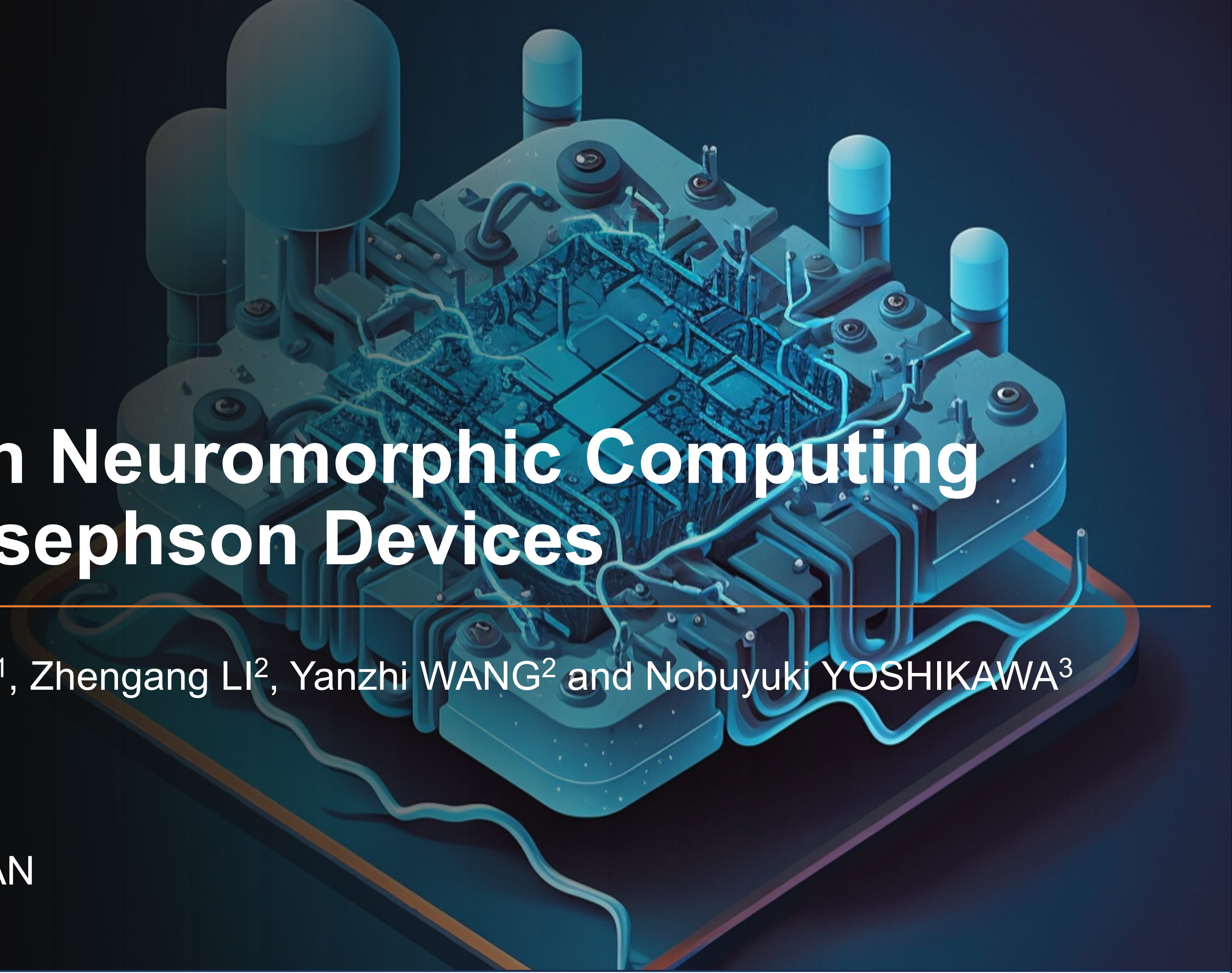


Recent Progress on Neuromorphic Computing Using Adiabatic Josephson Devices

Olivia CHEN[1], Tomoharu YAMAUCHI[1], Zhengang LI[2], Yanzhi WANG[2] and Nobuyuki YOSHIKAWA[3]

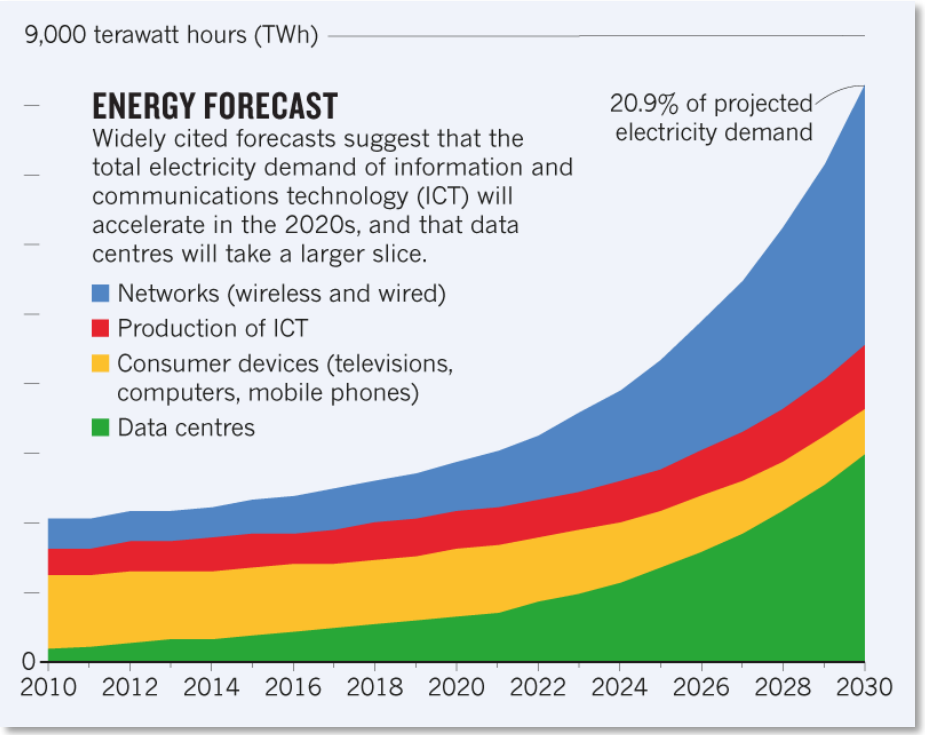
[1] Tokyo City University, JAPAN

[2] Northeastern University, USA

[3] Yokohama National University, JAPAN

IEEE-CSC, ESAS and CSSJ SUPERCONDUCTIVITY NEWS FORUM (global edition), Issue No. 55, January, 2024. Invited presentation given at EUCAS 2023, September 4, 2023, Bologna, Italy

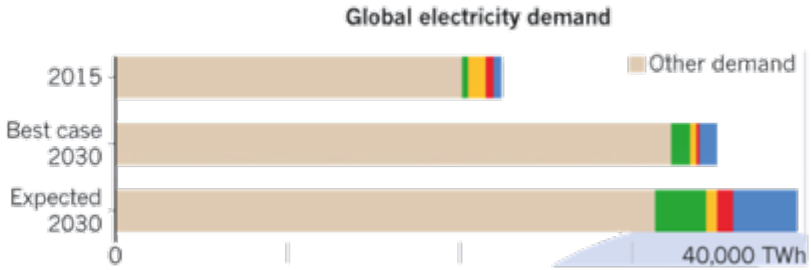
Motivation



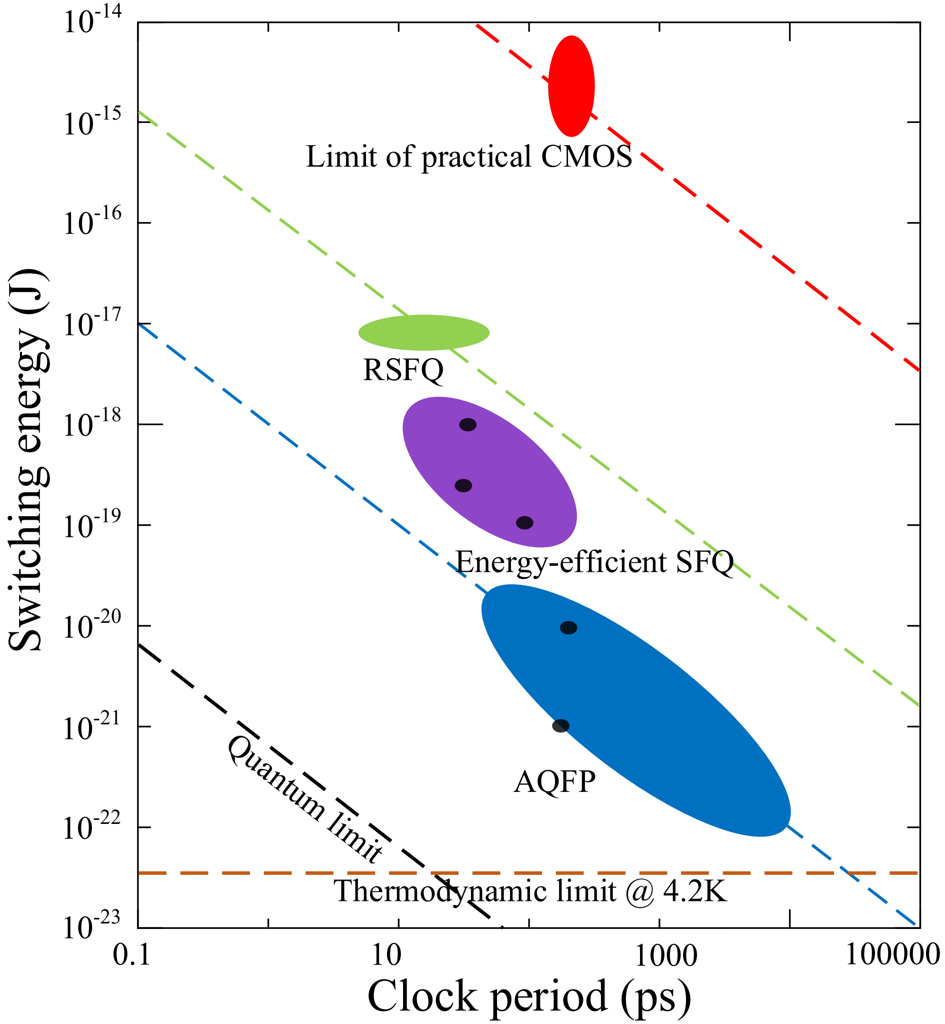
ENERGY FORECAST

Widely cited forecasts suggest that the total electricity demand of information and communications technology (ICT) will accelerate in the 2020s, and that data centres will take a larger slice.

20.9% of projected electricity demand



Global electricity demand



Back to 2018

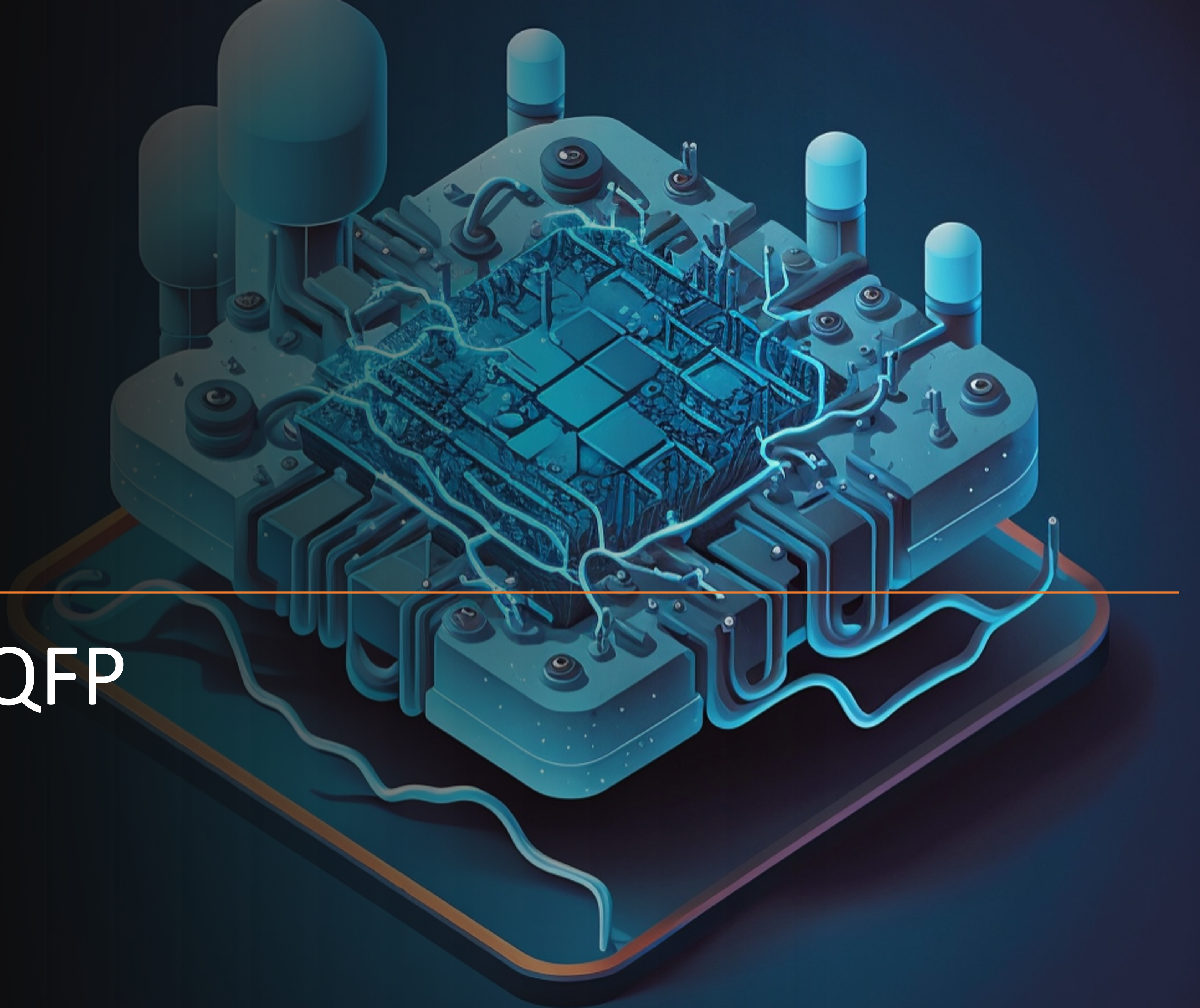
O. CHEN ***ASC 2018***, Young Professional Plenary, Seattle, US

- First digital circuit based machine learning acceleration attempt
- Using adiabatic quantum-flux-parametron (AQFP) devices

What has been achieved during the past 5 year?

Outline

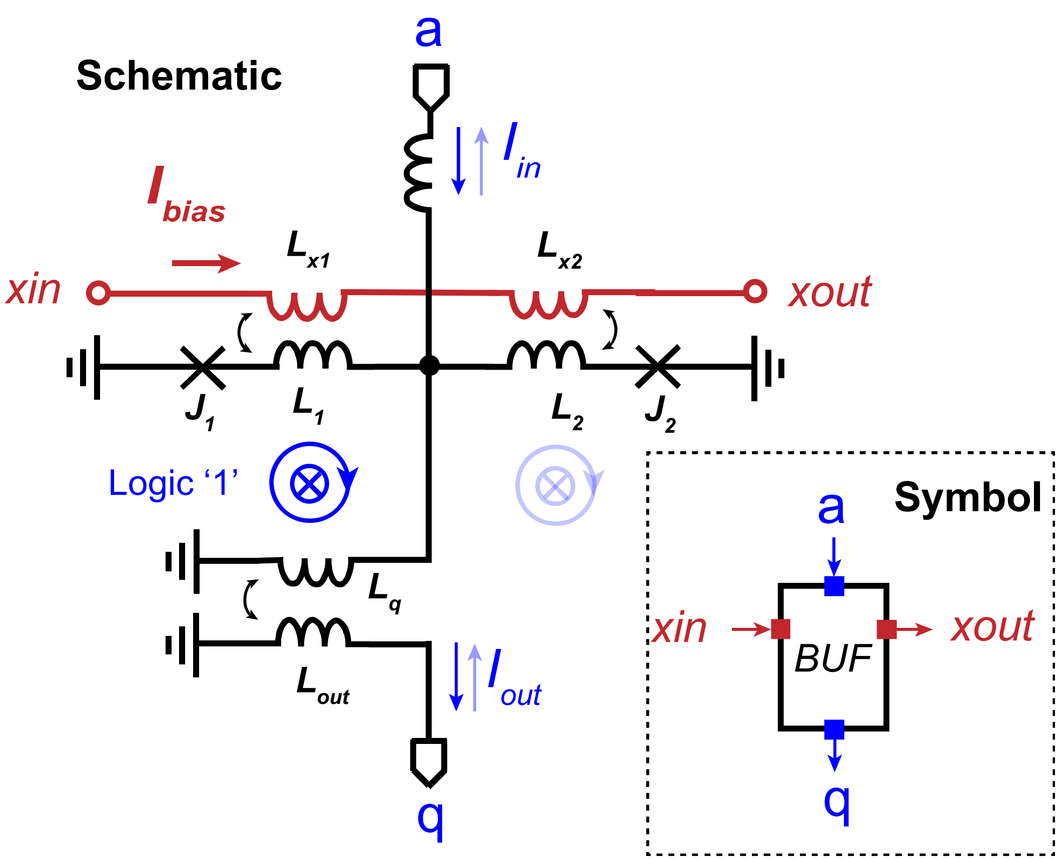
- Introduction
    - AQFP basics
    - Design methodology
- AQFP-based nerual network acceleration
    - Stochastic computing-based neural network design
    - Binarized neural network design
- Hardware-Algorithm Co-optimization
    - EDA-based circuit optimization
    - Hardware-oritented training optimization
    - Architecture optimizatition



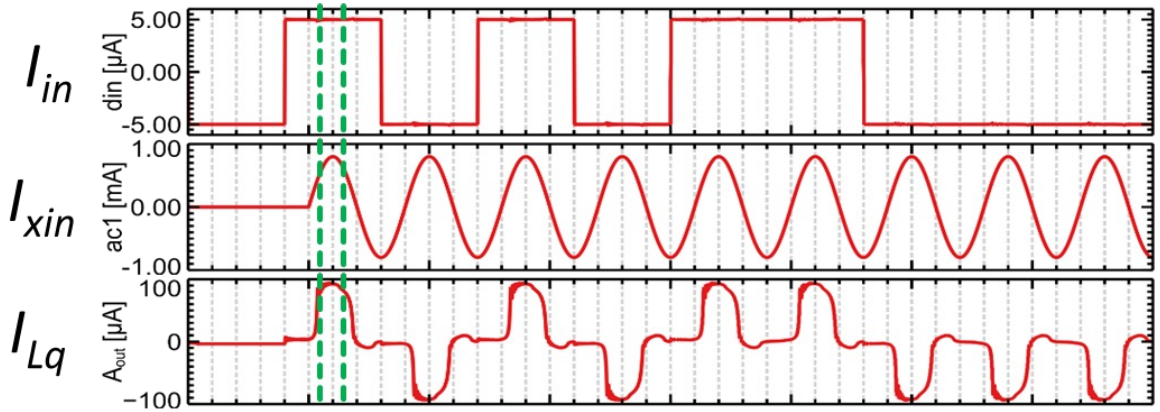
Introduction to AQFP

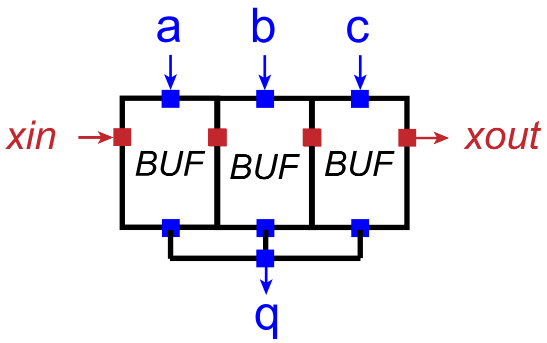
Adiabatic Quantum Flux Parametron (AQFP) Logic

Basic idea to resolve the static power issue in SFQ: replace DC with AC, static power -> 0

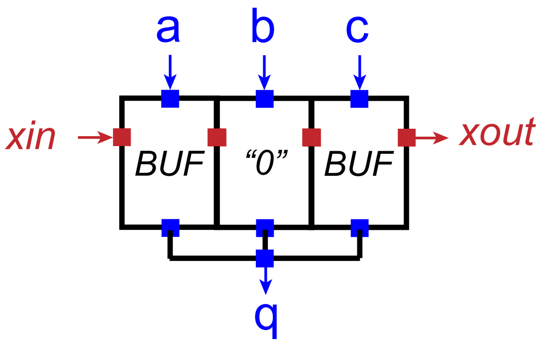


Schematic of an AC-biased buffer





Majority Logic



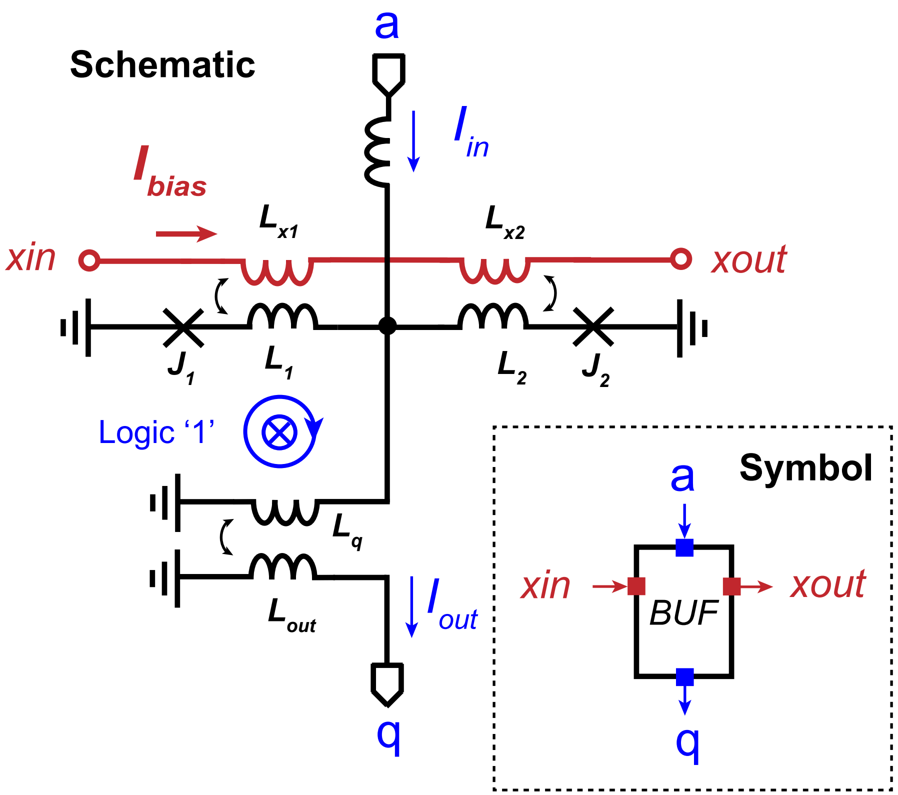
AND Logic

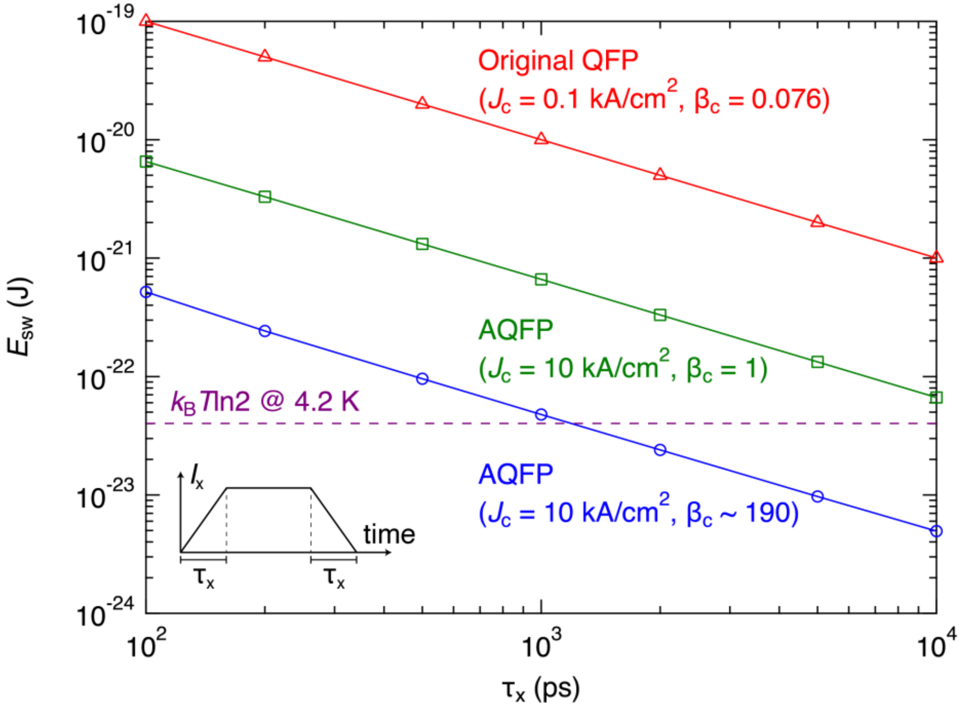
Possible majority-based logic representation instead of traditional AND-OR-INV

Adiabatic Quantum Flux Parametron (AQFP) Logic

Further optimization to resolve the static power issue in SFQ: parameter adjustment

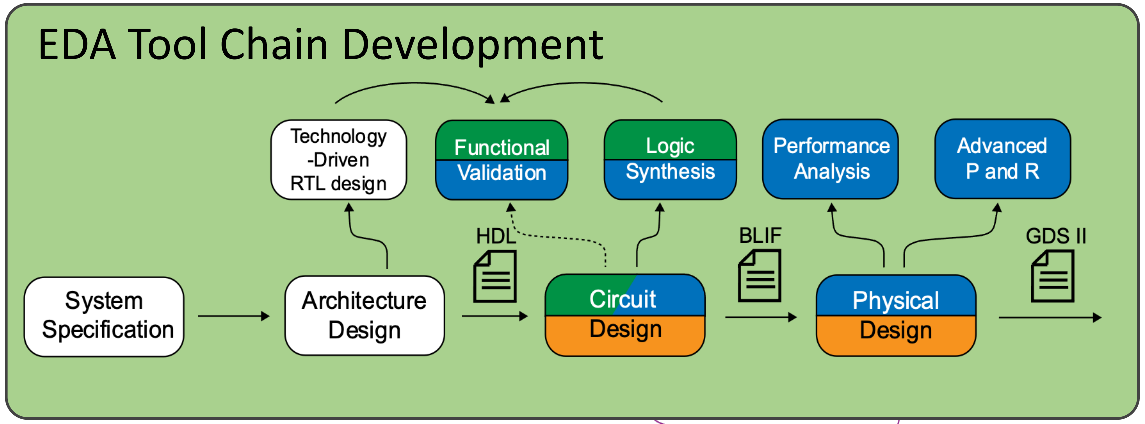
low I_c (50 μA), high J_c (10 kA/cm^2), high-βc (underdamped)



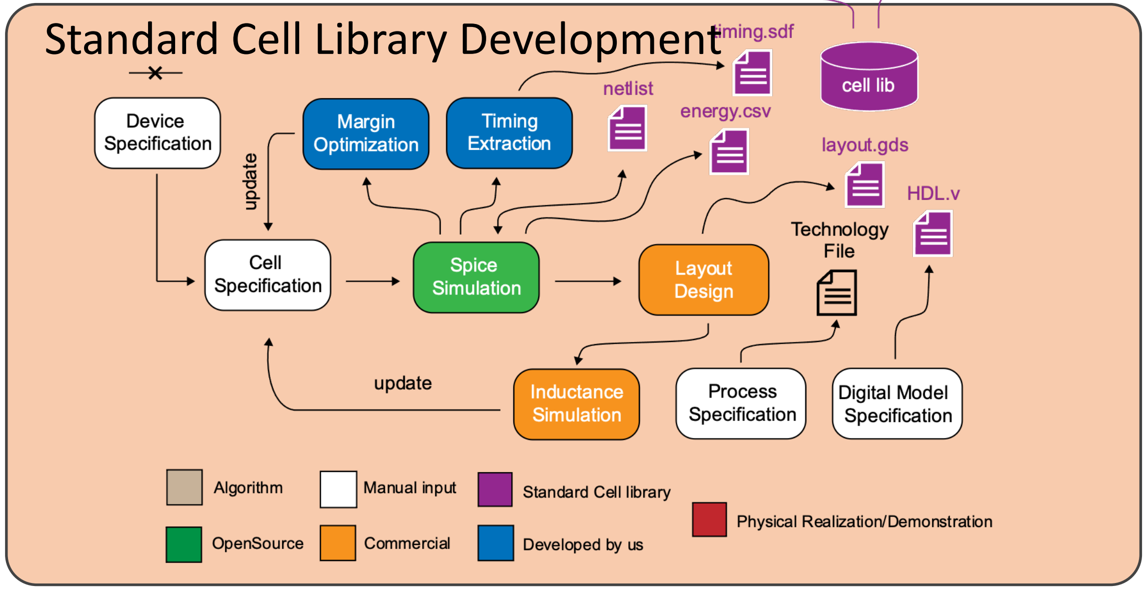


N. Takeuchi, et al., IEICE TRANS. ELECTRON., VOL.E105–C, NO.6 JUNE 2022

AQFP Circuit Design Methodology

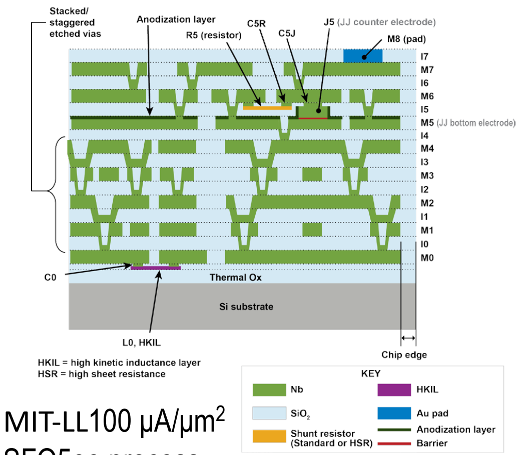


EDA Tool Chain Development

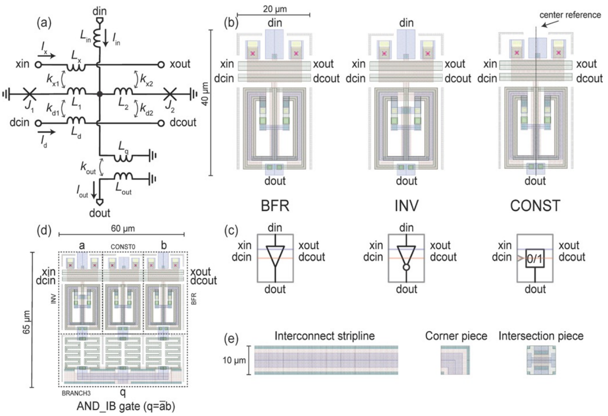


Standard Cell Library Development

▨ Algorithm □ Manual input 🟪 Standard Cell library 🟥 Physical Realization/Demonstration
🟩 OpenSource 🟧 Commercial 🟦 Developed by us



MIT-LL100 μA/μm^2
SFQ5ee process



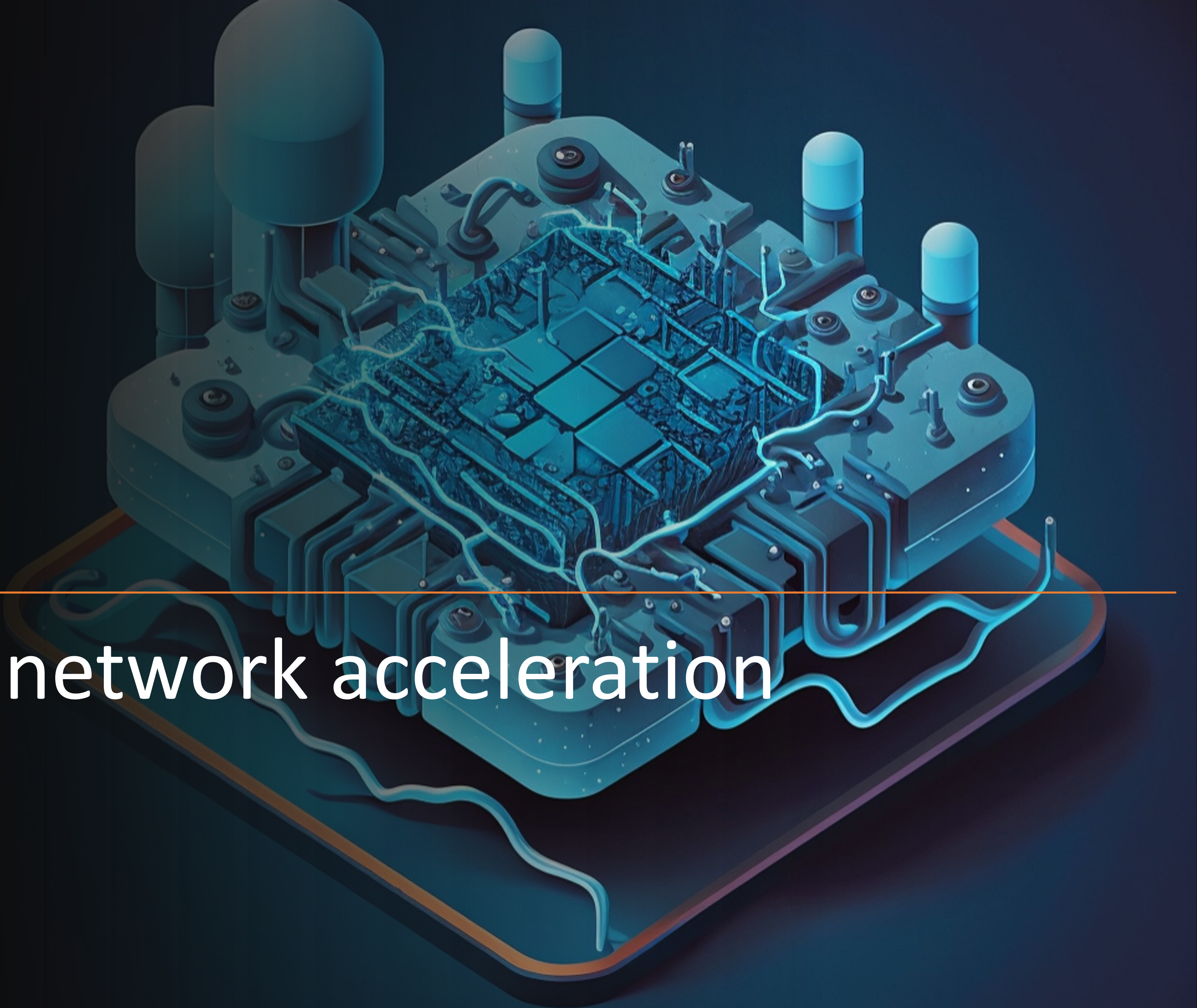
Cell library example

US IARPA SuperTools Program (2017 〜 2022)

- Standard cell library w/ more 80+ gates design for MIT-LL SFQ5ee process
- EDA tool chain for the designed cell library
  - Majority Logic Syntheis
  - Timing allignment optimization
  - Process tailored placement and routing
  - Probablistic power analysis

Yokohama National Univ.
Northeasten Univ., USC,
Stellenbosch Univ. (Coldflux Team)

C. J. Fourie et al., ***IEEE TAS*** 2023.

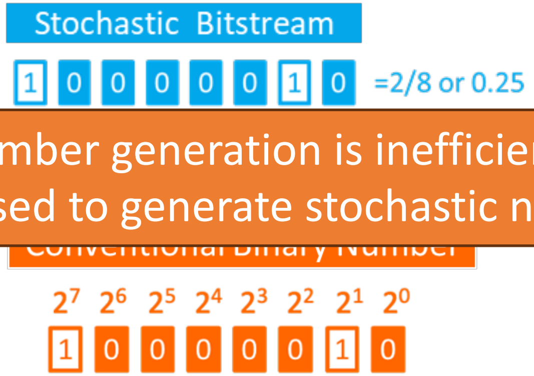


AQFP-based nerual network acceleration

AQFP for Stochastic Computing

Stochastic Computing

- Approximation computing
- Convert numbers to probability r
- N
- Extremely small hardware footprint
- Robust against errors



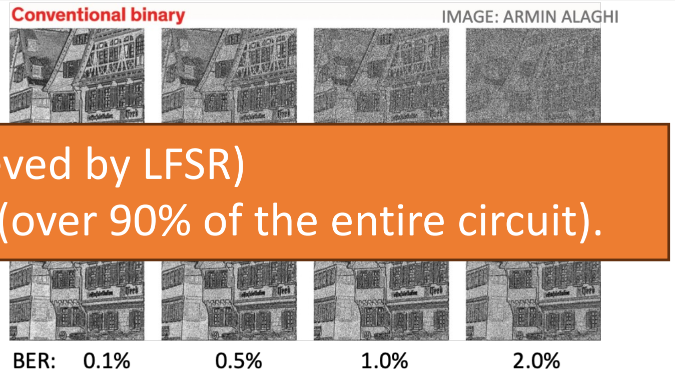
Stochastic Bitstream

1 0 0 0 0 0 1 0 =2/8 or 0.25

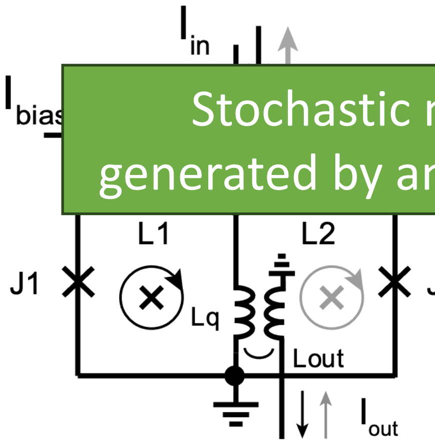
Conventional Binary Number

2^7 2^6 2^5 2^4 2^3 2^2 2^1 2^0

1 0 0 0 0 0 1 0

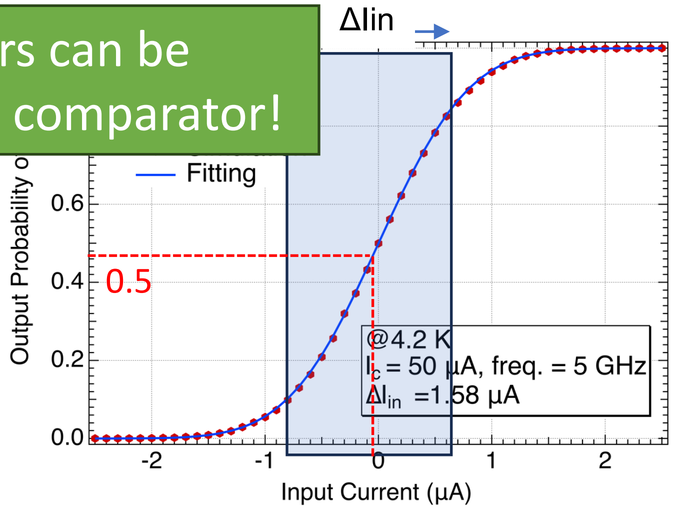


- In conventional CMOS, random number generation is inefficient (achieved by LFSR)
- Area issue by digital comparator used to generate stochastic numbers (over 90% of the entire circuit).

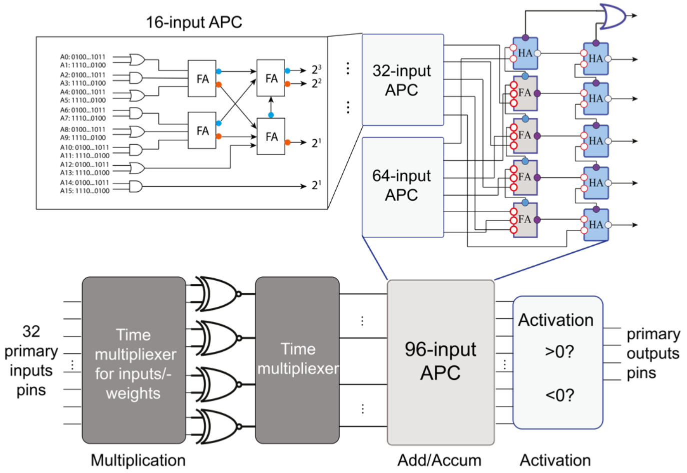


Stochastic numbers can be generated by an AQFP comparator!

Relationship curve between input current and output probability.



ASC 2018 Young Professional Plenary

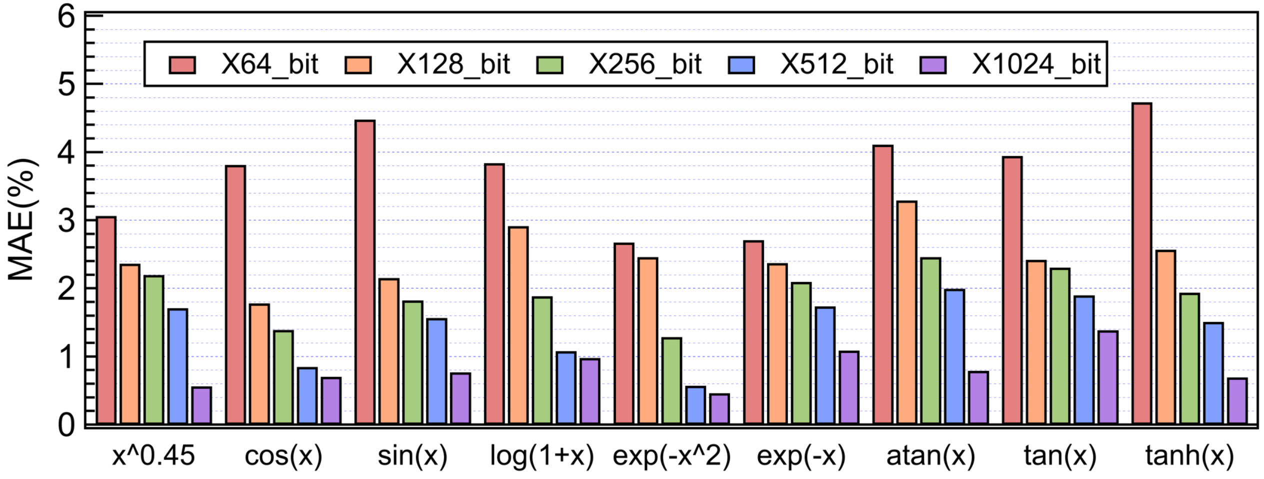


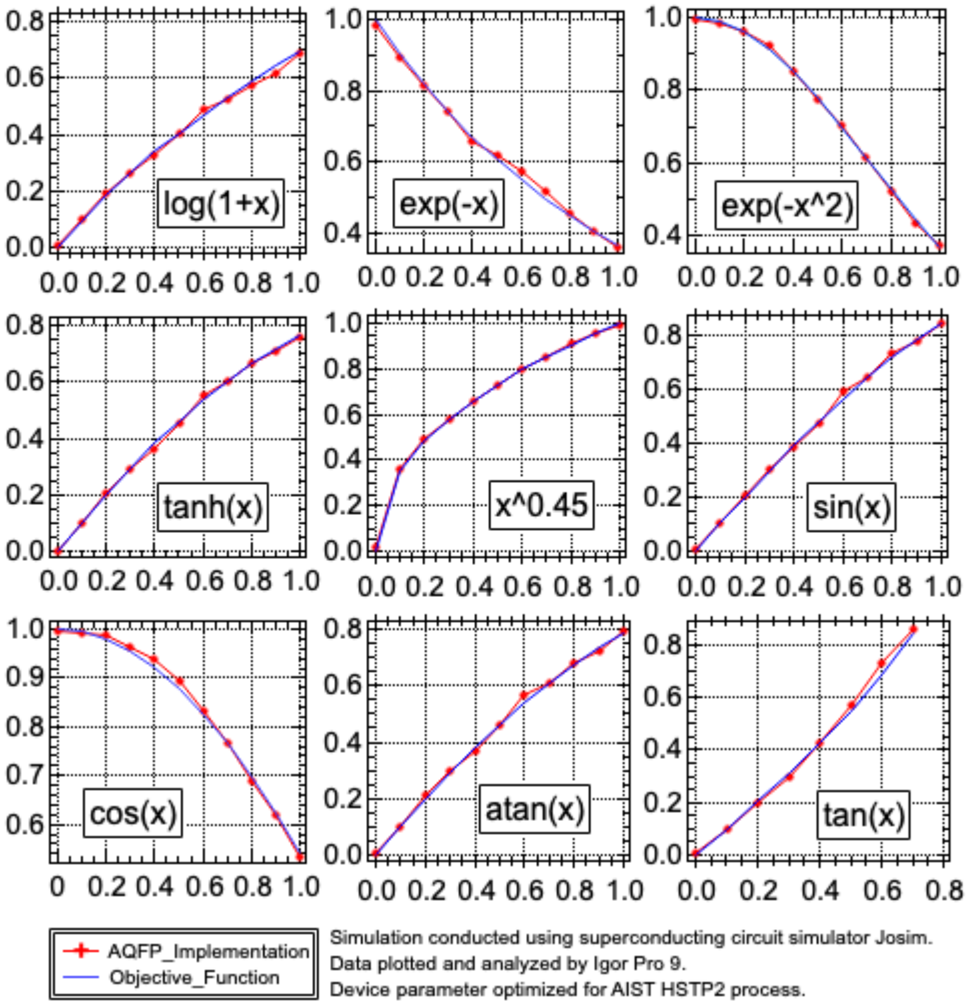
Stochastic computing based AQFP MAC circuit.

- Superconducting current comparator with symmetrical structure.
- When Iin = 0, it has random behavior due to thermal noise

Non-linear Function Approximation w/ SC + AQFP

- Berstein polynomials used for approximation
- 9 functions are employed for benchmarking
- Input and output normalized in range [0,1] for SC paradigm
- Bit-stream length from 64 to 1024 are tested
- Device parameter optimized for AIST HSTP process

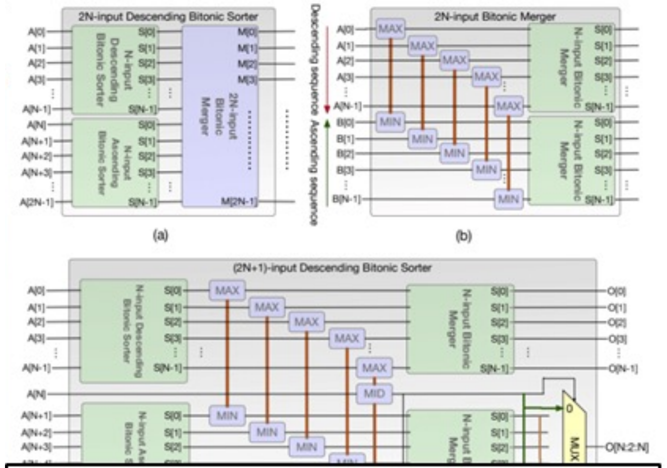




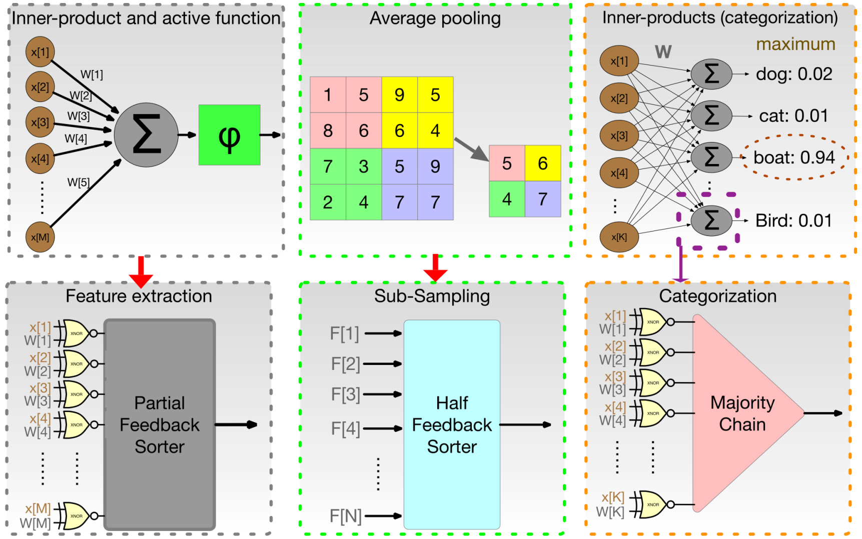
Simulation conducted using superconducting circuit simulator Josim.
Data plotted and analyzed by Igor Pro 9.
Device parameter optimized for AIST HSTP2 process.

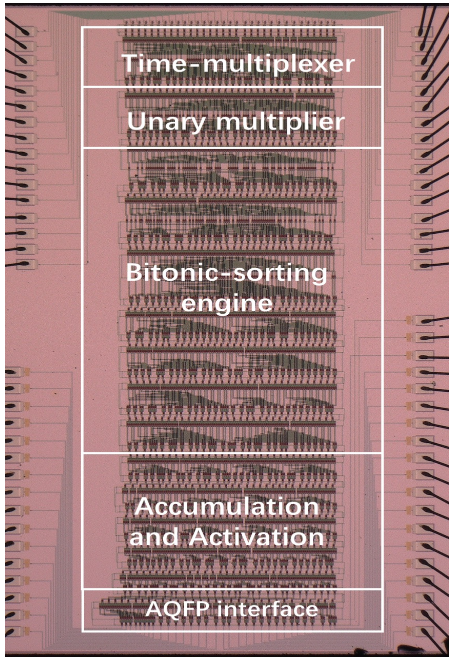
AQFP SC-based Neural Network

Collaboration w/ Northeastern University


2N-input Descending Bitonic Sorter
N-input Descending Bitonic Sorter
N-input Ascending Bitonic Sorter
2N-input Bitonic Merger
Descending sequence
Ascending sequence
(a)
2N-input Bitonic Merger
N-input Bitonic Merger
MAX
MIN
(b)
(2N+1)-input Descending Bitonic Sorter
N-input Descending Bitonic Sorter
N-input Bitonic Merger
MID
MUX

Bitonic−sortingを用いた畳込み演算





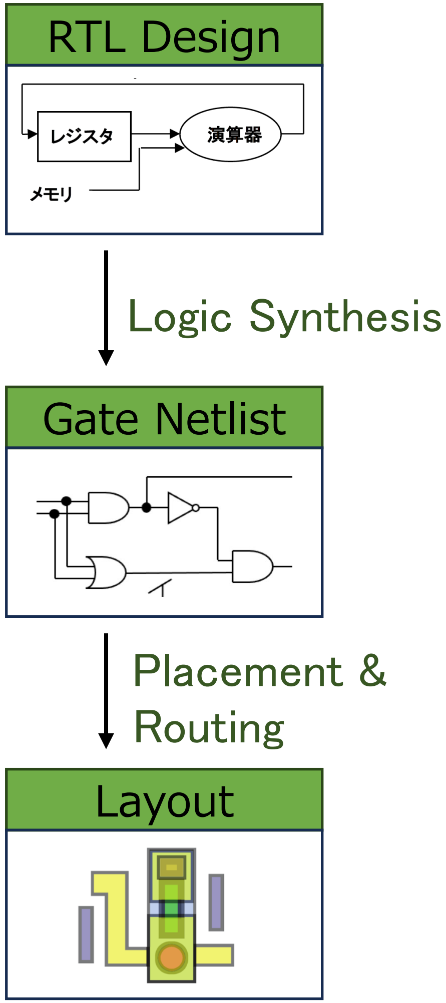
- Efficiency improvement of multiplication-sum + activation operation using Bitonic sorter structure
- Automated design using toolsets developed in 'SuperTools’
- Prototype circuit implementation and low-temperature operation demonstration

Tested on MNIST dataset

Network	Platform	Accuracy	Energy(μJ)	Throughput(images/ms)
SNN	Software	99.04%	–	–
	CMOS	97.35%	39.46	231
	AQFP	97.91%	5.606E-4	8305
DNN	Software	99.17%	–	–
	CMOS	96.62%	219.37	229
	AQFP	96.95%	2.482E-3	6667

R. Cai, O. Chen, et al., ***ISCA*** 2019
O. Chen et al., ***SOCC*** 2022

Circuit Optimization



RTL Design

Logic Synthesis

Gate Netlist

Placement & Routing

Layout

Majority Logic Synthesis

GLSVLSI'19 LUT-based cell mapping (R. Cai et al.)

DATE'23 Bayesian optimization considering latency and fanout (R. Fu et al.)

Timing Alignment

ICCD'20 Heuristic algorithms (R. Cai et al.)

ASPDAC'23 Gate count optimization using dynamic programming and approximate solutions of ILP (R. Fu et al)

Placement

IEEE TAS'16 GA-based placement (Murai et al.)

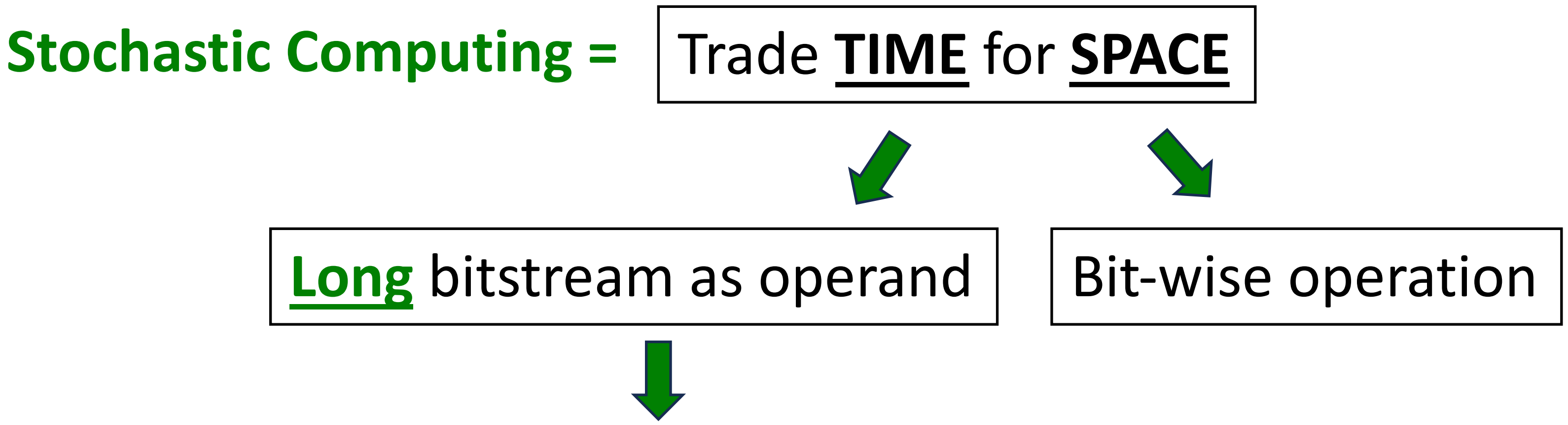
ICCAD'20 Analytical global placement and row-wise detailed placement (Y. Chang et al.)

DAC'22 Timing-aware placement using convex optimization (P. Dong et al.)

ICCAD'23 Placement optimization for delay-line clocking scheme (R. Fu et al.)

*Result under SuperTool Program

Now What is Problem?



Stochastic Computing = Trade **TIME** for **SPACE**

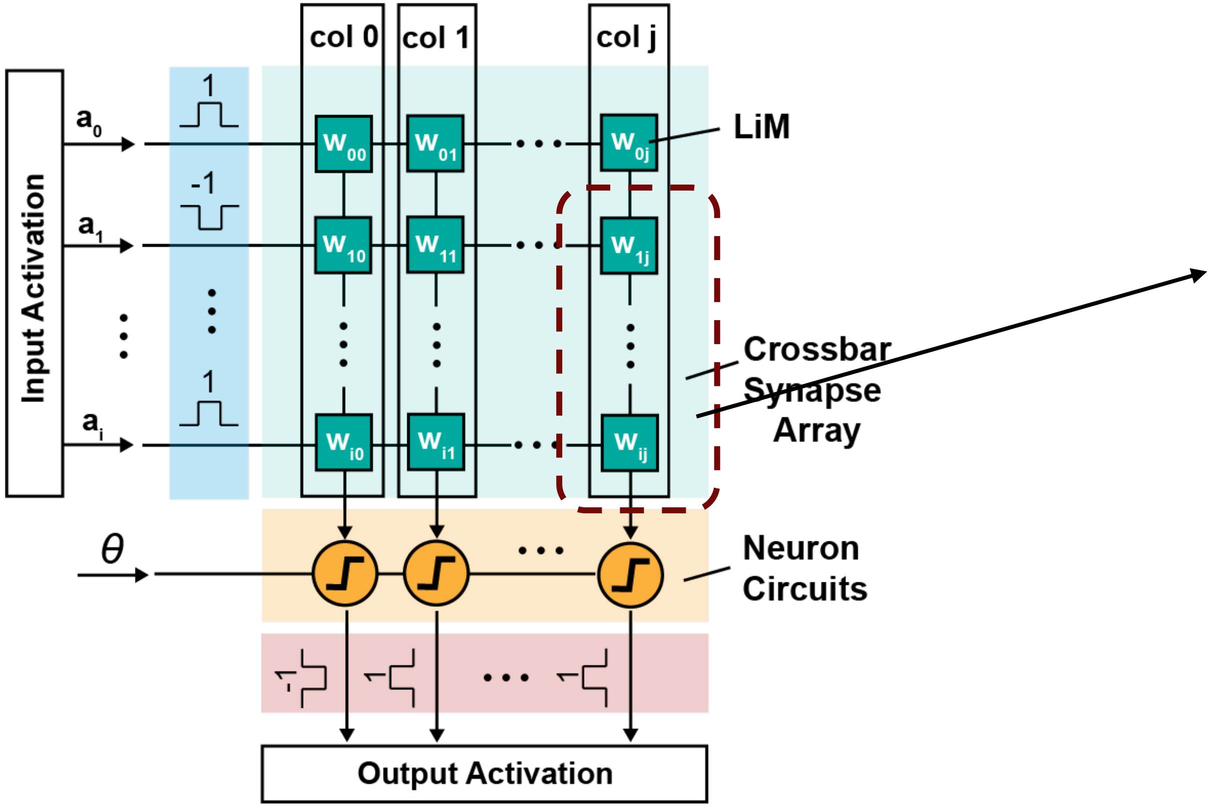
Long bitstream as operand

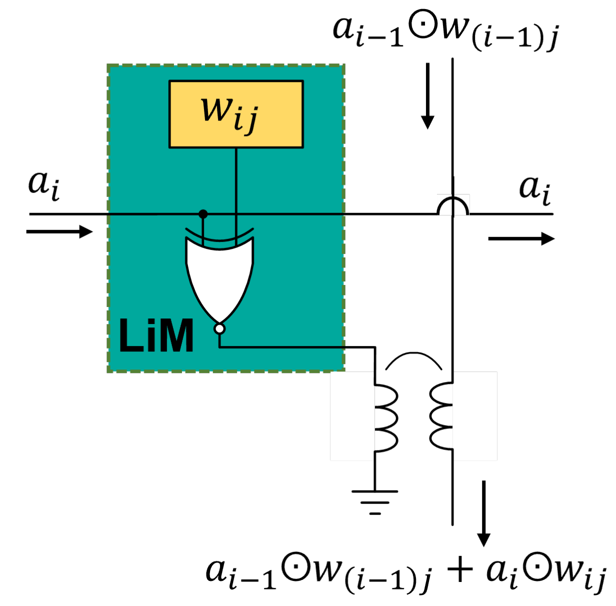
Bit-wise operation

1. To achieve reasonable accuracy for CIFAR-10 dataset, 2048 ～ 4096 expected

2. Eventhough applying data-level parallelism, still a processor-memory seperated stucture

Crossbar Architecture+ Binary Neural Network



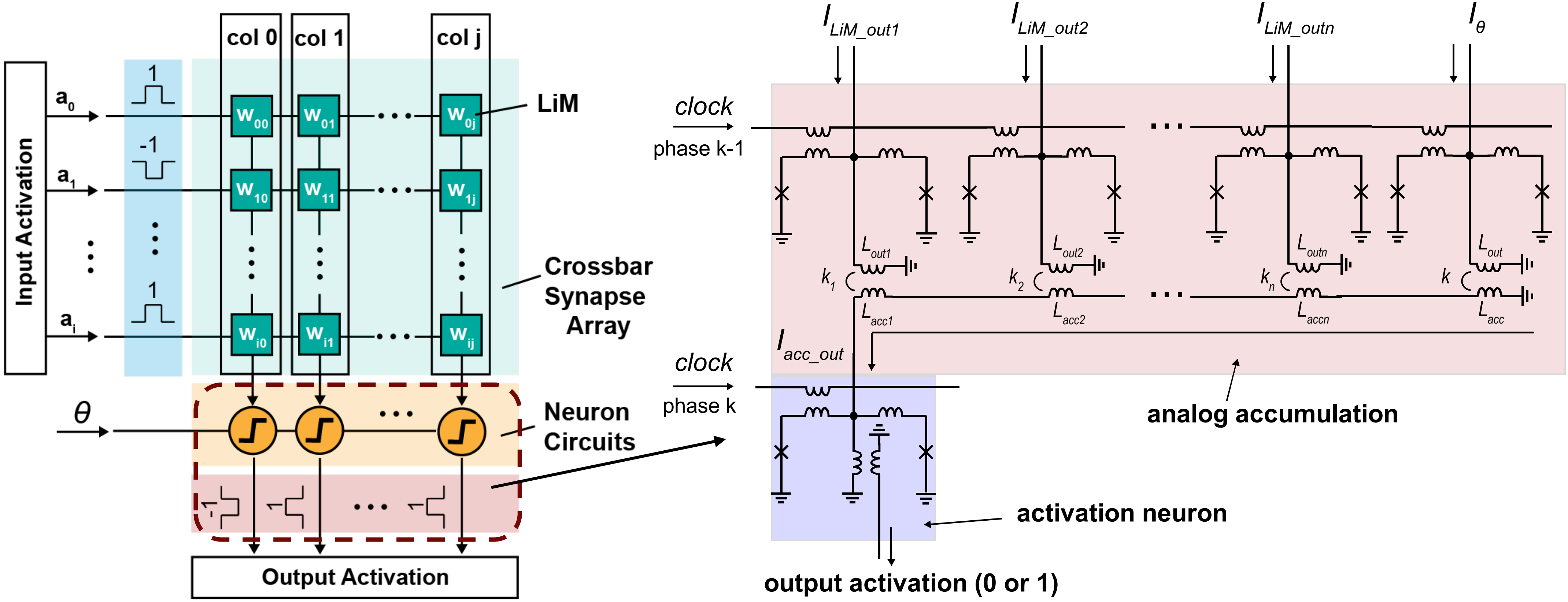


$a_{i-1} \odot w_{(i-1)j}$
w_{ij}
a_i
a_i
LiM
$a_{i-1} \odot w_{(i-1)j} + a_i \odot w_{ij}$

T. Yamauchi, O. Chen, et al., IEEE TAS, 2023.

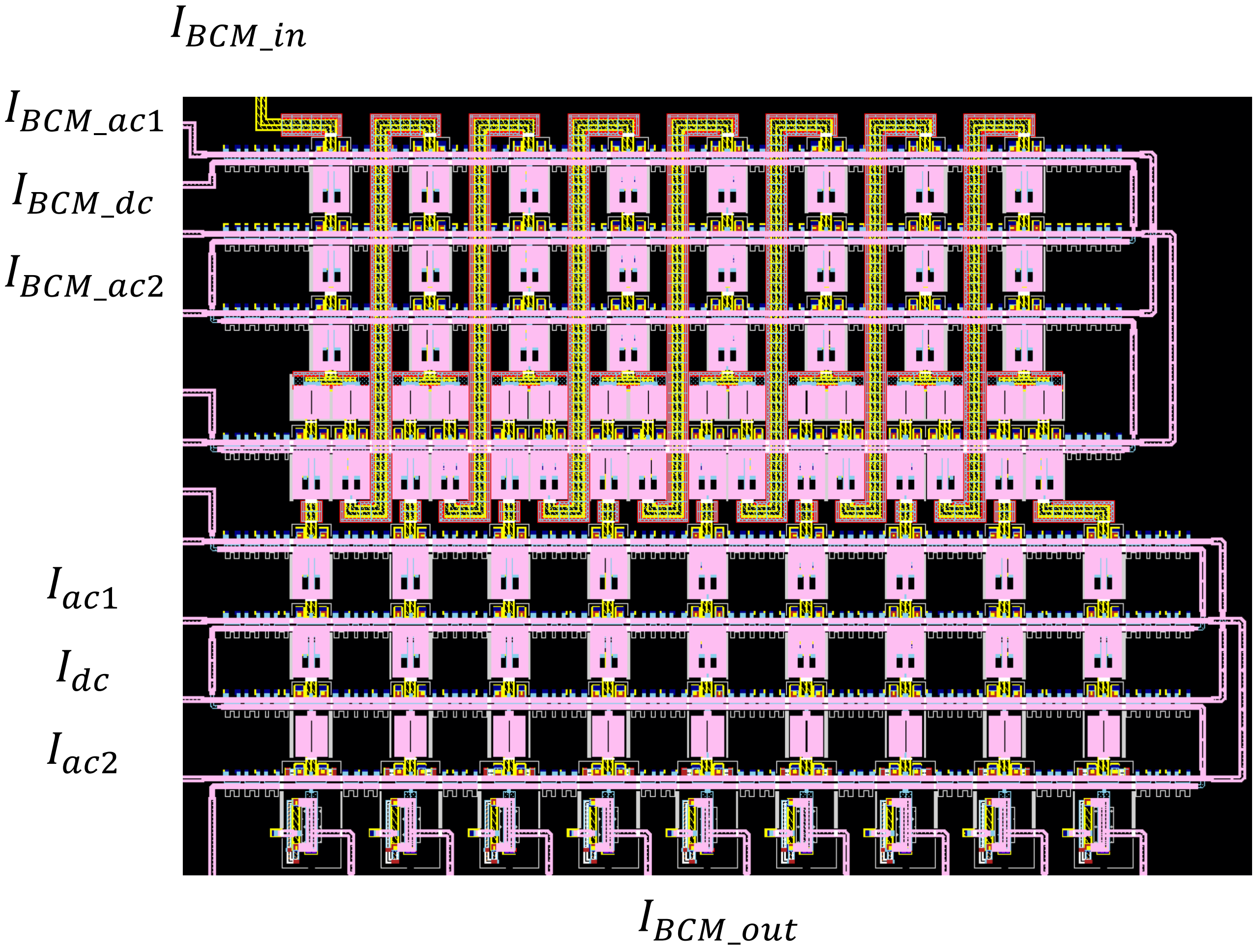
- Logic-in-memory cell design to perform binary multiplication w/ prestored 1-bit weight

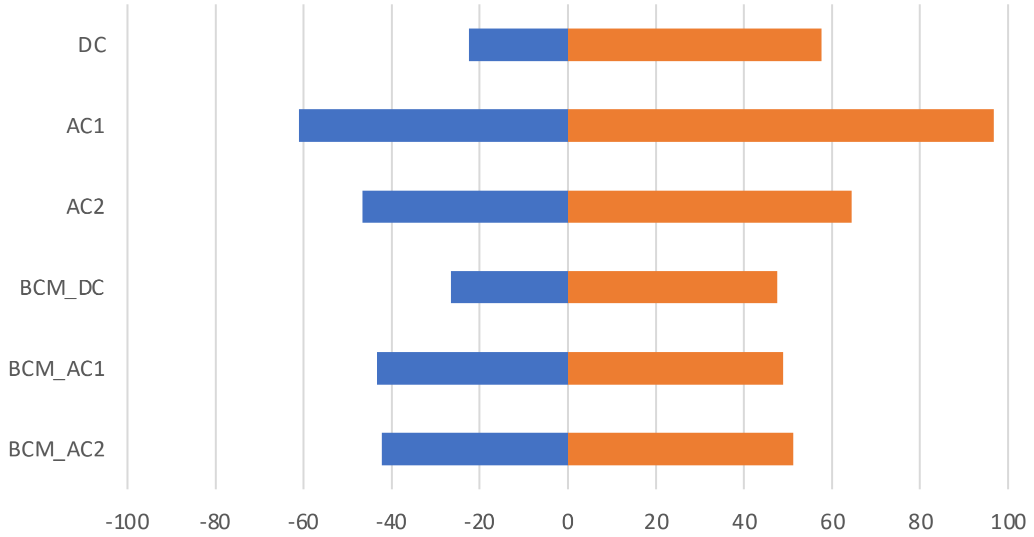
Analog Accumulation and AQFP Comparator-based Neuron



- Analog current accumulation for column summation
  (1 and 0 are represented by positive and negative current pulses in AQFP)
- Flux coupling /Current accumulated via superconducting inductance
- AQFP comparator servers as neuron to perform activation

Implementation Example of Memory



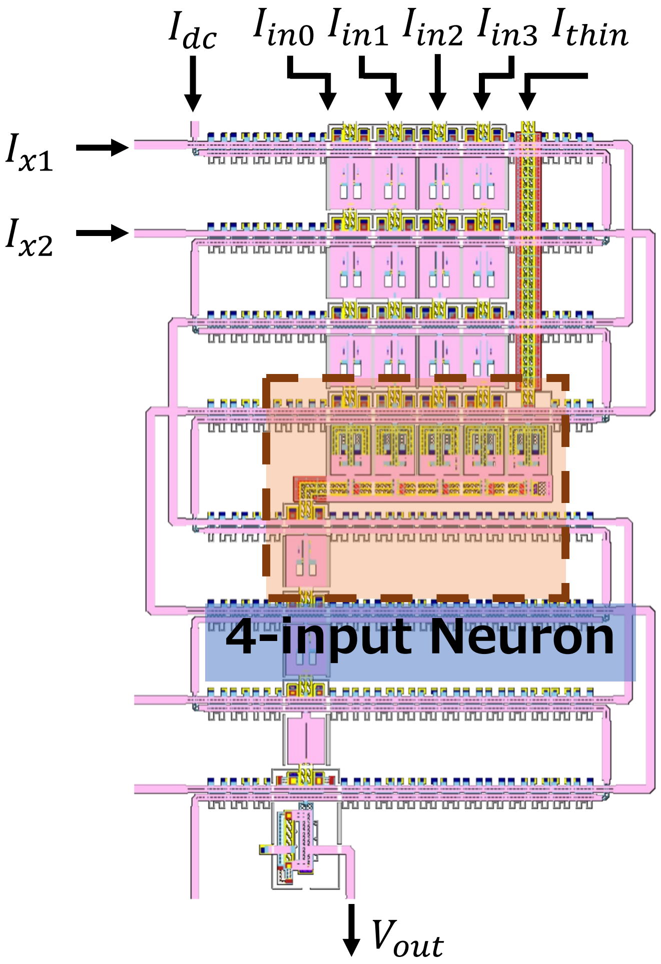


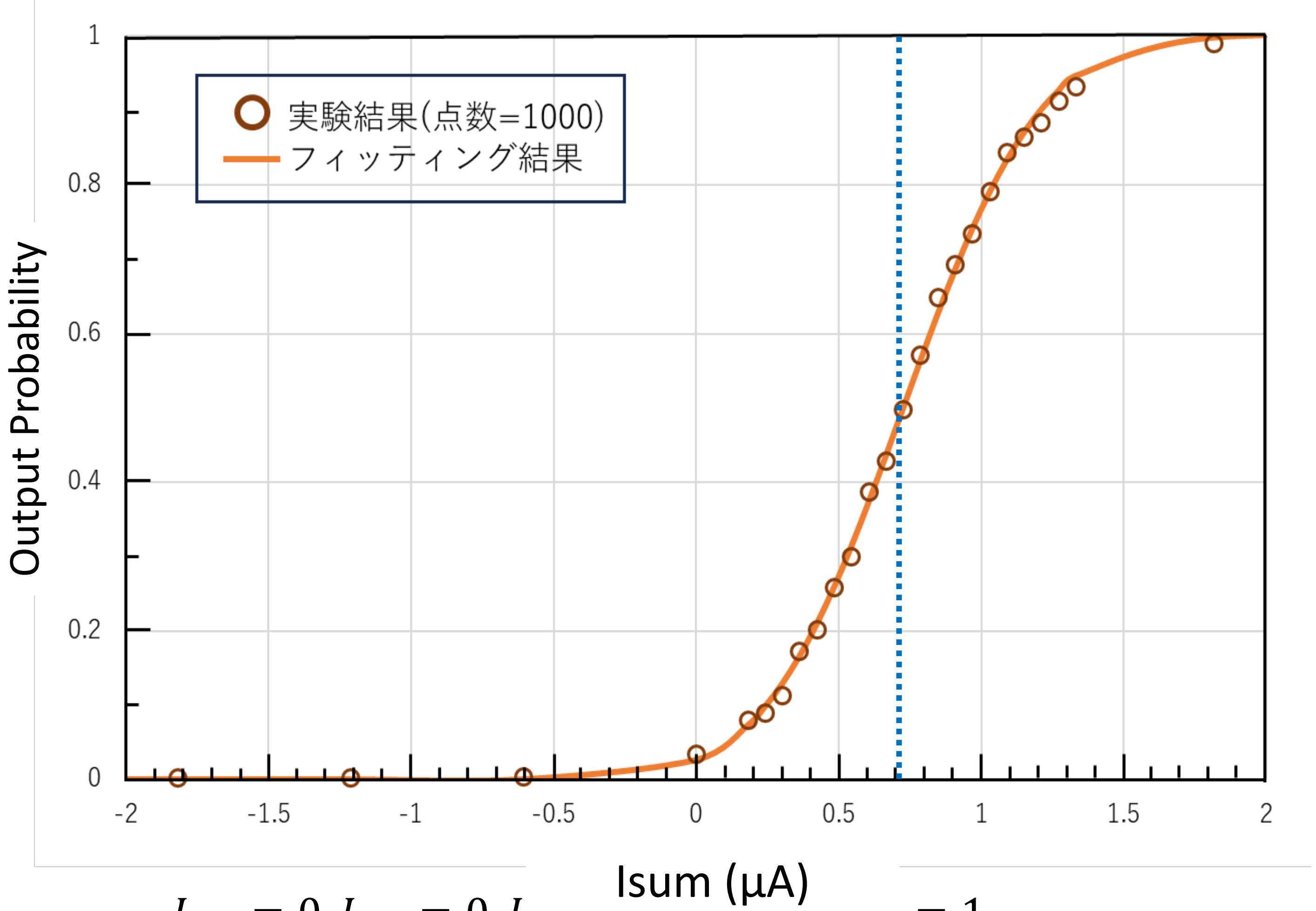
DC	-22.5%	57.5%
AC1	-61.1%	96.7%
AC2	-46.7%	64.4%
BCM_DC	-26.7%	47.5%
BCM_AC1	-43.3%	48.9%
BCM_AC2	-42.2%	51.1%

- Serial write_in parallel read out memory using delayed buffer chain

Superconducting Quantum Circuit Fabrication Facility

Implementation of an Example 4-input Neuron Circuit



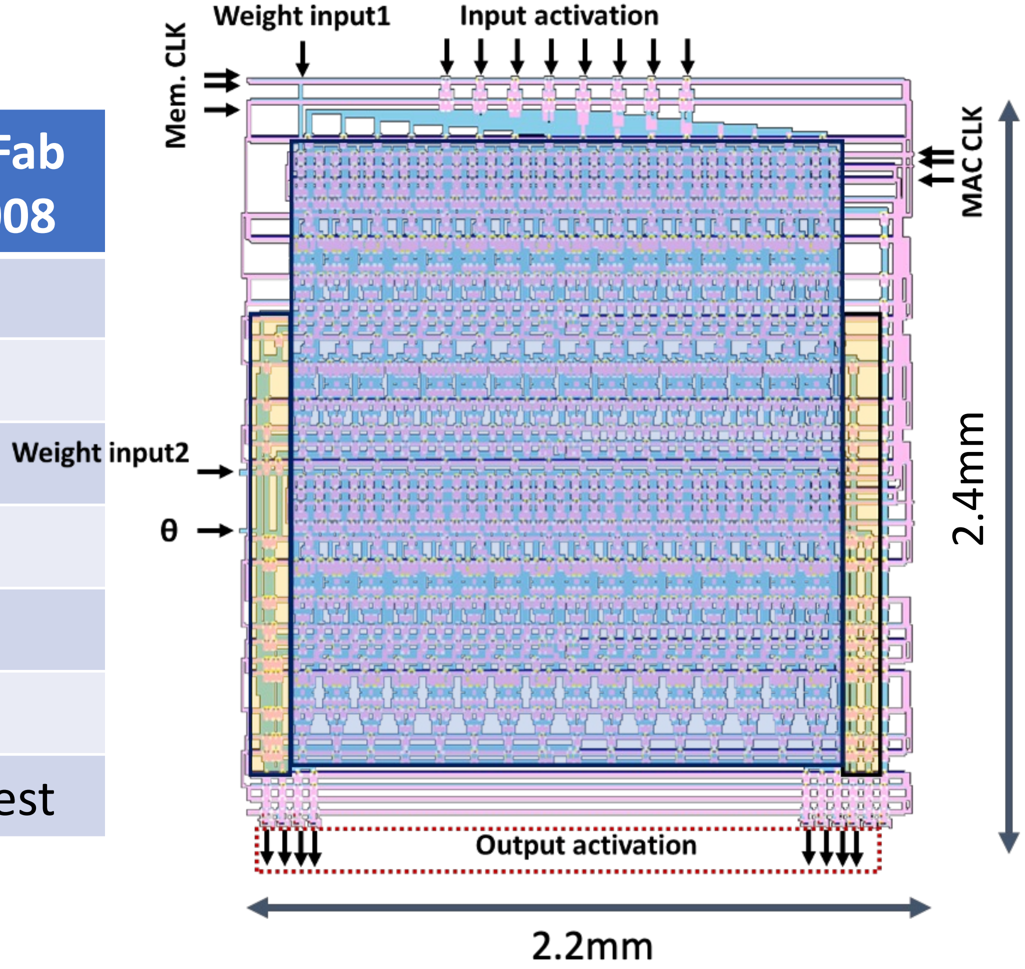


$I_{in0} = 0, I_{in1} = 0, I_{in2} = 1, I_{in3} = 1, I_{th} = 1$

Qufab Superconducting Quantum Circuit Fabrication Facility

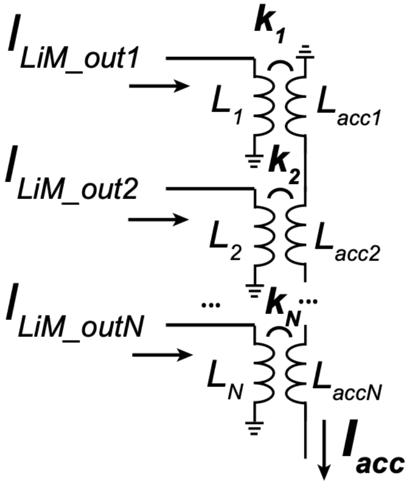
Module Implementation and Test Summary

Module Name	JJ Count	ASIT QuFab HSTPA 007	ASIT QuFab HSTPA 008
Offset XNOR	22	×	○
BCM (8-bit)	152	N/A	○
Neuron (4-in)	42	○	N/A
Neuron (8-in)	74	○	N/A
Neuron (16-in)	138	○	N/A
4x4 BNN	690	×	○
8x8 BNN	2236	N/A	Under Test

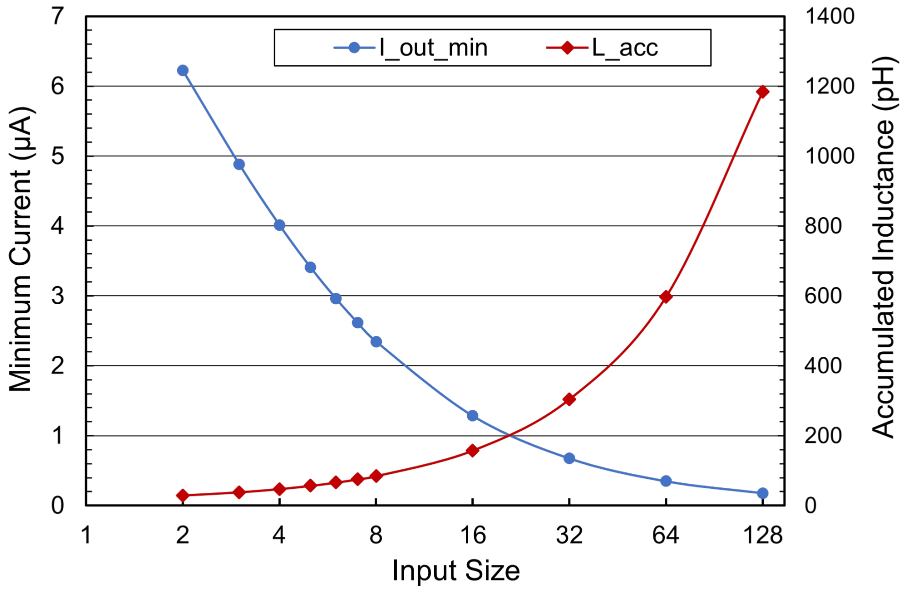


QuFab Superconducting Quantum Circuit Fabrication Facility

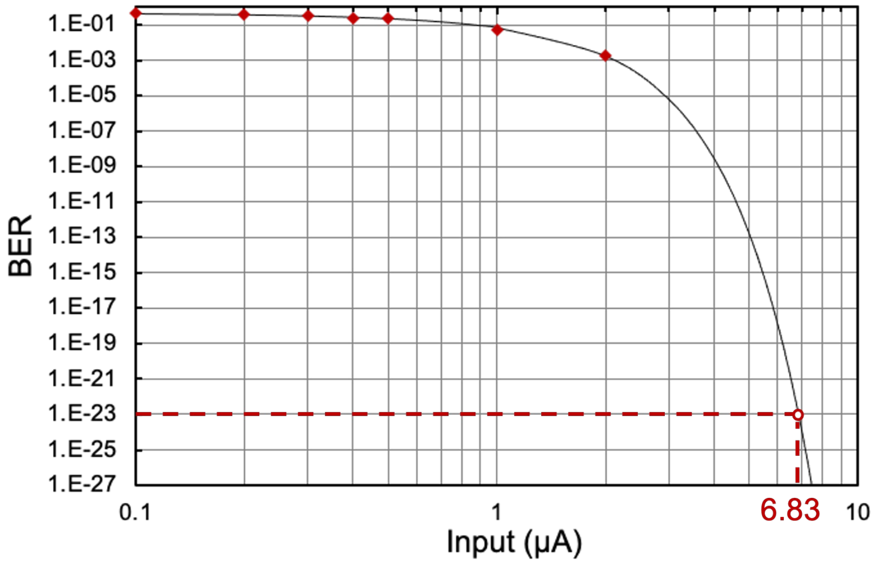
Chanllenges



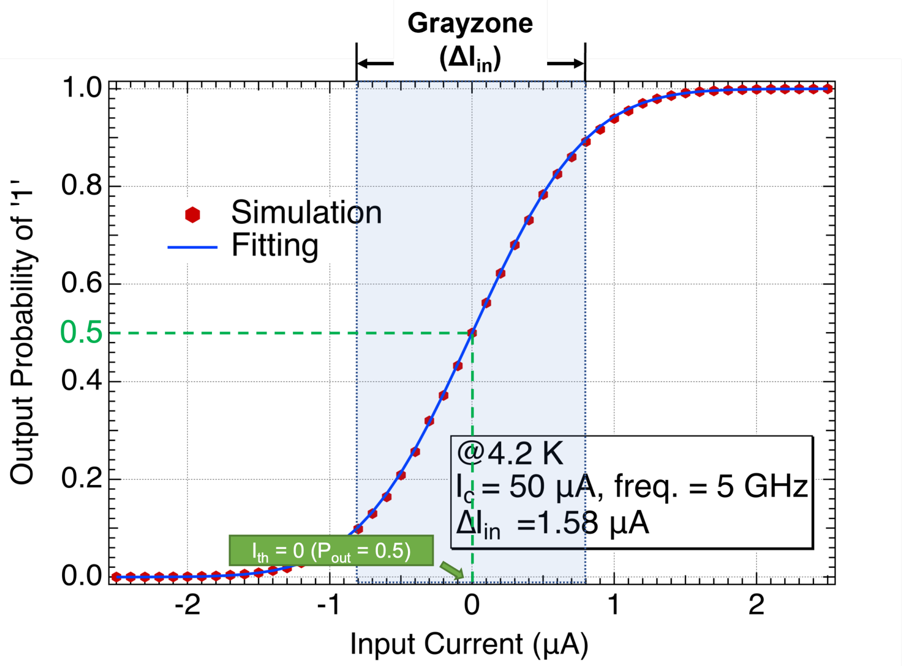
$I_{acc} \propto 1/\Sigma\, L_{acc}(n)$



- Accumulated current attenuate due to the large inductance for magnetic coupling.
- Neurons are not able to perform accumulation function.



- Minimum input current for an AQFP comparator to output digital information (BER~10e-23) would be ±6.83µA



O. Chen, et al., IEEE AICAS2023.

Algorithm-Hardware Co-Optimization

Review: Problems and Motivations

- Stochastic switching of AQFP neurons. (Not step function anymore)
- Current attenuation in AQFP-based crossbar related to the crossbar size.
- Software and hardware mismatch caused by stochastic switching and current attenuation.
- How to decide the operatable crossbar size?
- How to accumulate results of multiple crossbar efficiently?
- Hardware configurations work on both energy-efficiency and model accuracy!

Assessments on Stochastic Switching on AQFP Neuron and Crossbar Current

- Switching probability fitting using error function

$$P(I_{in}) = 0.5 + 0.5erf\left(\sqrt{\pi}\frac{(I_{in} - I_{th})}{\Delta I_{in}}\right)$$

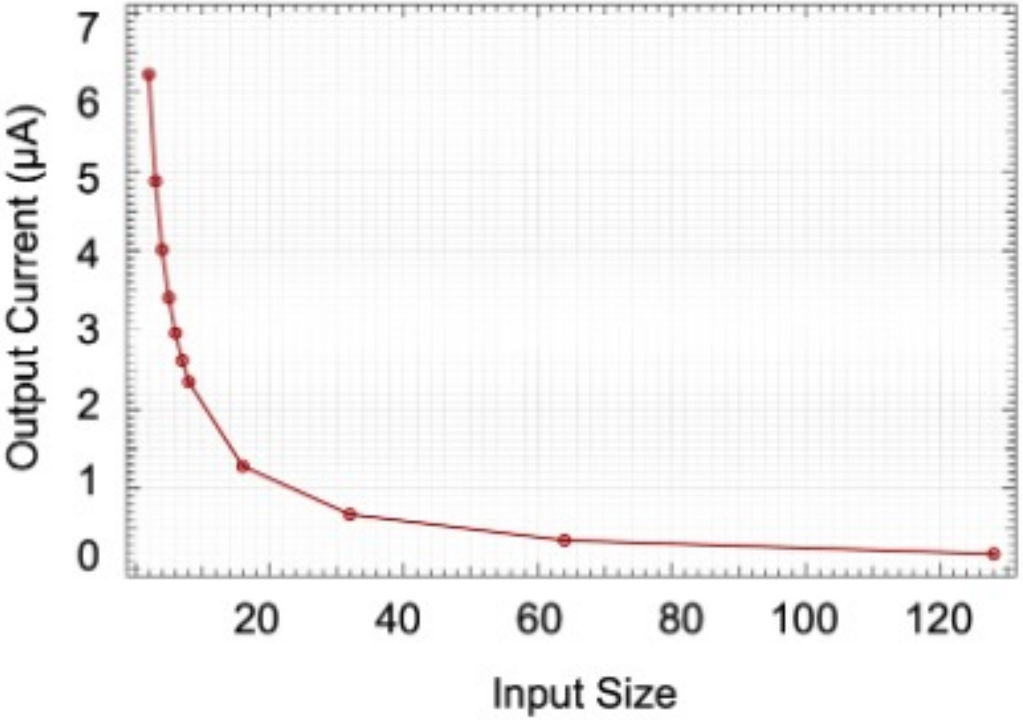
- Current attenuation is determined by the crossbar size C_s. Annealing function:

$$I_1(C_s) = A \cdot C_s^{-B},$$

- DNN value conversion

$$P_v\left(V_{in}\right) = 0.5 + 0.5\,\mathrm{erf}\left(\sqrt{\pi}\frac{(V_{in} - V_{th})}{\Delta V_{in}(C_s)}\right)$$

$$\Delta V_{in}(C_s) = \Delta I_{in}/I_1(C_s).$$



The relationship between output current representing value ‘1’ with crossbar synapse array size

Approach: Hardware-Algorithm Co-Optimization

Randomness-aware Binary Neural Network Training

- Non-deterministic activation mapping with AQFP neuron switching probability distribution

$$w_b = \operatorname{sign}(w_r) = \begin{cases} +1, & \text{if } w_r \geq 0, \\ -1, & \text{otherwise} \end{cases},$$

$$a_b = \operatorname{sign}(a_r) = \begin{cases} +1, & \text{with probability } P_v(a_r), \\ -1, & \text{with probability } 1 - P_v(a_r), \end{cases}$$

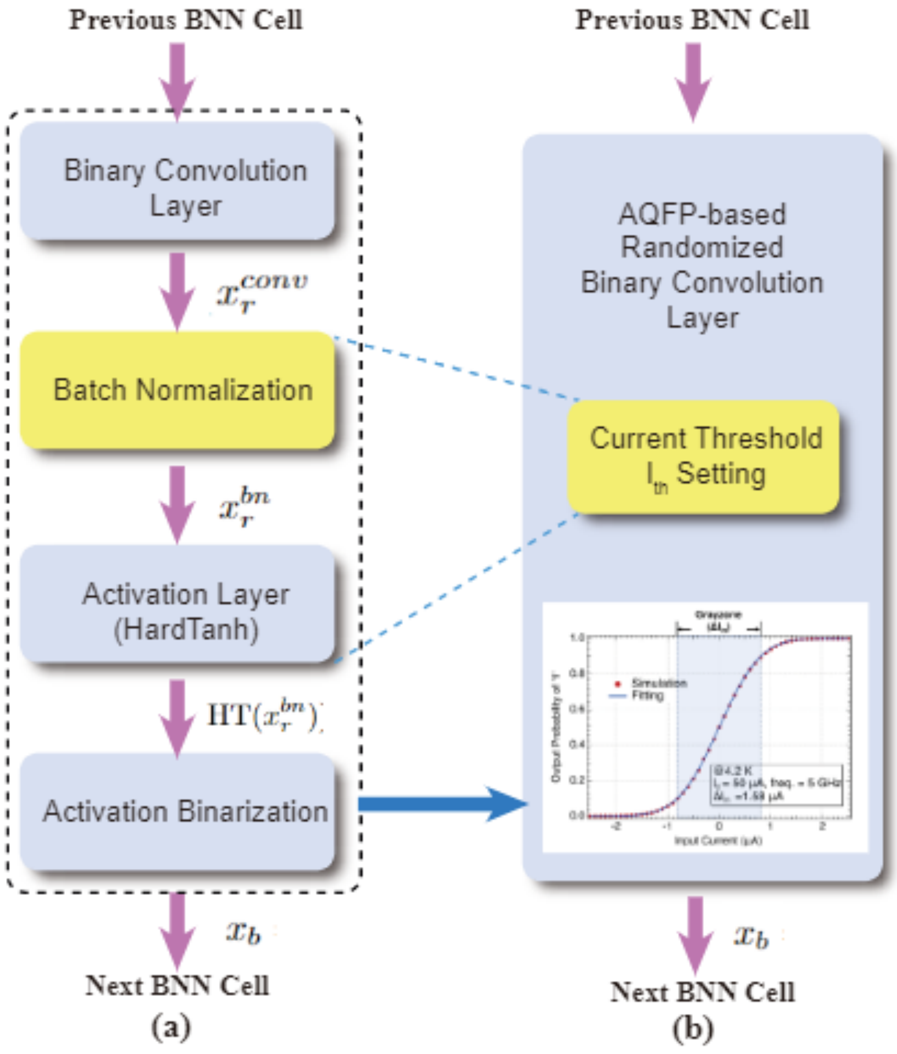
- Back propagation mapping

$$\begin{aligned} \frac{\partial \mathbb{E}(a_b)}{\partial a_r} &= \frac{\partial \operatorname{erf}\left(\sqrt{\pi}\frac{(a_r - V_{th})}{\Delta V_{in}(C_s)}\right)}{\partial a_r} \\ &= \frac{\partial \sqrt{\pi}\frac{(a_r - V_{th})}{\Delta V_{in}(C_s)}}{\partial a_r} \cdot \frac{2}{\sqrt{\pi}} e^{-\left(\sqrt{\pi}\frac{(a_r - V_{th})}{\Delta V_{in}(C_s)}\right)^2} \end{aligned}$$

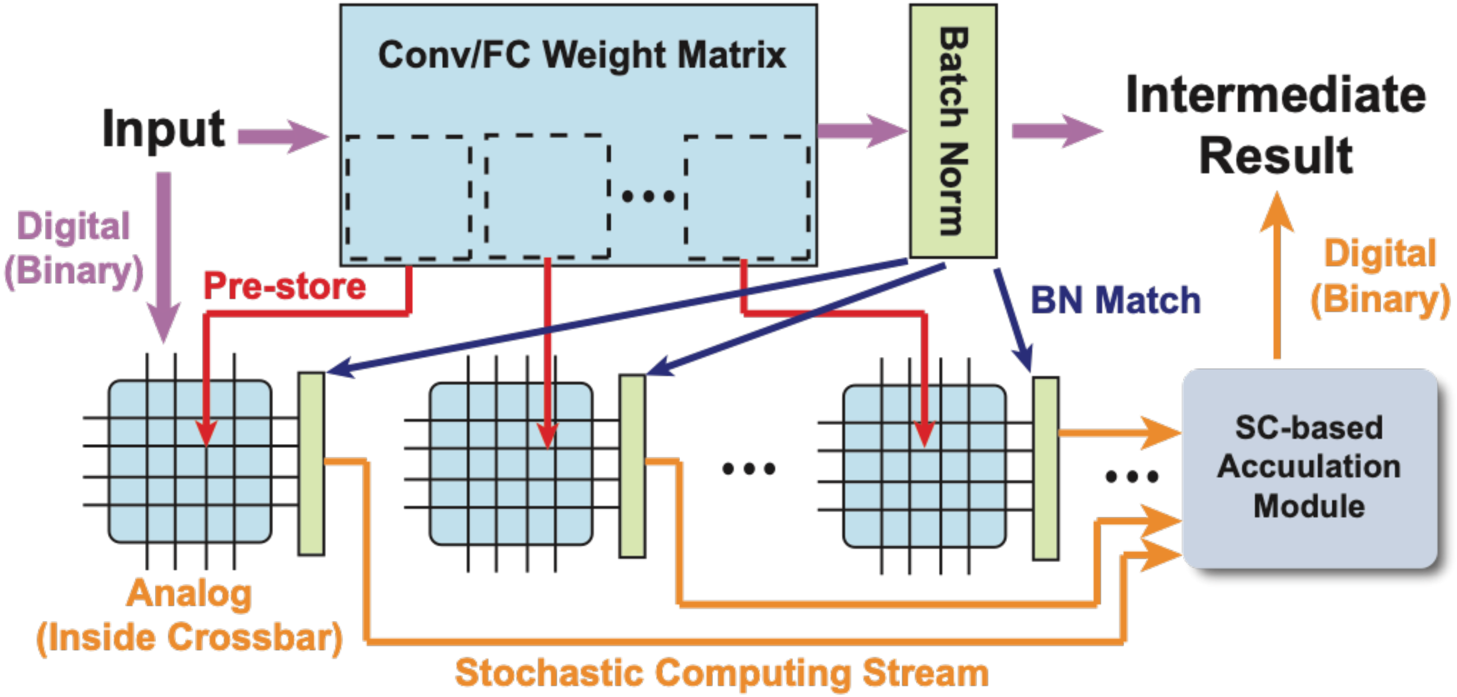
Batch Normalization Matching (Hardware Mapping)

- Batch norm mapping with analog threshold current input in AQFP

$$I_{th} = \left(-\frac{\beta\sqrt{\sigma^2 + \epsilon}}{\gamma \cdot \alpha} + \frac{\mu}{\alpha}\right) \cdot I_1(C_s).$$



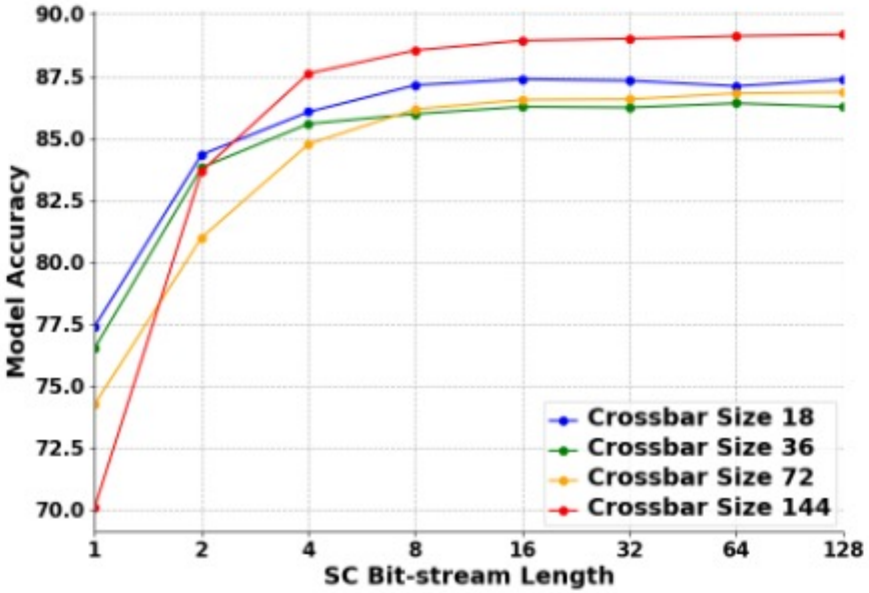
Hardware Design of AQFP-Based Randomized BNN Accelerator



- Stochastic Computing-based Accumulation Module Design
  - Crossbar may not be large enough for the whole filter’s computation in DNN. We use stochastic computing (SC) to accumulate the results from multiple crossbars.
  - Using AQFP neuron directly as the SC generator.
  - APC is used in the SC addition.

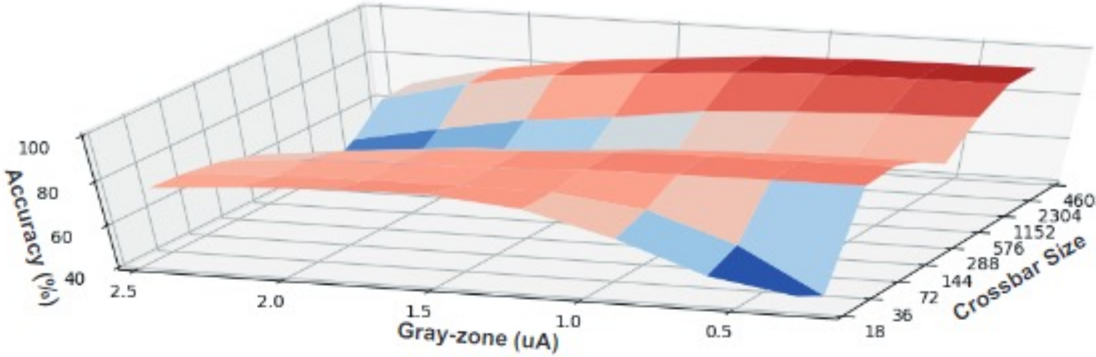
Hardware Configuration Optimization

1. Stochastic Computing Bit-stream Length Optimization:



- Relationship between SC bit-stream length with model accuracy.
- VGG-small trained on CIFAR-10 with 4 different crossbar sizes are deployed.

2. Optimization for Width of Grayzone ΔI_{in} and Crossbar Size C_s:



- Accuracy distribution in two demensions of Grayzone width and crossbar size.
- The stochastic bit-stream length used here is 1.

Experimental Results

Model accuracy on Cifar-10 dataset under different energy efficiency constraints.

Design	Scheme	Accuracy	Energy Efficiency w/o cooling (TOPS/W)	Energy Efficiency w/ cooling (TOPS/W)	Power (mW)	Throughput (image/s)
DNN (VGG-Small) [1]	Full-precision	92.5	0.28	-	-	-
IMB [2]	Binary	87.7	82.6	-	12.5	1.3
STT-BNN [3]	Binary	80.1	311	-	-	-
CMOS-BNN [4]	Binary	92	617	-	-	-
Ours (VGG-Small)	Binary	91.7	1.9×10^{5}	4.8×10^{2}	6.2×10^{-3}	2
Ours (VGG-Small)	Binary	90.6	3.8×10^{5}	9.5×10^{2}	6.3×10^{-3}	3.9
Ours (VGG-Small)	Binary	89.2	1.5×10^{6}	3.8×10^{3}	6.4×10^{-3}	15.2
Ours (VGG-Small)	Binary	87.4	6.8×10^{6}	1.7×10^{4}	7.6×10^{-3}	47.4
Ours (ResNet-18)	Binary	92.2	1.9×10^{5}	4.8×10^{2}	6.2×10^{-3}	2.2

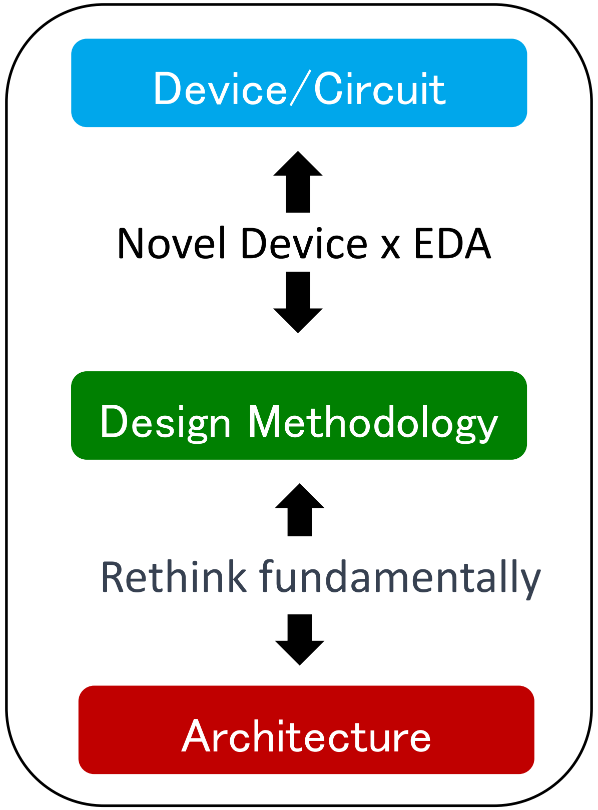
Design	Accuracy	Energy Efficiency (TOPS/W)	
		w/o cooling	w/ cooling
SyncBNN [5]	98.4	36.6	36.6
RSFQ [5]	97.9	2.4×10^{3}	8.1
ERSFQ [5]	97.9	1.5×10^{4}	50
SC-AQFP [6]	96.9	9.8×10^{3}	24.5
Ours	98.1	1.5×10^{6}	3.8×10^{3}

Comparison with RSFQ-JBNN, ERSFQ-JBNN, CMOS-based SyncBNN, SC-AQFP, and our implementation (MLP) on MNIST Dataset.

[1] Y. Chen, MICRO 2014.
[2] H. Kim, ASPDAC 2019.
[3] T. N. Pham, IEEE ETCS, 2022.
[4] P. C. Knag, IEEE JSCC, 2020.
[5] R. Fu, IEEE TCAD, 2022.
[6] R. Cai, ISCA 2019.

Will be presented at ***MICRO 2023***, October, Toronto, Canada

Summary



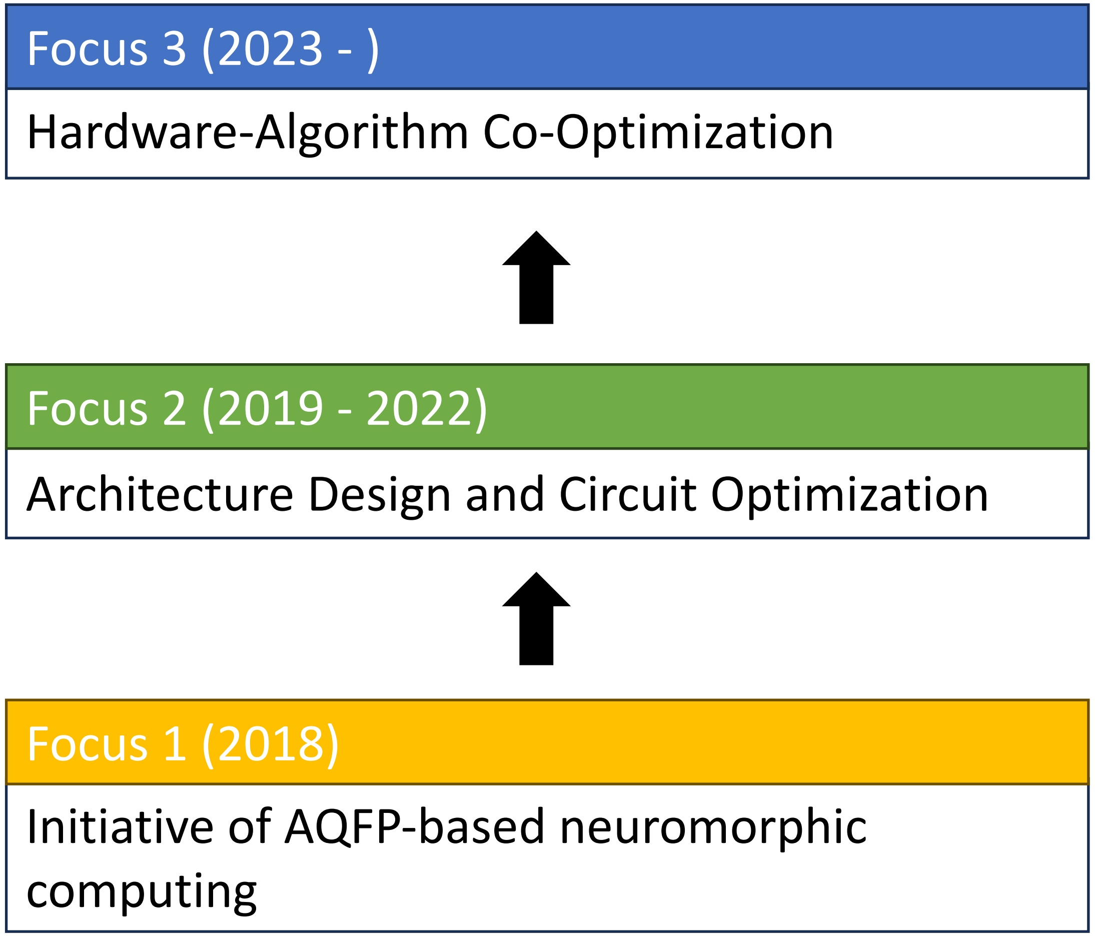
Device/Circuit

Novel Device x EDA

Design Methodology

Rethink fundamentally

Architecture



Focus 3 (2023 -)

Hardware-Algorithm Co-Optimization

Focus 2 (2019 - 2022)

Architecture Design and Circuit Optimization

Focus 1 (2018)

Initiative of AQFP-based neuromorphic computing

Thanks to Dr. Chris Ayala, Dr. Naoki Takeuchi, Dr. Tsung-Yi Ho, Mr. Wenhui Luo, Coldflux Team Members, Sponsors

創発的研究支援事業
Fusion Oriented REsearch for disruptive Science and Technology

PRESTO
SAKIGAKE

科研費
KAKENHI