IEEE CSC & ESAS SUPERCONDUCTIVITY NEWS FORUM (global edition), November 2019.
Invited presentation 3-RE-C-4 given at ISEC, 28 July-1 August 2019, Riverside, USA.

Design and Implementation of a Bitonic Sorter-Based DNN Using Adiabatic Superconducting Logic

O. Chen[1], T. Tanaka[1], R. Cai[2], Y. Wang[2] and N. Yoshikawa[1]

[1] Yokohama National University

[2] Northeastern University

Current Neural Networks

- Massive data required

Network	Model Size
LeNet-5	60,000
AlexNet	60M
BERT	340M

High-performance computing plays the key role (Data center, work station)

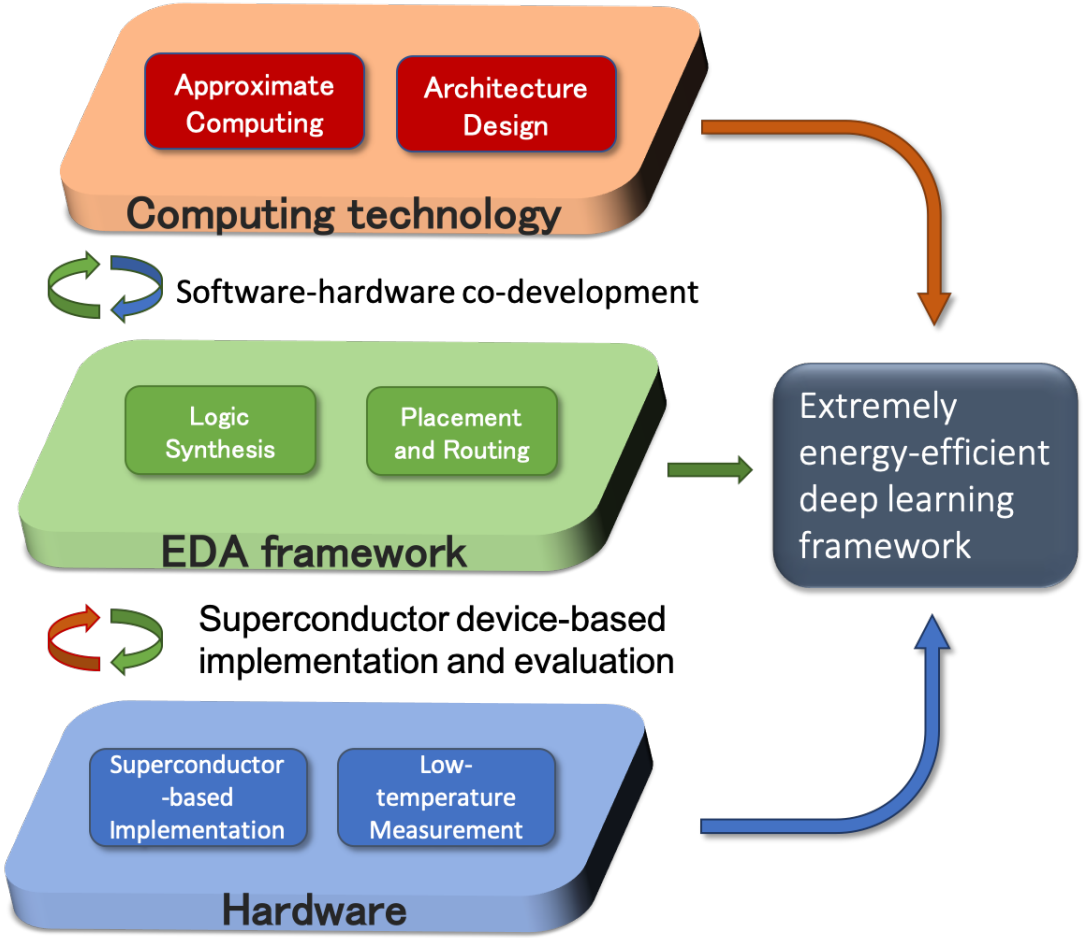
- Massive power consumed: （10% of nation's power consumption）

Facebook Data Center (Lulea, Sweden)

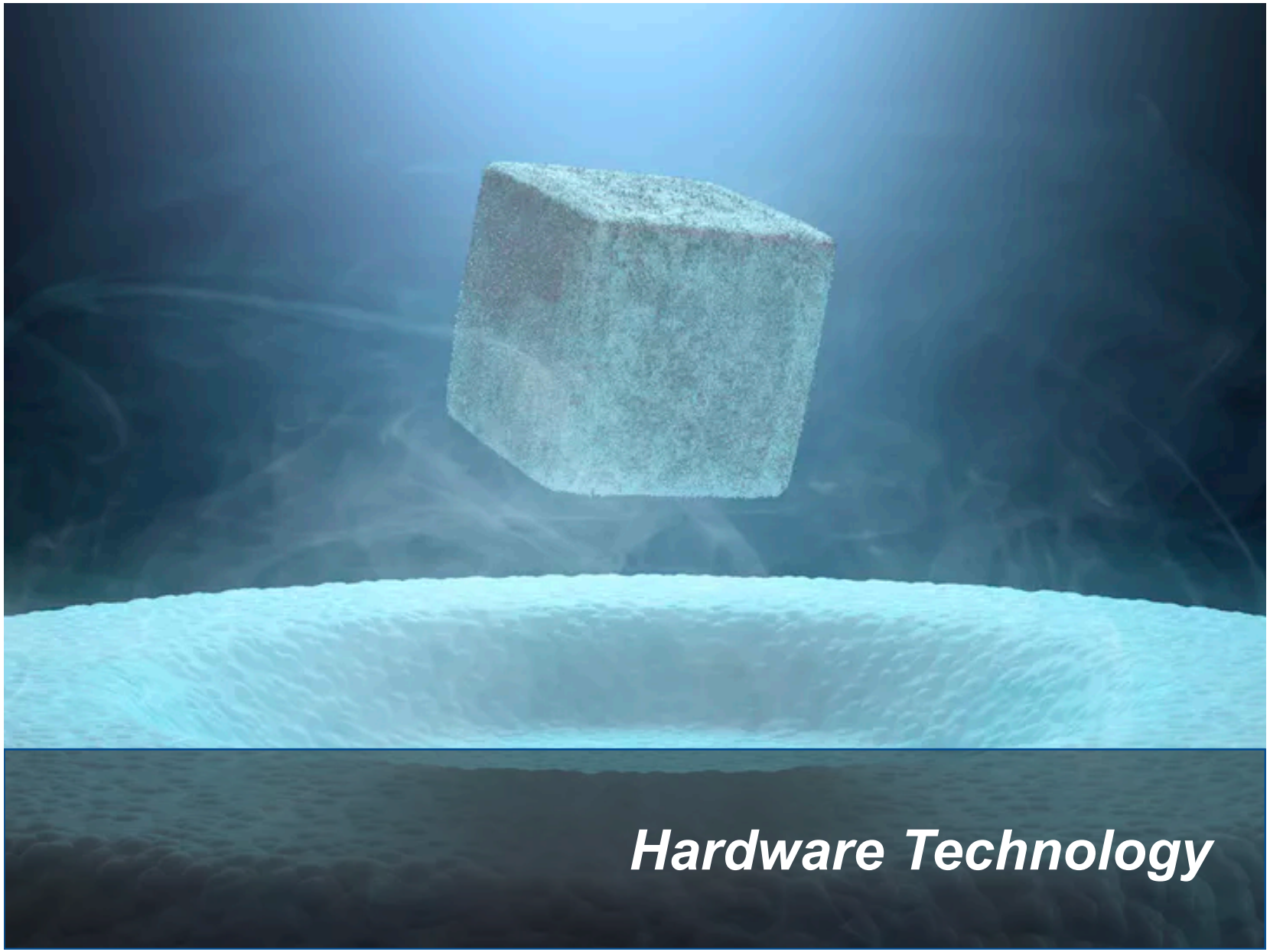
Performance: 27-51 PFLOP/s
Power consumption: 84 MW (average)
Small-scale power plant equivalent

Breakthrough in energy efficiency!

Overview of the Proposed Work



IEEE CSC & ESAS SUPERCONDUCTIVITY NEWS FORUM (global edition), November 2019.
Invited presentation 3-RE-C-4 given at ISEC, 28 July-1 August 2019, Riverside, USA.

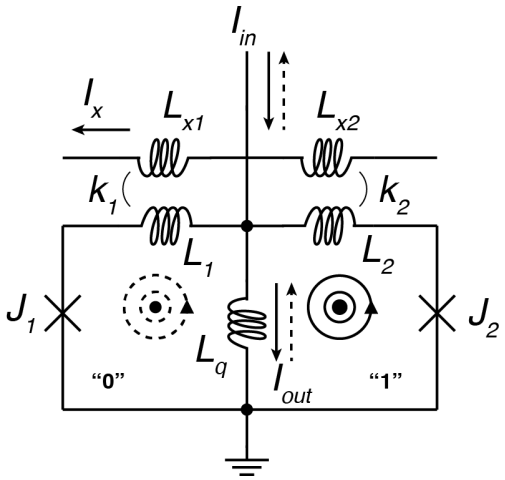


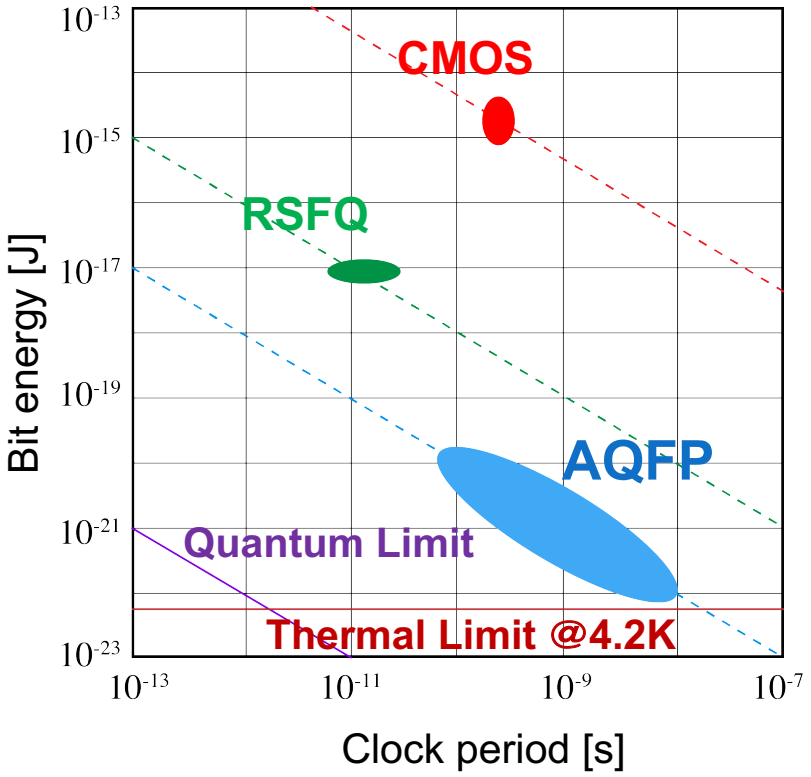
IEEE CSC & ESAS SUPERCONDUCTIVITY NEWS FORUM (global edition), November 2019.
Invited presentation 3-RE-C-4 given at ISEC, 28 July-1 August 2019, Riverside, USA.

Hardware: Adiabatic Quantum Flux Parametron（AQFP）

Operational principle:

- Based on QFP (Goto et al.)
- Adopt adiabatic version of QFP (Takeuchi et al.)
  - (10^2-10^3 less energy dissipation comparing to QFP)

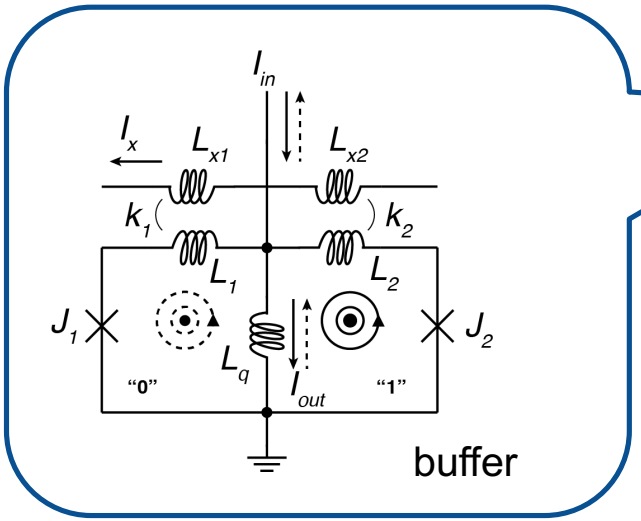




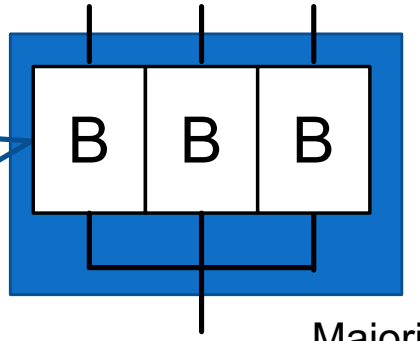
Ideal for the proposed neural network

N. Takeuchi et al., Supercond. Sci. Technol. 28, 015003 (2015).5

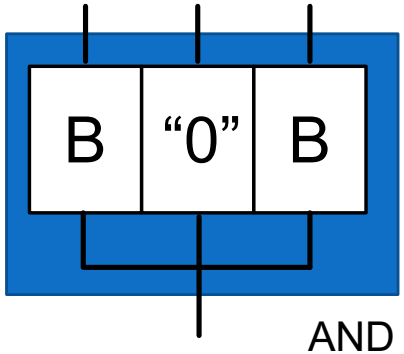
Design Guidelines for AQFP



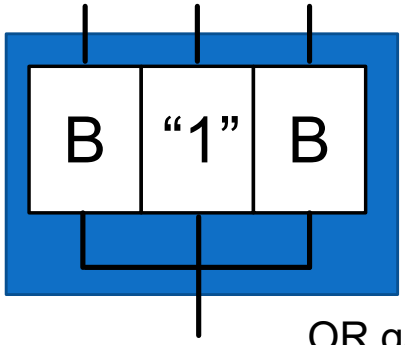
buffer



Majority gate

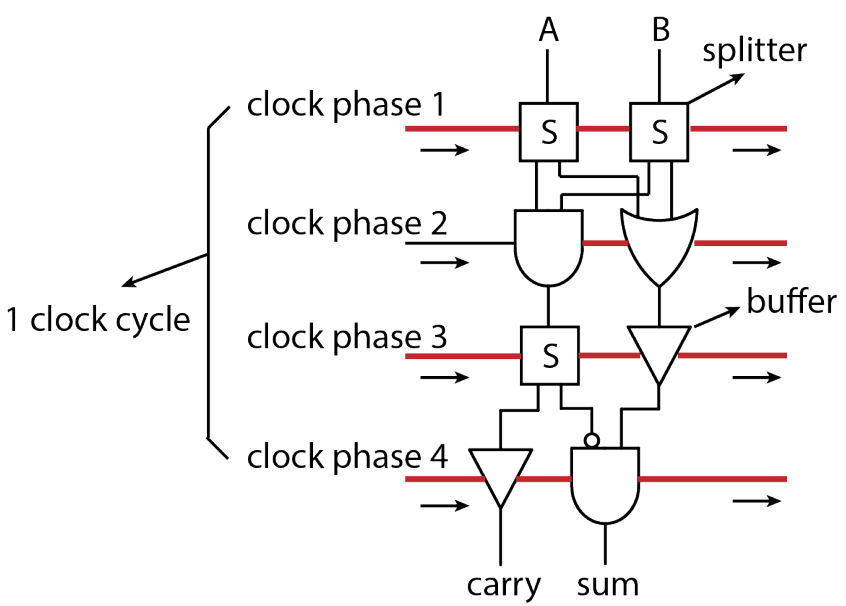


AND gate



OR gate

Data Propagation in AQFP

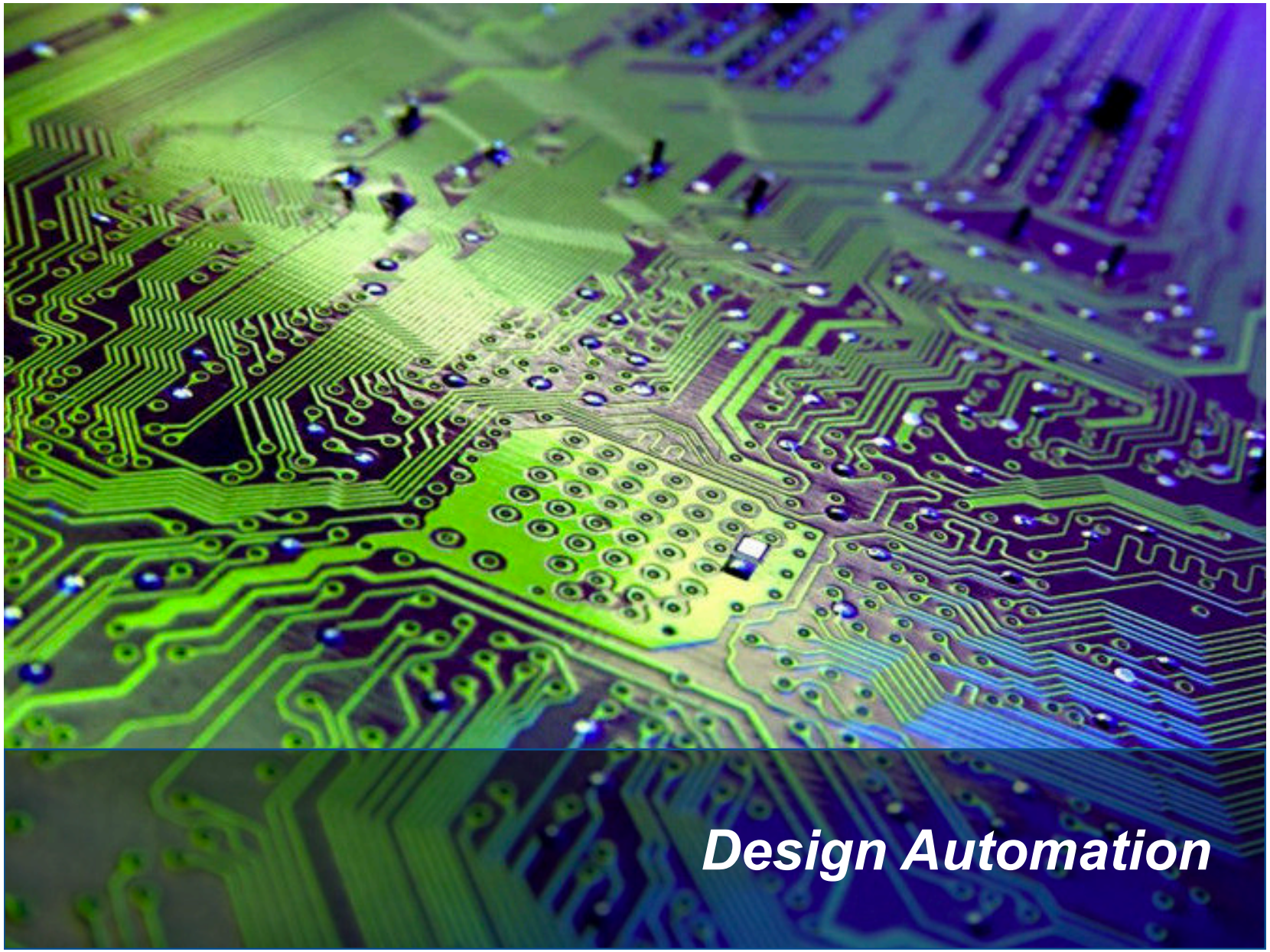


1-bit adder in AQFP

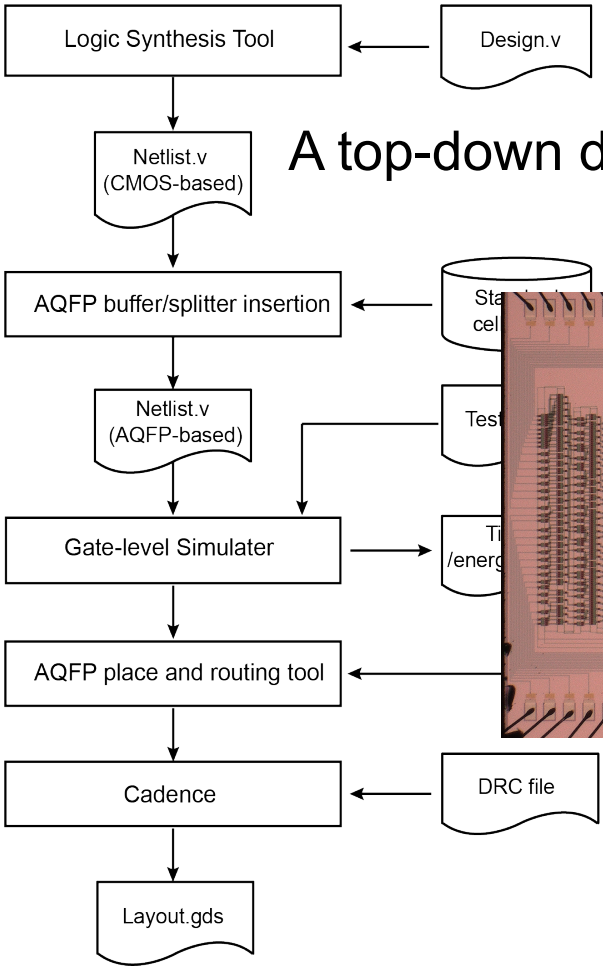
Data propagation direction

- Both combinational and sequential AQFP logic gates are driven by AC-power.
- The AC-power also servers as clock to synchronize gates outputs.
- Data propagation in AQFP requires neighboring clock signals overlapping.
- All inputs to any gate mush have the same delay from the primary inputs.

IEEE CSC & ESAS SUPERCONDUCTIVITY NEWS FORUM (global edition), November 2019.
Invited presentation 3-RE-C-4 given at ISEC, 28 July-1 August 2019, Riverside, USA.

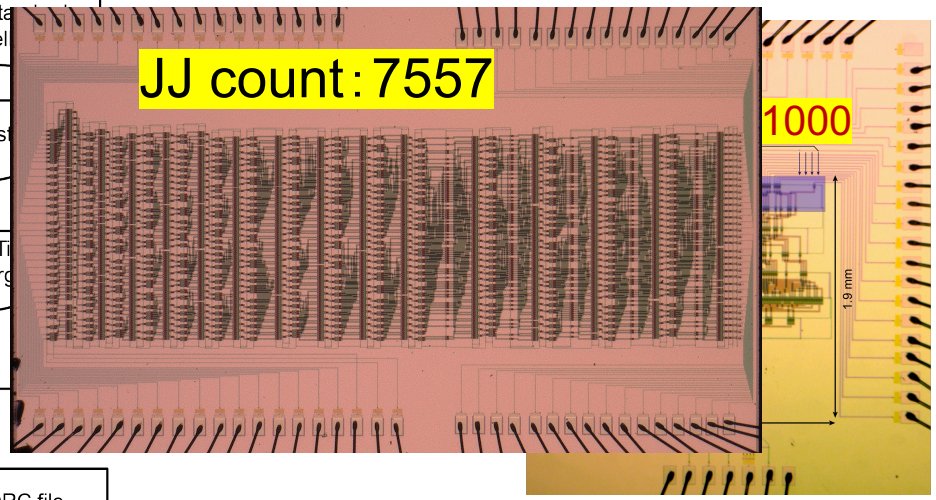


Design Automation: AQFP EDA Framework



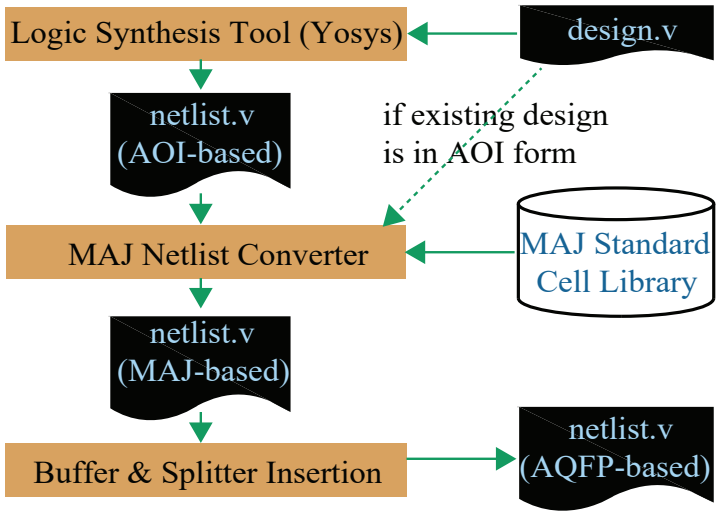
A top-down design flow

The first auto-designed chip: 8 channel MUX



O.Chen et al., IEEE Trans. Appl. Supercond. 29, 5 (2019).

Majority Synthesis and Buffer Optimization

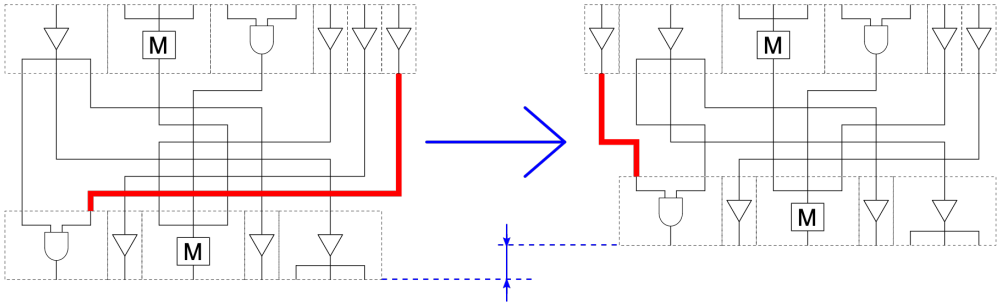


- AQFP MAJ = AND/OR
- Covert AOI to MAJ netlist when necessary
- Buffers and splitter insertions performed after target AOI converted to a MAJ netlist

Logic	AOI		MAJ w/ buffer optimization	
	JJ count	JJ level	JJ count	JJ level
C6288 (ISCAS)	78,246	180	25,870	94
32-bit RISC-V ALU	75,458	172	25,752	84

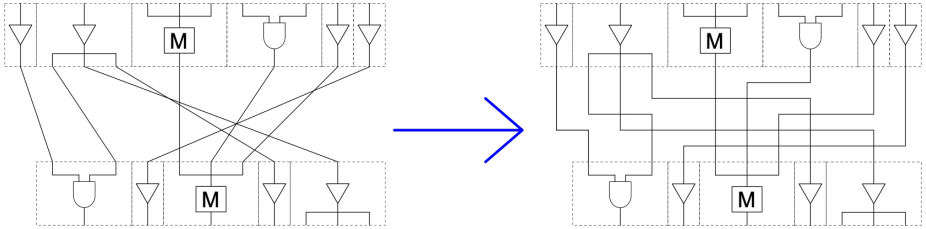
Placement and Routing

- Placement
  - Genetic Algorithm
  - Decided cells combination to minimize area of circuit and length of interconnections.

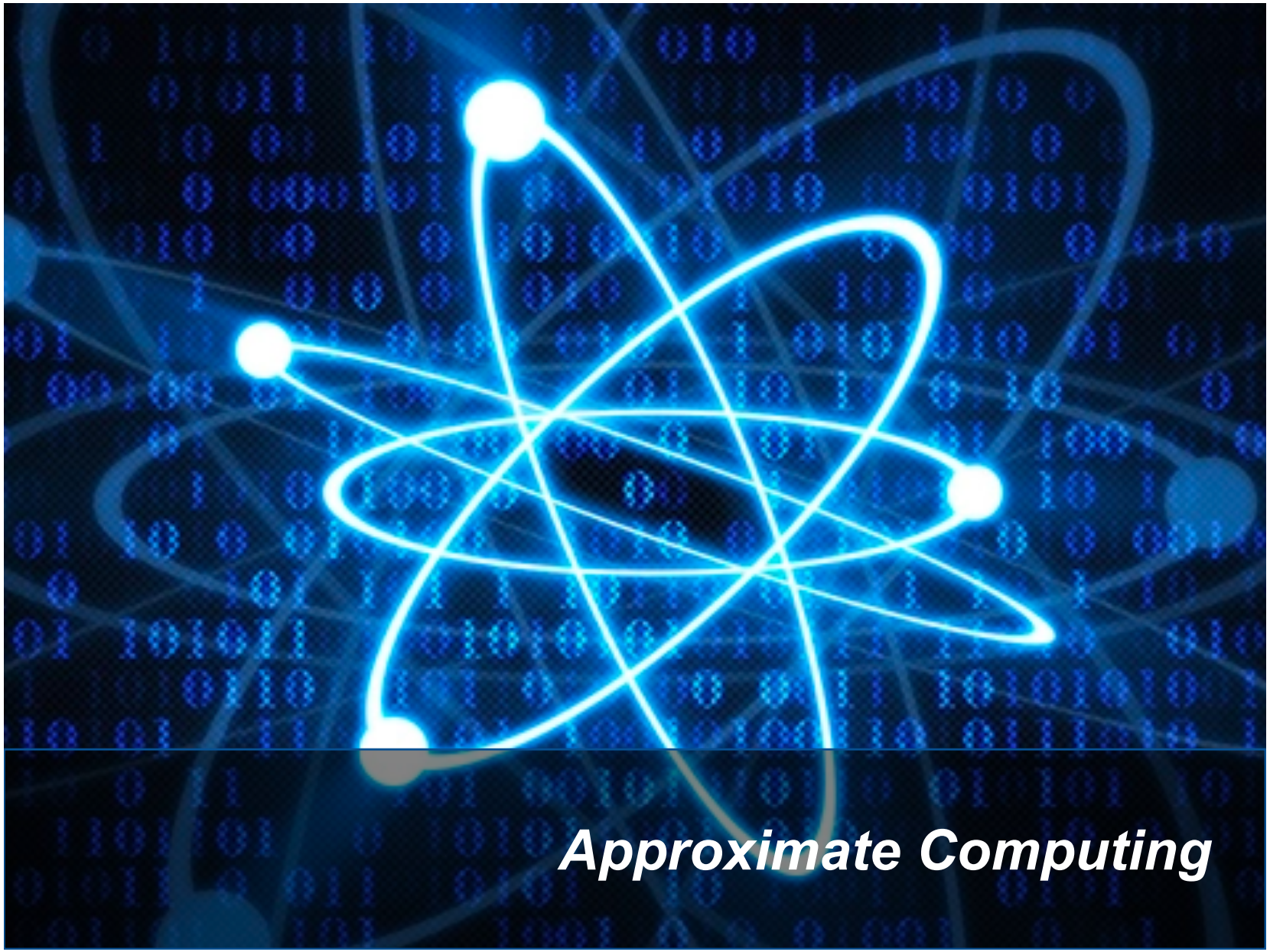


- Routing
  - Left Edge Algorithm
  - Convert connection information between cells into actual layout.

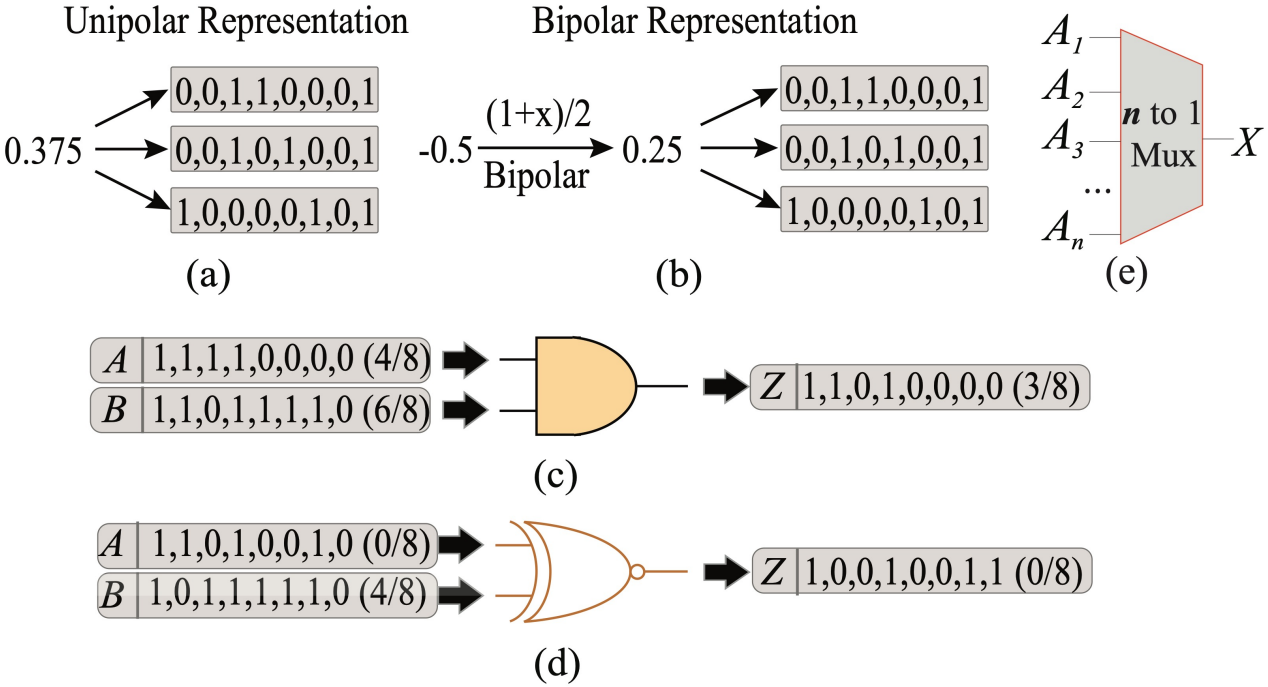
Tanaka et al., *IEEE TAS*, 10.1109/TASC.2019.2900220



IEEE CSC & ESAS SUPERCONDUCTIVITY NEWS FORUM (global edition), November 2019.
Invited presentation 3-RE-C-4 given at ISEC, 28 July-1 August 2019, Riverside, USA.



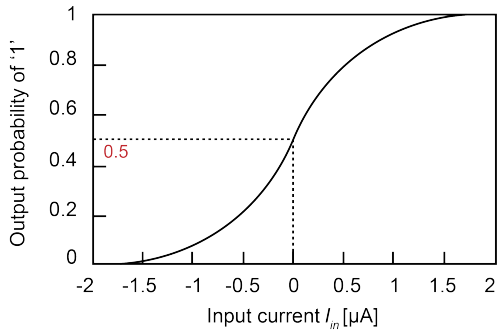
Stochastic Computing



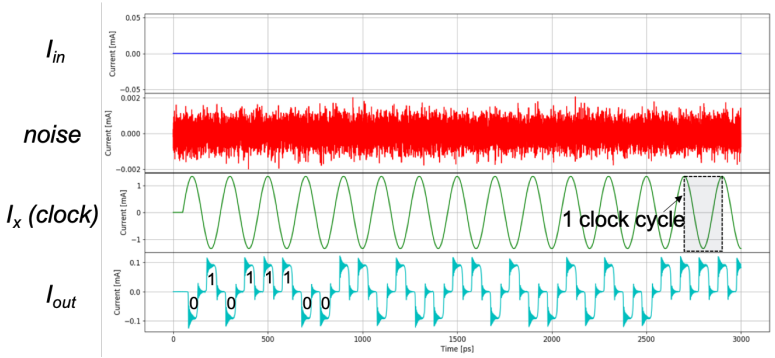
- Stochastic computing (SC) is a paradigm that represents a number, by counting the number of ones in a bit-stream.
- Compatible with the deep-pipelining nature of AQFP
- Low hardware resource utilization.

Stochastic Number Generation in AQFP

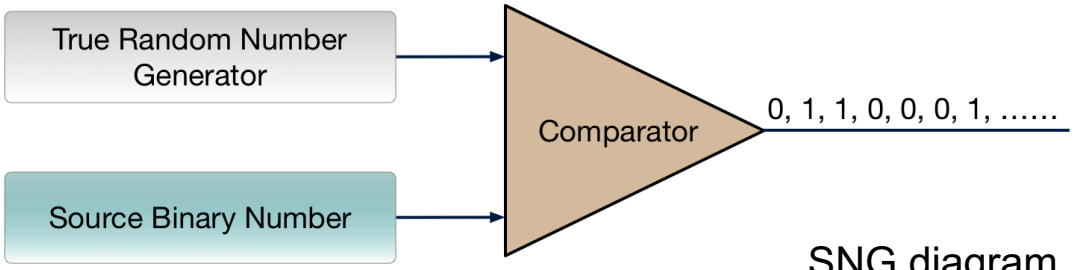
- AQFP offers one bit true RNG based on one bit buffer.
- SNG can be implemented with RNG and comparator.



Output probability of a buffer w/o input



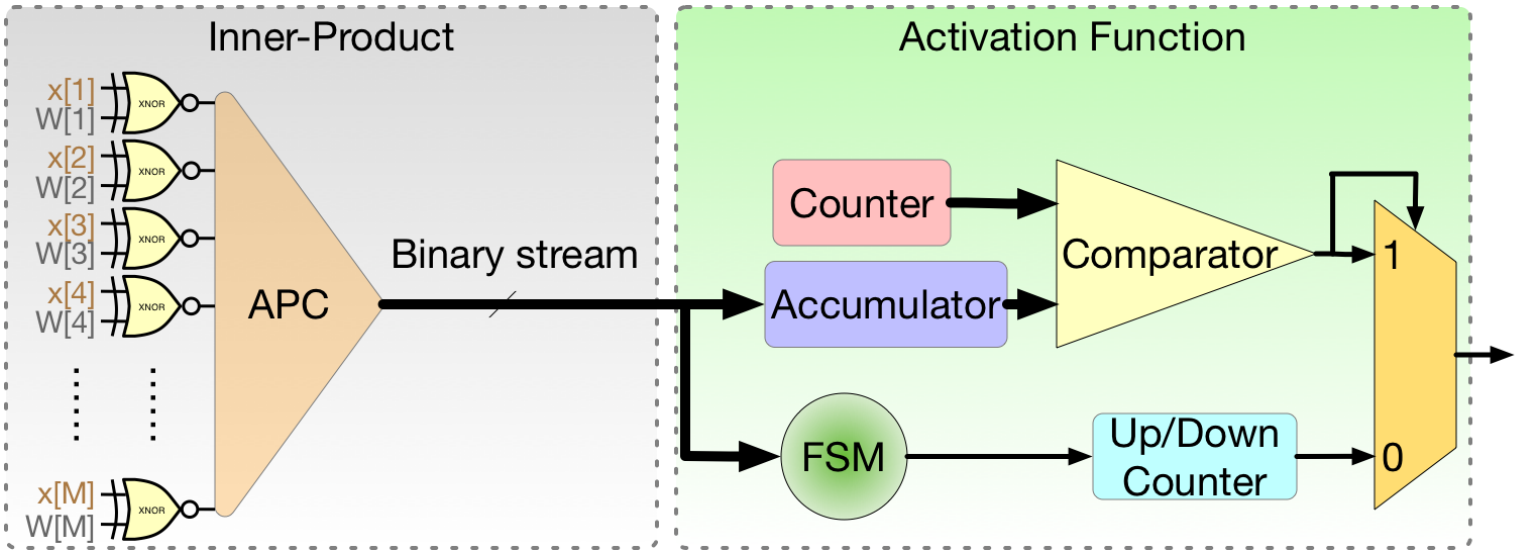
Simulation of a 1-bit RNG



SNG diagram

Will be presented at EUCAS'19

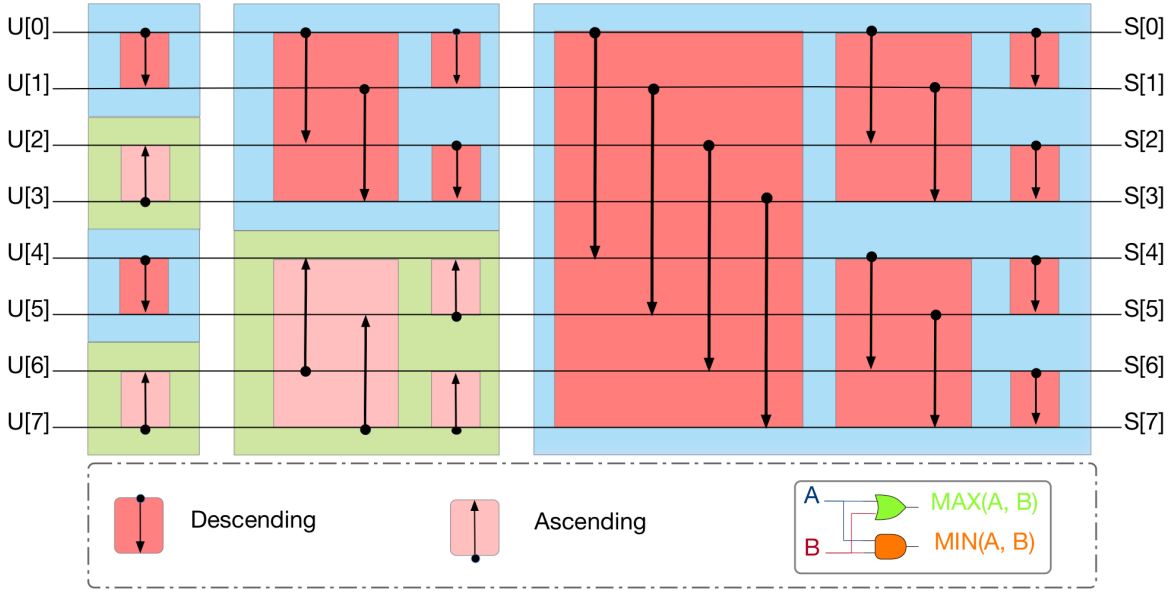
SC-based DNN Framework Using CMOS



- Adder tree for accumulation
- FSM for activation function
- Not ideal for AQFP technology

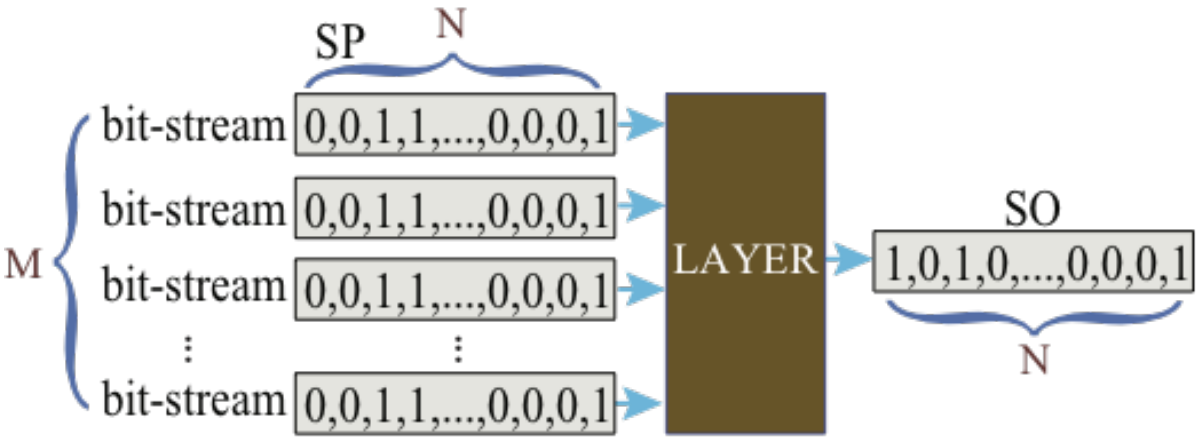
Bitonic Binary Sorter

- Efficient sorting design
- Ideal for AQFP:
  - All signals having the same delay.
  - All signals are duplicated at each stage



SC Blocks Design Approach

- SP is an input binary matrix. M is the number of inputs.
- SO in an output binary vector. N is the stream length.
- Convert input-output function to SC domain.
- How many 1s should be generated according to the number of 1s in the input matrix.



Feature Extraction Block

Output function: $$\mathbf{SO} = clip(\sum_{i=1}^{M} SP_i, -1, 1)$$

Translate to SC domain: $$\frac{2\times\sum_{i=1}^{N}\mathbf{SO}_i - N}{N} = clip(\frac{2\times\sum_{i=1}^{N}\sum_{j=1}^{M}\mathbf{SP}_{i,j} - N\times M}{N}, -1, 1)$$

Factor all by **N:** $$\sum_{i=1}^{N}\mathbf{SO}_i = clip(\sum_{i=1}^{N}\sum_{j=1}^{M}\mathbf{SP}_{i,j} - \frac{M-1}{2}\times N, 0, N)$$

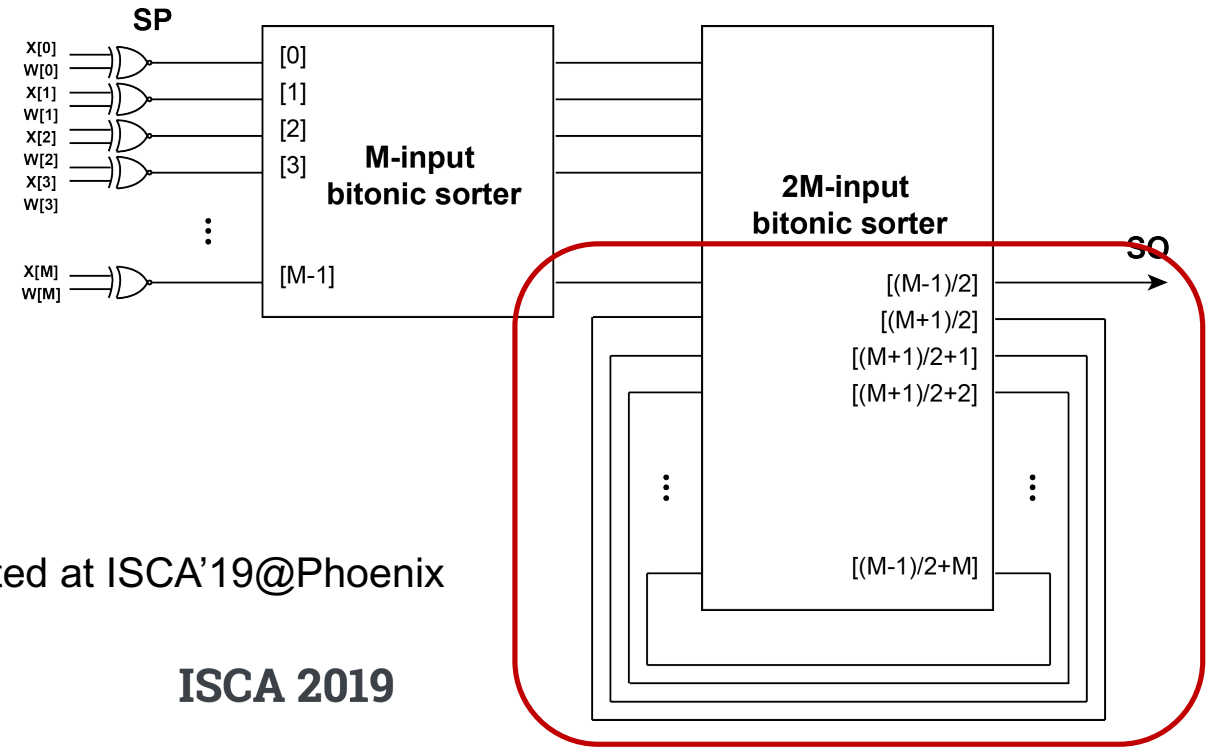
For each bit generation cycle *j*:

$$\mathbf{SO}_i = clip(\sum_{i=1}^{n}\sum_{j=1}^{M}\mathbf{SP}_{i,j} - \frac{M-1}{2}\times N, 0, N) - \sum_{i=1}^{n-1}\mathbf{SO}_i$$

$$= \sum_{i=1}^{n-1}(\sum_{j=1}^{M}\mathbf{SP}_{i,j} - \frac{M-1}{2} - \mathbf{SO}_i) + \sum_{j=1}^{M}\mathbf{SP}_{n,j} - \frac{M-1}{2}$$

Feature Extraction Block

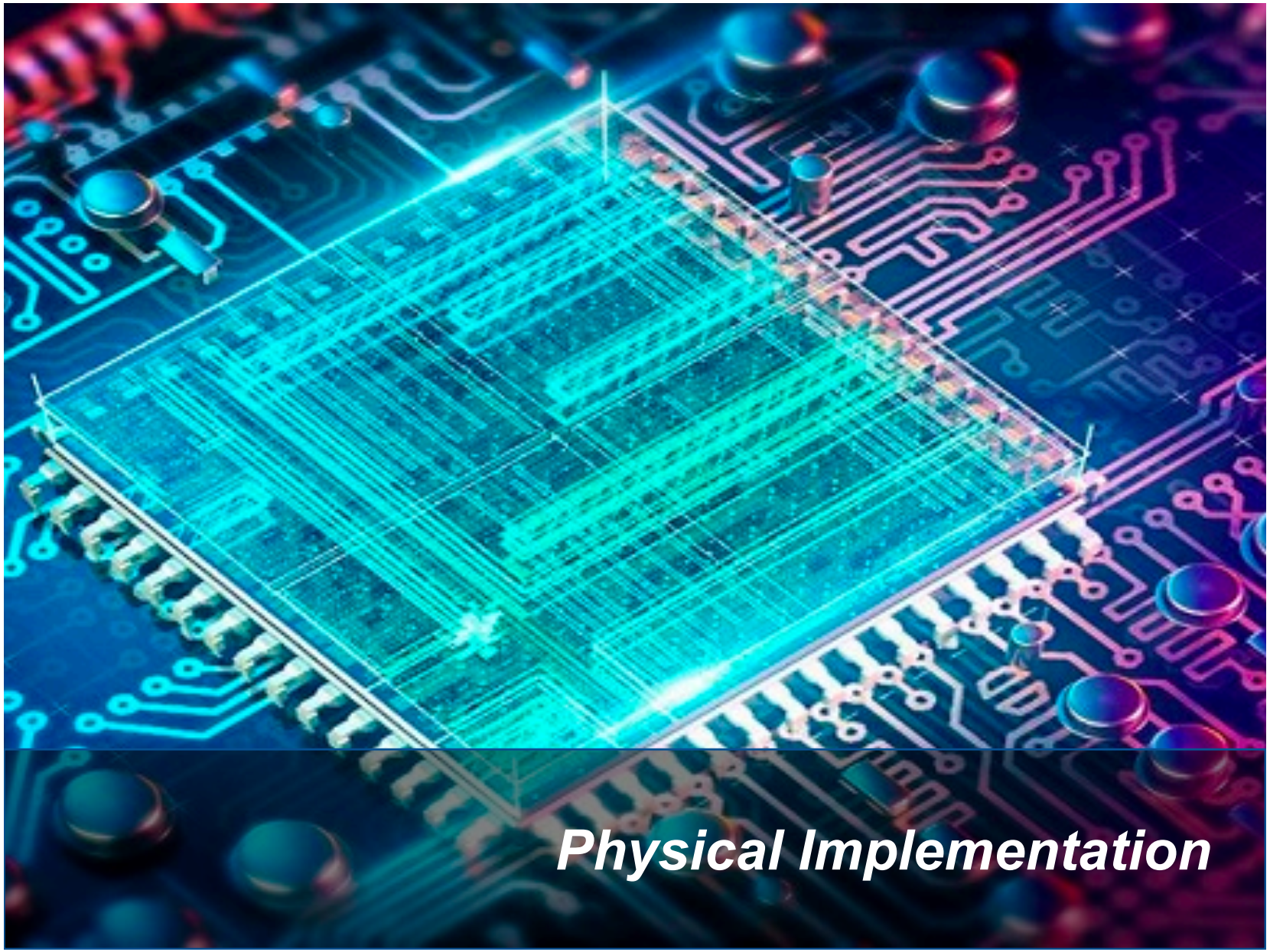
- Bitonic sorter based design with partial feedback



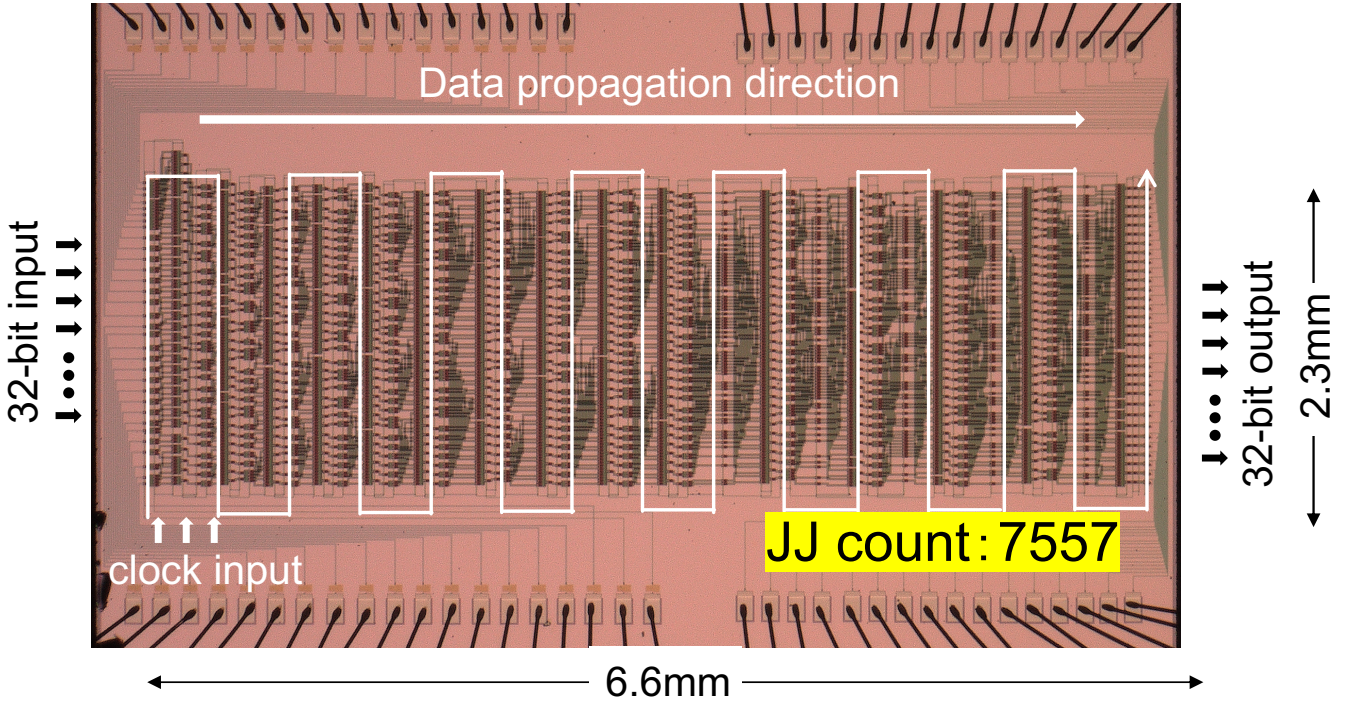
Presented at ISCA'19@Phoenix

ISCA 2019

IEEE CSC & ESAS SUPERCONDUCTIVITY NEWS FORUM (global edition), November 2019.
Invited presentation 3-RE-C-4 given at ISEC, 28 July-1 August 2019, Riverside, USA.

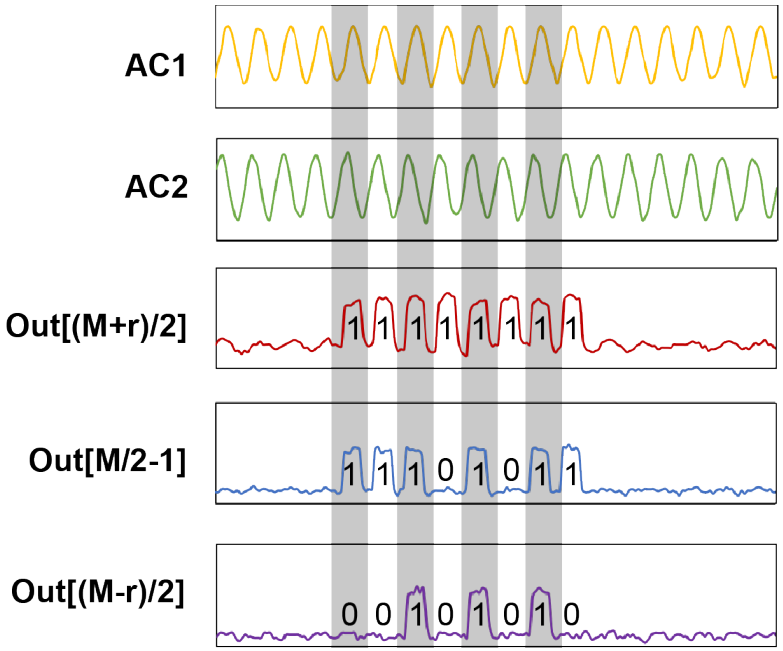


Hardware Implementation of a 32-bit Sorter

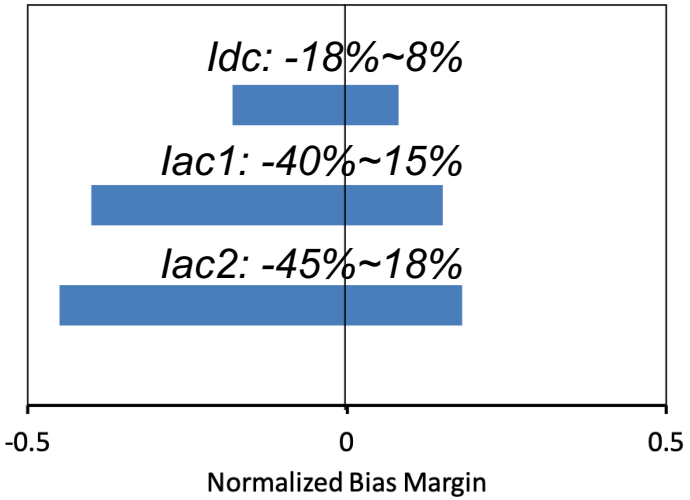


- Largest auto-designed AQFP circuit
- Fabricated with AIST 10kA/cm^2 HSTP

Low Speed Test Results @100kHz



M=32, r=30



Flux trapping observed, all 32 outputs cannot be generated simultaneously.

Count of 1s in Inputs	17	17	32	16	32	16	31	17

Summary

- A framework for AQFP-based DNN has been established.
  - Stochastic computing
  - Design automation
  - Hardware implementation

Technology	Power (μm)	Delay (ps)	EPC (fJ)
TSMC 12nm	18.449	0.35	6.4572
TSMC 28nm	34.141	1.49	50.8701
TSMC 40nm	61.967	2.57	159.2552
AQFP HSTP	--	--	0.0049

- Future works
  - Yield analysis
  - TNN implementation

Acknowledgment

Office of the Director of National Intelligence
IARPA
BE THE FUTURE

SuperTools Program

日本学術振興会
Japan Society for the Promotion of Science

JSPS KAKENHI Grant Number 19K15041

CRAVITY
AIST

The circuits were fabricated in the Clean Room for Analog-digital superconductivity (CRAVITY) of National Institute of Advanced Industrial Science and Technology (AIST) with the highspeed standard process (HSTP).

NSF

National Science Foundation Grant No. OISE-1854213