上海科技大学
ShanghaiTech University

Energy-Efficienct AI Acceleration through Approximation

ShanghaiTech University
School of Information Science and Technology
Speaker: Dr. Siting Liu

立志成才 报国裕民

Outline

- Motivation
- Background
  - Approximation by quantization
- Hardware approximation
  - Stochastic computing (SC)
  - SC for neural network (NN) inference
  - SC for NN training
- Summary

Motivation

```
sitingliu — python — 64×20
(base) sitingliu@Sitings-MacBook-Air ~ % python
Python 3.9.13 (main, Aug 25 2022, 18:29:29)
[Clang 12.0.0 ] :: Anaconda, Inc. on darwin
Type "help", "copyright", "credits" or "license" for more inform
ation.
>>> 0.1+0.2
```

Motivation

sitingliu — python — 64×20

```
(base) sitingliu@Sitings-MacBook-Air ~ % python
Python 3.9.13 (main, Aug 25 2022, 18:29:29)
[Clang 12.0.0 ] :: Anaconda, Inc. on darwin
Type "help", "copyright", "credits" or "license" for more inform
ation.
>>> 0.1+0.2
0.30000000000000004
>>>
```

Number Representation

- Fractions
  - Floating-point numbers (IEEE 754 standard single precision)

1 sign bit	8 bits	23 bits
S	Exponent	Significand/Mantissa

$$(-1)^{S} \times (1.\mathrm{Significand})_2 \times 2^{\mathrm{Exponent}-127}$$

  - Limited precision & rounding introduce accuracy loss

$$
\begin{array}{r}
1.11111111 \\
+\,0.000011111111 \\
\hline
1.111111111111
\end{array}
$$

Motivation

- Digital systems are inherently inaccurate;
- AI applications can tolerate plenty inaccuracy;
  - Resilience from application and computation patterns.

App. no single correct answer

- E.g. recommendation systems, search query, etc.
- Optimization problems such as neural network training.

Computation patters

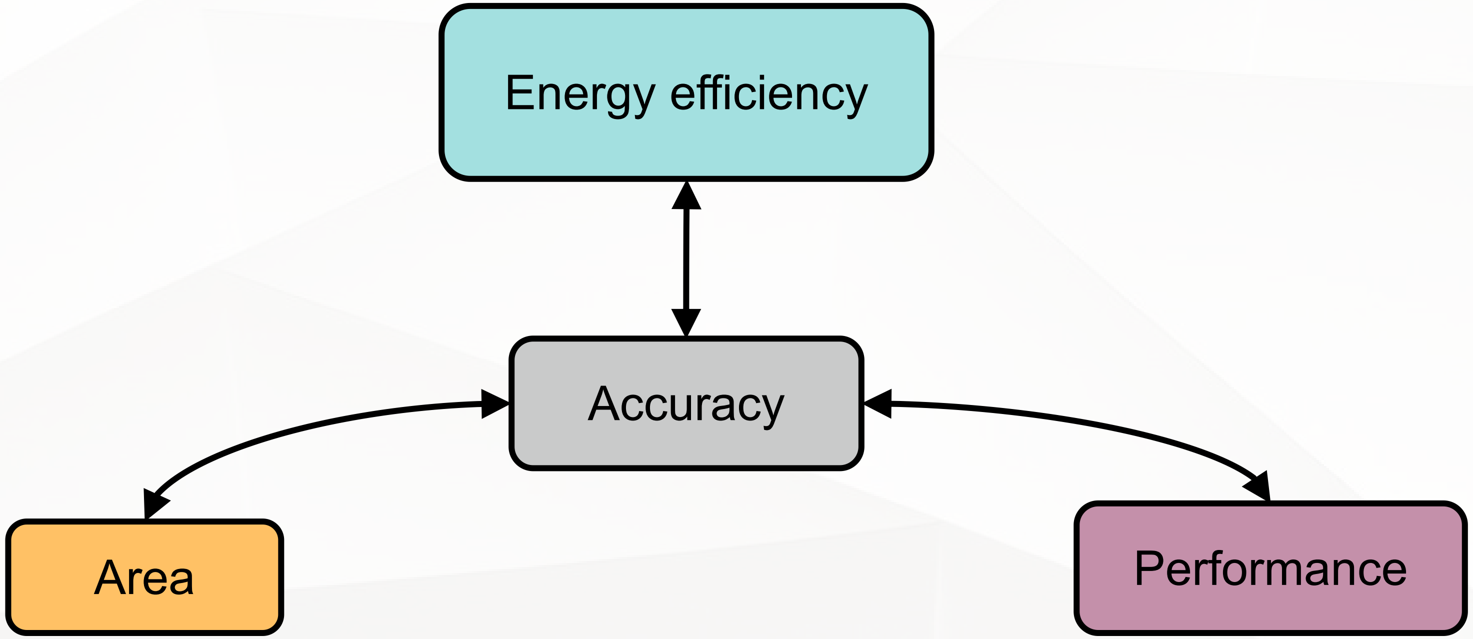
- Map complex features (images, text) to several classes.
- Nonlinear functions have strong saturation effects.

Motivation

上海科技大学 ShanghaiTech University

- Digital systems are inherently inaccurate;
- AI applications can tolerate plenty inaccuracy;
- Trade accuracy for energy efficiency, performance, area, etc. through

Approximation



Approximation at Different Levels

Software/model

- Pruning
- Quantization
- Distillation
- Low-rank approximation
- ... ...

Hardware/architecture

- Analog computing
- Approximate arithmetic circuits
- Stochastic computing
- ... ...

Approximation at Different Levels

Software/model

- Pruning
- **Quantization**
- Distillation
- Low-rank approximation
- ... ...

Hardware/architecture

- Analog computing
- Approximate arithmetic circuits
- **Stochastic computing**
- ... ...

Quantization——Motivation

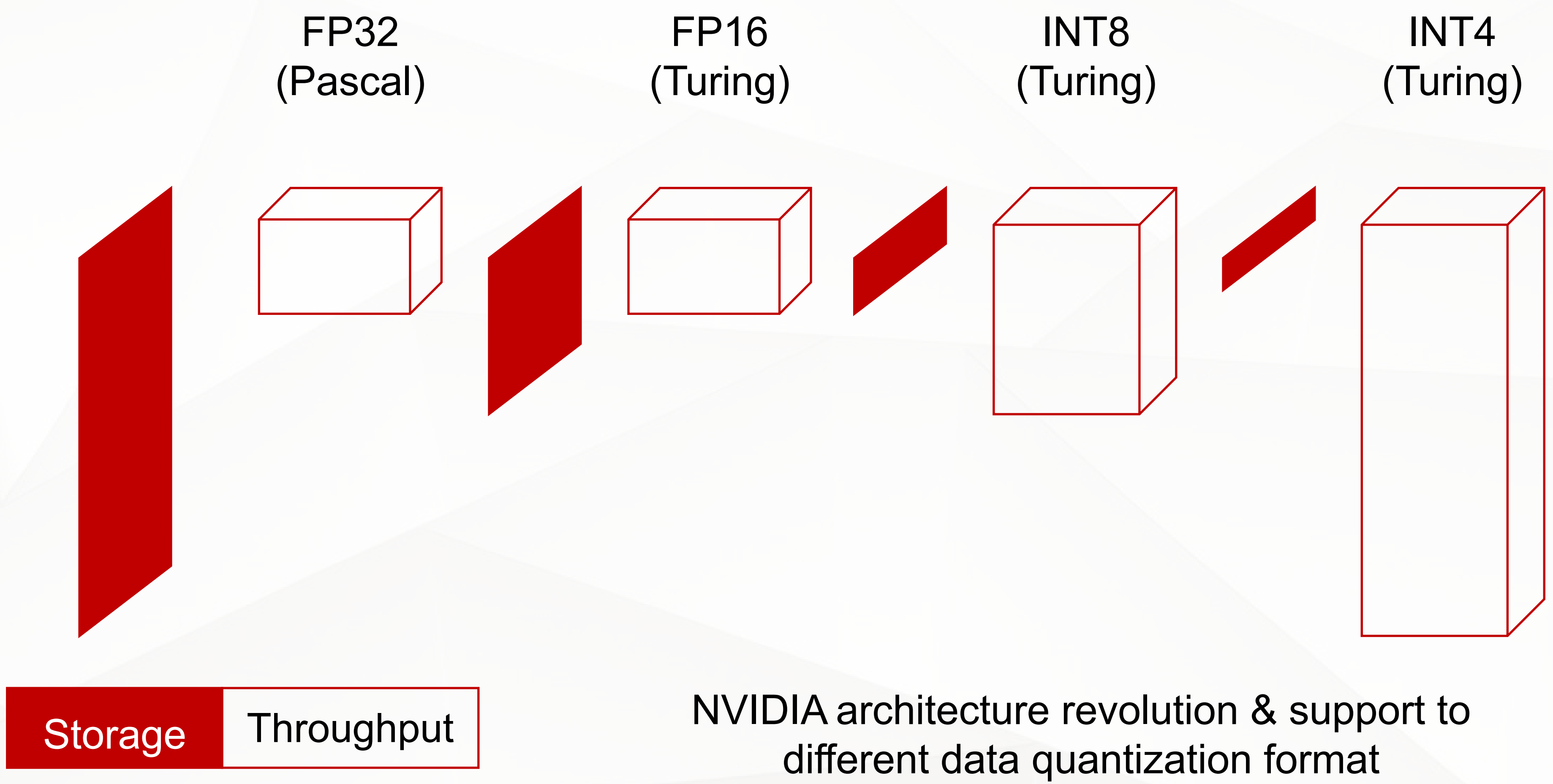
- Rough energy numbers (in 45-nm technology node) from "Computing's Energy Problem, M. Horowitz, ISSCC, 2014"

INT	
ADD	
8 bit	0.03 pJ
32 bit	0.1 pJ
MULTI	
8 bit	0.2 pJ
32 bit	3 pJ

FP	
FADD	
16 bit	0.4 pJ
32 bit	0.9 pJ
FMULTI	
16 bit	1 pJ
32 bit	4 pJ

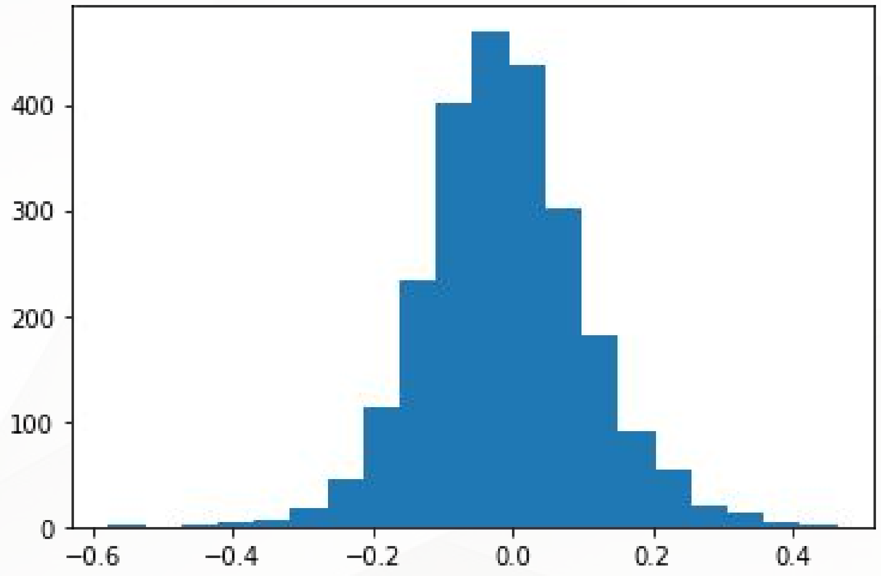
Memory	
Cache	(64 bit)
8 KB	10 pJ
32 KB	20 pJ
1 MB	100 pJ
DRAM	1.3-2.6 nJ

Quantization——Hardware Implication

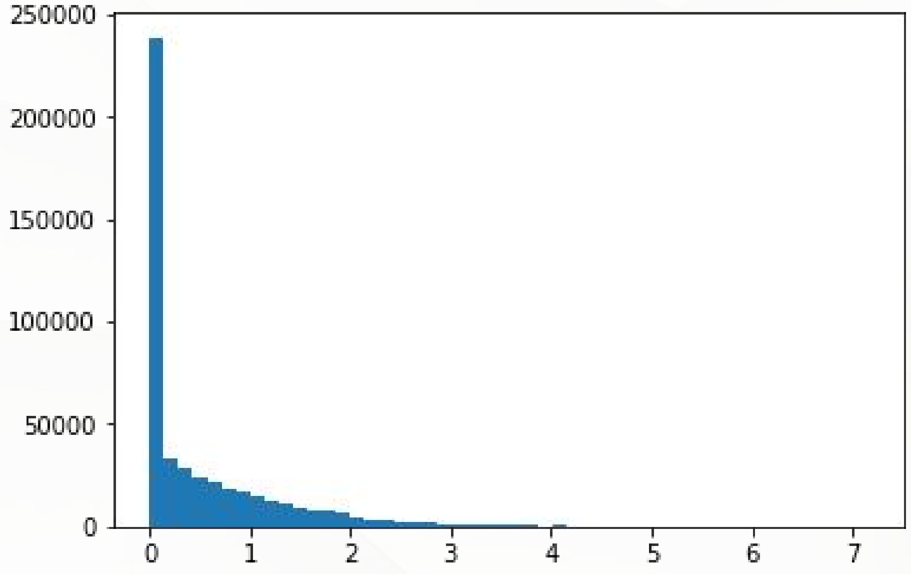


NVIDIA architecture revolution & support to different data quantization format

Quantization——Conventinoal Method



Weight distribution @ CNN layer

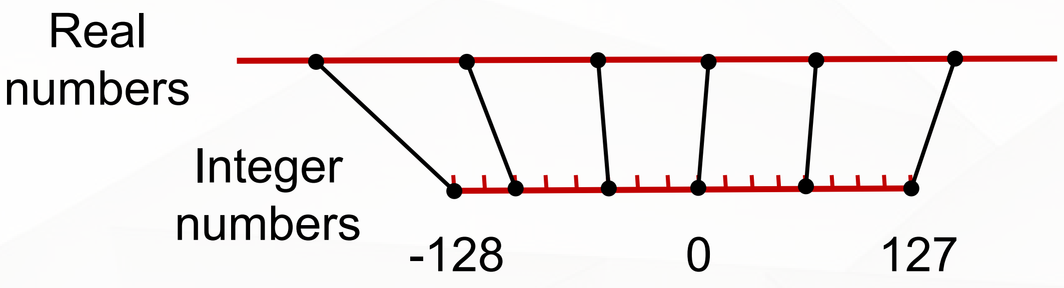


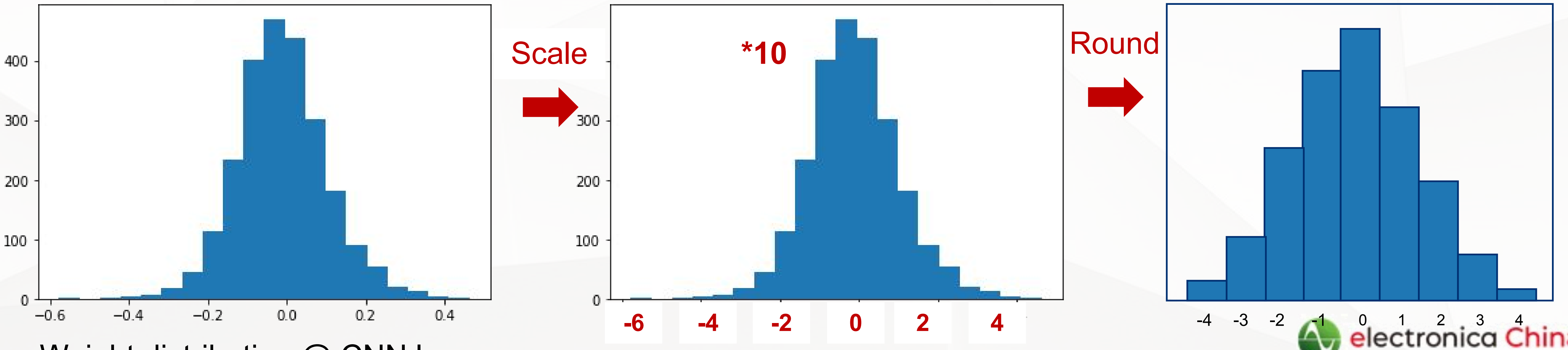
Act. distribution @ CNN layer

Scale, (shift) and round

Quantization——Scale & Round

- Floating-point → integer arithmetics

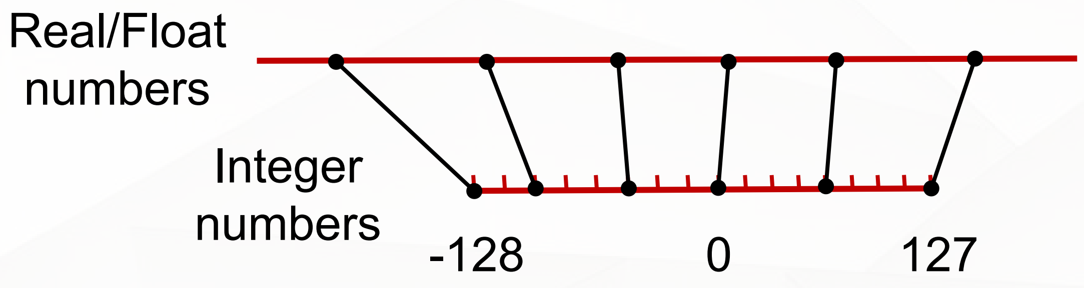




Weight distribution @ CNN layer

Quantization——Arithmetic Change

- Floating-point → integer arithmetics



$$\mathbf{x} * \mathbf{w} = \sum x_i * w_i \quad \Rightarrow \quad (s_x \mathbf{x}) * (s_w \mathbf{w}) = \sum (s_x x_i) * (s_w w_i)$$

$$\text{round}(s_x \mathbf{x}) * \text{round}(s_w \mathbf{w}) = \sum \text{round}(s_x x_i) * \text{round}(s_w w_i)$$

Quantization—Push the Limits

- Floating-point → INT4/FP8/FP4/Binary
  - But difficult to build a single piece of hardware to support all formats with different bit widths
- Another dimension: progressive precision in stochastic computing

$(\text{xxxx xxxx})_2 \longrightarrow$ x x x x x x ...

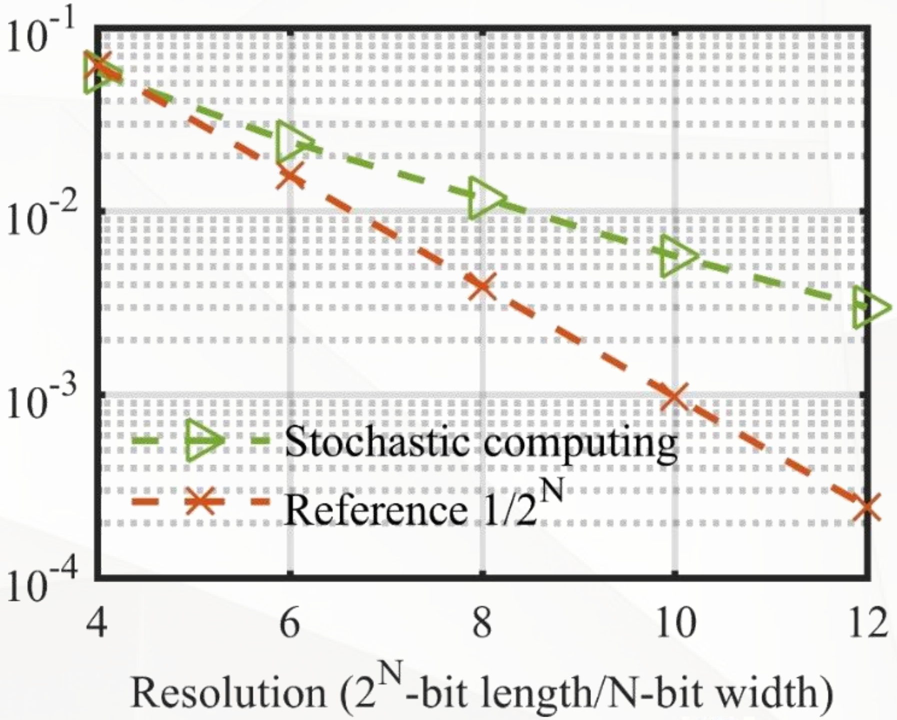
Stochastic computing (SC) uses serial binary bits to represent a value.

Stochastic Computing & Progressive Precision

- Stochastic computing (SC) employs probability to encode a value.

1 0 1 1 0 0 1 0 ⟶ **4/8 = 0.5**

- Since each bit is generated randomly, increasing the sequence length improves the accuracy.

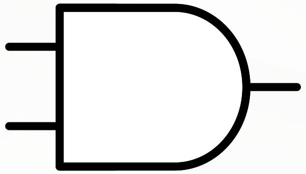


Stochastic Computing—Basic Elements

- Simple logics gates perform complex arithmetics

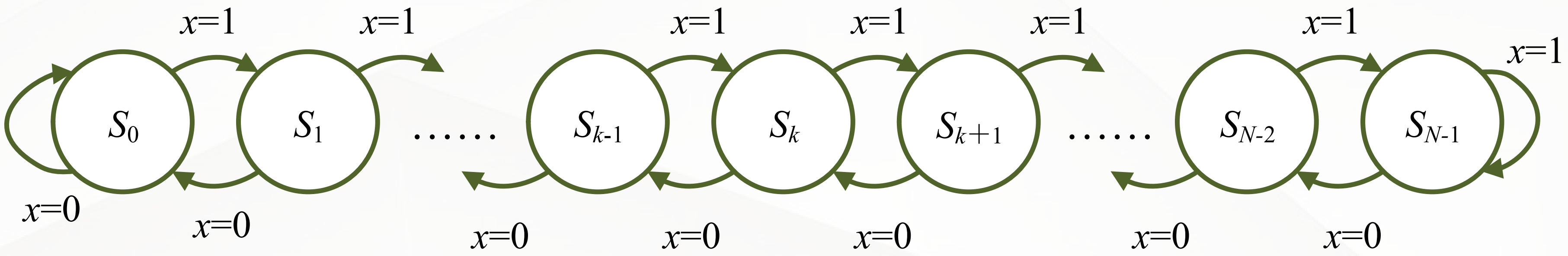
(0.5) 1 0 1 1 0 0 1 0

(0.5) 0 1 1 0 0 1 1 0



0 0 1 0 0 0 1 0 (0.25)

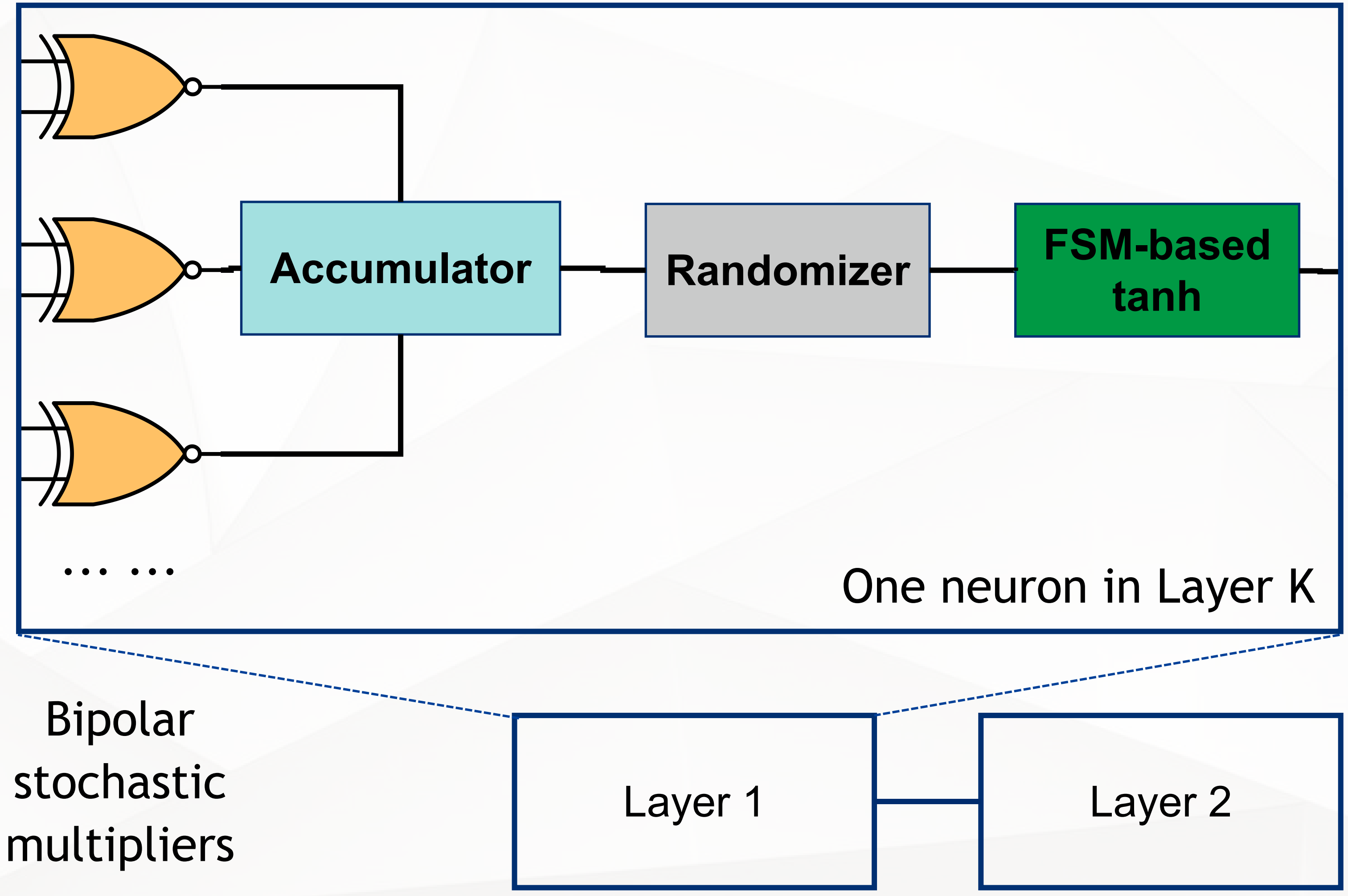
Stochastic multiplier (AND gate)



FSM implements tanh function in SC

An SC based-Neural Network Implementation

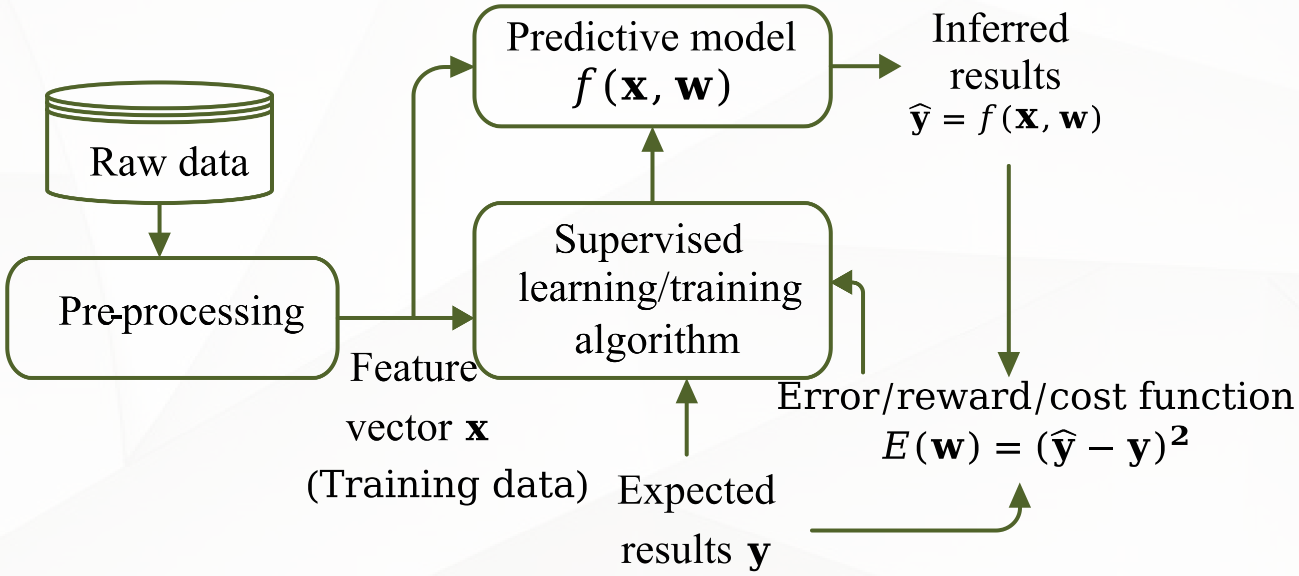
上海科技大学
ShanghaiTech University

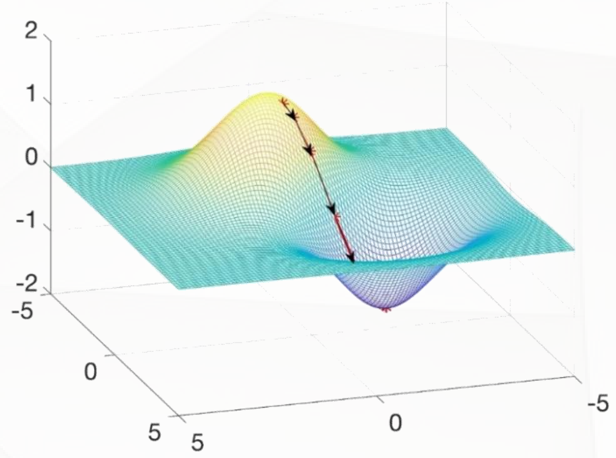


Model	Bit flip rate	Accuracy
MLP	0	97.77
SC-MLP	0	97.71
	1	97.33
	5	96.92
	10	94.84

Stochastic Computing for NN Training

- Gradient descent (with momentum) as optimizer



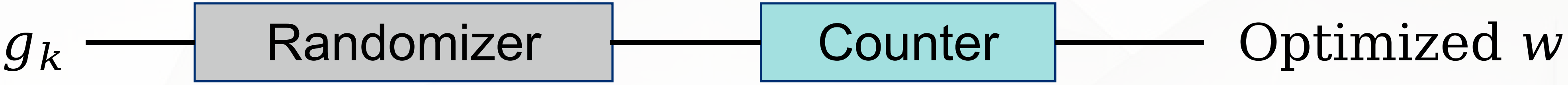


Gradient descent (GD) searching for local minimum.

- The optimization result is an accumulation of multiple steps of the gradients $g_i = \nabla E(\mathbf{w}_i)$.

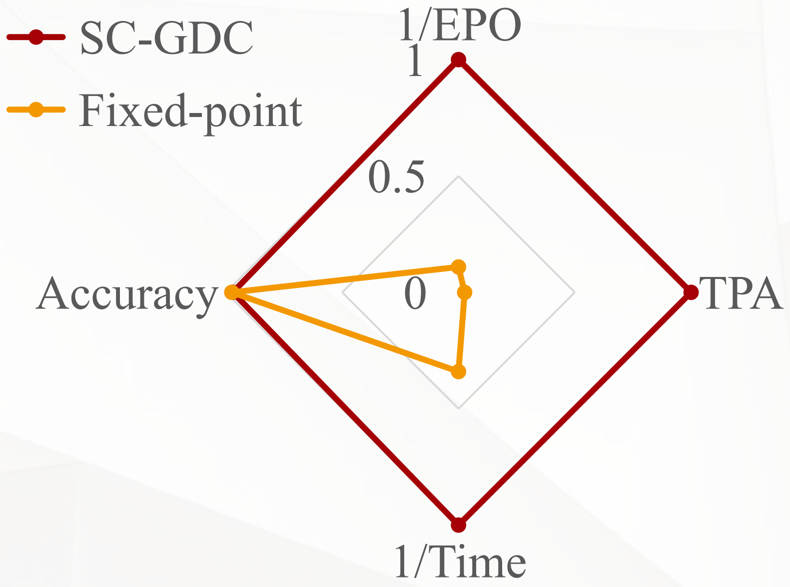
SC-GDC Design

- Implement the iterative accumulation $w = \sum g_k$ in SC.
- g_k is stochastically quantized to -1/0/+1 and accumlated by a counter.

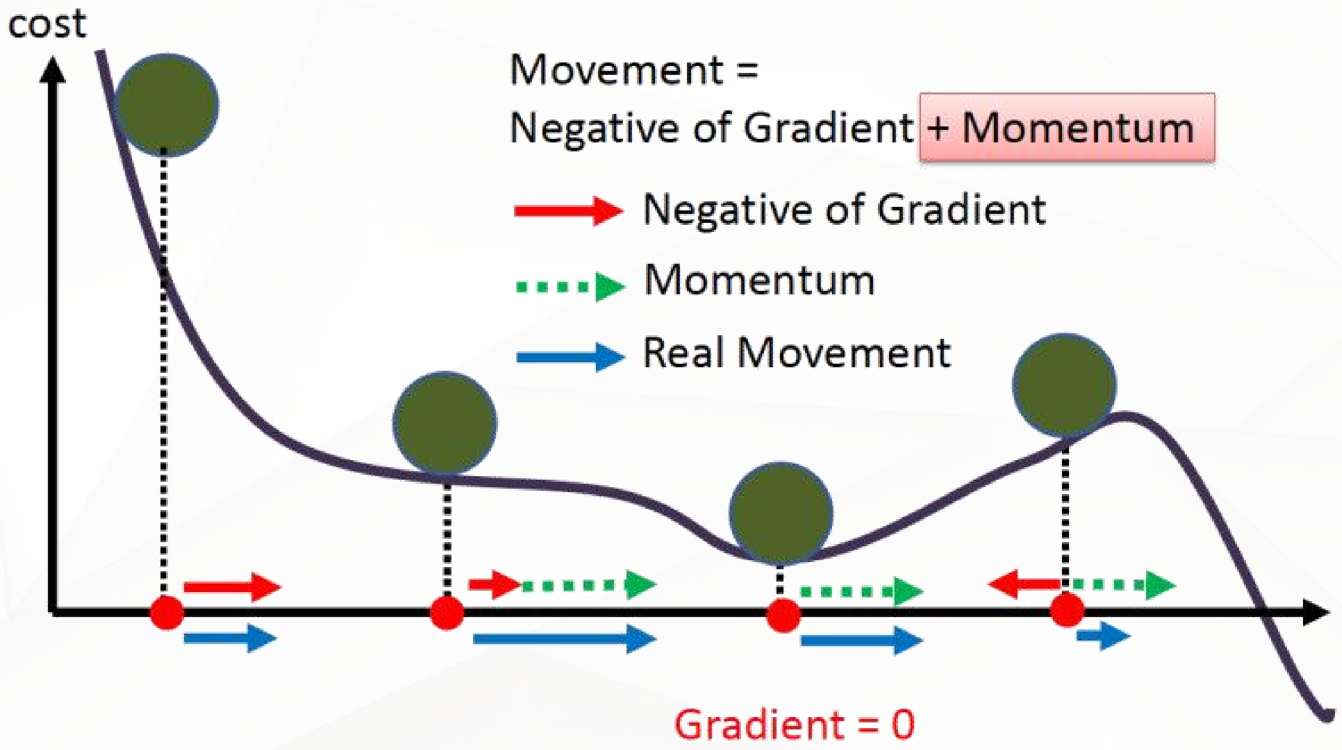


Hardware evaluation of the signed SC-GDC array training a 784-128-128-10 NN

Metrics	SC-GDCs	Fixed-point
Step size	2^{-10}	2^{-10}
Epochs	20	20
Min. time (ns)	1.6×10^6	4.7×10^6
EPO (fJ)	1.2×10^7	1.1×10^8
TPA (images/$s/\mu m^2$)	5.7×10^1	1.5
Aver. test Accu.	97.04%	97.49%



SC-GDM Design

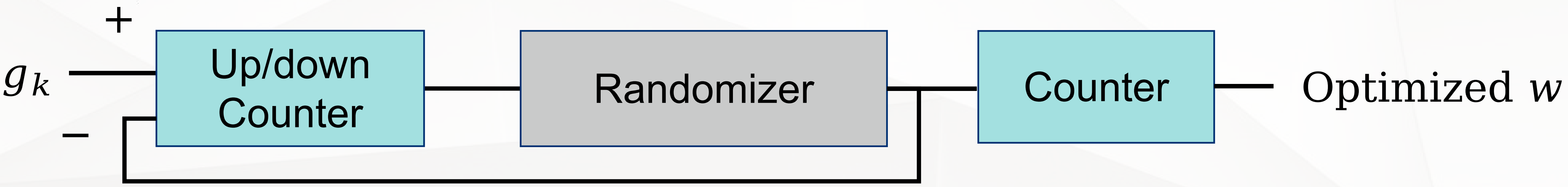


Vanilla GD

$$w_i = w_{i-1} - g_i$$

GD+Momentum

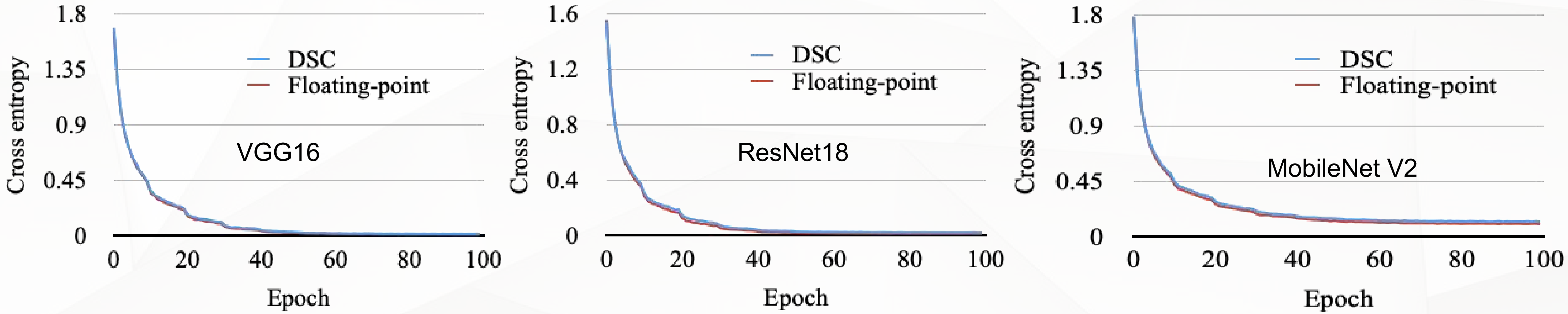
$$v_i = \beta v_{i-1} + g_i$$
$$w_i = w_{i-1} - \mu v_i$$



Experiments & Results

上海科技大学 ShanghaiTech University

- Can train more complex NN architecture such as VGG16, ResNet18 and MobileNet V2 with CIFAR10 dataset.



Test accuracy (%)	VGG16	ResNet18	MobileNetV2
SC-GDM	90.23	91.36	88.51
Floating-point	90.55	91.85	88.82

Take-aways

上海科技大学 ShanghaiTech University

- With numerical approximation, the computation efficiency can be improved;
- Stochastic computing (SC) is a numerical approximation technique that represents a value by the probability of a binary bit stream;
- SC is able to achieve bit-flip-resilient NN inference and energy-efficient and high-performance training;
- It achieves a higher energy efficiency and performance compared with traditional computing paradigm.

Thanks for your attention!

上海科技大学
ShanghaiTech University

Q & A