AI 2025
SUMMER SCHOOL
ai.uni-jena.de

# Using Stochastic Computing with Adjustable Sequence Length to optimize energy consumption of embedded Neural Networks

**Niclas Starost**
**Technische Universität Ilmenau**

## Abstract

Using **neural networks** (NN) on embedded/mobile devices is most of the time infeasible due to energy consumption, performance and the resultant latency.

One way of reducing the computational costs is by exchanging **floating point** (FP) operations with **stochastic computing** (SC) operations that can be implemented very efficient in hardware and also tolerate errors by definition.

In this Poster, the methods of stochastic computing is introduced and how they are integrated into NNs. Also one new way of reducing the sequence length after training while preserving sufficient accuracy is shown.

## Basics of Stochastic Computing (SC)

Unlike deterministic computing, SC uses a bitstream representing a stochastic sequence where each bit follows a Bernoulli distribution of a fraction p, where:

$$P(1)=p, P(0)=1-p, p\in[0,1]$$

**Example**: The fraction $\frac{2}{8}$ could produce this bitstream '**00100010**' with sequence length 8. With this representation, complex arithmetic operations can then be calculated with simple logical operations:

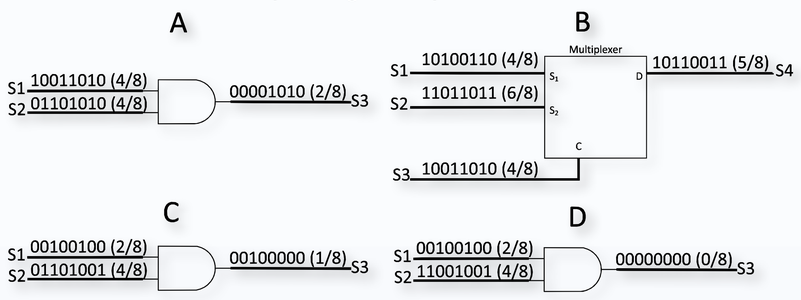


Fig. 1: Stochastic Logic Circuits
**A**: AND gate used for multiplication of fractions.

$$S_3=P(S_3)=P(S_1)P(S_2)=S_1*S_2$$

**B**: MUX used for adding fractions.

$$S_4=P(S_3)P(S_1)+(1-P(S_3))P(S_2)$$
$$S_4=\frac{P(S_1)+P(S_2)}{2}=\frac{S_1+S_2}{2}, P(S_3)=0.5$$

**C:** Uncorrelated bitstreams result in accurate calculations. **D:** Correlated bitstreams result in inaccurate calculations. Source: [1]

Thus, the accuracy of calculations heavily depends on the randomness of generated bitstreams. That means the core element of SC is the **stochastic number generator** (SNG) and also the most energy consuming part.

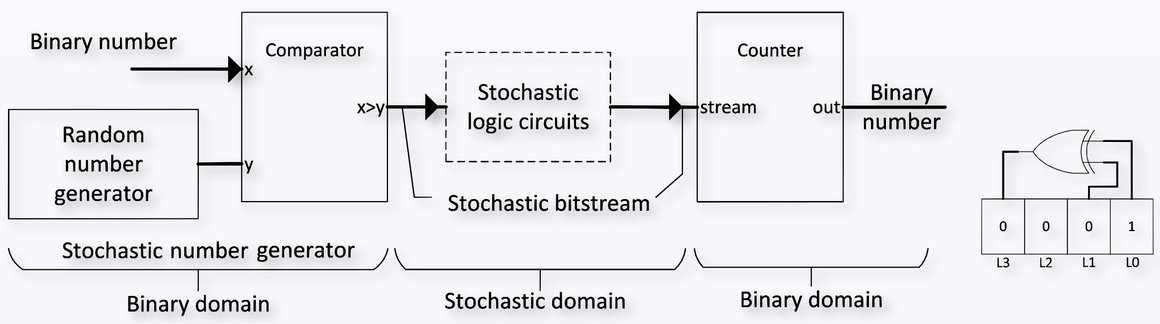


Fig. 2: Stochastic computing process and components Source: [1]




Fig. 3:
RNG with LFSR
Source: [1]

As **random number generator** (RNG) usually a **linear-feedback shift register** (LFSR) which is simple and can be implemented very efficiently in hardware.
As a comparator there are binary and **weighted binary generator** (WBG).
For converting it back, a simple flip-flop counter can be used to count the ones.

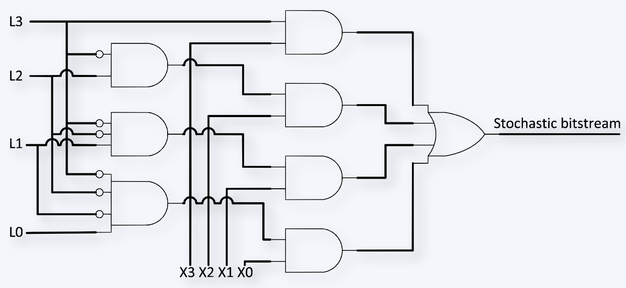


Fig. 4: WBG generating a stochastic bitstream when inserted a random number and the binary number representing the fraction
Source: [2]

## SC in Neural Networks

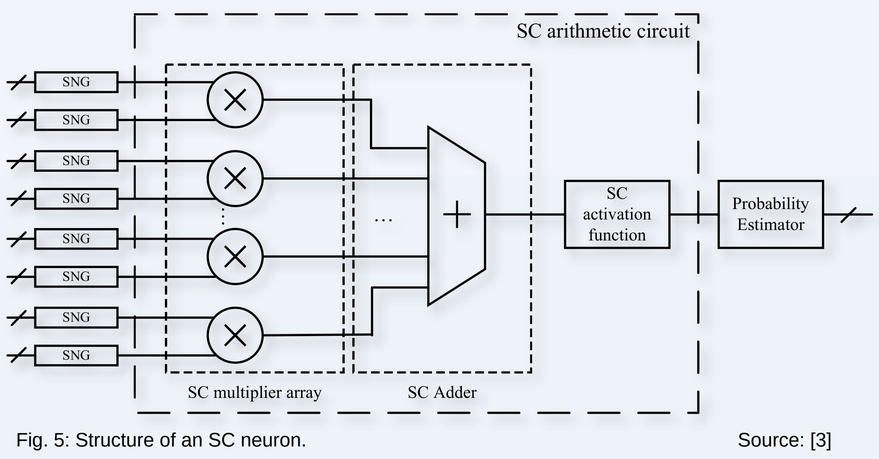


Fig. 5: Structure of an SC neuron. Source: [3]

A neuron consists of an array of SNGs for converting inputs/weights/biases into the SC domain. Those then are multiplied/added like a usual neuron except the operations are simpler in the SC domain.
Then adjusted activation functions equivalent to normal NN are used. Afterwards the SC sequence is converted back to the binary domain with a **probability estimator** (PE).

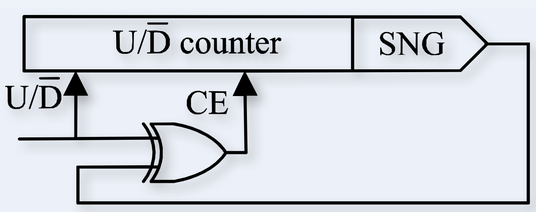


Fig. 6: **Adaptive digital element** (ADDIE) PE for converting the stochastic sequence back to a binary one.
Source: [3]

## Activation Functions in SC

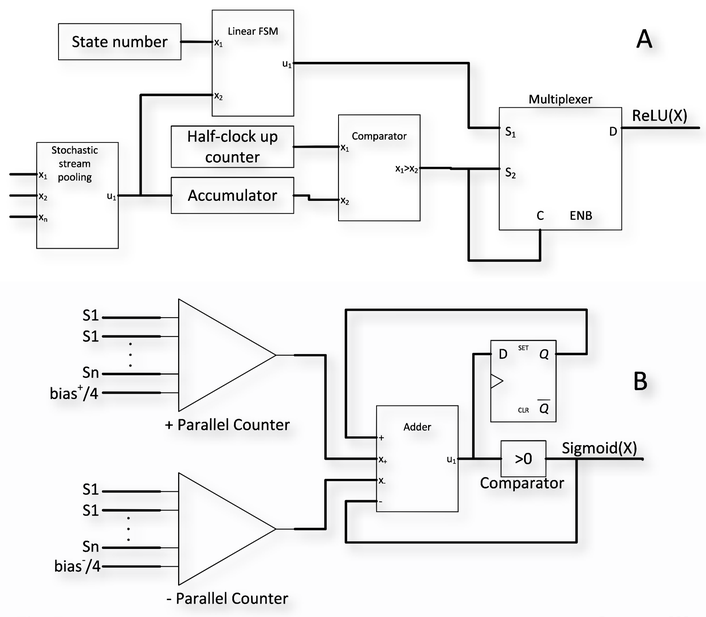


Fig. 7:
**A:** ReLU for SC **B:** Sigmoid for SC
Source: [2]

ReLU can be implemented with a different approach but requires a **finite state machine** (FSM).
An approximated Sigmoid function on the other hand can be implemented with simpler components.

$$\frac{1}{1+\exp(-x)}\approx\begin{cases}1, x>2\\ \frac{1}{2}+\frac{1}{4}*x, -2\le x\le 2\\ 0, x<-2\end{cases}$$

$$A=\frac{1}{4}*\sum_{positive} P*Q+\frac{1}{2}+\frac{+bias}{4},$$
$$B=\frac{1}{4}*\sum_{negative} P*Q+\frac{-bias}{4},$$
$$A-B=\frac{1}{2}+\frac{1}{4}\left(\sum P*Q+bias\right)$$

## Comparison FP and SC NN

| CNN Model | Platform | Year | Method | Area [mm²] | Power [W] | Accuracy [%] | Energy efficiency [images/J] |
|---|---|---|---|---|---|---|---|
| LeNet-5 | CPU | 2009 | Software | 263 | 156 | 99.17 | 4.2 |
| | GPU | 2011 | Software | 520 | 202.5 | 99.17 | 3.2 |
| | ASIC | 2016 | SC 256bit | 36.4 | 3.53 | 98.26 | 221,287 |
| | ASIC | 2018 | SC 128bit | 22.9 | 2.6 | 99.07 | 1,231,971 |
| AlexNet (last second layer) | CPU | 2009 | Software | 263 | 156 | - | 0.9 |
| | GPU | 2011 | Software | 520 | 202.5 | - | 2.8 |
| | ASIC | 2018 | SC 128bit | 24.7 | 1.9 | - | 1,326,400 |

## Adjustable Sequence Length (ASL)

The introduced structure can reduce the sequence length by truncating the random number or the input sequence directly without retraining the model.
Thus, the trained network can get optimized for energy efficiency afterwards.
To further improve the accuracy, a Sobol sequence is used for the random number input.

The sequence length in the earlier layers needs to be preserved while the later layer can be truncated without losing much accuracy.

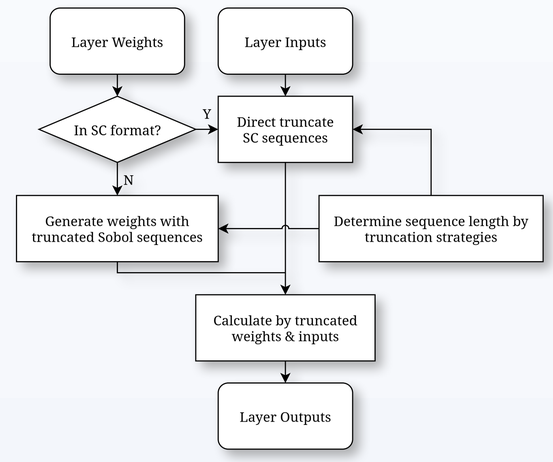


Fig. 8: Flowchart of ASL truncation Source: [1]

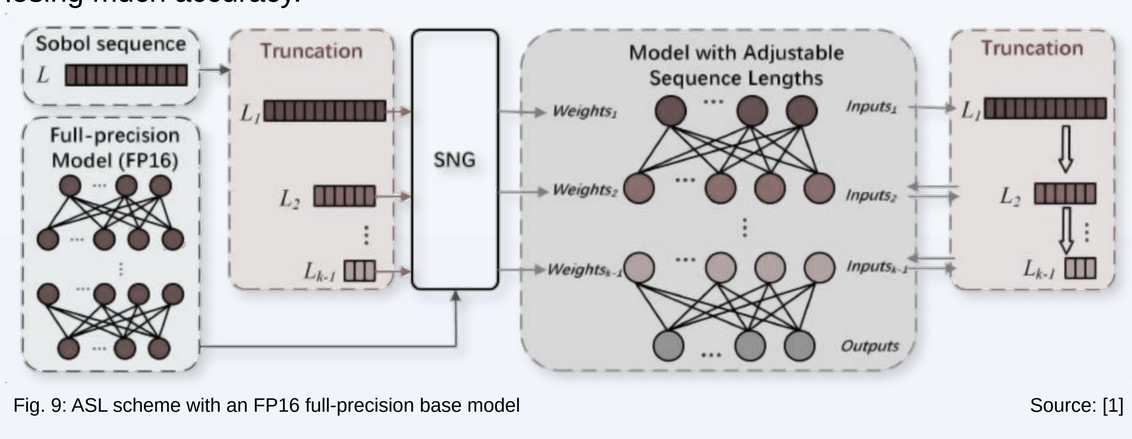


Fig. 9: ASL scheme with an FP16 full-precision base model Source: [1]

## Energy Consumption with ASL

| Dataset | Layer Size | Baseline Accuracy | Energy [mJ] | Number of Cycles | Accuracy Loss | Savings in energy [%] | Savings in latency [%] |
|---|---|---|---|---|---|---|---|
| Fashion MNIST | 784-1024-1024-512-256-10 | 91.98% | 3.38 | 1797 | 0.083 | 49.83 | 64.94% |
| SVHN | 1024-1024-1024-512-256-10 | 90.42% | 4.68 | 1925 | 0.040 | 43.64 | 62.44% |
| CIFAR10 | 1024-1024-1024-512-256-10 | 64.86% | 4.70 | 1989 | 0.095 | 43.38 | 61.19% |

## References

[1] Wang, Ziheng, et al. "Energy-Efficient Stochastic Computing (SC) Neural Networks for Internet of Things Devices With Layer-Wise Adjustable Sequence Length (ASL)." IEEE Internet of Things Journal (2025).

[2] Lee, Yang Yang, and Zaini Abdul Halim. "Stochastic computing in convolutional neural network implementation: a review." `PeerJ. Computer science vol. 6 e309. 9 Nov. 2020, doi:10.7717/peerj-cs.309

[3] Y. Liu, S. Liu, Y. Wang, F. Lombardi and J. Han, "A Survey of Stochastic Computing Neural Networks for Machine Learning Applications," in IEEE Transactions on Neural Networks and Learning Systems, vol. 32, no. 7, July 2021, doi: 10.1109/TNNLS.2020.3009047.

[4] Mehlin, Vanessa, Sigurd Schacht, and Carsten Lanquillon. "Towards energy-efficient deep learning: An overview of energy-efficient approaches along the deep learning lifecycle." arXiv preprint arXiv:2303.01980 (2023).

[5] Matthew Carrano, Director: Scott Koziol, Ph.D., Combining Machine Learning with Stochastic Computing

Contact: nist42@mailbox.org