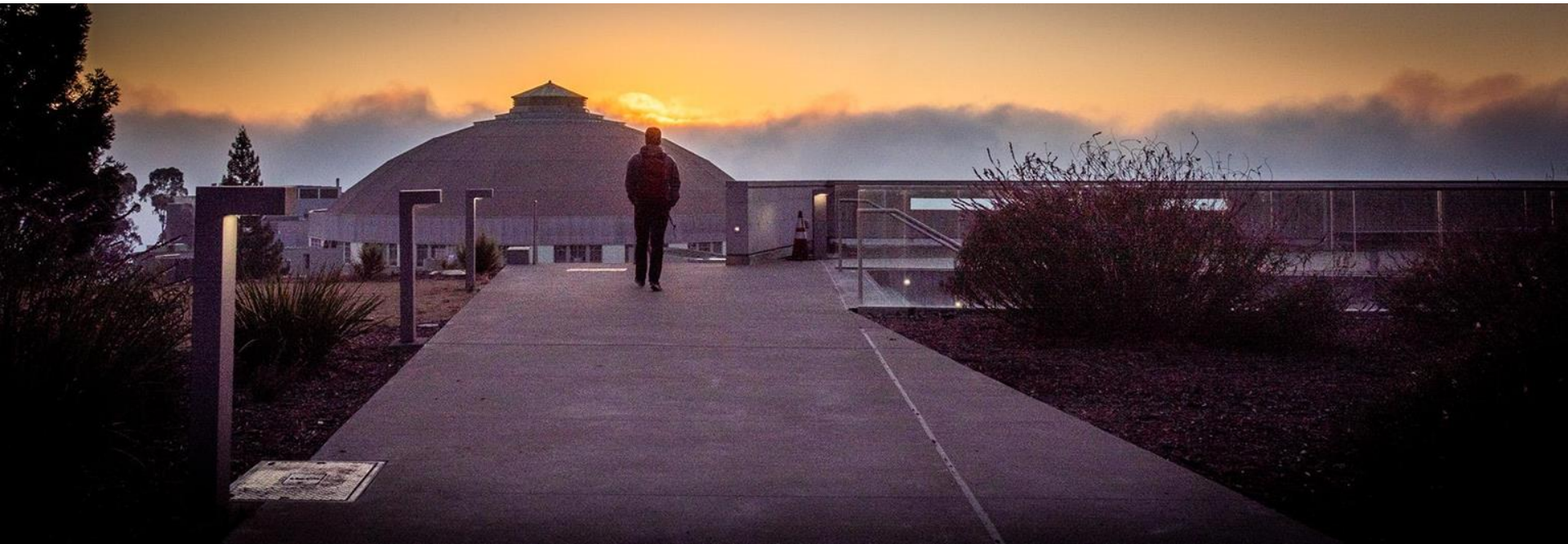
BERKELEY LAB
Bringing Science Solutions to the World

U.S. DEPARTMENT OF ENERGY
Office of Science

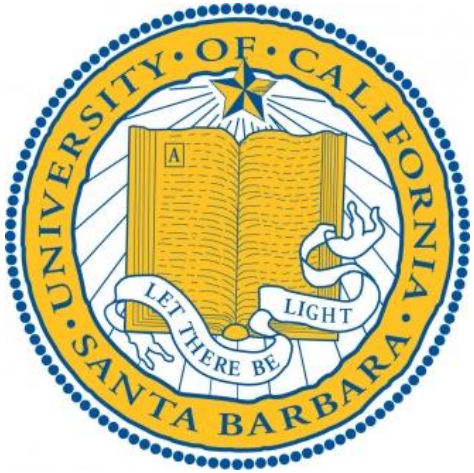
Digital Computation in Cryo-Cooled Environments

Presenter: George Michelogiannakis
Applied Math and Computational Research Division, LBNL
mihelog@lbl.gov



The Team

LBNL and UCSB team. Funded by ARO since late 2019.





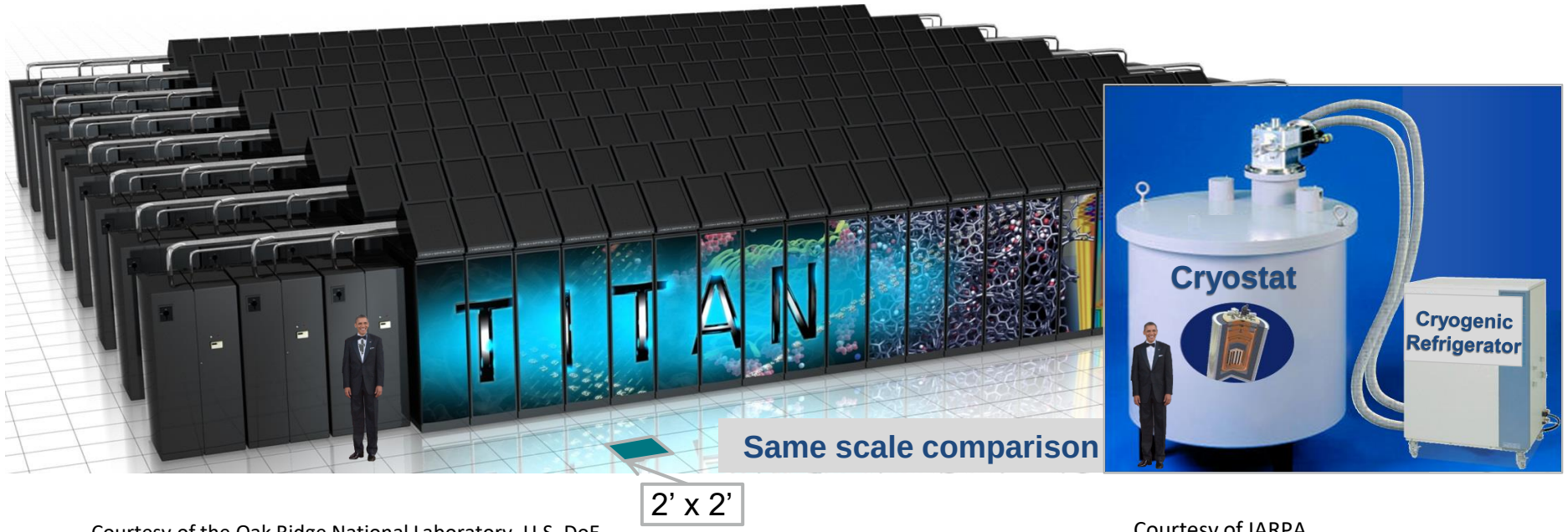







Background & Motivation

S Holmes, “Superconductor Computing: A Possible Future 2”, SGAR 2022



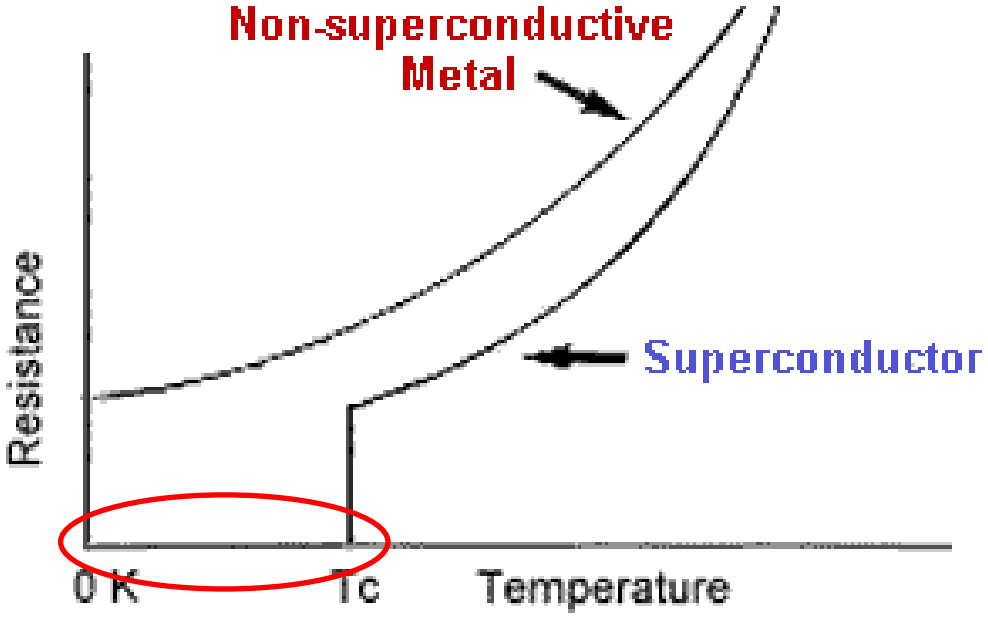
Courtesy of the Oak Ridge National Laboratory, U.S. DoE

Courtesy of IARPA

	Titan at ORNL	Superconductor Supercomputer	
Performance	**17.6** PFLOP/s (#2 in world*)	**20** PFLOP/s	~1x
Memory	**710** TB (0.04 B/FLOPS)	**5** PB (0.25 B/FLOPS)	7x
Power	**8,200** kW avg. (not included: cooling, storage memory)	**80** kW total power (includes cooling)	0.01x
Space	**4,350** ft^2 (404 m^2, not including cooling)	~**200** ft^2 (19 m^2, includes cooling)	0.05x
Cooling	Additional power, space and infrastructure required	All cooling shown	

Superconductivity

Resistance of mercury and other materials drops to near zero below ~4K

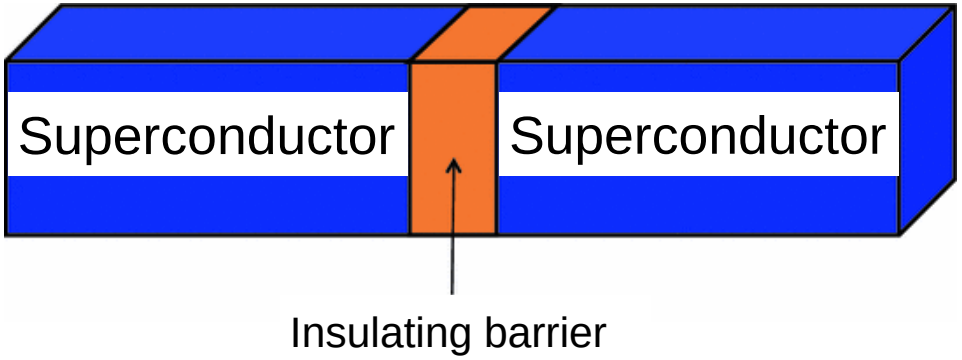


Gallardo et al, “Superconductivity observation in a (CuInTe 2) 1-x (NbTe) x alloy with x=0.5”

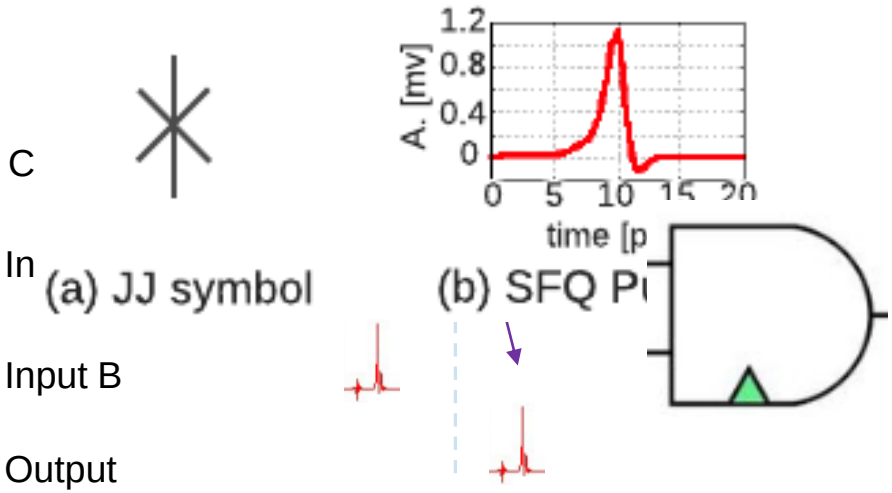
RSFQ Basics

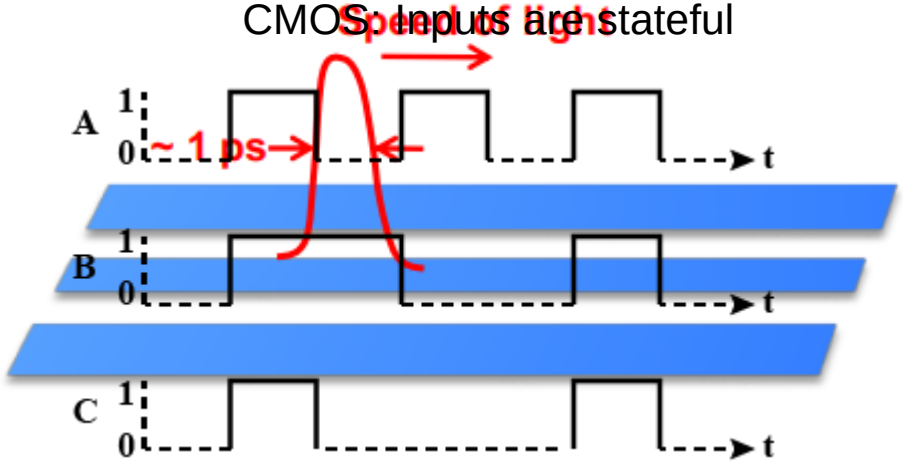
mV, pS pulses. Presence of a pulse encodes “1”. Absence encodes “0”.

- Josephson Junction. Current passes through a barrier with no resistance up to a critical current. Then resistive
- Superconducting to resistive state produces a pulse to the output
- As a result, classic binary gates have to be clocked



CMOS: Inputs are stateful

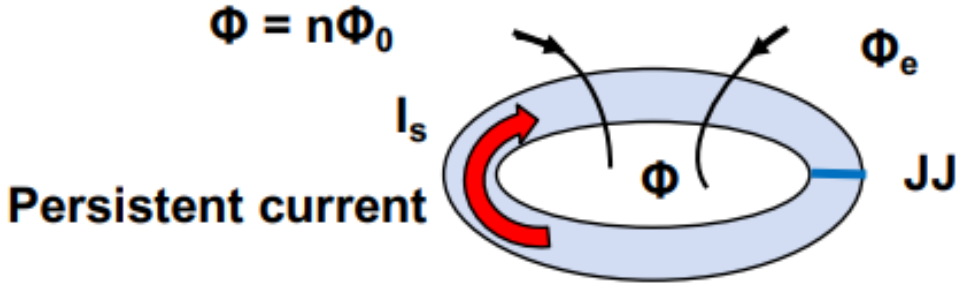




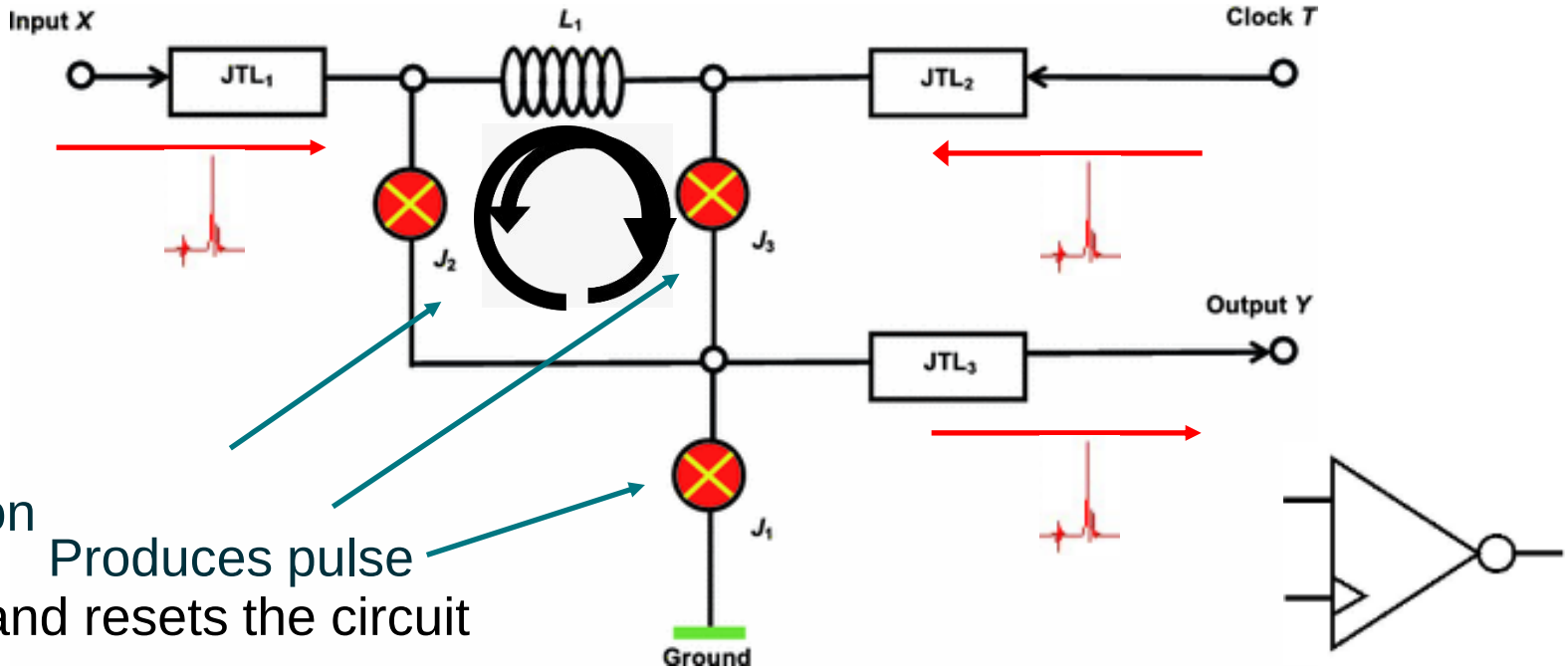
Johnson, “Superconducting Microelectronics for Next-Generation Computing”, 2018 MIT R&D conference
Khanna, “Rapid Single Quantum Flux (RSFQ) Logic”, Springer Integrated Nanophotonics 2016

JJs in a Loop to Compute

- JJ in a superconducting loop allows switching



Inverter:



Inverts direction

Produces pulse and resets the circuit

Johnson, "Superconducting Microelectronics for Next-Generation Computing", 2018 MIT R&D conference
Khanna, "Rapid Single Quantum Flux (RSFQ) Logic", Springer Integrated Nanophotonics 2016

CMOS-Inspired RSFQ Architectures

An example: 8x8 multiplier. Binary encoding. Every gate is clocked

- 15 pipeline stages. 8-bit signed inputs.
- Binary encoding means 8 wires per operand and 16 for product
- 17,488 number of JJs. Approx 1,700 gates. **All clocked** except for splitters and mergers

~14 mm^2

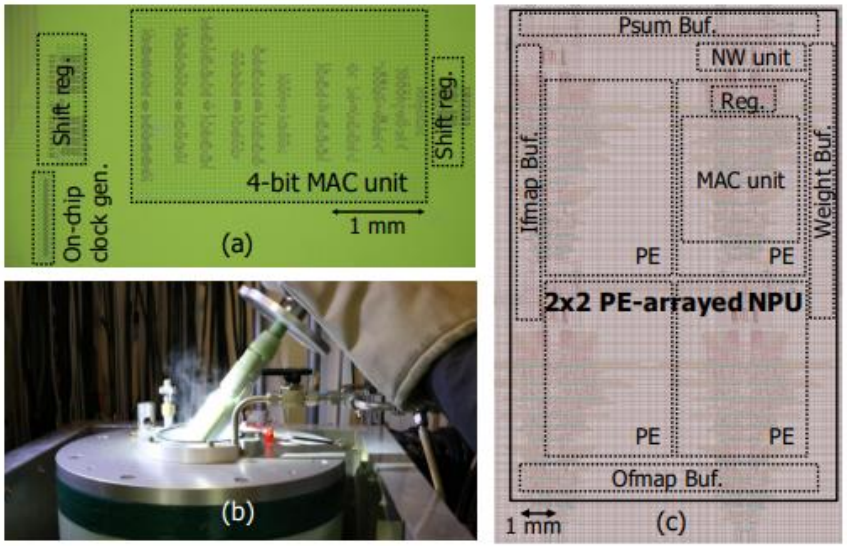
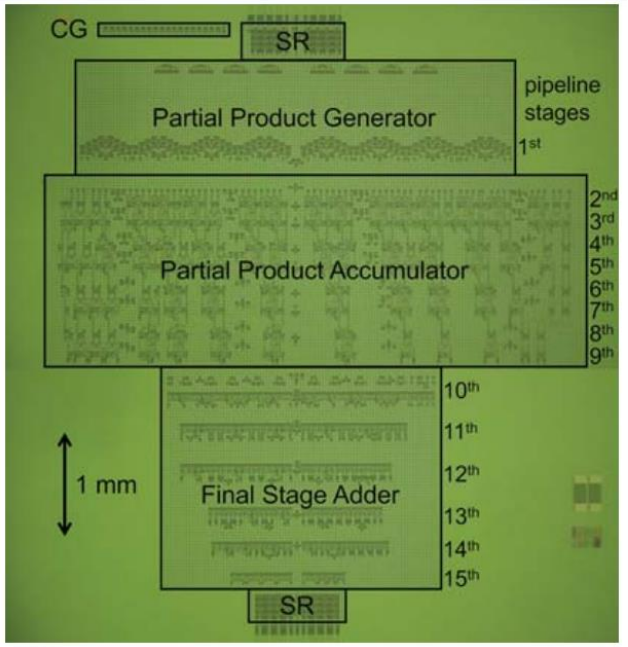


Fig. 12. Model validation setup (a) Chip microphotograph of 4-bit MAC unit (b) 4 K measurement setup (c) Layout of the 2 × 2 PE-arrayed NPU

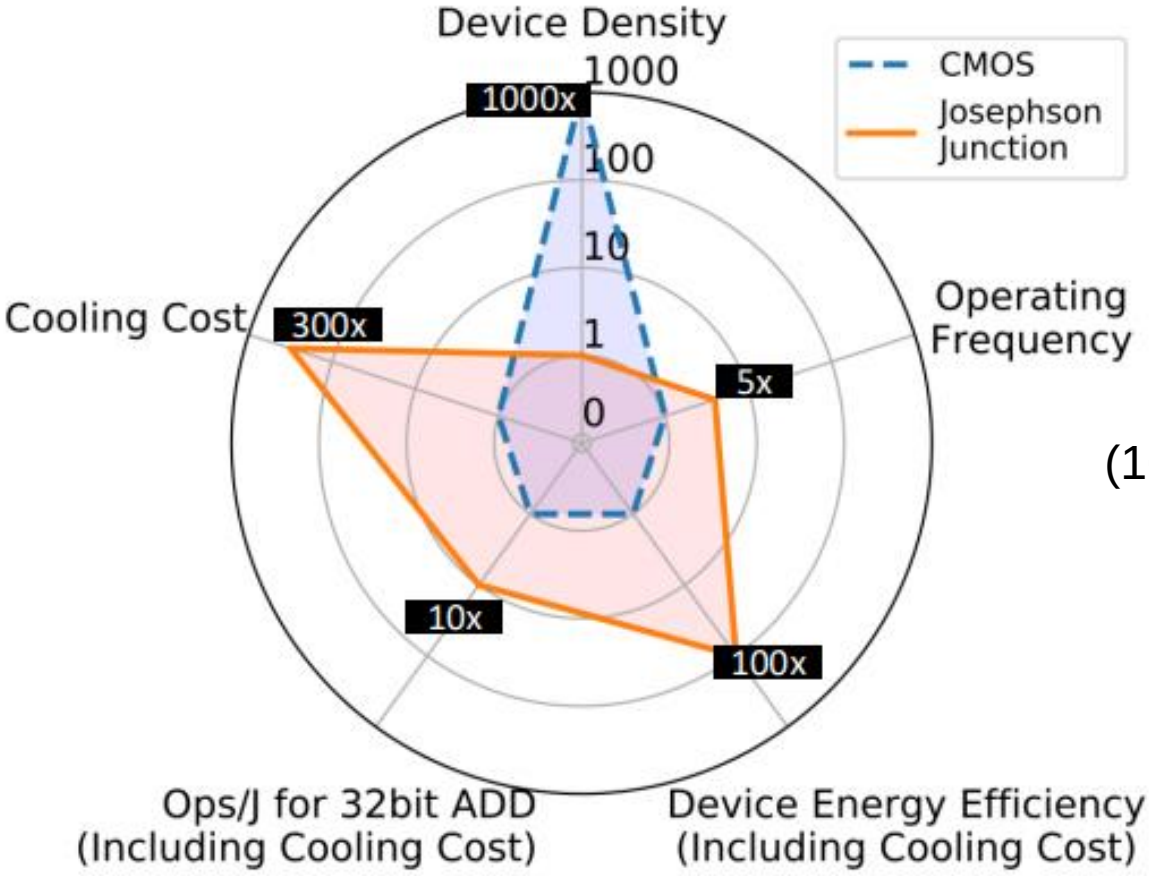


I Nagaoka et al., “A 48GHz 5.6mW Gate-Level-Pipelined Multiplier Using Single-Flux Quantum Logic”, ISSCC 2019

Ishida et al., “SuperNPU: An Extremely Fast Neural Processing Unit Using Superconducting Logic Devices”, MICRO 2020

Technology Comparison Versus CMOS

Device density is the major problem



(16nm CMOS)

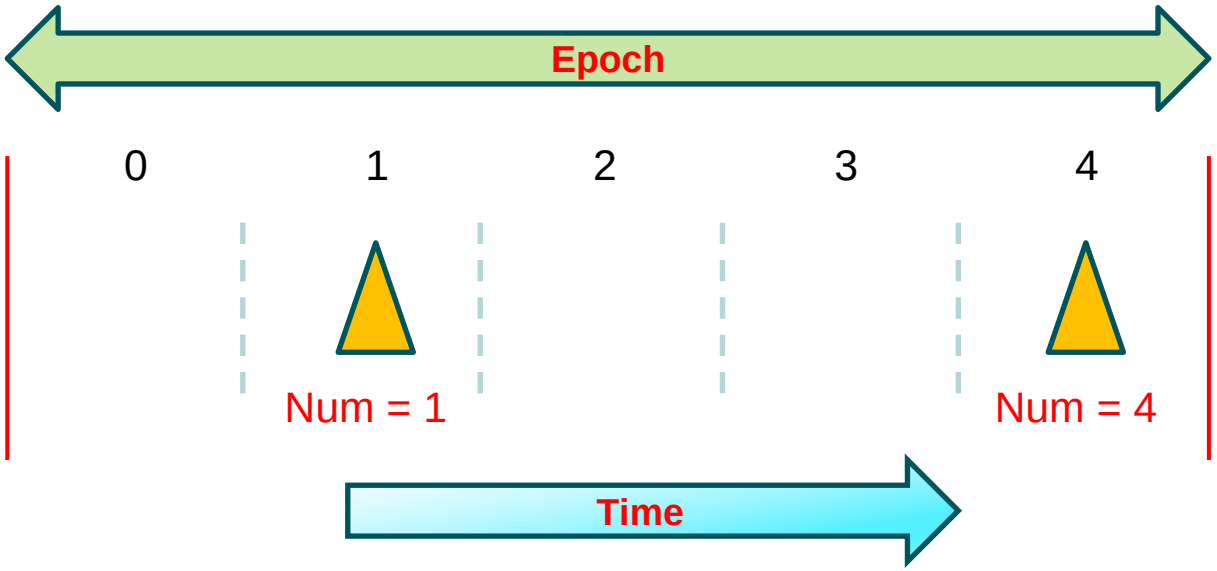
Tannu et al, “A case for superconducting accelerators”, CF 2019

Temporal Computing

Data Encoding in Race Logic

An epoch contains N time slots. A pulse in time slot “I” encodes the value “I”

- Epochs repeat
- Each pulse represents an equivalent 2^N binary number (N = NumTimeSlots)



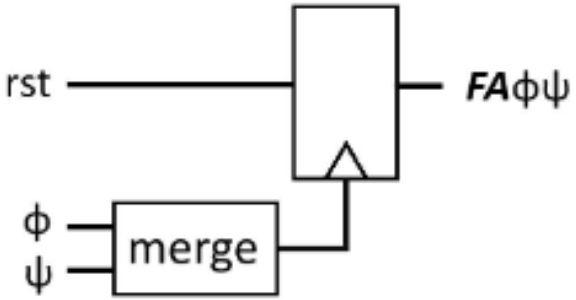
G Tzimpragos et al., “A computational temporal logic for superconducting accelerators”, ASPLOS 2020

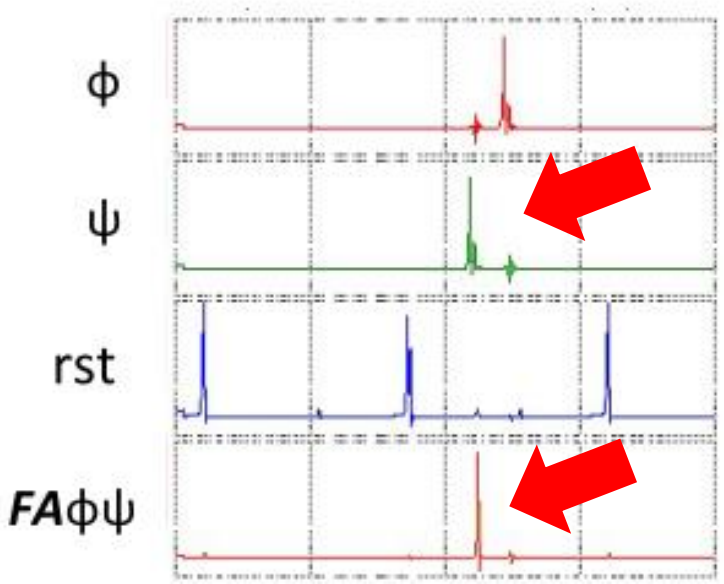
First Arrival – The MIN Function

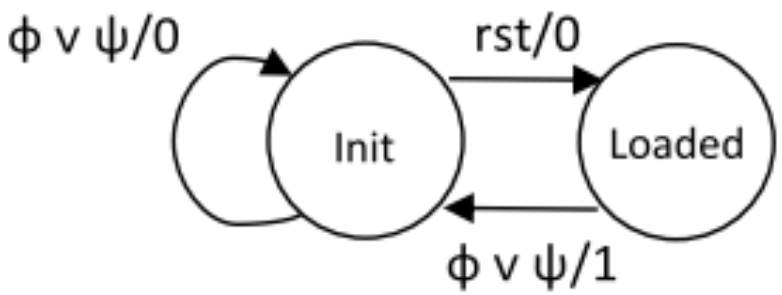
First incoming pulse causes an output pulse. Has a reset. Gate is stateful

MIN(φ,ψ)

DFF's clock input repurposed





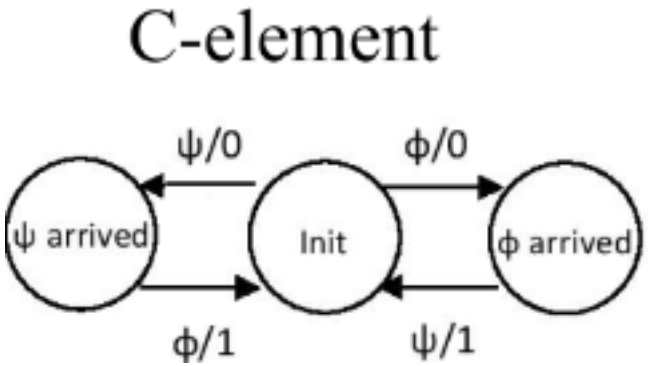


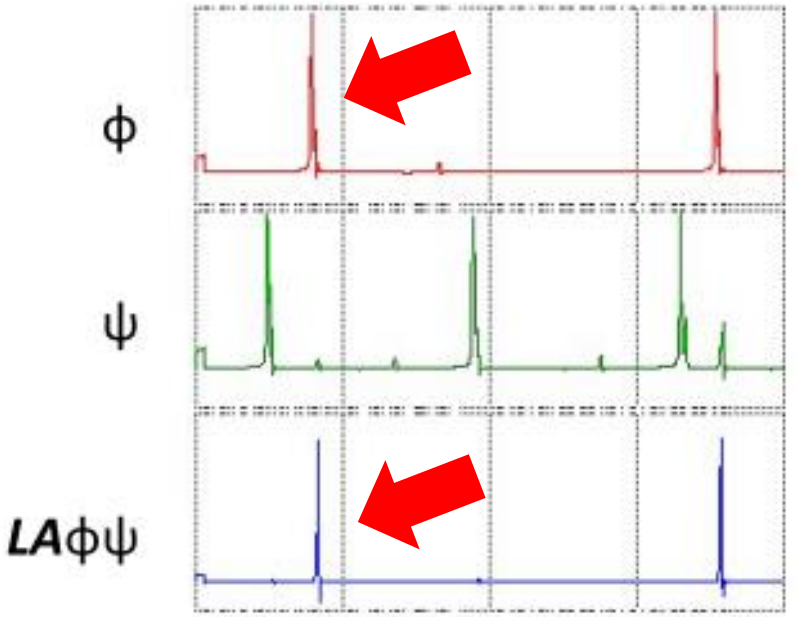
G Tzimpragos et al., "A computational temporal logic for superconducting accelerators", ASPLOS 2020

Last Arrival – The MAX Function

Last incoming pulse causes an output pulse. Also stateful

MAX(φ,ψ)

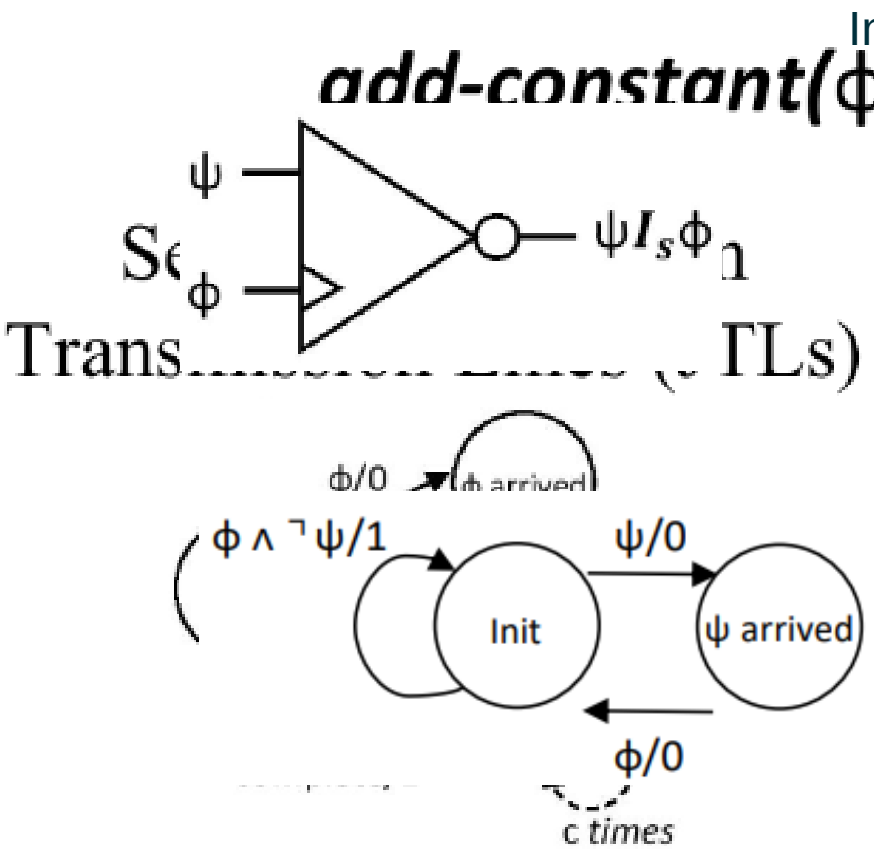


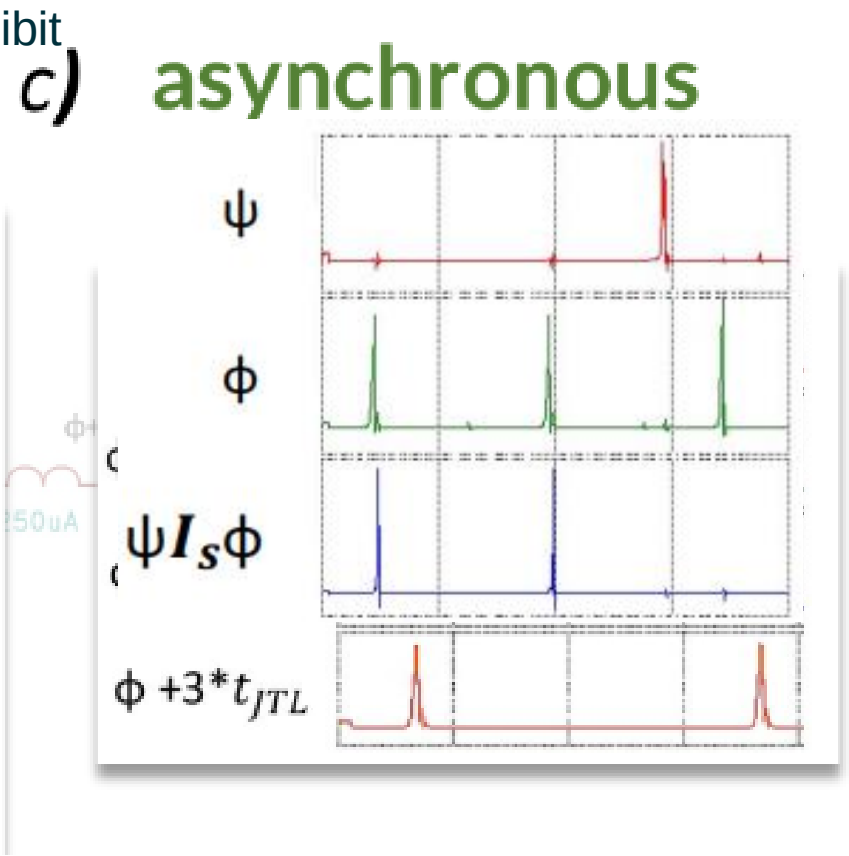


G Tzimpragos et al., “A computational temporal logic for superconducting accelerators”, ASPLOS 2020

Delay and Coincidence Gates

Constant delay is an addition. Coincidence gate is optional

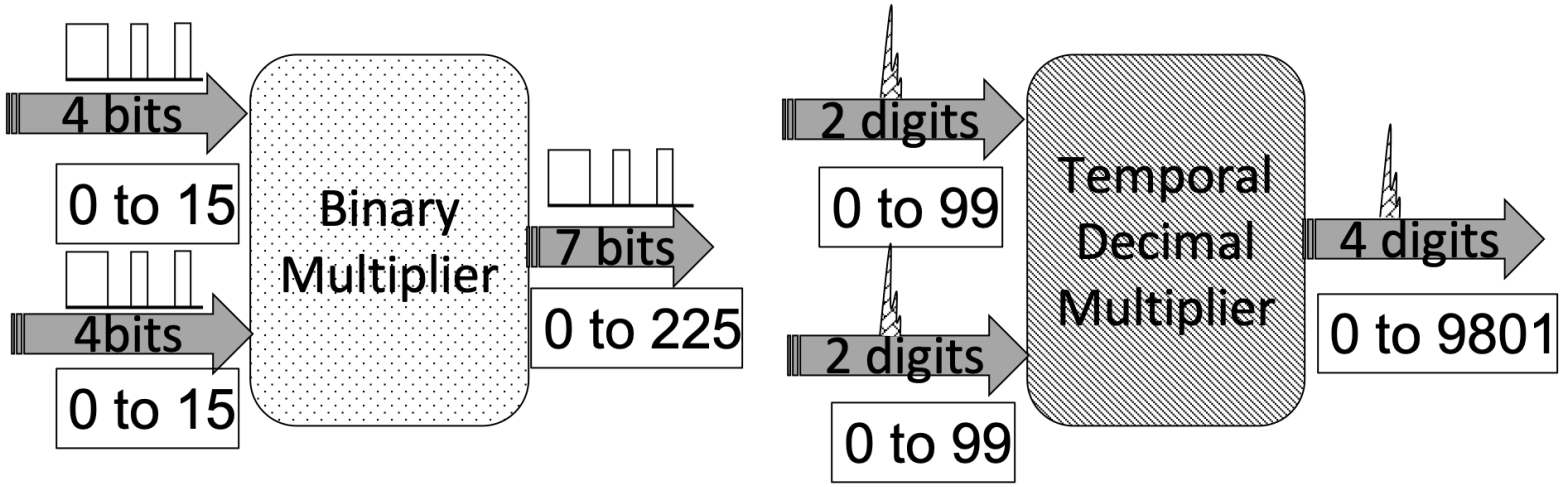




G Tzimpragos et al., “A computational temporal logic for superconducting accelerators”, ASPLOS 2020

Temporal Decimal Multiplier

Reduced datawidth: 3.2x for FFT and 2.67x for Deepbench. JJ reduction by approx. half



Input Width	2x Reduction
Output Width	1.75 x Reduction
Input Data Representation	6.6 x higher
Output Data Representation	43.5 x higher

D Vasudevan et al., “Efficient Temporal Arithmetic Logic Design for Superconducting RSFQ Logic”, IEEE Transactions on Applied Superconductivity 2023 (in press)

Hyperdimensional Computing

Online learning. Resilient to noise. Dimensionality of vectors in the thousands

$Vector Length (N)$	1000	2000	4000	8000	10000
Chip power (μW)	1.95	3.82	7.56	15.01	18.75
Cooling power (mW)	0.77	1.51	2.98	5.93	7.40
Combined (mW)	0.77	1.51	2.99	5.93	7.42

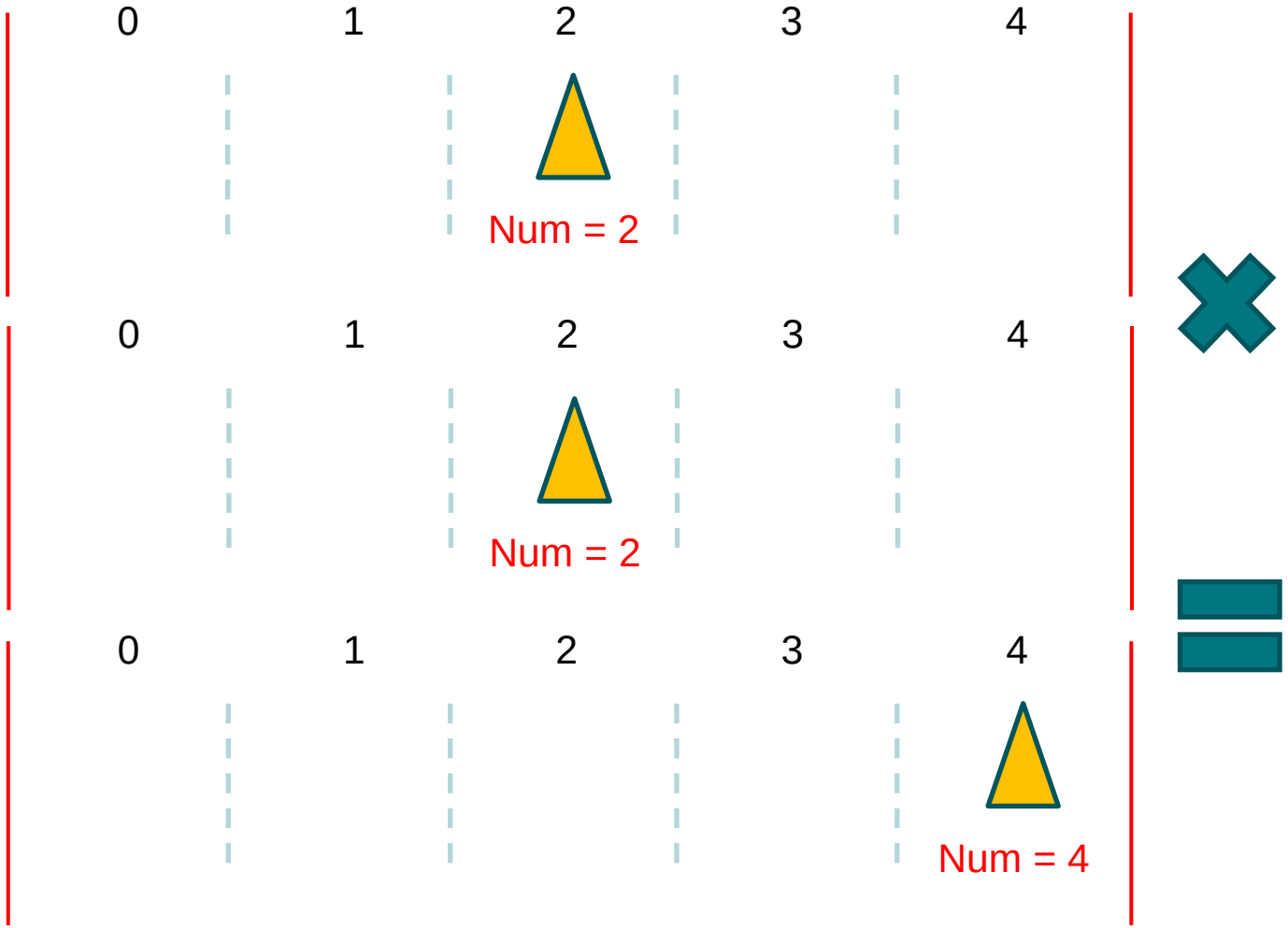
$Vector Length (N)$	1000	2000	4000	8000	10000
Thput (M enc ops/) (EM)	22.24	22.24	22.24	22.24	22.24
Thput (M searches/s) (ASM)	25.94	14.42	7.69	3.99	3.21
Thput (MCO/s) (overall)	22.24	14.42	7.69	3.99	3.21

K Huch et al., “Superconducting Hyperdimensional Associative Memory Circuit for Scalable Machine Learning”, IEEE Transactions on Applied Superconductivity 2023 (under review)

Hybrid Data Representation

Problem Statement: Reduce Cost of RL Arithmetic

For instance, multiplying two race logic pulses



Instead: Unipolar and Bipolar Race Logic

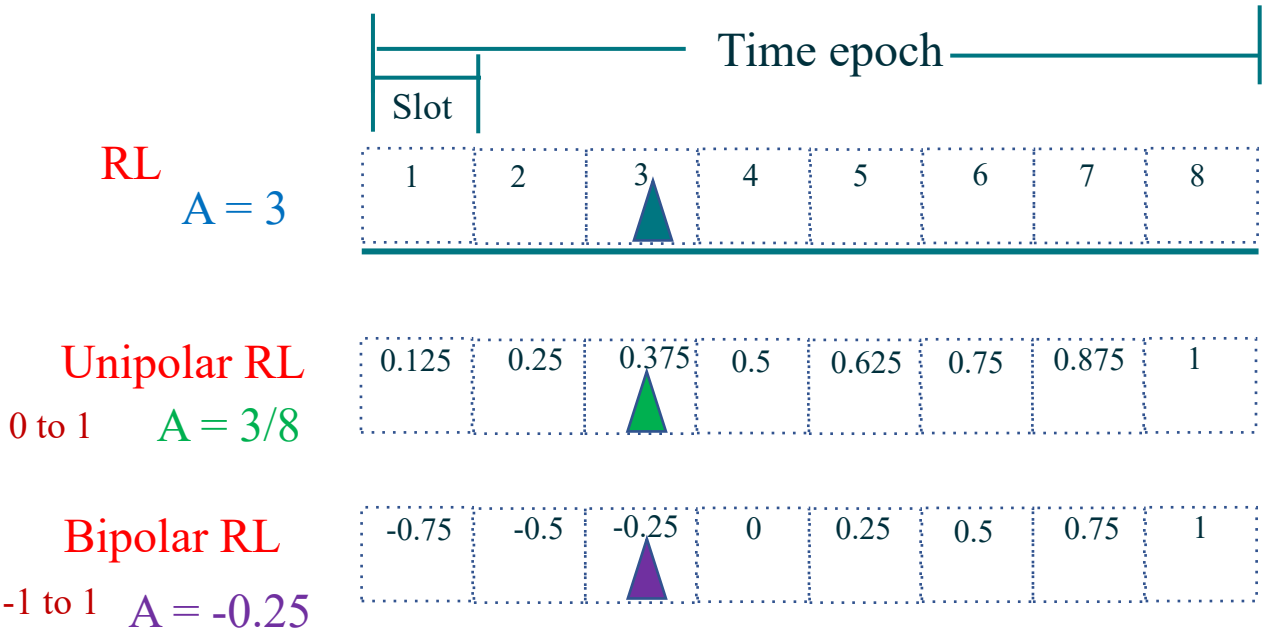
Changing the range of representation to [0,1] (unipolar)

or [-1,1] (bipolar)

To obtain bipolar representation

$$N_{max} = 8$$

$$A_b = 2A_u - 1$$



P Gonzalez-Guerrero et al., “Temporal and SFQ pulse-streams encoding for area-efficient superconducting accelerators”, ASPLOS 2022

Pulse Train Operands

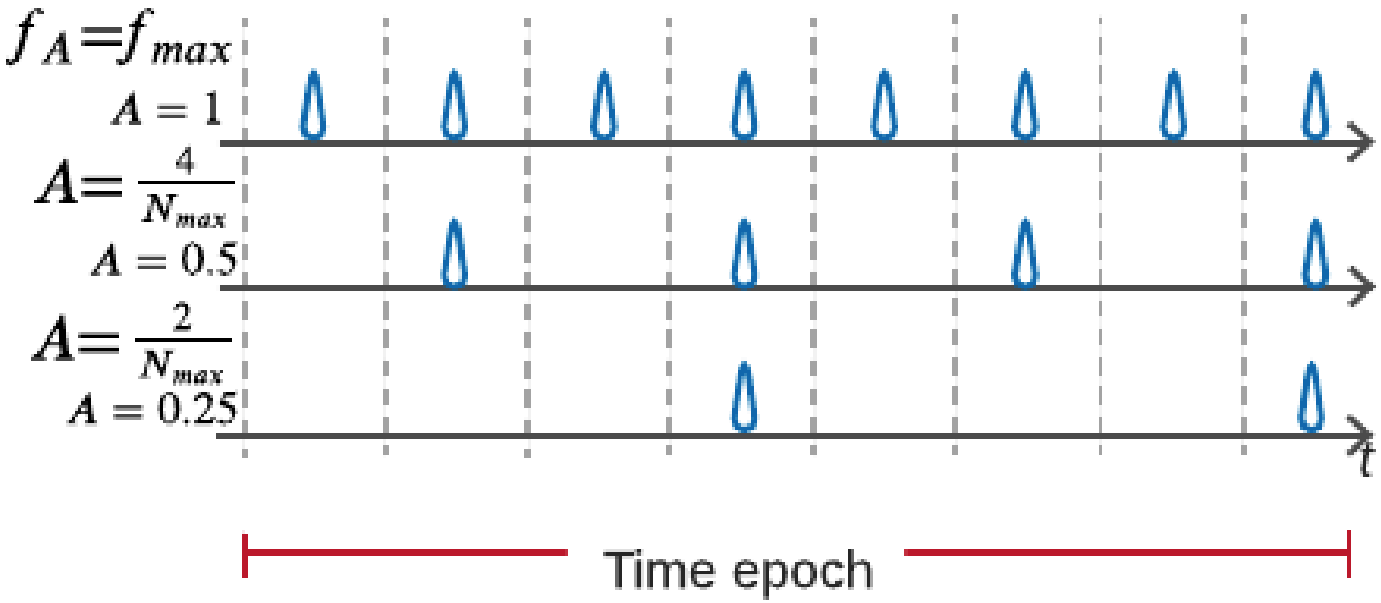
Maps a value to the number of pulses. “1” is for the maximum number of pulses

f_{max} $N_{max} = 8$

$A = n/N_{max}$

To obtain bipolar representation (not shown)

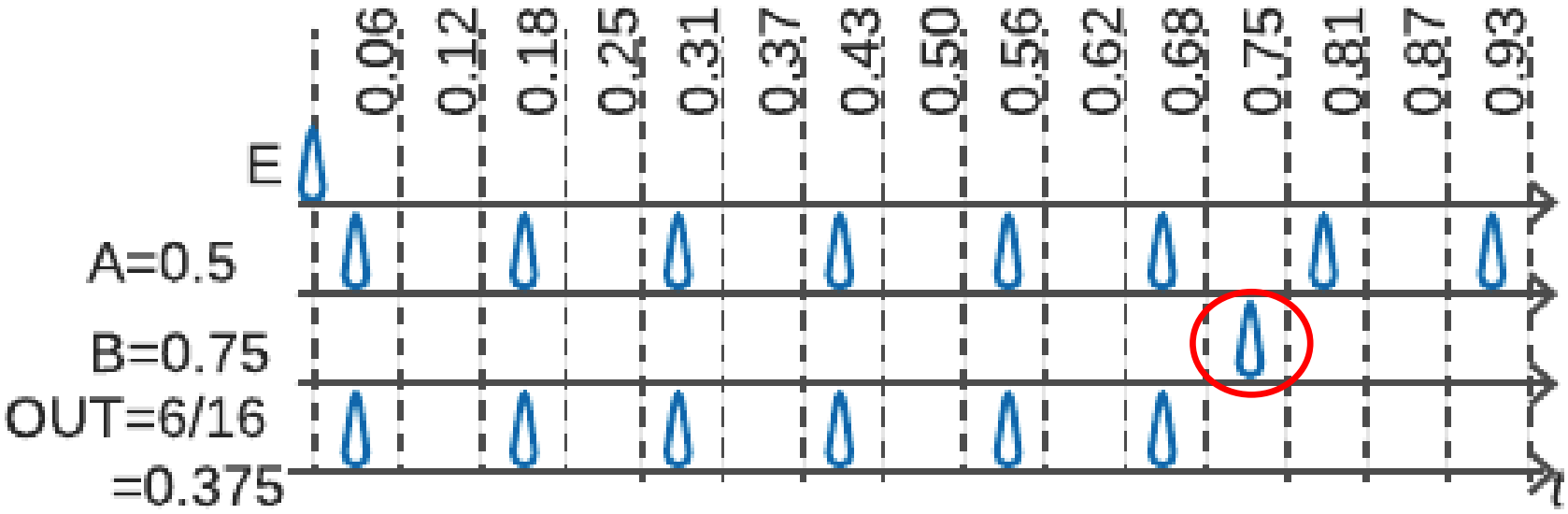
$A_b = 2A_u - 1$



P Gonzalez-Guerrero et al., “Temporal and SFQ pulse-streams encoding for area-efficient superconducting accelerators”, ASPLOS 2022

U-SFQ: Race Logic and Pulse Stream Operands

This shows a multiplication. The output is a pulse train



Arrival of RL pulse B suppresses future pulses at output

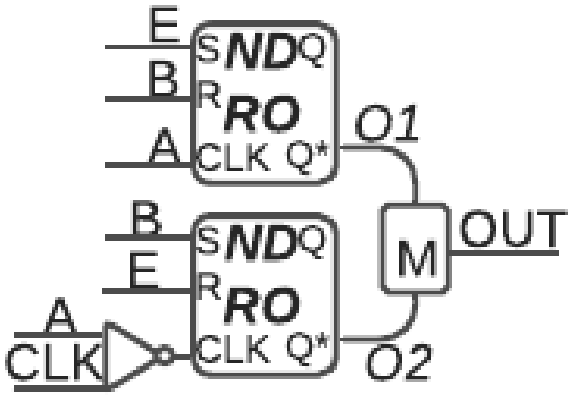
P Gonzalez-Guerrero et al., “Temporal and SFQ pulse-streams encoding for area-efficient superconducting accelerators”, ASPLOS 2022

Multiplication With Just One or Two Cells

Essentially a CMOS XNOR
The bipolar multiplier for stochastic computing

Before “B”, pulses in “A” pass
After “B”, the complement of “A” pass
The output is their merge

Unipolar SFQ multiplier



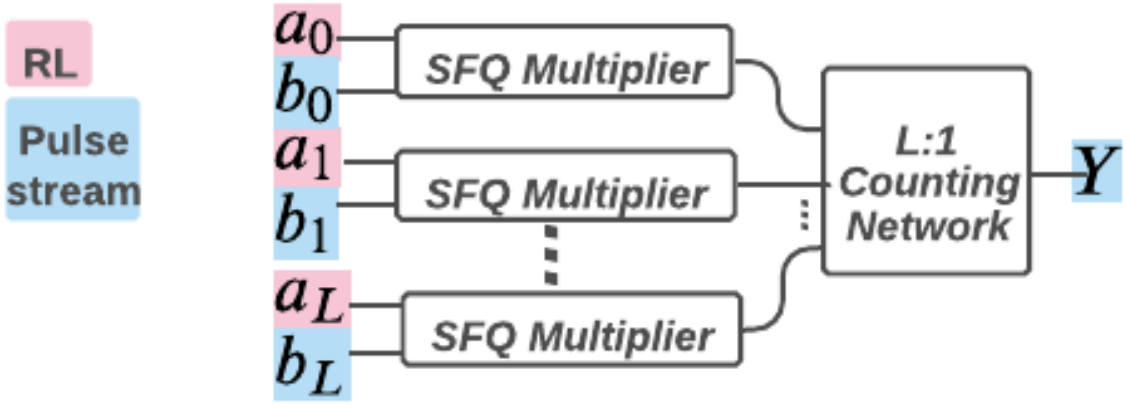
Bipolar SFQ multiplier

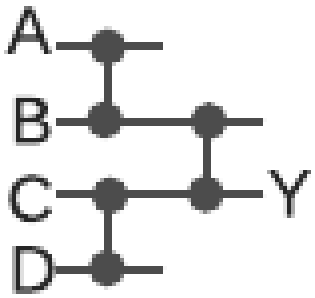
“Clk” denotes time slot boundaries

P Gonzalez-Guerrero et al., “Temporal and SFQ pulse-streams encoding for area-efficient superconducting accelerators”, ASPLOS 2022

Multiply-Accumulate Unit

Final result is a pulse stream





11x – 200x less area

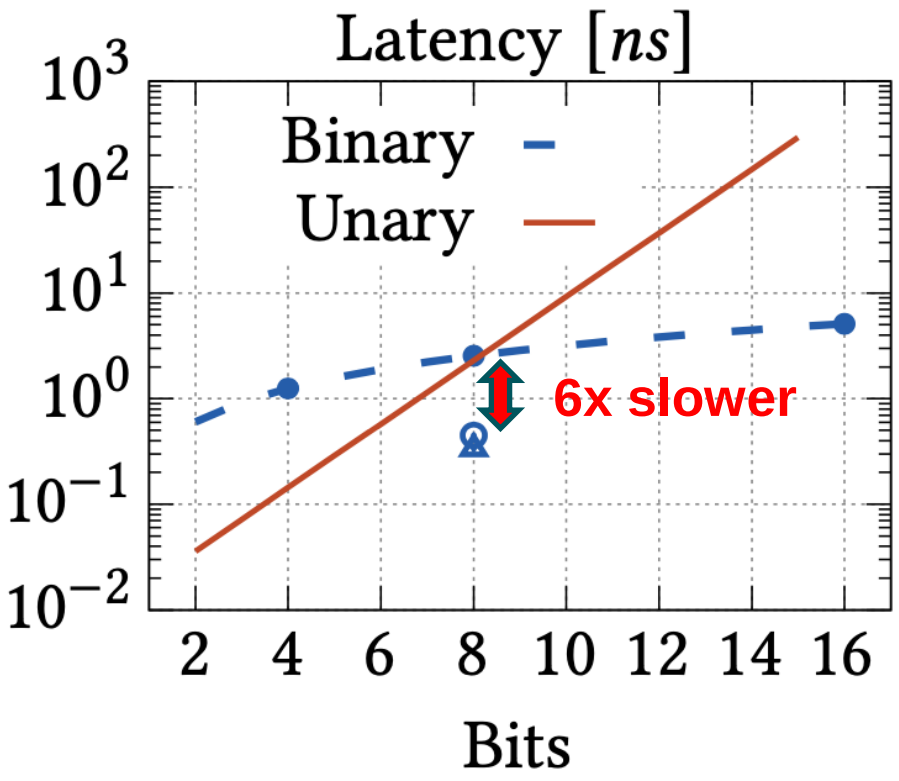
$$y = a \circ b = \sum_{i=0}^{L-1} a[i]b[i]$$

P Gonzalez-Guerrero et al., “Temporal and SFQ pulse-streams encoding for area-efficient superconducting accelerators”, ASPLOS 2022

U-SFQ Multiplier Exposes an Area-Latency Tradeoff

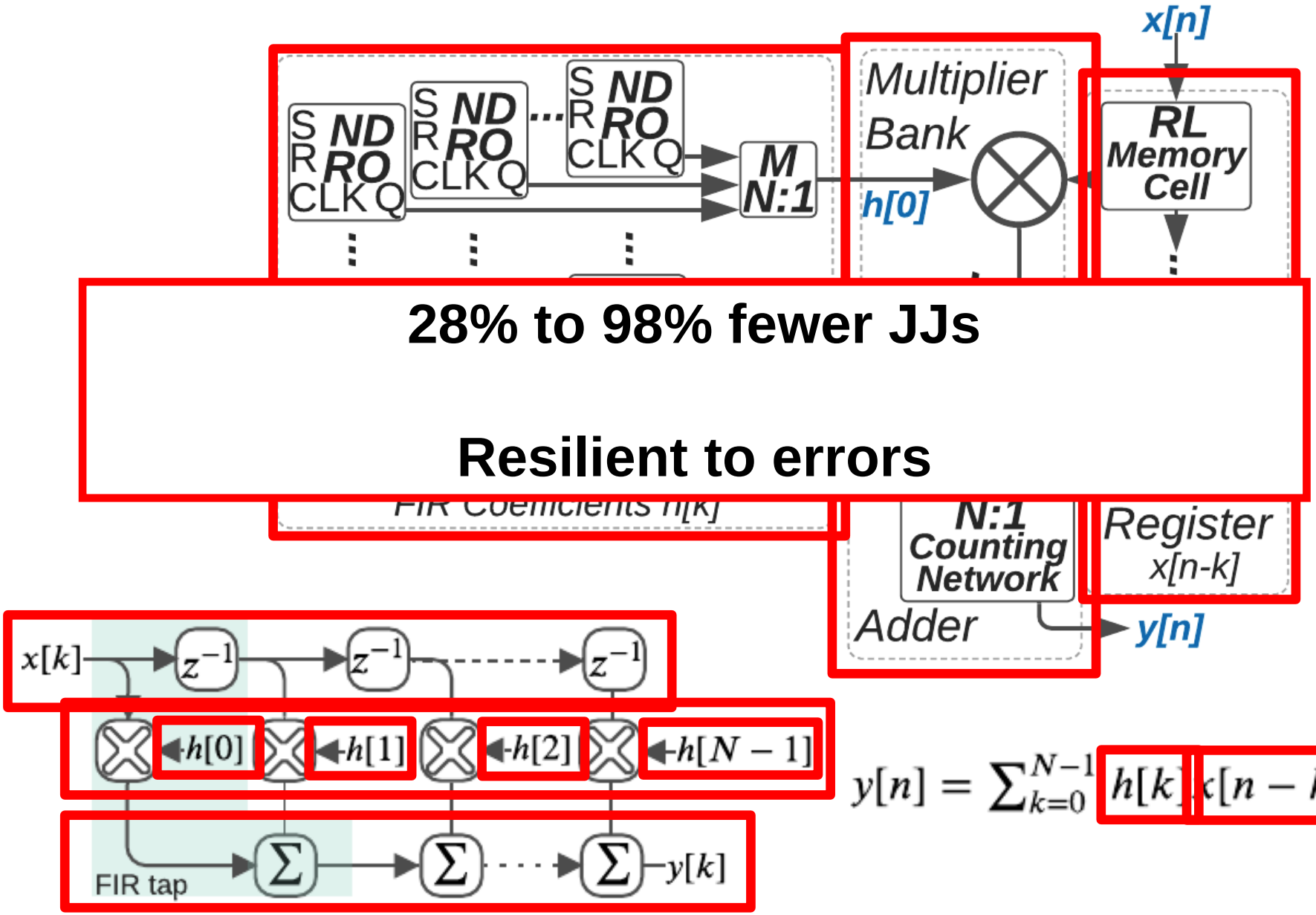
A fundamental tradeoff in race logic compute circuits.

U-SFQ provides higher performance over area



P Gonzalez-Guerrero et al., “Temporal and SFQ pulse-streams encoding for area-efficient superconducting accelerators”, ASPLOS 2022

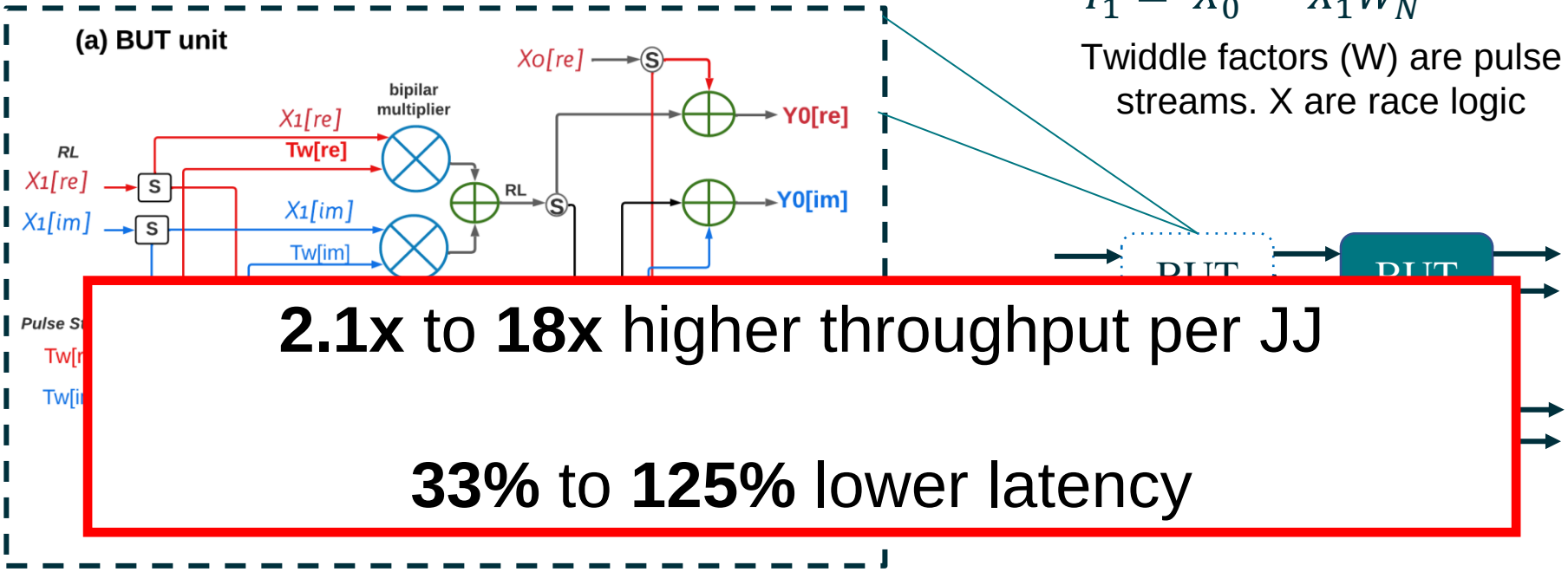
Finite Impulse Response (FIR) Filter



Fast Fourier Transform (FFT)

$$Y_0 = X_0 + X_1 W_N^2$$
$$Y_1 = X_0 - X_1 W_N^2$$

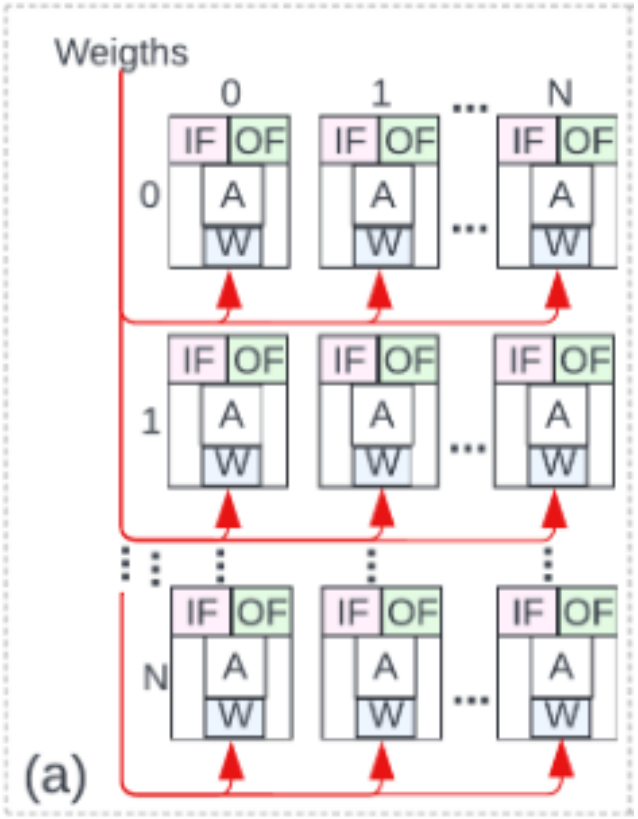
Twiddle factors (W) are pulse streams. X are race logic

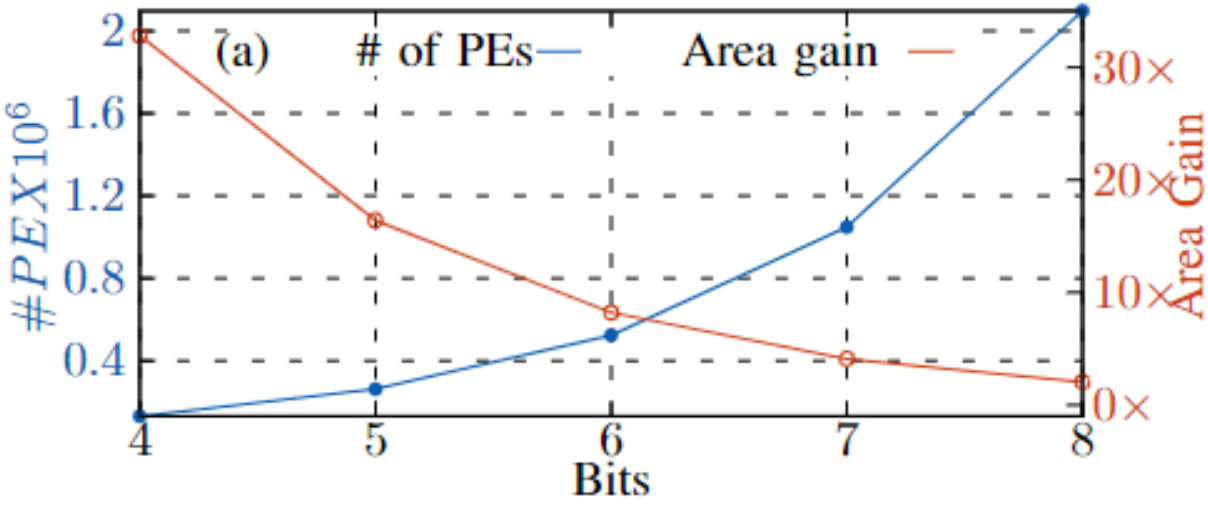


MG Bautista et al., “Superconducting Digital DIT Butterfly Unit for Fast Fourier Transform Using Race Logic”, NEWCAS 2022

Convolutional Neural Networks (CNNs)

2D mesh of processing elements with input feature output feature, and weight buffers

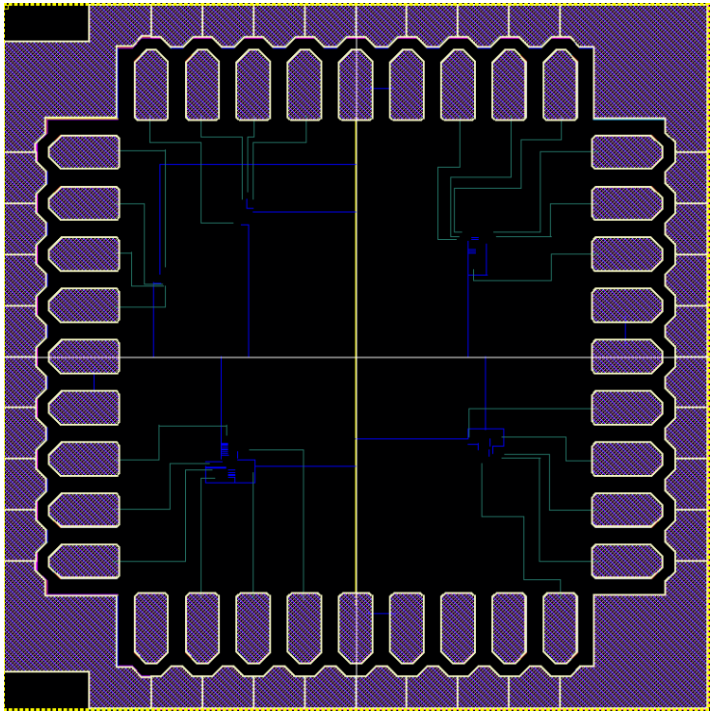


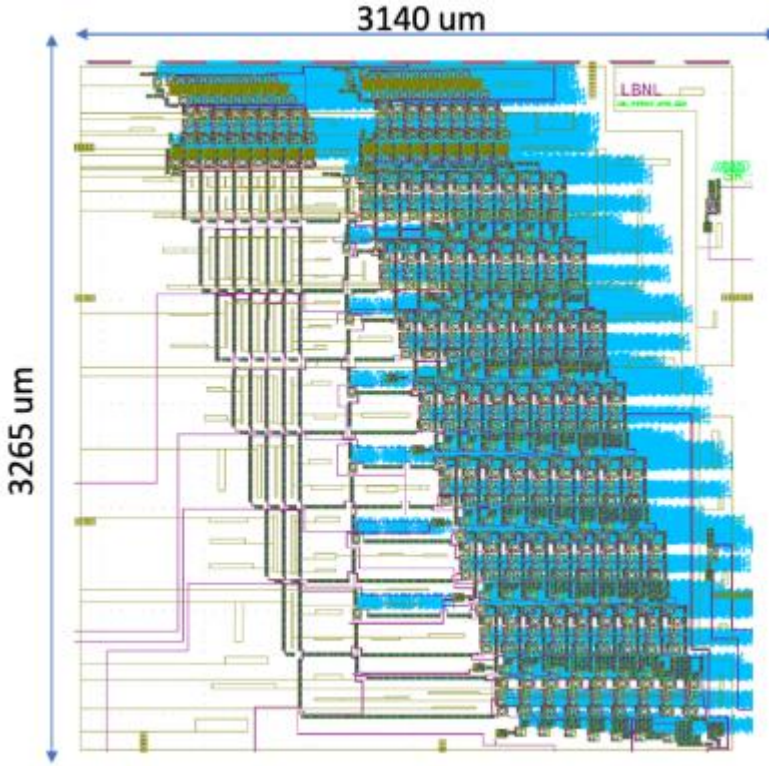


P Gonzalez et al., “An Area Efficient Superconducting Unary CNN Accelerator”, ISQED 2023

Chip Tapeouts With MIT Lincoln Lab

- Have manufactured seven 5x5 mm^2 test chips
- Predominantly scaled down versions of our various circuits
- Lack of mature EDA tools increases risk
- Chip on the right: five small circuits from FFT and FIR designs



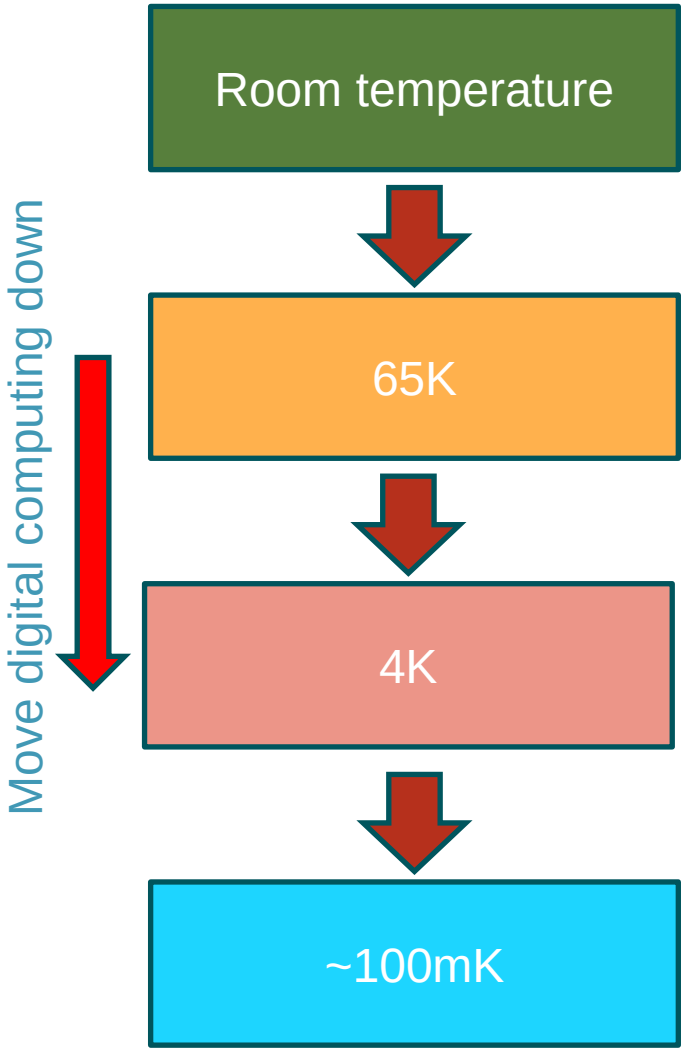


Conclusions and Thoughts

Opportunities For Sensors

Move compute closer to the cryogenic sensor

- Question: What kind of digital computing would you move closer to the instrumentation/sensor?
- Benefits:
  - More compute per unit power
  - Less data movement
  - Can replace other expensive components
- Each level offers different challenges/opportunities:
  - Room: Conventional CMOS. Expensive to move data to and from the cryo environment
  - 65K: Can use cryo-CMOS for denser memory
  - 4K: Majority of RSFQ circuits
  - 100mK: More noise in circuit and tighter power limits. But race logic/pulse trains are a good fit



Superconducting Digital Computing

- Significant value in co-designing with underlying technology
  - Abstraction re-use not necessarily productive
- Work remains to build the other related layers (e.g., EDA tools, design methodologies). Should engage experts from multiple disciplines
  - The ecosystem is not complete
- With improvements in device density and cooling, superconducting digital computing will become more attractive

List of Publications

- G Michelogiannakis et al., “SRNoC: A Statically-Scheduled Circuit-Switched Superconducting Race Logic NoC”, IPDPS 2021
- MG Bautista et al., “Superconducting Shuttle-flux Shift Buffer for Race Logic”, MWCAS 2021
- P Gonzalez et al., “Temporal and SFQ pulse-streams encoding for area-efficient superconducting accelerators”, ASPLOS 2022
- MG Bautista et al., “Superconducting Digital DIT Butterfly Unit for Fast Fourier Transform Using Race Logic”, NEWCAS 2022
- MG Bautista et al., “Superconducting Shuttle-Flux Shift Register for Race Logic and its Applications”, IEEE Transactions on Circuits and Systems, 2022
- D Lyles et al., “PaST-NoC: A Packet-Switched Superconducting Temporal NoC”, IEEE Transactions on Applied Superconductivity 2023
- D Vasudevan et al., “Efficient Temporal Arithmetic Logic Design for Superconducting RSFQ Logic”, IEEE Transactions on Applied Superconductivity 2023 (in press)
- P Gonzalez et al., “An Area Efficient Superconducting Unary CNN Accelerator”, ISQED 2023
- K Huch et al., “Superconducting Hyperdimensional Associative Memory Circuit for Scalable Machine Learning”, IEEE Transactions on Applied Superconductivity 2023 (under review)


?
A