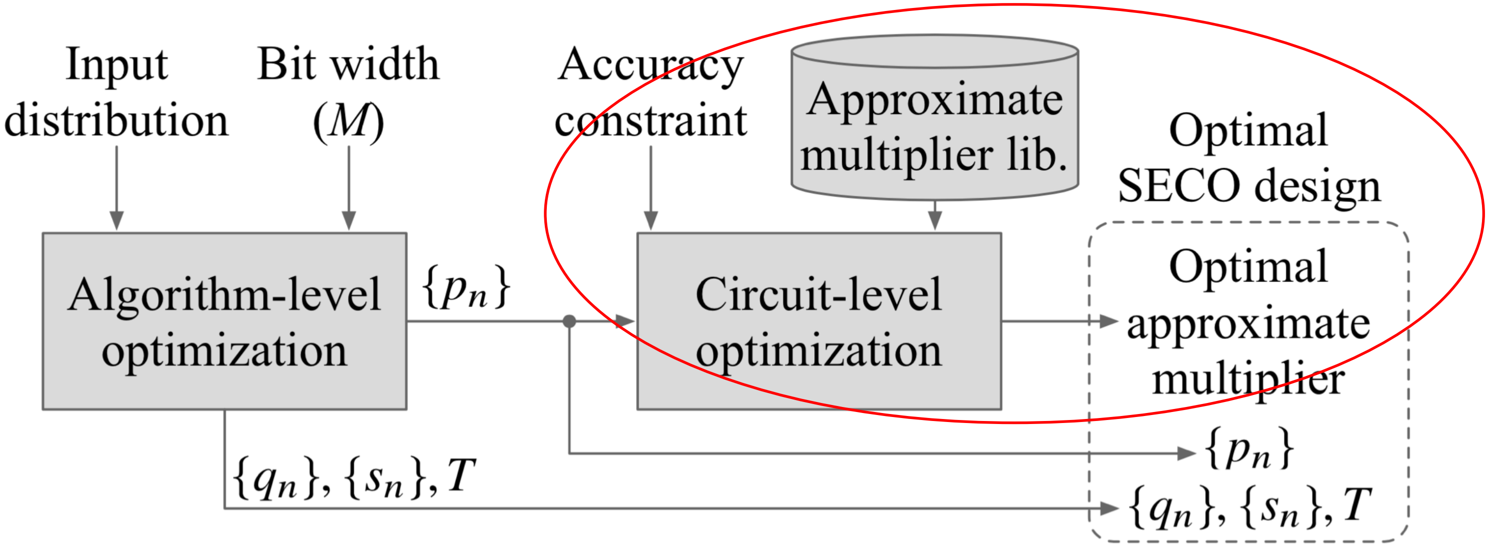
1. multiplication skipping
2. approximate division
3. double-sided expansion



Department of Electrical and Computer Engineering
UNIVERSITY OF WISCONSIN–MADISON

Cross-layer optimization – Flow

- ➢ Algorithm level
  - Find the best parameters
- ➢ Circuit level
  - Find the best approximate multiplier from verified library

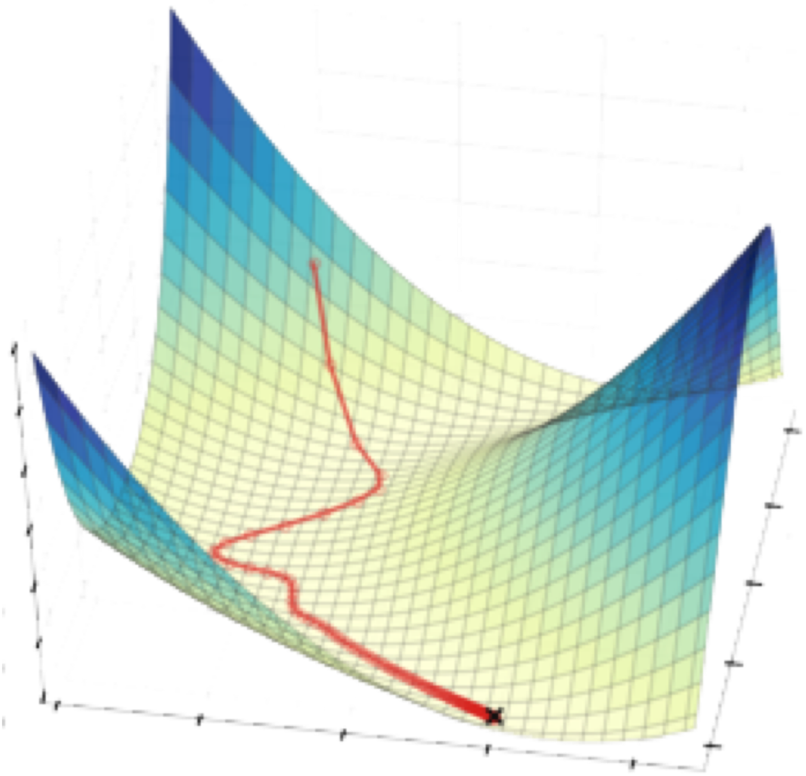


V. Mrazek, EvoApproxSb: Library of approximate adders and multipliers for circuit design and benchmarking of approximation methods, 2017

Cross-layer optimization – Algorithm

- Algorithm level
  - Optimized greedy search: discrete gradient descent
  - Large discrete parameter space
  - Regard output error as gradient
  - Choose the parameter with the least error at each order

$\{p_n\}$	0,	1,	2,	3,	4,	5,	5
$\{s_n \cdot q_n\}$	0,	0,	1,	2,	-3,	4,	5



Department of Electrical and Computer Engineering
UNIVERSITY OF WISCONSIN–MADISON

Cross-layer optimization – Algorithm

- Circuit level
  - Input distribution aware
  - Weighted output error

$$\overline{WMRE} = \sum_{m=0}^{2^M-1} P_m \cdot \left(\frac{\overline{\exp}(x_m)}{\exp(x_m)} - 1 \right)$$

  - Select proper multipliers from the verified library via error prediction
  - Exhaustive profiling on selected approximate multipliers

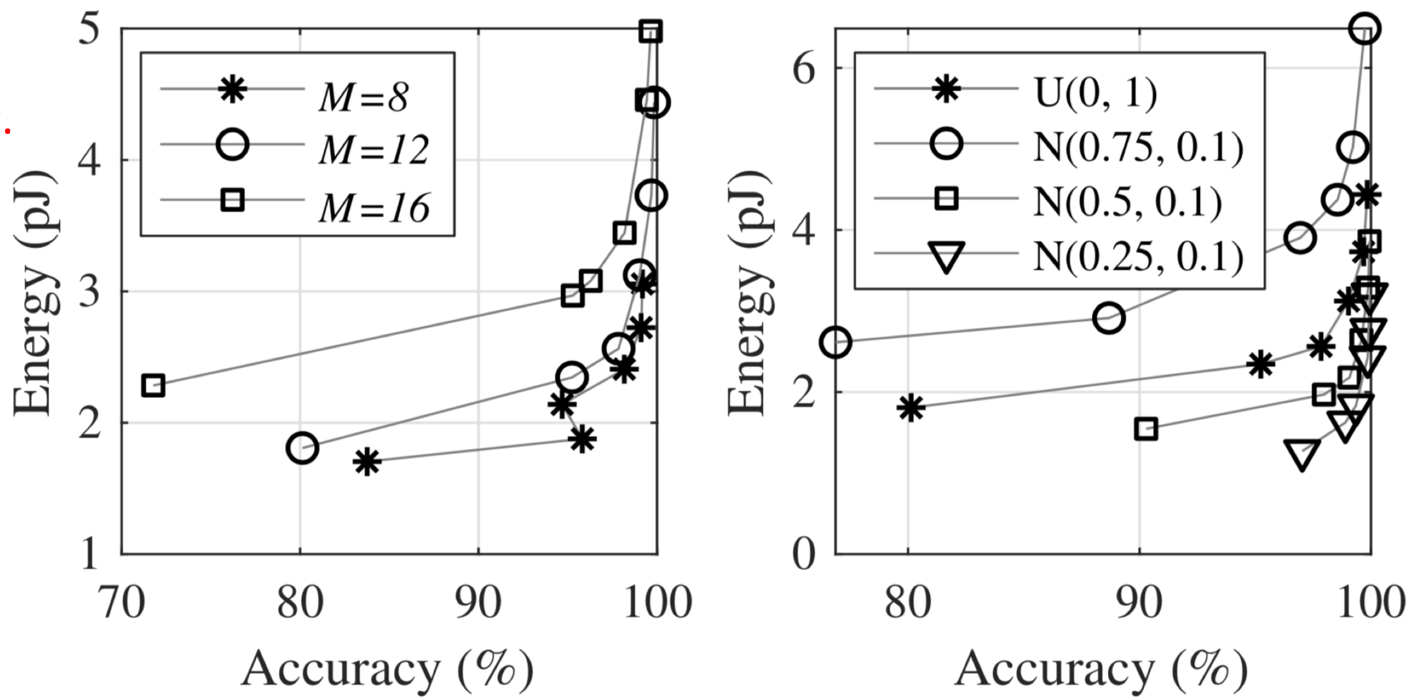
Outline

- ❑ Background
- ❑ Approximate Taylor series
- ❑ Cross-layer optimization
- ❑ **Performance evaluation**
- ❑ Case study on AdEx neuron
- ❑ Conclusion
- ❑ Future work

Performance evaluation – Hardware

- Static Design-time optimization
  a) Varying bitwidth M
  b) Varying input distribution (U~uniform, N~ Gaussian)

Each point refers to
a unique accuracy budget.



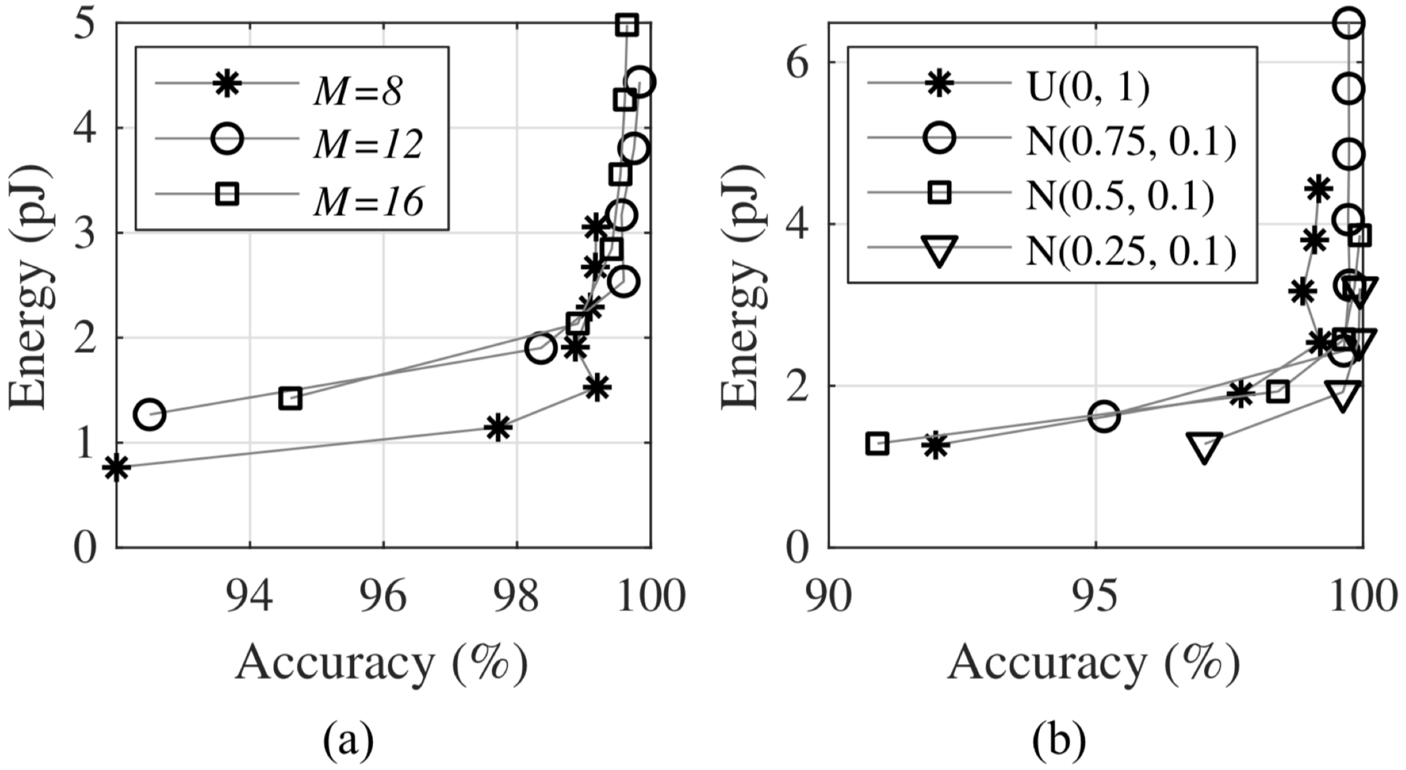
(a) (b)

Department of Electrical and Computer Engineering
UNIVERSITY OF WISCONSIN–MADISON

Performance evaluation – Hardware

- Dynamic run-time energy-accuracy scaling
  a) Varying bitwidth M
  b) Varying input distribution (U~uniform, N~Gaussian)

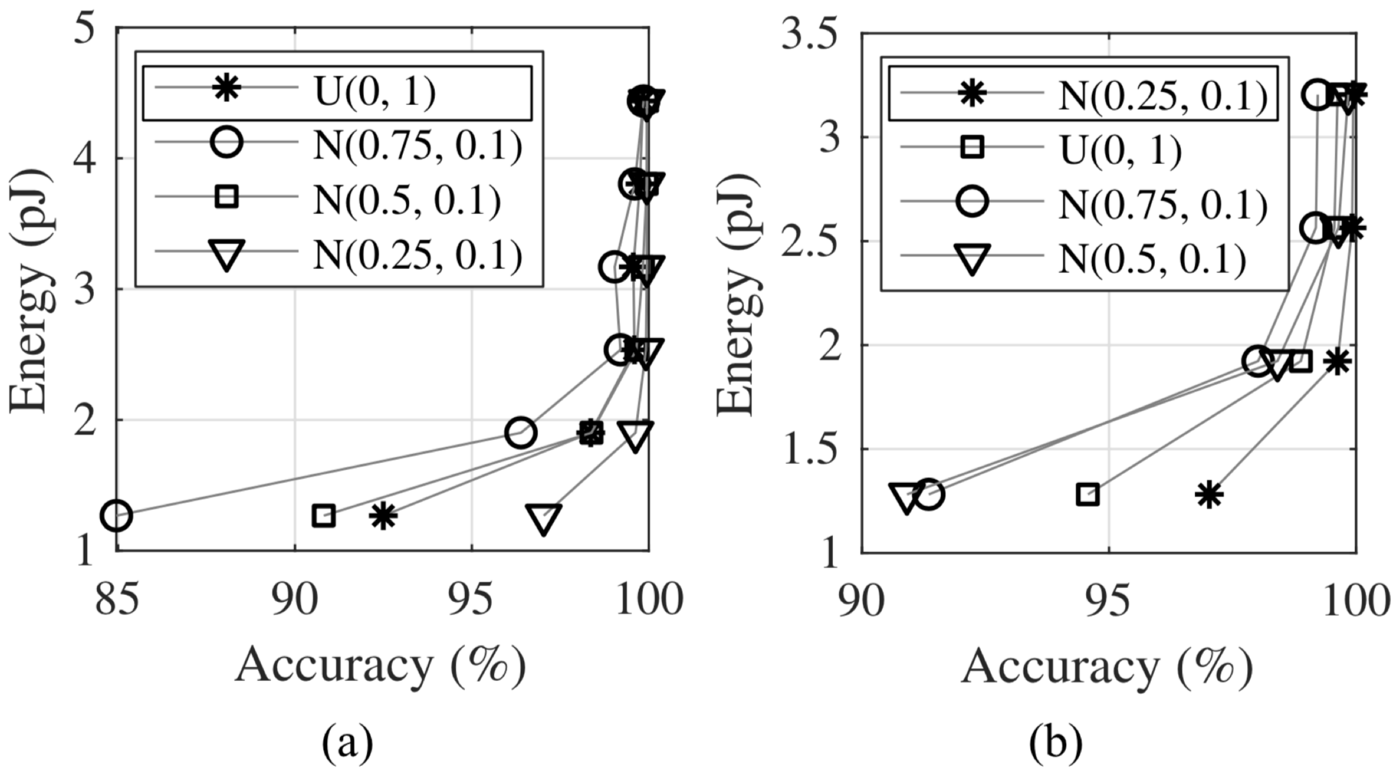
Each point refers to a cycle during computing. Right points have more terms.



Performance evaluation – Hardware

- Input variation affects dynamic run-time scaling
  a) Different inputs to the circuit for uniform~U(0, 1)
  b) Different inputs to the circuit for Gaussian~N(0.25, 0.1)

Each point refers to a cycle during computing. Right points have more terms.



Performance evaluation – Hardware

- Synthesized with Design Compiler
  - TSMC 45 nm vs STM 65 nm
  - Largest accuracy drop between 99.7% to 99.1%
  - 23% of original power
  - 5.4% of original area
  - 17.5% of original latency

Design	Accuracy const. (%)	Latency (ns)	Area (μm^2)	Power (mW)	Energy (pJ)
SECO	99.7	17.5	1,118	0.223	3.73
	99.1	20	611	0.136	2.72
	98.2	20	517	0.120	2.41
	95.2	20	378	0.094	1.88
	83.8	20	328	0.085	1.70
[1]	99.997	100	20,700	0.959	95.9

[1] P. Nilsson et al., "Hardware implementation of the exponential function using taylor series," in NORCHIP, 2014.

Department of Electrical and Computer Engineering
UNIVERSITY OF WISCONSIN–MADISON

Outline

- ❑ Background
- ❑ Approximate Taylor series
- ❑ Cross-layer optimization
- ❑ Performance evaluation
- ❑ **Case study on AdEx neuron**
- ❑ Conclusion
- ❑ Future work

Case study on AdEx neuron

- Adaptive Exponential (AdEx) Neuron Simulation
  - Key component in brain simulation
  - Fires spike after membrane potential crosses threshold
  - Differential equations model injection current and membrane potential.

$$C\frac{dV}{dt} = -g_L(V - E_L) + g_L \cdot \Delta_T \cdot \exp\left(\frac{V - V_T}{\Delta_T}\right) + I - w,$$

$$\tau_w \frac{dw}{dt} = a(V - E_L) - w,$$

$$\text{if } V > 0 \text{ then } \begin{cases} V \rightarrow V_r, \\ w \rightarrow w_r = w + b, \end{cases}$$

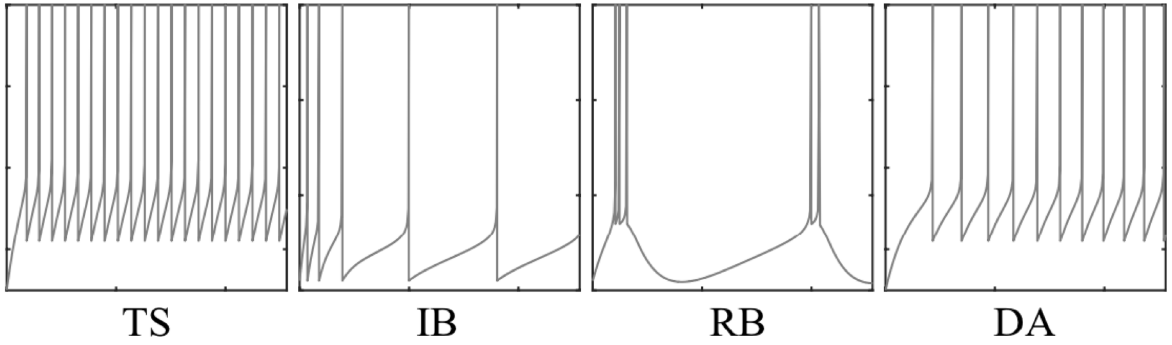
Department of Electrical and Computer Engineering
UNIVERSITY OF WISCONSIN–MADISON

Case study on AdEx neuron

- Spiking metrics
  - Timing error
    - Percent error of spike response time
  - Value error
    - Normalized root mean square deviation

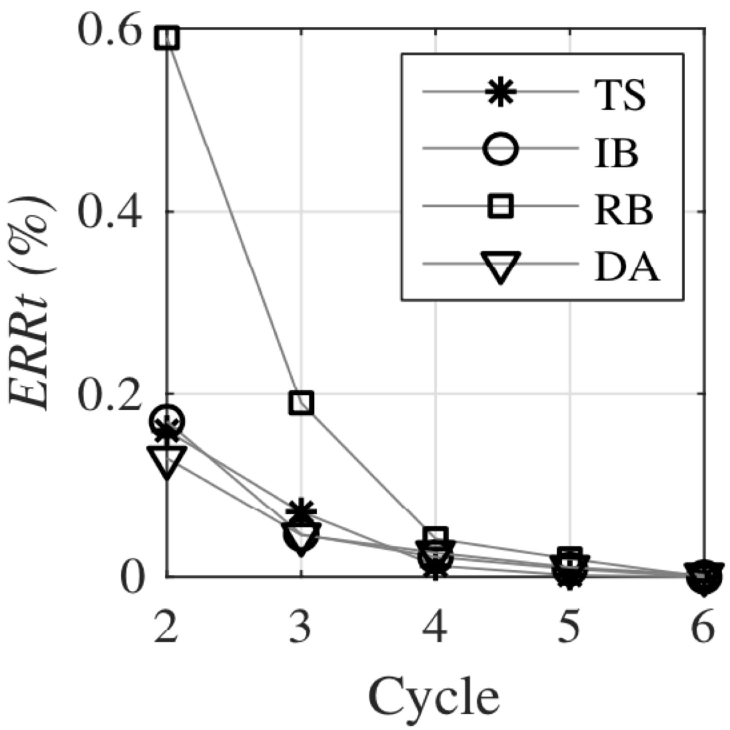
$$ERRt = \left| \frac{\Delta t_p - \Delta t_o}{\Delta t_o} \right| \times 100,$$

$$NRMSD = \sqrt{\frac{\sum_{i=1}^{n} (vp(i) - vo(i))^2}{n}} \cdot \frac{1}{v_{max} - v_{min}},$$

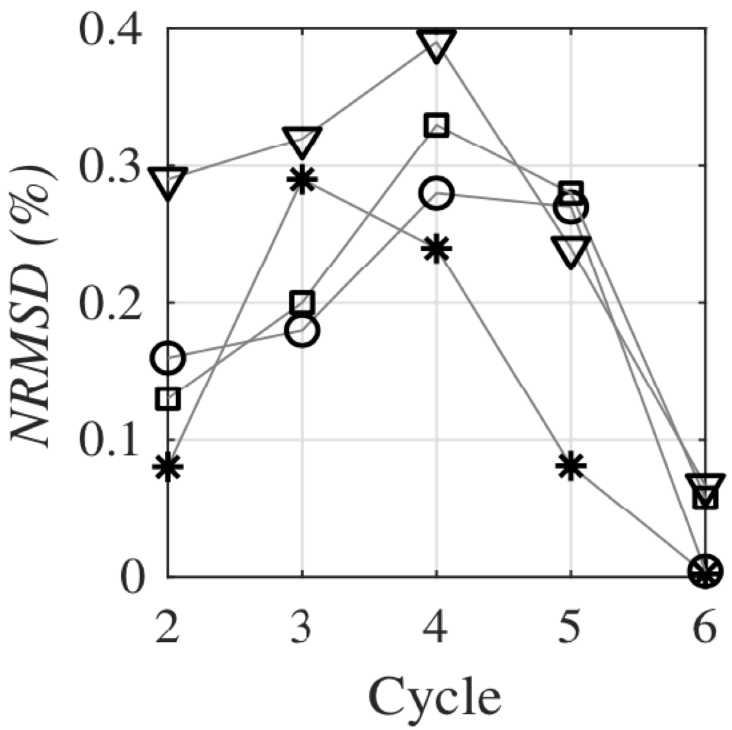


Department of Electrical and Computer Engineering
UNIVERSITY OF WISCONSIN–MADISON

Case study on AdEx neuron



Time Error



Value Error

Department of Electrical and Computer Engineering
UNIVERSITY OF WISCONSIN–MADISON

Outline

- ❑ Background
- ❑ Approximate Taylor series
- ❑ Cross-layer optimization
- ❑ Performance evaluation
- ❑ Case study on AdEx neuron
- ❑ **Conclusion**
- ❑ Future work

Department of Electrical and Computer Engineering
UNIVERSITY OF WISCONSIN–MADISON

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**

Outline

- ❑ Background
- ❑ Approximate Taylor series
- ❑ Cross-layer optimization
- ❑ Performance evaluation
- ❑ Case study on AdEx neuron
- ❑ Conclusion
- ❑ **Future work**

Further work

- Create **full processing unit** with combined approximate computing methods
- Evaluate on **full neural network**
- Explore **Binary Expansion** opposed to Taylor Series expansion

Department of Electrical and Computer Engineering
UNIVERSITY OF WISCONSIN–MADISON

Thank you!
Q & A

Di Wu, Tianen Chen, Chienfu Chen, Oghenefego Ahia, Joshua San Miguel, Mikko Lipasti, and Younghyun Kim
University of Wisconsin-Madison

Department of Electrical and Computer Engineering
UNIVERSITY OF WISCONSIN–MADISON