Automated Synthesis of Mixed-signal ML Inference Hardware under Accuracy Constraints

Kishor Kunal, Jitesh Poojary, S Ramprasath,
Ramesh Harjani, Sachin S. Sapatnekar

University of Minnesota

Minneapolis, MN

UNIVERSITY OF MINNESOTA

Driven to Discover®

Machine learning is changing lives

AlphaGo

bridge-global.com/blog

Traffic prediction

nytimes.com/2012

Self-driving car

insights.dice.com/2019

Virtual assistants

Winner is AlphaGo!

AlphaGo uses 1920 CPUs and 280 GPUs

https://cdn.ndtv.com

20 Watts

1 vs 50000

Image: Reuters/Pawel Kopczynski

1MW

Another Way Of Looking At Lee Sedol vs AlphaGo · Jacques Mattheij

Bringing smartness in your hands

miro.medium.com

Require energy efficient models!

Agenda

- Motivation
- Mixed Signal Optimization for low power ML (MiSOML)
  - Common ML operations in analog domain
  - ML model performance with noise
  - ILP formulation
- Results
- Conclusion

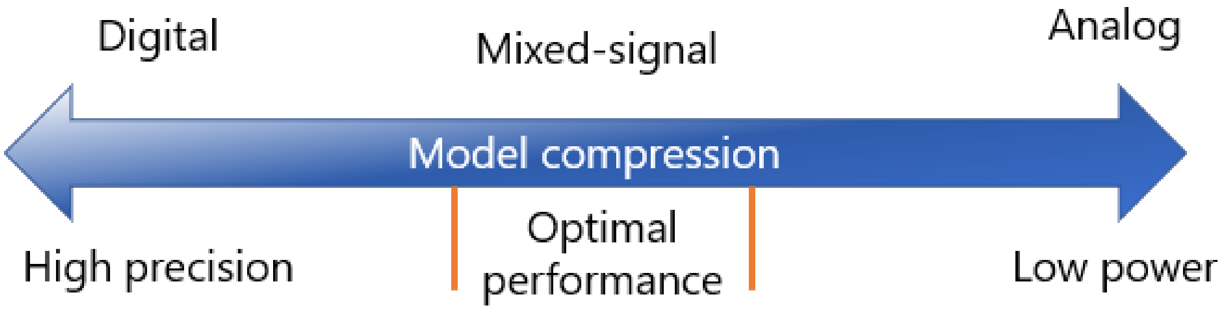
Efforts in reducing energy of ML models

Digital design optimization

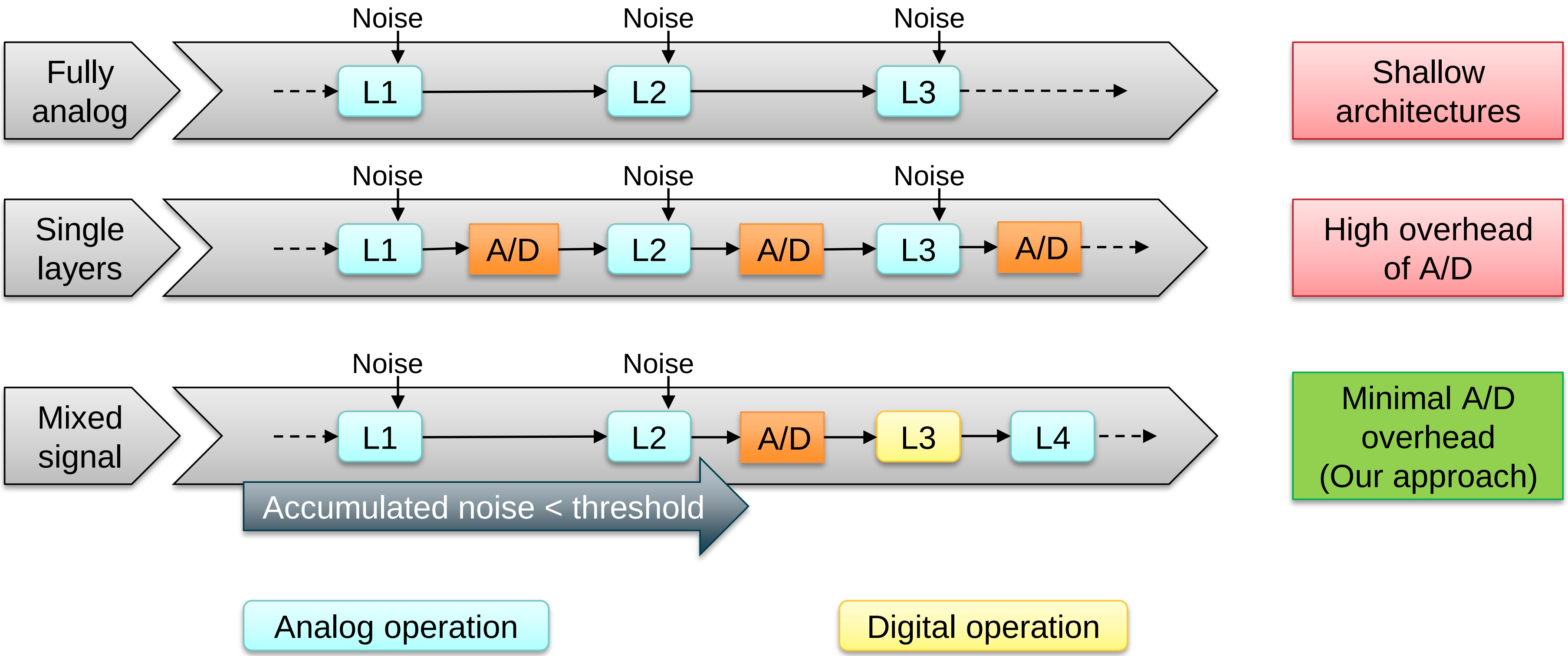
- Model compression
- Quantization
- Approximate computations

New contenders-analog designs

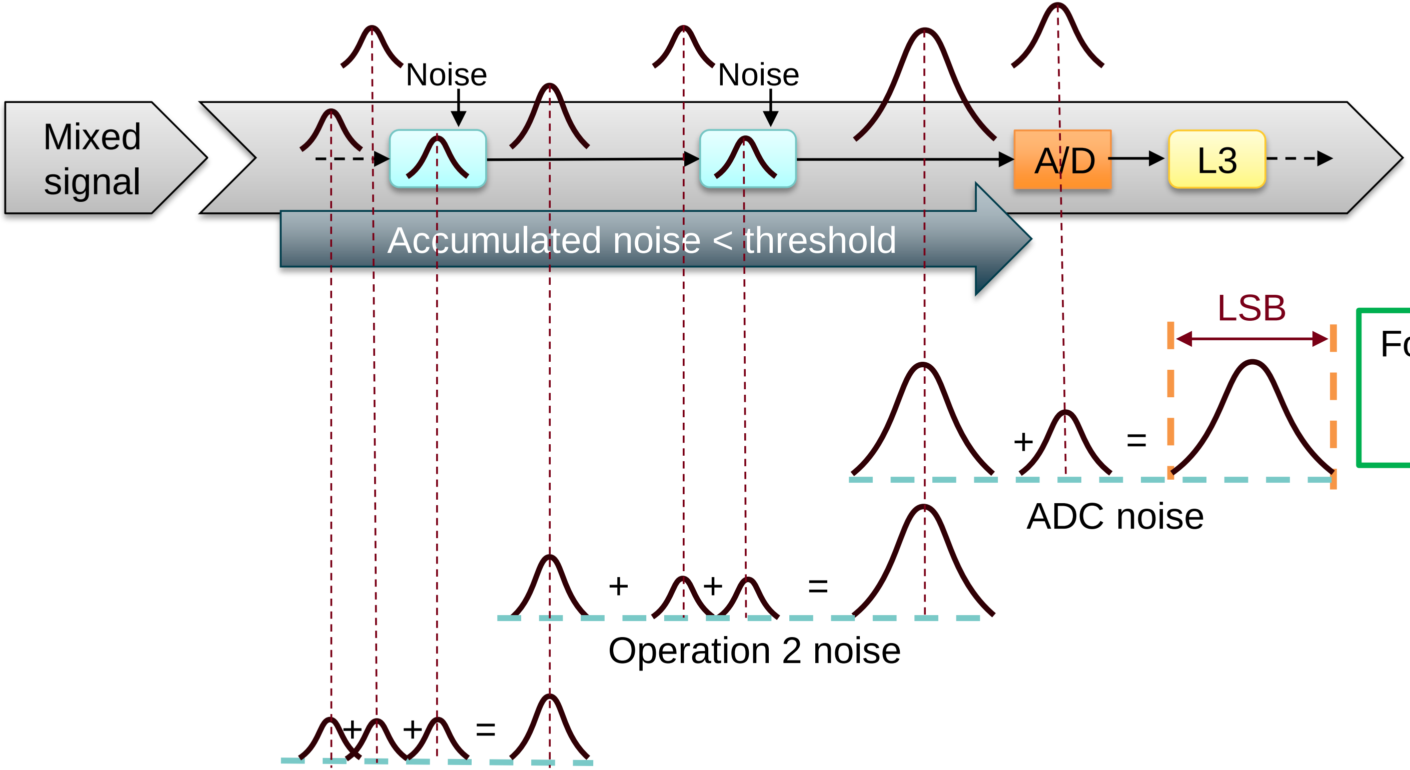
- In-Memory computation
- **Analog computation**



Challenges of analog computation



Error propagation

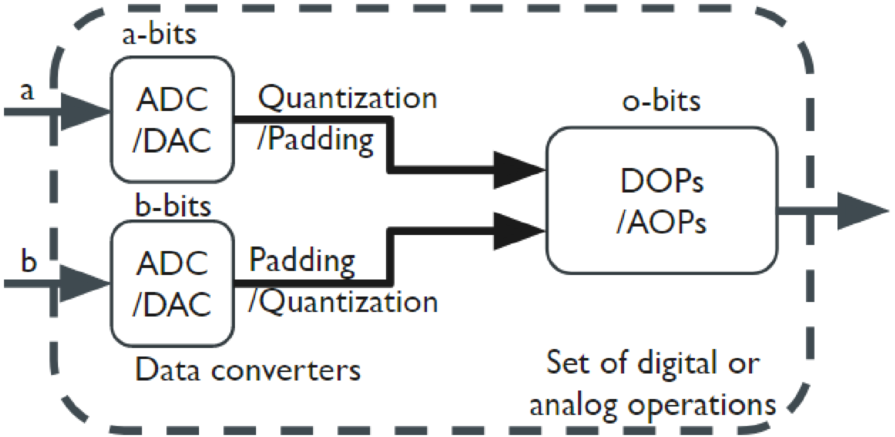


For signal restoration
SNR_{ADC}> 0.5LSB
(threshold)

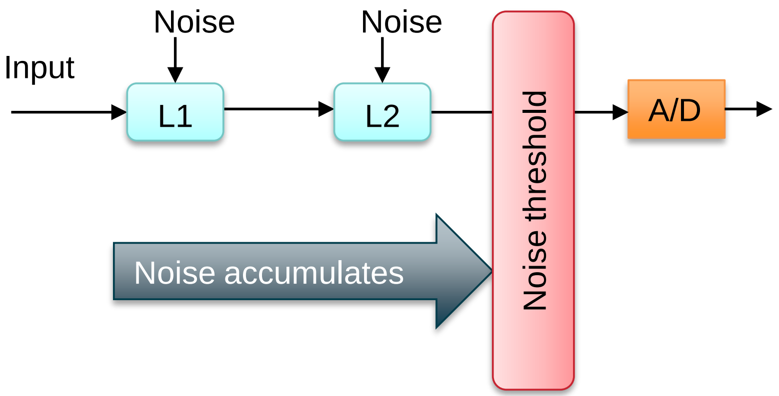
Agenda

- Motivation
- Mixed Signal optimization for low power ML (MiSOML)
  - Common ML operations in Analog domain
  - ML model performance with noise
  - ILP formulation
- Results
- Conclusion

MiSOML: building block



MiSOML unit

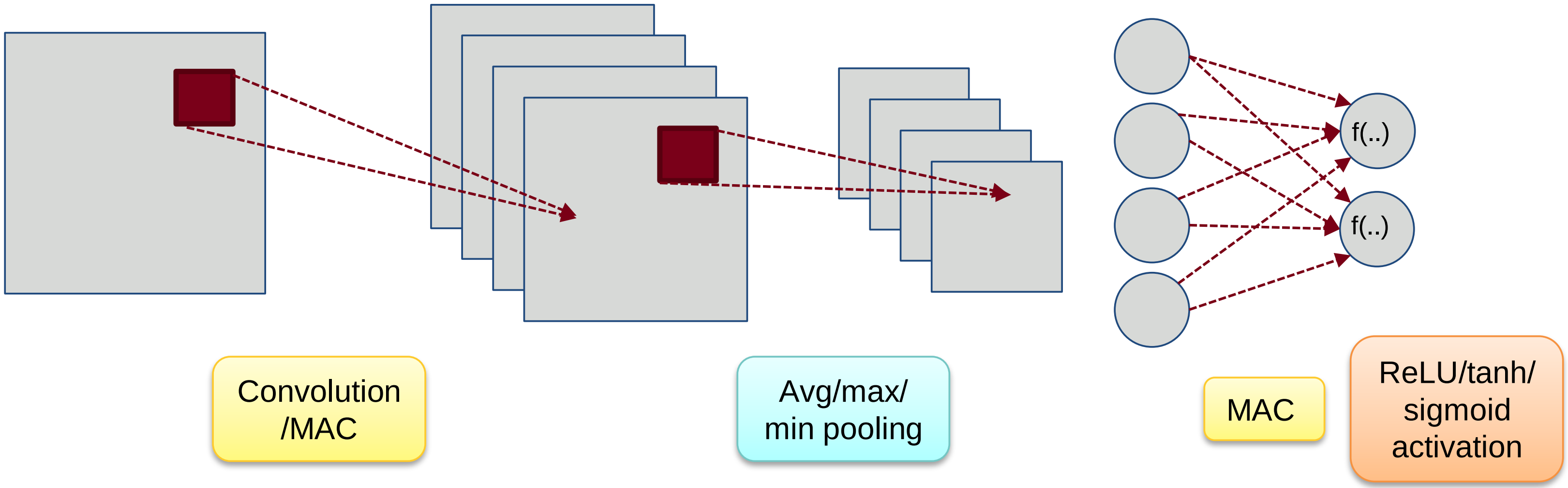


Basic idea

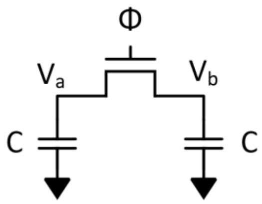
Agenda

- Motivation
- Mixed Signal optimization for low power ML (MiSOML)
  - Common ML operations in Analog domain
  - ML model performance with noise
  - ILP formulation
- Results
- Conclusion

Common ML operations

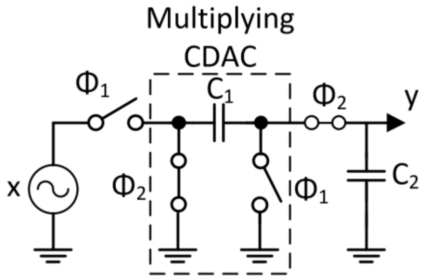


ML operations in analog domain



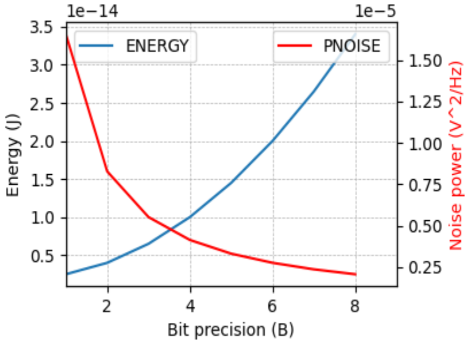
$$E_{\mathrm{DAC}} = 2^B \alpha (C_u V_{DD}^2 + E_{switch})$$
$$N_{\mathrm{DAC}} = kT/(2^B C_u)$$

B. Sadhu, JSSC 2013

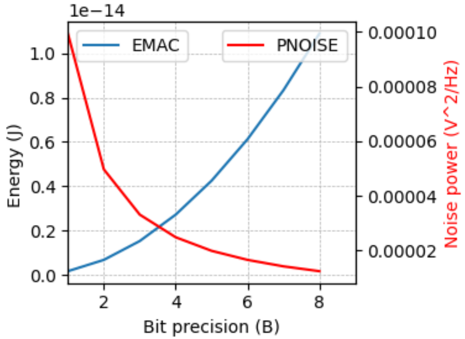


$$E_{AMAC} = 4E_{switch} + (C_1 + C_2)V_{DD}^2$$
$$N_{AMAC} \approx kT/C_2;$$

E.H.Lee, JSSC 2017



Energy and noise vs precision

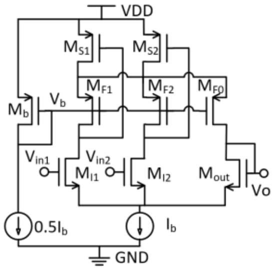


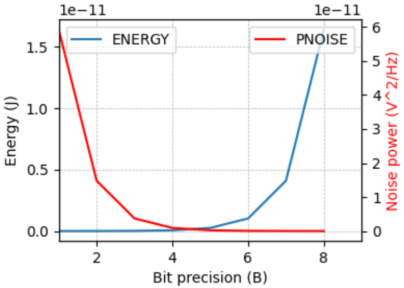
Energy and noise vs precision

<u>Addition/subtraction</u>

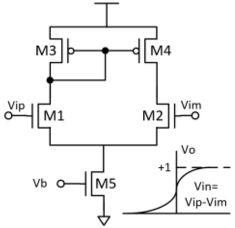
<u>Multiply-accumulate (MAC)</u>

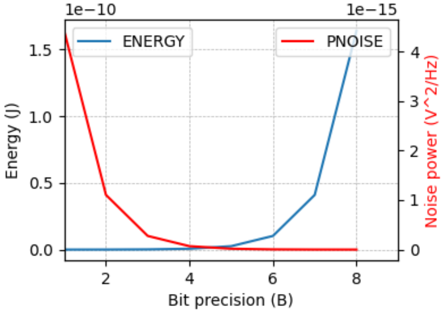
Other ML operations in analog domain





J. Choi, MDPI sensors journal 2020

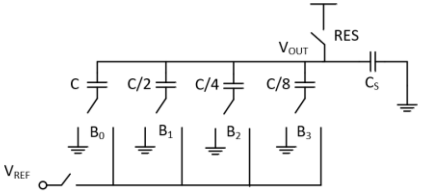


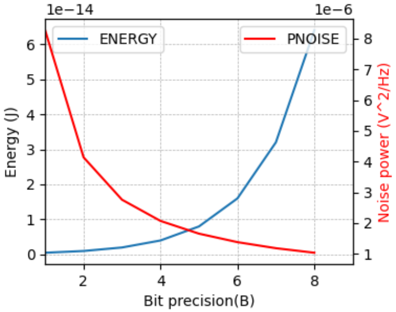


S. Sadasivuni, Scientific reports, 2022

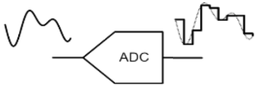
Max-pooling

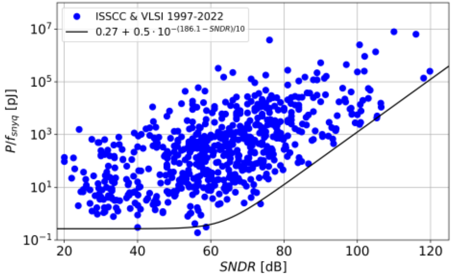
sigmoid/tanh activation





B. Razavi, 2001





https://github.com/bmurmann/ADC-survey

Capacitive DAC

SAR ADC

Energy and noise exhibit exponential relationship with precision

Summary: Energy lookup table

Operations	Analog Energy (fJ)							
Precision	**1-bit**	**2-bit**	**3-bit**	**4-bit**	**5-bit**	**6-bit**	**7-bit**	**8-bit**
ADC	100	200	300	400	501	624	747	885 [9]
Addition	2.5	4.0	6.5	10.0	14.5	20.0	26.5	34.0 [12]
DMAC	0.2	0.7	1.5	2.7	4.2	6.1	8.3	10.8 [9]
Max-pool	1	4	16	64	2.6e2	1.0e3	4.1e3	1.6e4 [18]
sigmoid	1	4	16	64	2.7e2	1.0e3	4.1e3	1.6e4 [19]

Operations	Digital Energy (fJ)							
Precision	**1-bit**	**2-bit**	**3-bit**	**4-bit**	**5-bit**	**6-bit**	**7-bit**	**8-bit**
DAC	0.5	1.0	2.0	4.0	8.0	16.0	32.0	64.0 [16]
Addition	9.5	12.5	15.4	18.3	21.2	24.1	27.0	30.0 [14]
DMAC	9.7	18.7	33.8	55.0	82.1	115.4	154.6	200 [14]
Max-pool	3.7e2	7.4e2	1.1e3	1.5e3	1.9e3	2.2e3	2.6e3	2.9e3 [14]
sigmoid	2.3e3	3.1e3	3.9e3	4.7e3	5.5e3	6.2e3	7.0e3	7.8e3 [15]

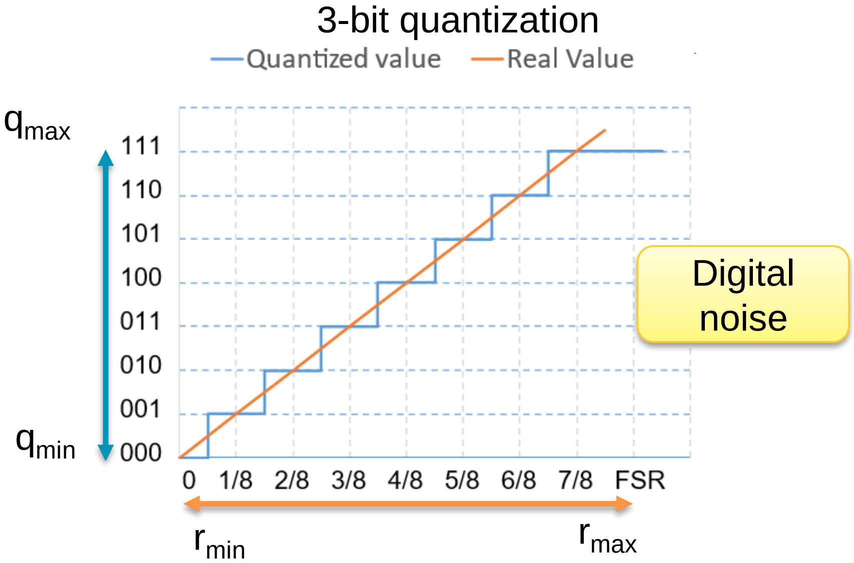
Signal conversion are costly

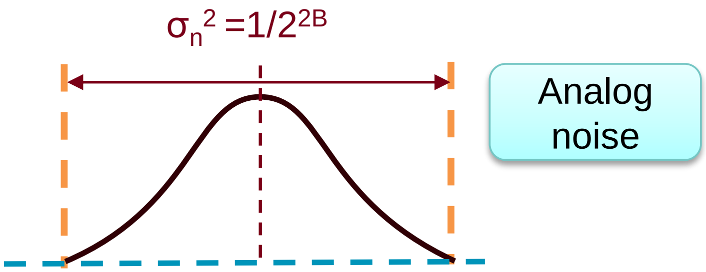
Agenda

- Motivation
- Mixed Signal optimization for low power ML (MiSOML)
  - Common ML operations in Analog domain
  - ML model performance with noise
  - ILP formulation
- Results
- Conclusion

Noise variance

- Let B = equivalent bit precision
- Quantization noise
  - Noise = real value/ step size – quantized value
  - Noise variance,$\sigma_q^2 \sim= 1/2^{2B}$
- Analog signal noise
  - White Gaussian noise
  - Noise variance, $\sigma_n^2 = 1/2^{2B}$

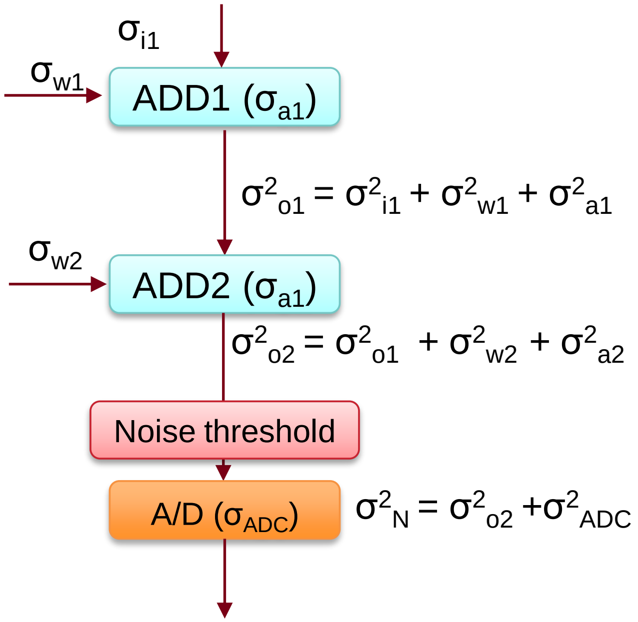




Modeling analog noise propagation

- Operations:
  - Addition: $\sigma^2 = \sigma_1^2 + \sigma_2^2$
  - Multiplication: $\sigma^2 = \sigma_1^2 \sigma_2^2$
  - ReLU: $\sigma^2 = \sigma_1^2 / 2$
  - Sigmoid: $\sigma^2 \approx \frac{\arctan\left(\sqrt{1+0.59\sigma_1^2}\right)}{\pi} - \frac{1}{4}$
- Noise threshold: $\sigma_N^2 < N_T$

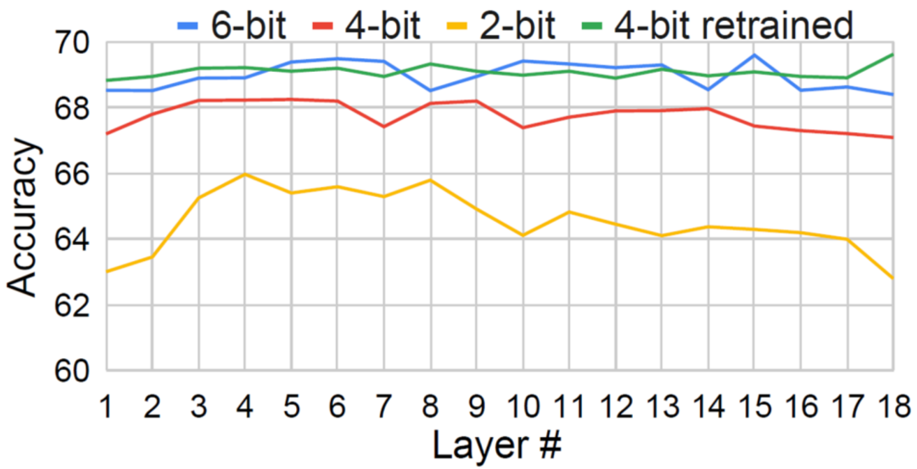
Noise accumulation across multiple layers



ResNet18 performance with noise

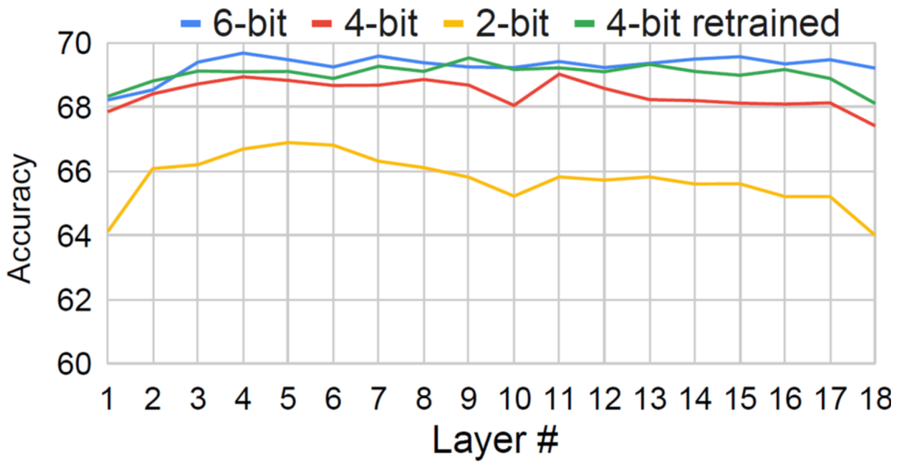
<u>Quantization noise</u>

Baseline ResNet18 FP32 =69.57%



<u>Analog Gaussian noise</u>

Baseline ResNet18 FP32 =69.57%



ResNet18 is more resilient to analog noise than to quantization loss

Agenda

- Motivation
- Mixed Signal optimization for low power ML (MiSOML)
  - Common ML operations in Analog domain
  - ML model performance with noise
  - ILP formulation
- Results
- Conclusion

Integer linear programming (ILP) optimization

Objective

$$min \sum_{j=1}^{M} \left(E(O_j^t) + S_j^t N(O_j^t)\right) + E_{ADC}^{total} + E_{DAC}^{total}$$

Layer operation energy
+ noise × layer noise sensitivity
+ domain switching energy

Energy

$$E(O_j^t) = \sum_{B=1}^{8} \left(I_j^{A_B^t} \cdot E(A_B^t) + I_j^{D_B^t} \cdot E(D_B^t) \right)$$

$$E_{ADC}^{total} = \sum_{j=1}^{M} \sum_{B=1}^{8} I_j^{ADC_B} \cdot E(ADC_B)$$

Energy calculation

Noise

$$\left(\bigwedge_{k \in P} I_k^{A_{B_k}^t} \cdot \sum_{k \in P} N(A_{B_k}^t) \right) \cdot I_l^{D_{B_l}^t} \leq N_T$$

Noise calculation

Constraint

$$\sum_{B=1}^{8} \left(I_j^{A_B^t} + I_j^{D_B^t} \right) = 1$$

$$I_j^{\mathrm{ADC}_B} = I_j^{A_B^t} \wedge \left(\bigvee_{k \in FO(j)} \left(\bigvee_{B=1}^{8} I_k^{D_B^k} \right) \right)$$

0–1 indicator variable:
Analog or digital?

Integer linear programming optimization

Objective

$$min \sum_{j=1}^{M} \left(E(O_j^t) + S_j^t N(O_j^t)\right) + E_{ADC}^{total} + E_{DAC}^{total}$$

Energy

$$E(O_j^t) = \sum_{B=1}^{8} \left(I_j^{A_B^t} \cdot E(A_B^t) + I_j^{D_B^t} \cdot E(D_B^t)\right)$$

$$E_{ADC}^{total} = \sum_{j=1}^{M} \sum_{B=1}^{8} I_j^{ADC_B} \cdot E(ADC_B)$$

Noise

$$\left(\bigwedge_{k \in P} I_k^{A_{B_k}^t} \cdot \sum_{k \in P} N(A_{B_k}^t)\right) \cdot I_l^{D_{B_l}^t} \leq N_T$$

Constraint

$$\sum_{B=1}^{8} \left(I_j^{A_B^t} + I_j^{D_B^t}\right) = 1$$

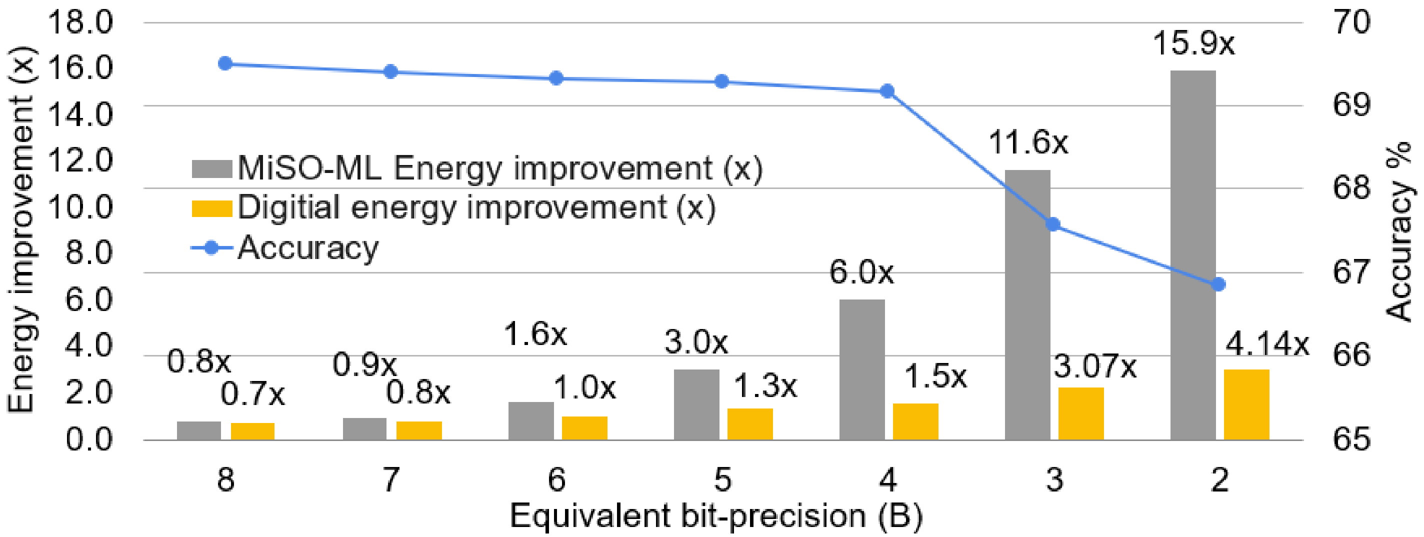
$$I_j^{\mathrm{ADC}_B} = I_j^{A_B^t} \wedge \left(\bigvee_{k \in FO(j)} \left(\bigvee_{B=1}^{8} I_k^{D_B^k}\right)\right)$$

t	Operator type; $t \in \mathcal{O}$
O_j^t	j^{th} operator in the network of type t.
$ADC_B(DAC_B)$	B-bit ADC (DAC); $1 \leq B \leq 8$
$D_B^t(A_B^t)$	B-bit precision digital(analog) operator of type t
$I_j^X \in \{0,1\}$	If X is A_B^t (D_B^t), this indicator variable chooses variant X of j^{th} operator, O_j^t; else if X is ADC_B (DAC_B), this indicates the presence(absence) of an $ADC_B(DAC_B)$ at output of j^{th} operator, O_j^t.
$E(X)(N(X))$	Energy consumed (Noise generated) by an operator X in Table I.
$E_{\mathrm{ADC}}^{\mathrm{total}}(E_{\mathrm{DAC}}^{\mathrm{total}})$	Total energy consumed by all ADCs (DACs).
S_t^j	Noise sensitivity for j^{th} operator of type t.
$FO(X)$	the fan-out vertices of operator X
N_T	Noise threshold beyond which successive analog operations has to be succeeded by a digital operation.

Agenda

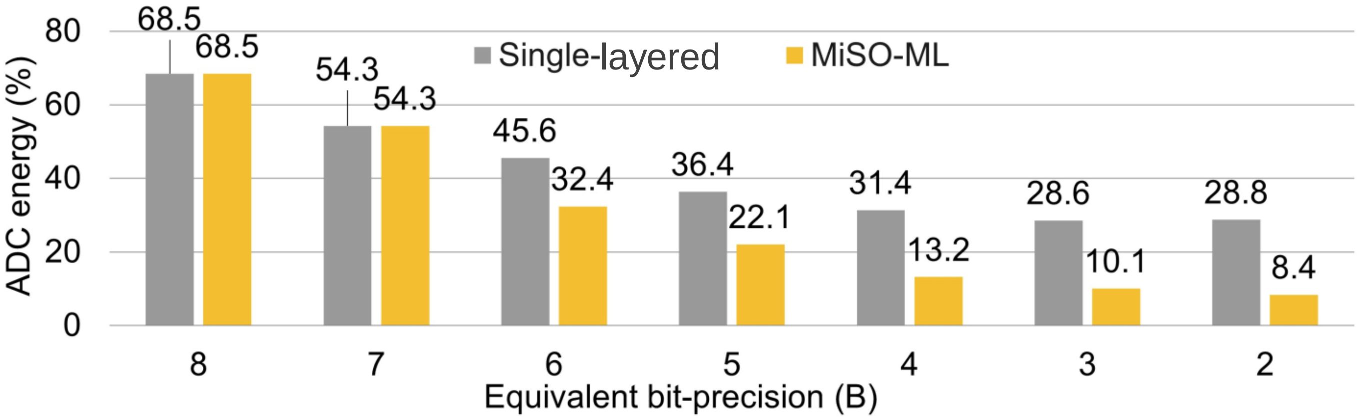
- Motivation
- Mixed Signal optimization for low power ML (MiSOML)
  - Common ML operations in Analog domain
  - ML model performance with noise
  - ILP formulation
- Results
- Conclusion

Energy improvement vs. accuracy trade-off



Energy vs. accuracy trade-off for different noise threshold constraints for ResNet18

ADC energy vs. noise threshold constraints



ADC energy vs. noise threshold constraints for ResNet18
**Gray columns indicate ADC inserted after each layer

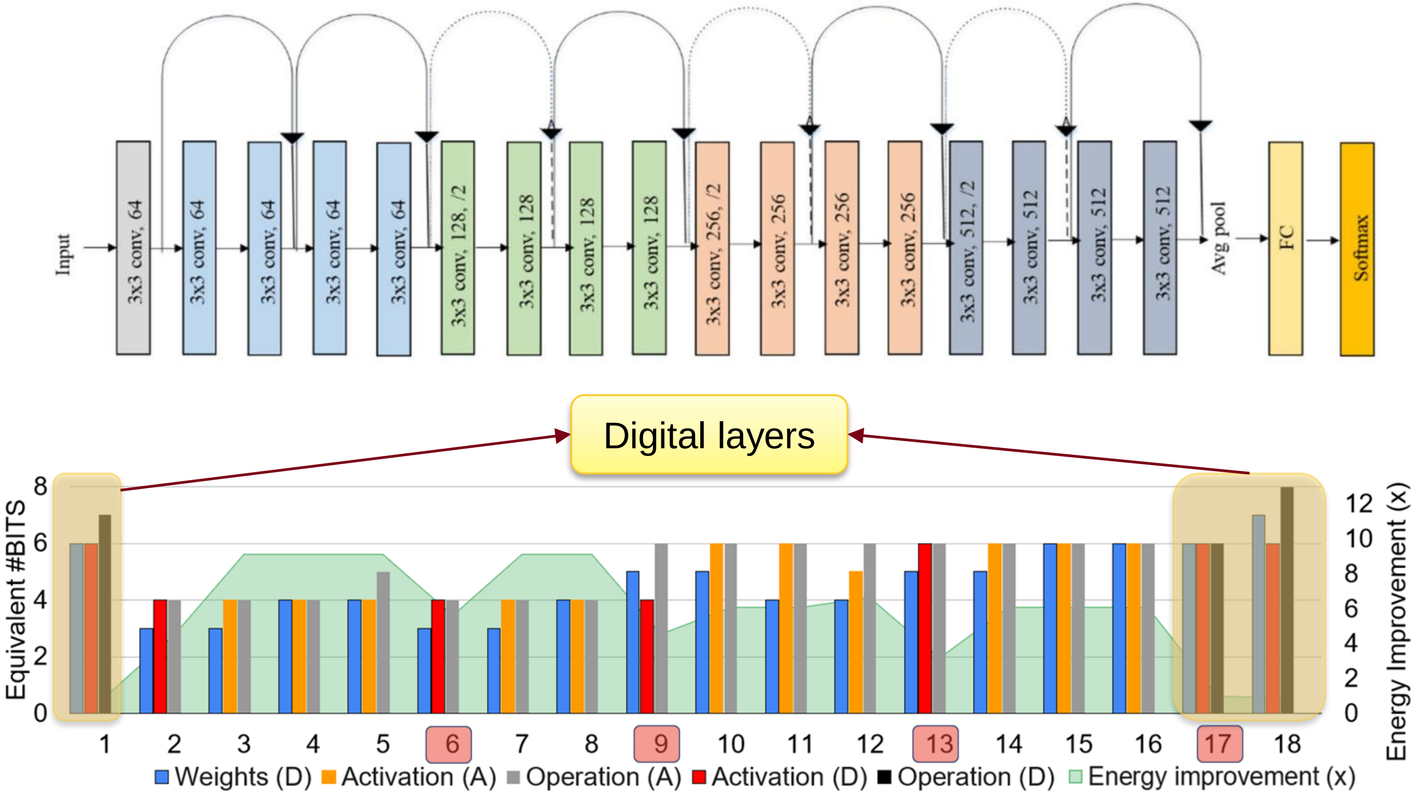
MiSOML: Energy reduction-ResNet18

- Mixed signal hardware is more energy efficient than digital solutions
- Noise threshold of 4-bit provides best energy for nominal accuracy loss

Precision (D) = Digital, (A) = Analog	**Accuracy**		**Energy** **(Improvement)**
	Inference	*Fine tuning*	
FP32 (D)	69.57%	–	–
8B Activation, 8B Weight (D)	69.49%	69.49%	25.62μJ (1.00×)
6B Activation, 6B Weight (D)	68.43%	69.21%	18.74μJ (1.37×)
4B Activation, 4B Weight (D)	66.82%	68.66%	12.32μJ (2.08×)
2B Activation, 2B Weight (D)	62.14%	65.38%	6.18μJ (4.14×)
MiSO-ML (4/6B Activation, 4/6B Weight (mix))	**67.14%**	**69.16%**	**3.14μJ (8.16×)**

8x energy efficiency

Mapping ResNet18 architecture to hardware



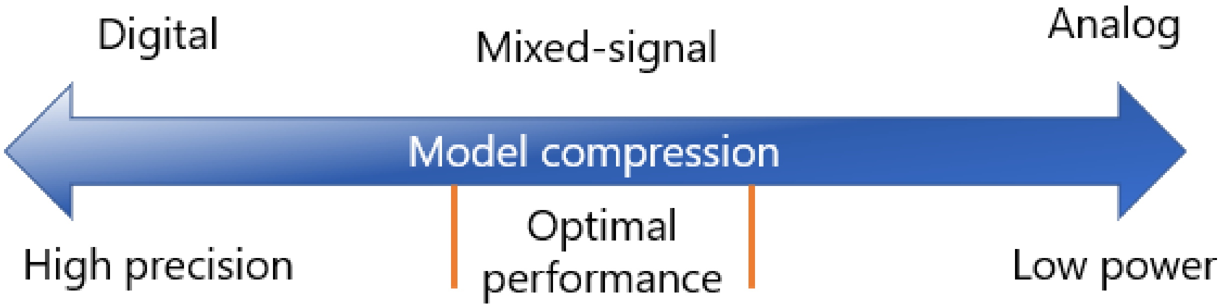
MiSOML: Energy reduction-ML architectures

Dataset	ML Architecture	Accuracy		Energy/inference	
MNIST	SVM (Accuracy = 94.23%)	MiSO-ML	92.31%	0.16µJ	(3.16×)
ImageNet	ResNet18 (Accuracy = 69.57%)	BitBlade [6]	68.66%	15.8µJ	(1.64×)
		MiSO-ML	**69.16%**	**3.14**µJ	**(8.16×)**
Imagenette	VGG16 (Accuracy = 92.32%)	MiSO-ML	90.51%	8.48µJ	(7.69×)
	GoogLeNet (Accuracy = 93.23%)	MiSO-ML	91.46%	2.56µJ	(6.25×)
	ResNet101 (Accuracy = 96.55%)	MiSO-ML	93.76%	5.08µJ	(9.12×)
CIFAR-10	ResNet20 (Accuracy = 91.12%)	AA-ResNet [7]	80.90%	**0.60**µJ	(5.76×)
		MiSO-ML	**87.25%**	0.66µJ	(5.52×)
	ResNet110 (Accuracy = 93.55%)	MiSO-ML	92.55%	2.89µJ	(5.56×)

5x – 8x energy efficiency

Summary

- Mixed signal optimization for ML inference
  - Modeled energy and noise for analog operations
  - ILP-based optimization
  - Analog circuits provide 5x-8x energy efficiency



Thank You