Temporal and SFQ Pulse-Streams Encoding for Area-Efficient Superconducting Accelerators

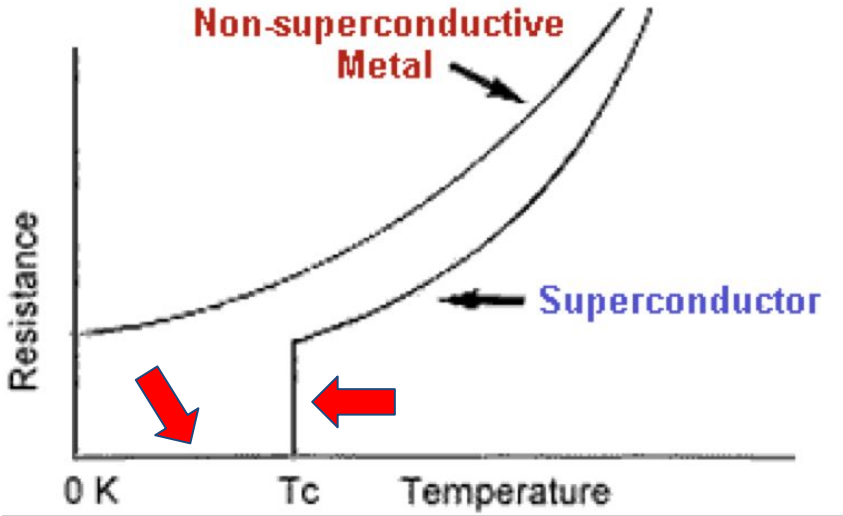
Lead author: Patricia Gonzalez-Guerrero

Speaker: George Michelogiannakis

Computer Architecture Group
Computational Research Division

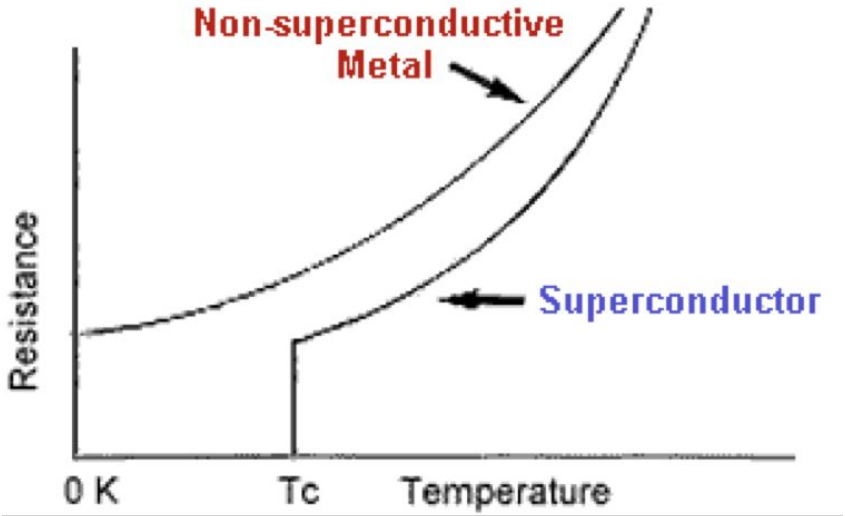
ASPLOS 2022

Superconductivity is a good candidate for the future of computing

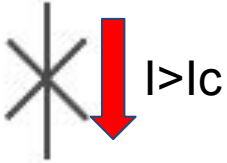


Gallardo et al, "Superconductivity observation in a (CuInTe 2) 1-x (NbTe) x alloy with x=0.5"

Superconductivity is a good candidate for the future of computing

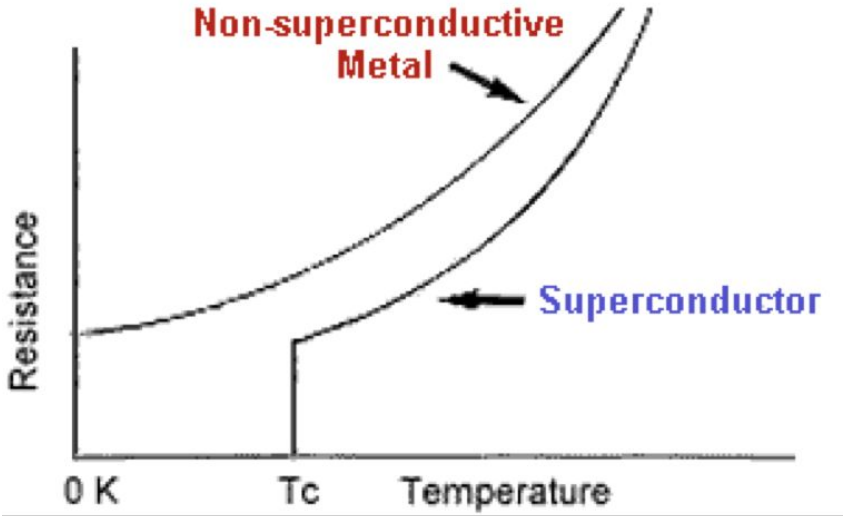


Gallardo et al, "Superconductivity observation in a (CuInTe 2) 1-x (NbTe) x alloy with x=0.5"

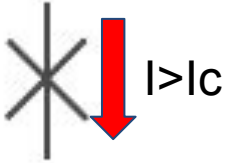


Josephson Junction (JJ)

Superconductivity is a good candidate for the future of computing

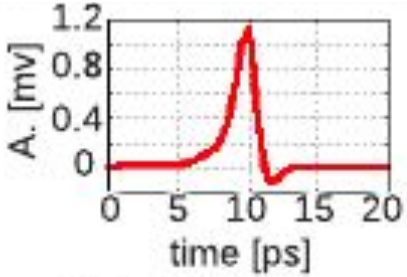


Gallardo et al, "Superconductivity observation in a (CuInTe 2) 1-x (NbTe) x alloy with x=0.5"



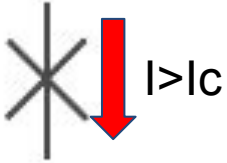
Josephson Junction (JJ)

Single Flux Quantum (SFQ) pulse



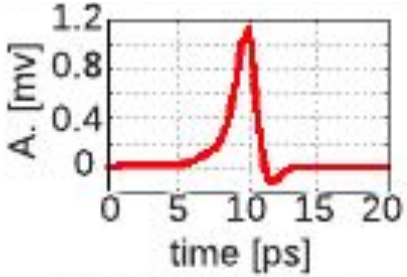
Superconductivity is a good candidate for the future of computing

- Logic Families:
  - Rapid Single Flux Quantum (RSFQ)
  - Energy Efficient Rapid Flux Quantum (ERSFQ)
- 10X-100X faster operation frequency than CMOS.
  - 20GHz - 80GHz
- ~10X lower active energy consumption than CMOS.



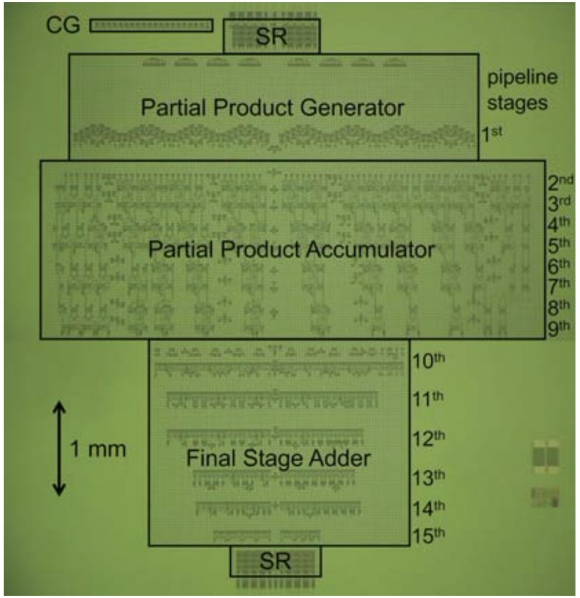
Josephson Junction (JJ)

Single Flux Quantum (SFQ) pulse



Is superconductivity a good candidate for the future of computing, though?

- Extremely constrained area density.
- Manufacturing limitations restrict the number of JJs per chip to ~30000 JJs.
- Current hardware accelerators such as Google's TPU require up to ~64000 multipliers.

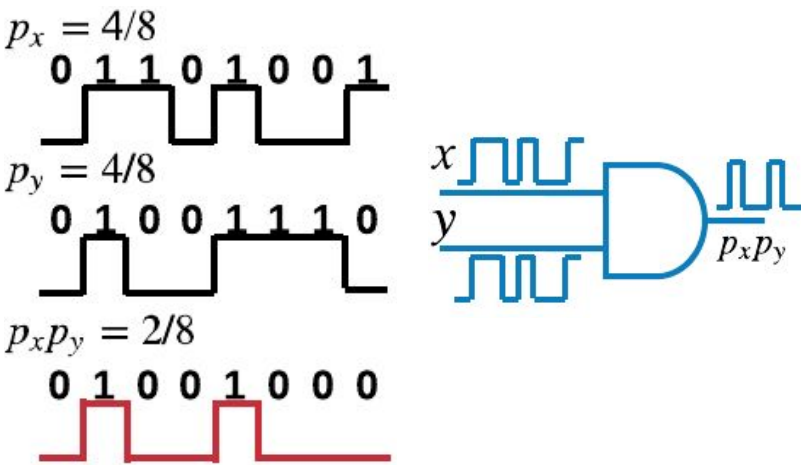


ISSCC'2019.

State of the art RSFQ <u>Binary</u> Multiplier **~17000 JJs**

Unary encoding may address the area density challenge

CMOS stochastic computing



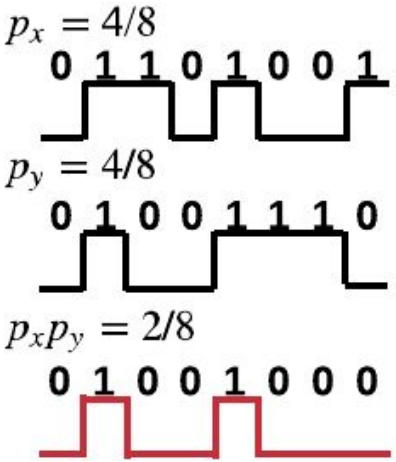
Good for arithmetic!

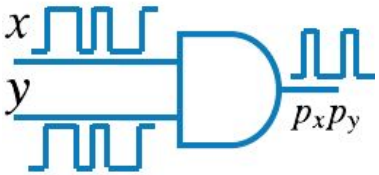
Unary encoding may address the area density challenge

Advait Madhavan, et al. Race Logic: A Hardware Acceleration for Dynamic Programming Algorithms. In ISCA '14

CMOS stochastic computing

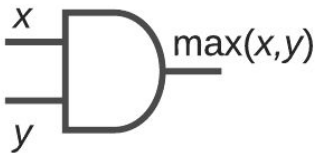
Many wires per variable from binary get replaced with a single wire

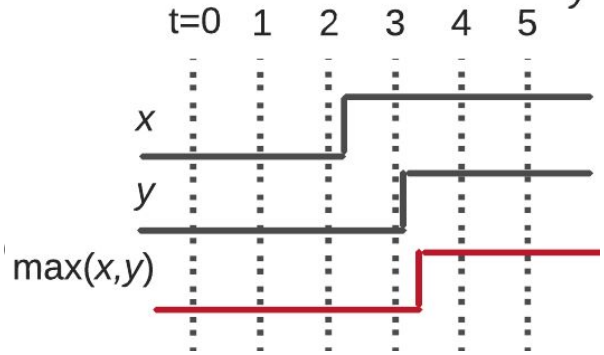




Good for arithmetic!

CMOS Race-Logic



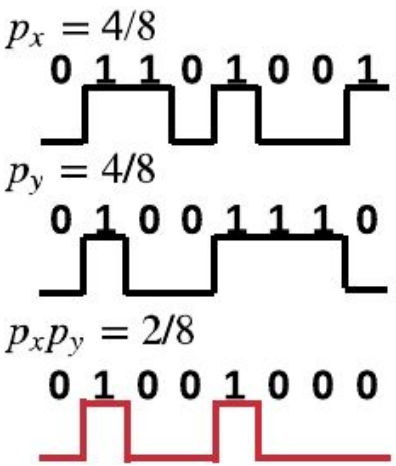


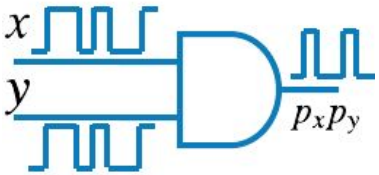
Good for dynamic programing algorithms!

Unary encoding may address the area density challenge

Advait Madhavan, et al. Race Logic: A Hardware Acceleration for Dynamic Programming Algorithms. In ISCA '14

CMOS stochastic computing

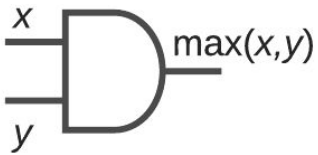


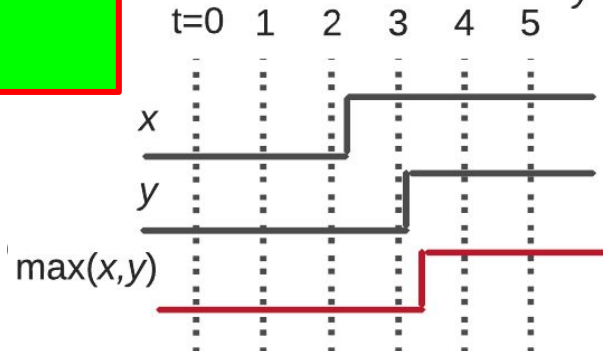


+90% savings in area

Good for arithmetic!

CMOS Race-Logic



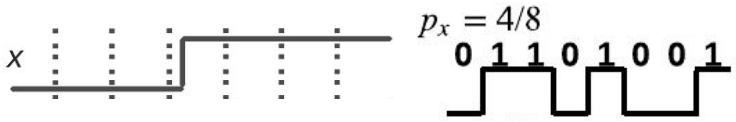


Good for dynamic programing algorithms!

Can we leverage the best of pulse-streams and race logic to build an area-efficient superconducting unary computing architecture?

Outline

1. Superconducting Unary SFQ encoding



CMOS

2. Building Blocks

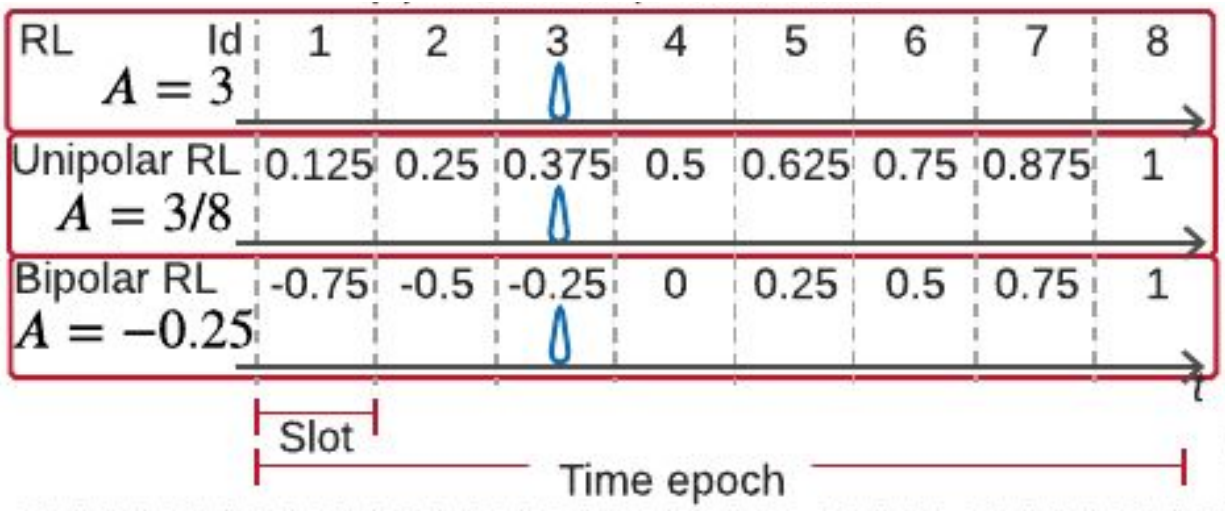
BINARY MULTIPLIER AND ADDER

3. Superconducting Accelerators

- Processing Element (PE) array for CGRAs or ANN
- Dot Product Unit
- Finite Impulse Response Filter

In superconducting race logic data is mapped to the time the SFQ pulse arrives

Georgios Tzimpragos et al. 2020. A Computational Temporal Logic for Superconducting Accelerators. In ASPLOS ’20

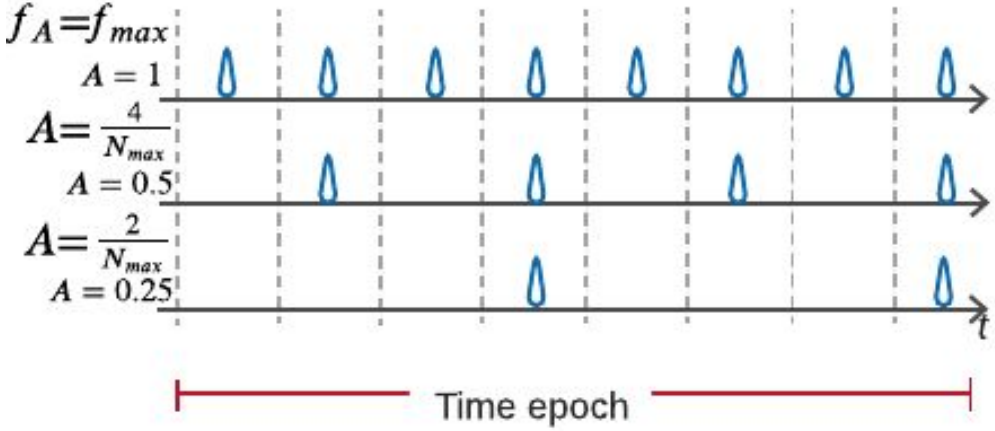


$$N_{max} = 8$$

$$A_b = 2A_u - 1$$

To obtain bipolar representation

In a superconducting pulse-stream data is mapped to the frequency of the SFQ pulses



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

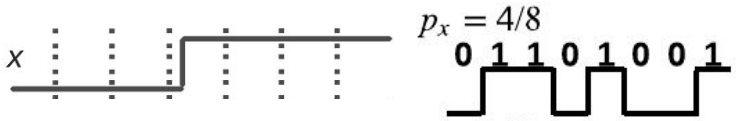
$$A = n/N_{max}$$

$$A_b = 2A_u - 1$$

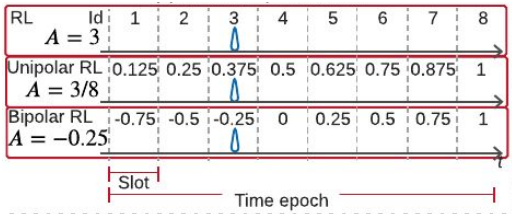
To obtain bipolar representation

Outline

1. **Superconducting Unary encoding**



CMOS

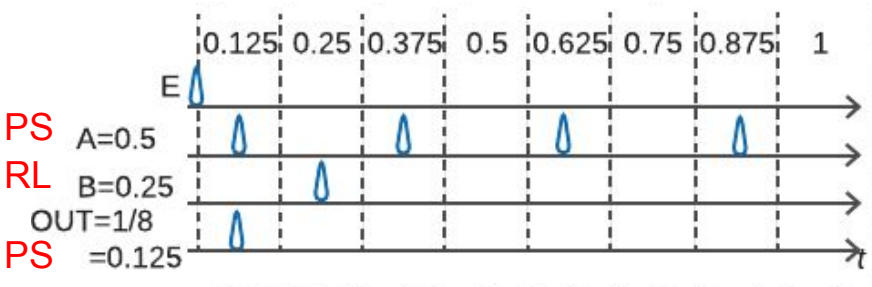


2. **Building Blocks**

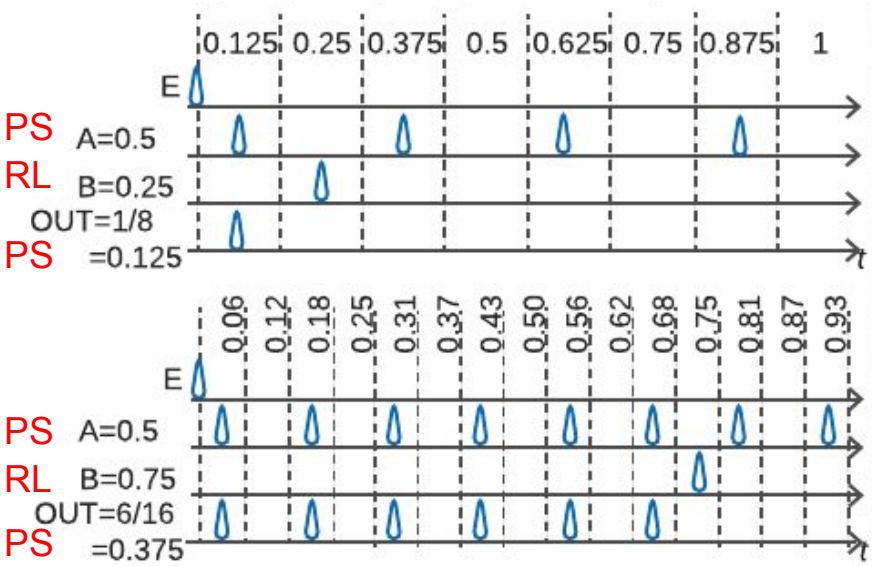
BINARY MULTIPLIER AND ADDER

3. **Superconducting Accelerators**
   - Processing Element (PE) array for CGRAs or ANN
   - Dot Product Unit
   - Finite Impulse Response Filter

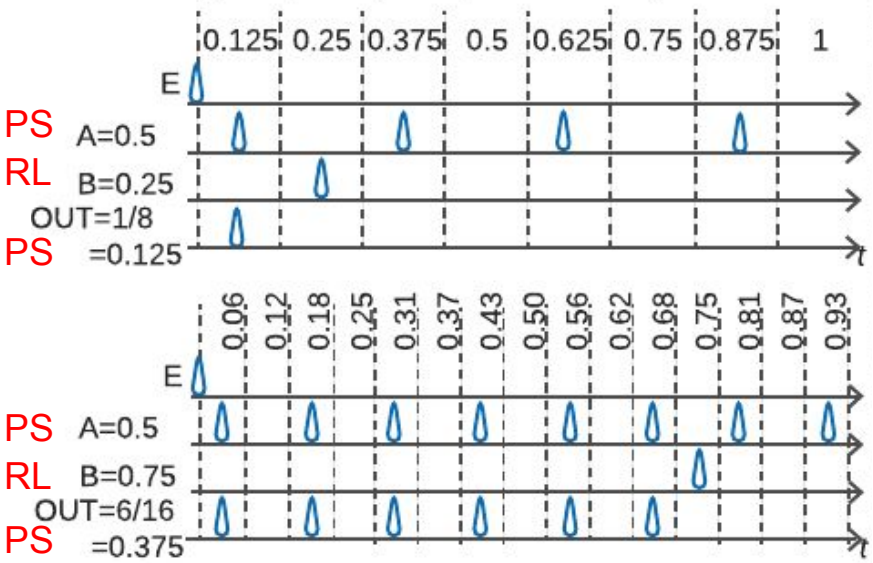
Building Blocks: Multipliers



Building Blocks: Multipliers

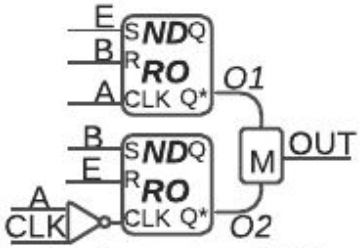


Building Blocks: Multipliers



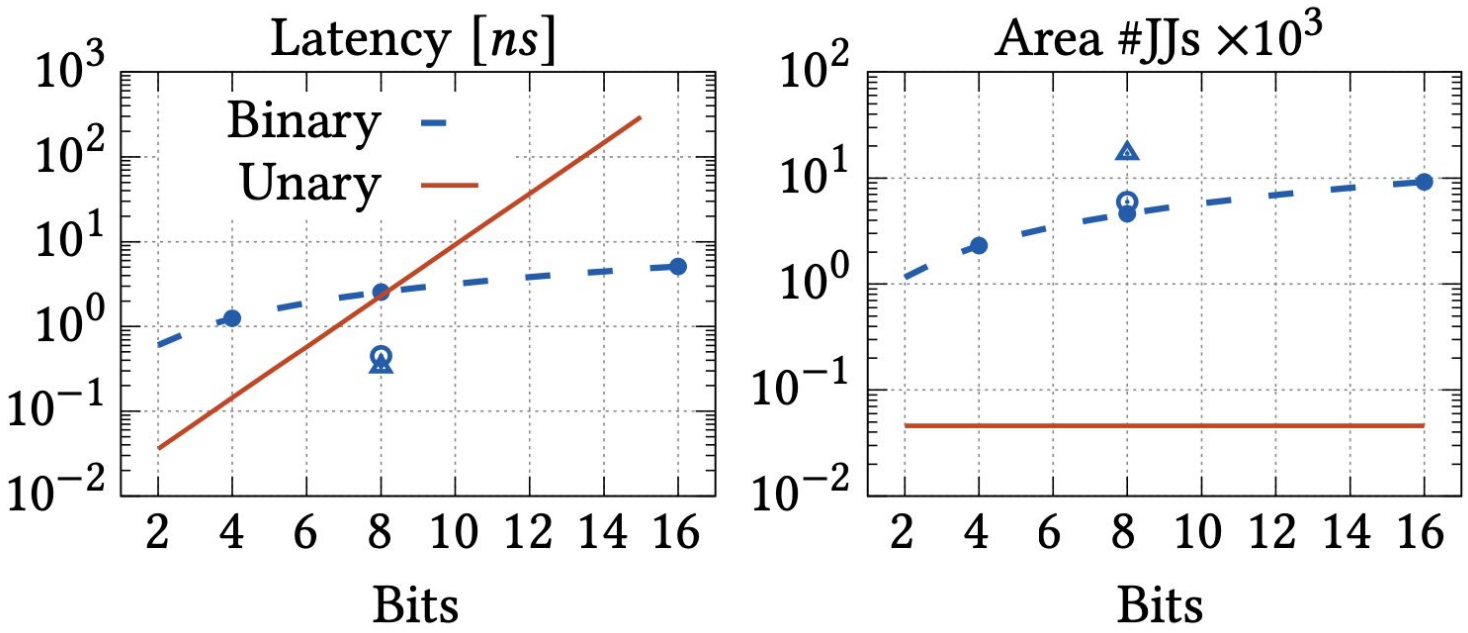
Essentially a CMOS XOR

Unipolar SFQ multiplier



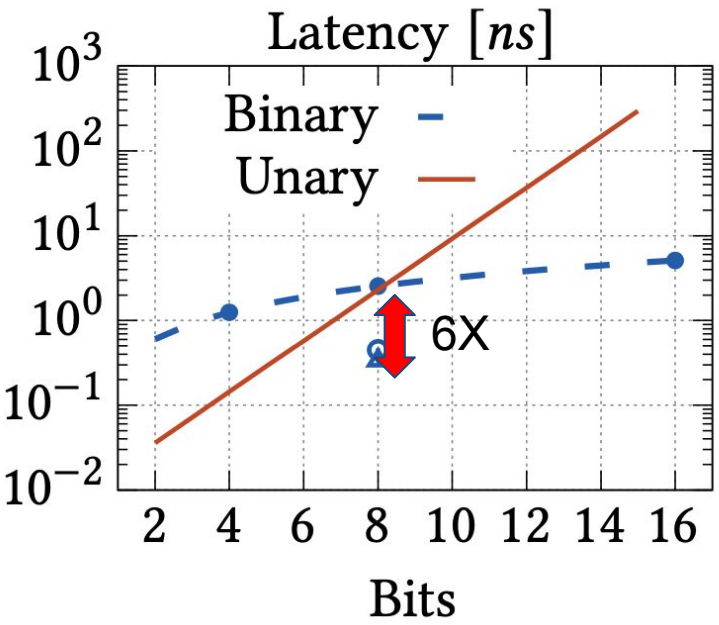
Bipolar SFQ multiplier

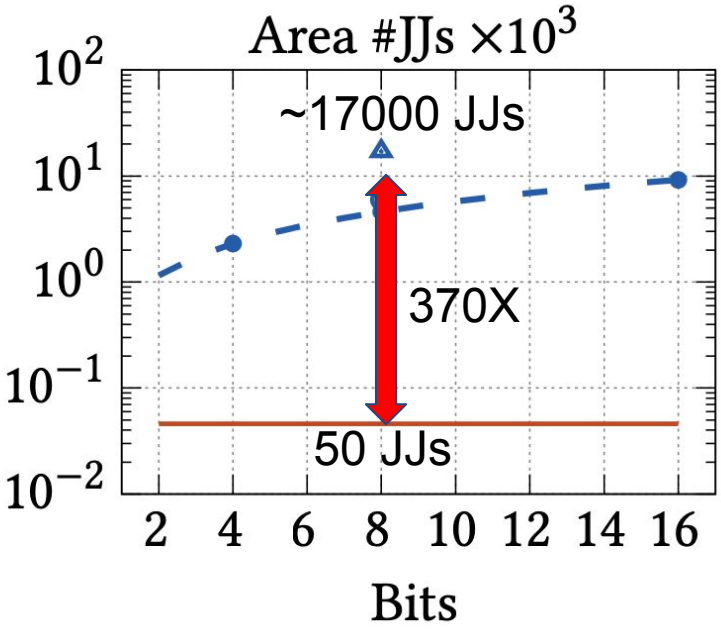
The superconducting unary multiplier exposes an area-latency tradeoff



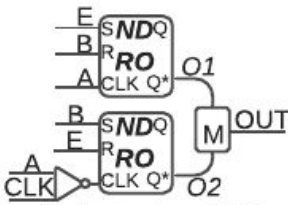
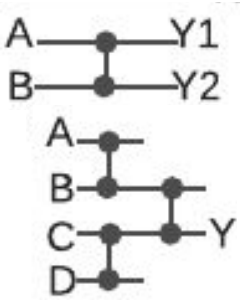
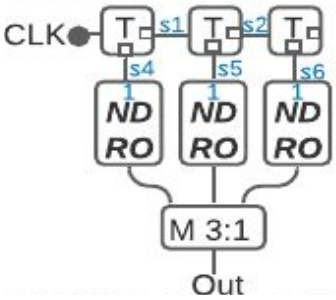
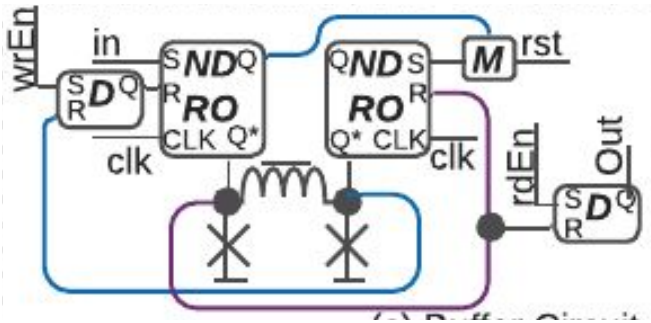
Bipolar unary multiplier

The superconducting unary multiplier exposes an area-latency tradeoff



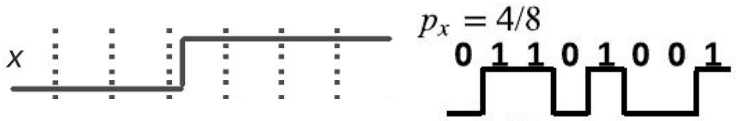


The building blocks for the superconducting unary architecture

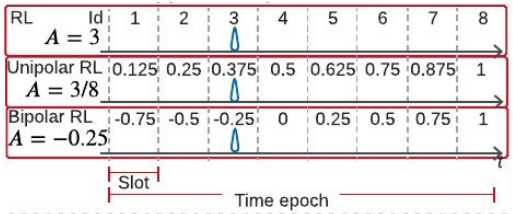
Multiplier	Adder	Pulse-stream generator	Memory cell for RL
			
25X-370X Area Savings	**11X-200X Area Savings**	**10% Area penalty**	**2.5X-1.3X Area penalty**

Outline

1. Superconducting Unary encoding

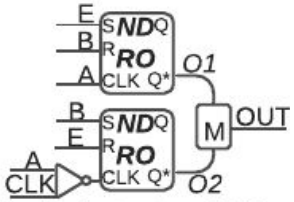


CMOS



2. Building Blocks

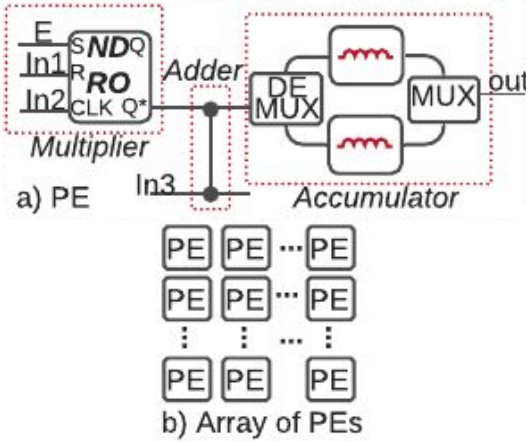
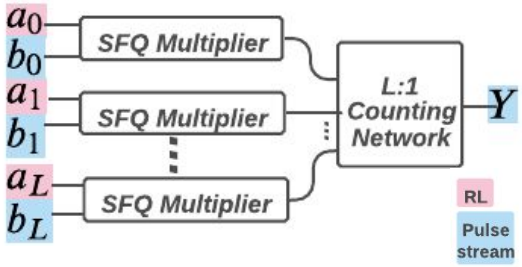
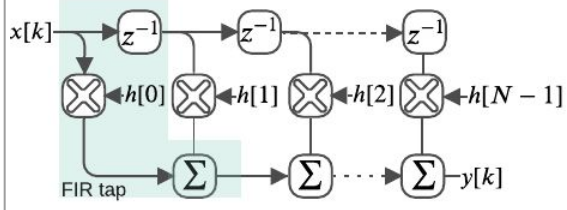
BINARY MULTIPLIER AND ADDER



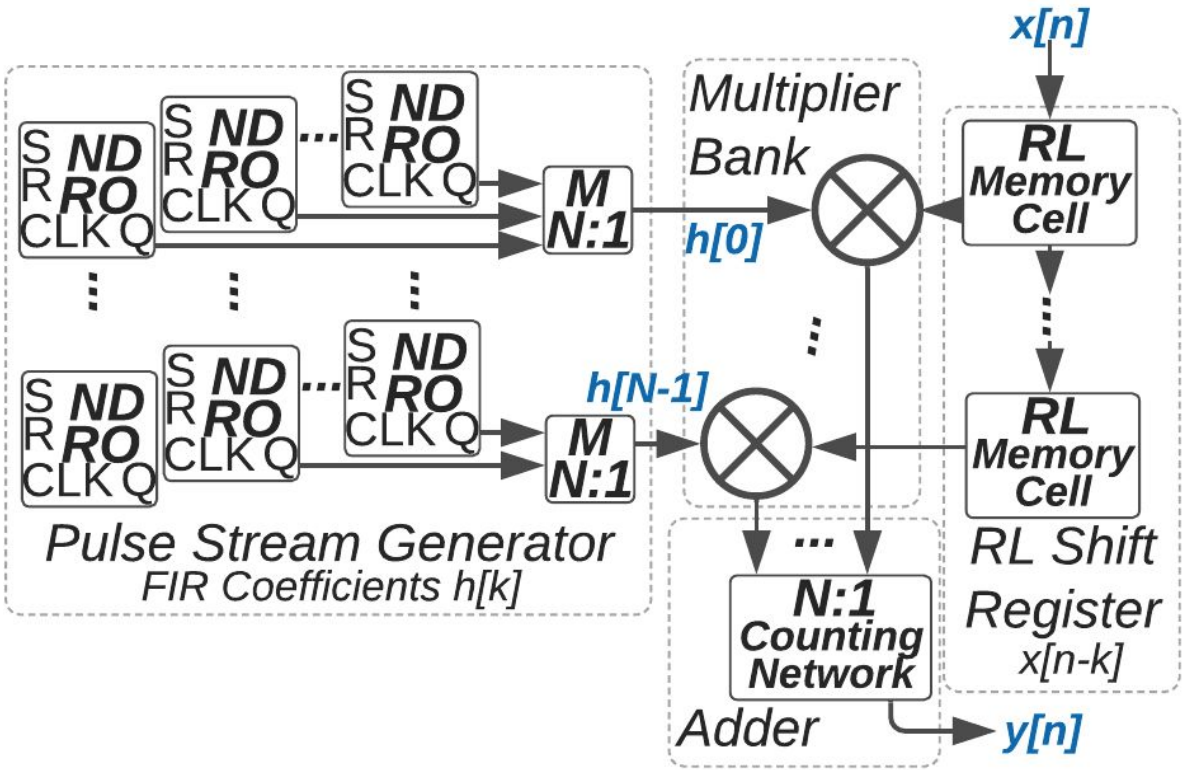
3. Superconducting Accelerators

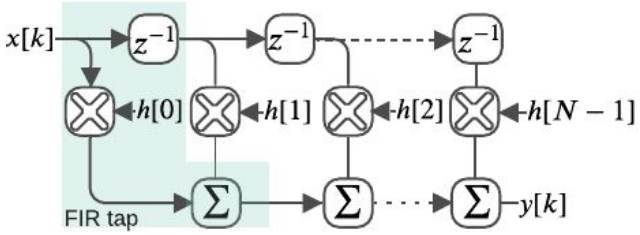
- Processing Element (PE) array for CGRAs or ANN
- Dot Product Unit
- Finite Impulse Response Filter

Evaluate the proposed superconducting architecture with three applications

Processing Element for CGRA and Neural Networks	Dot Product Unit for DSP and Neural Networks	Programmable Finite Impulse Response Filter
	 $y = a \circ b = \sum_{i=0}^{L-1} a[i]b[i]$	 $y[n] = \sum_{k=0}^{N-1} h[k]x[n-k]$

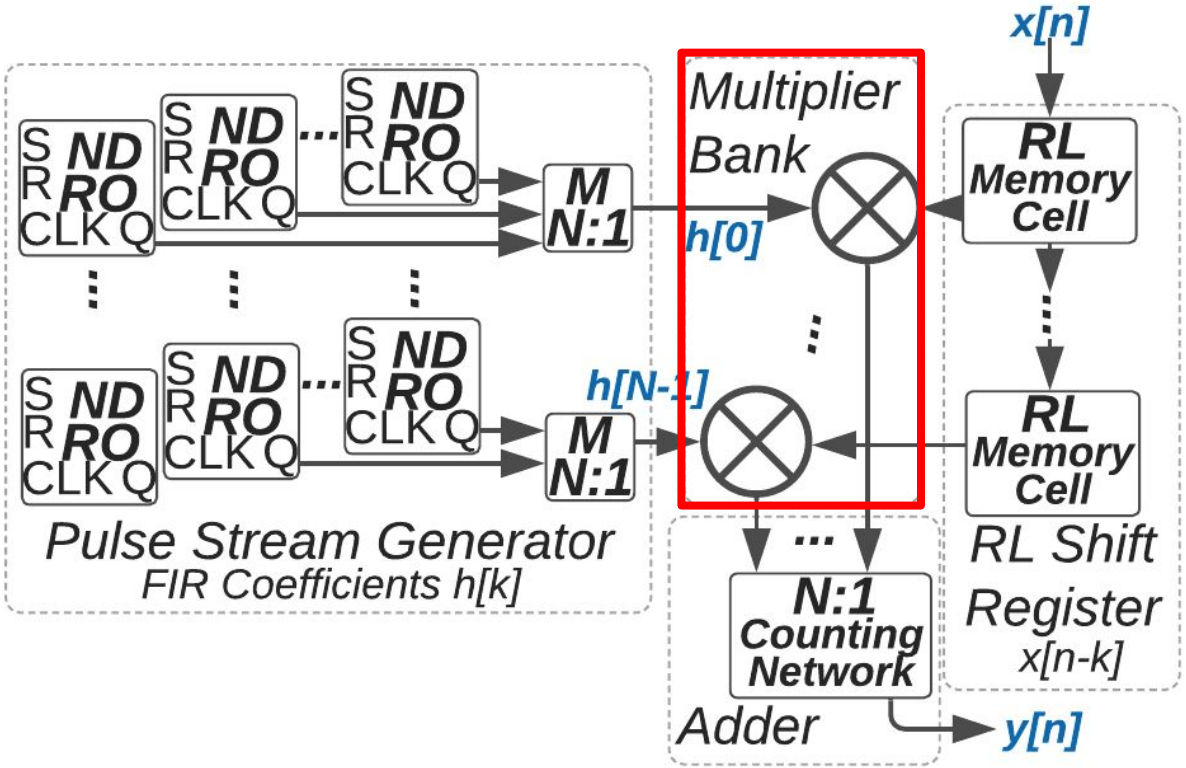
The FIR Filter accelerator shows the strengths and challenges of the proposed unary architecture

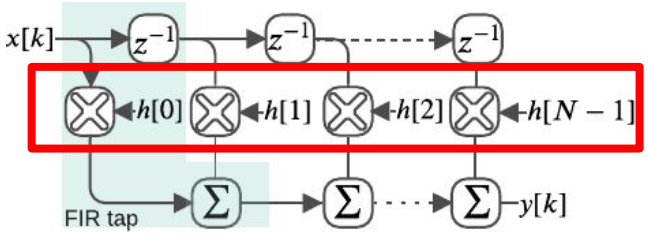




$$y[n] = \sum_{k=0}^{N-1} h[k]x[n-k]$$

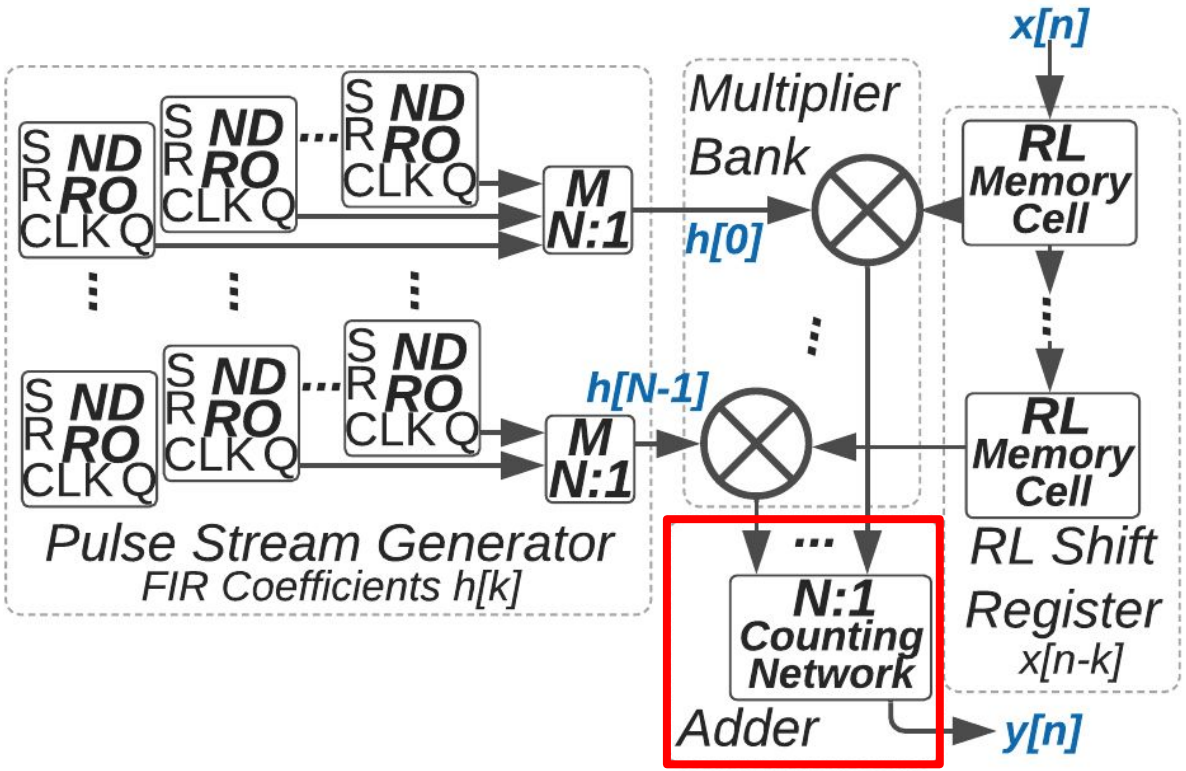
The FIR Filter accelerator shows the strengths and challenges of the proposed unary architecture

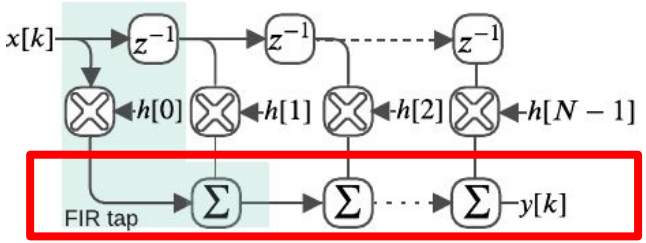




$$y[n] = \sum_{k=0}^{N-1} h[k]x[n-k]$$

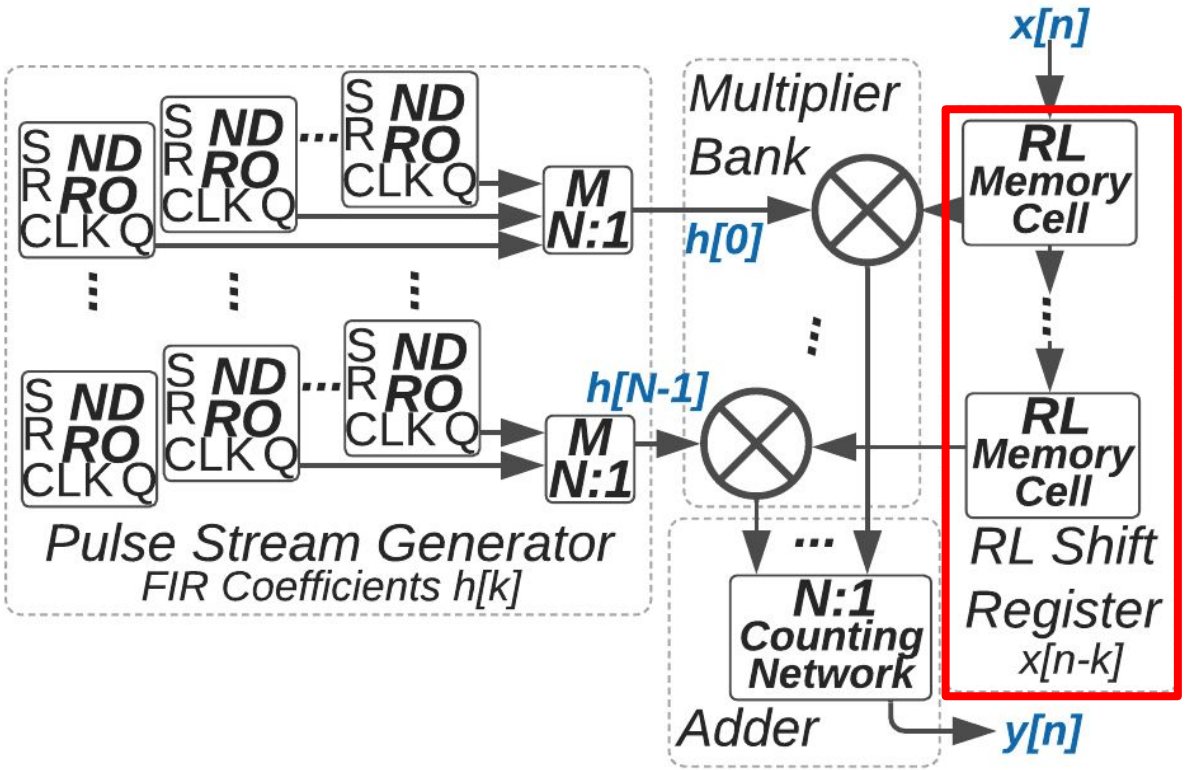
The FIR Filter accelerator shows the strengths and challenges of the proposed unary architecture

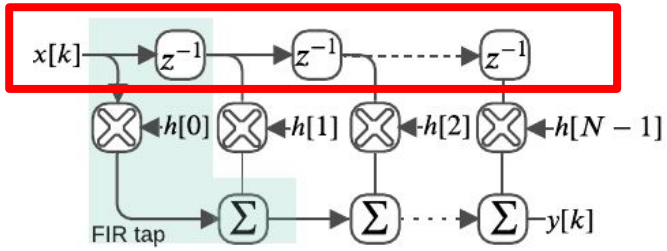




$$y[n] = \sum_{k=0}^{N-1} h[k]x[n-k]$$

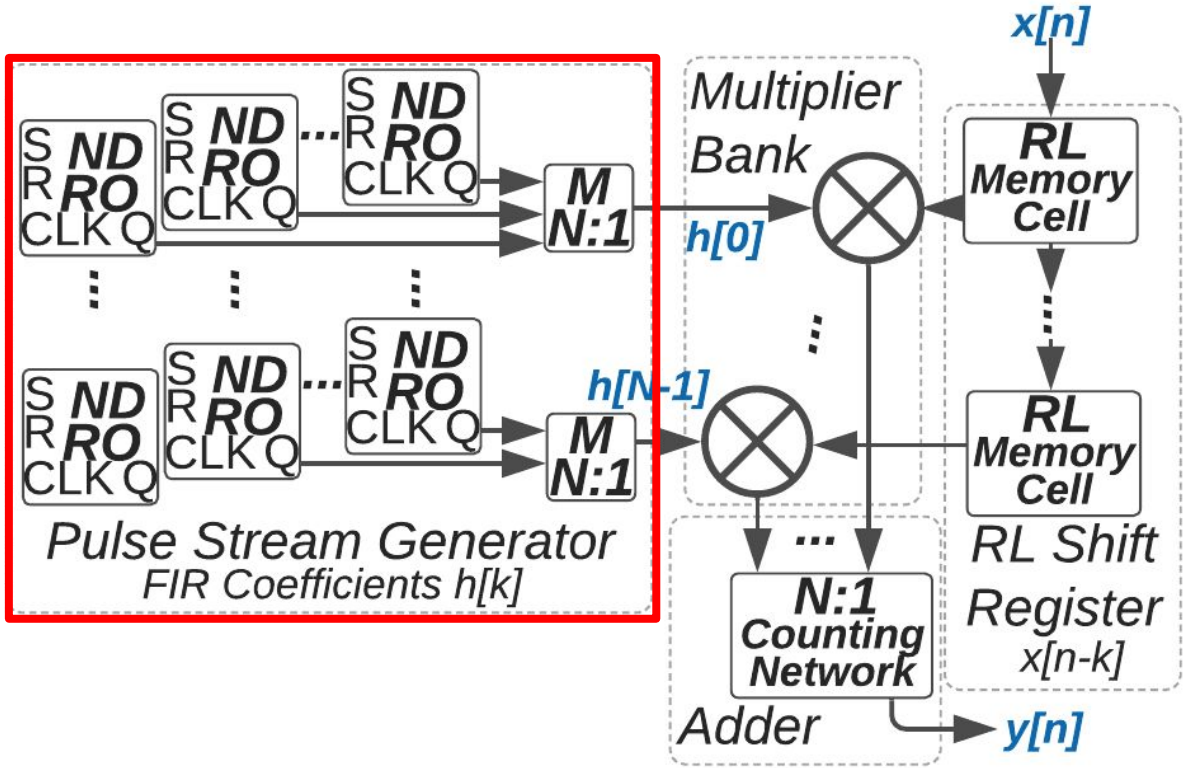
The FIR Filter accelerator shows the strengths and challenges of the proposed unary architecture

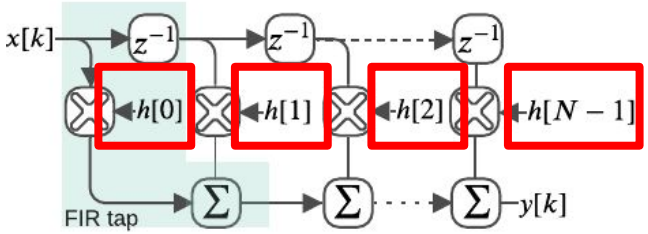




$$y[n] = \sum_{k=0}^{N-1} h[k]\,x[n-k]$$

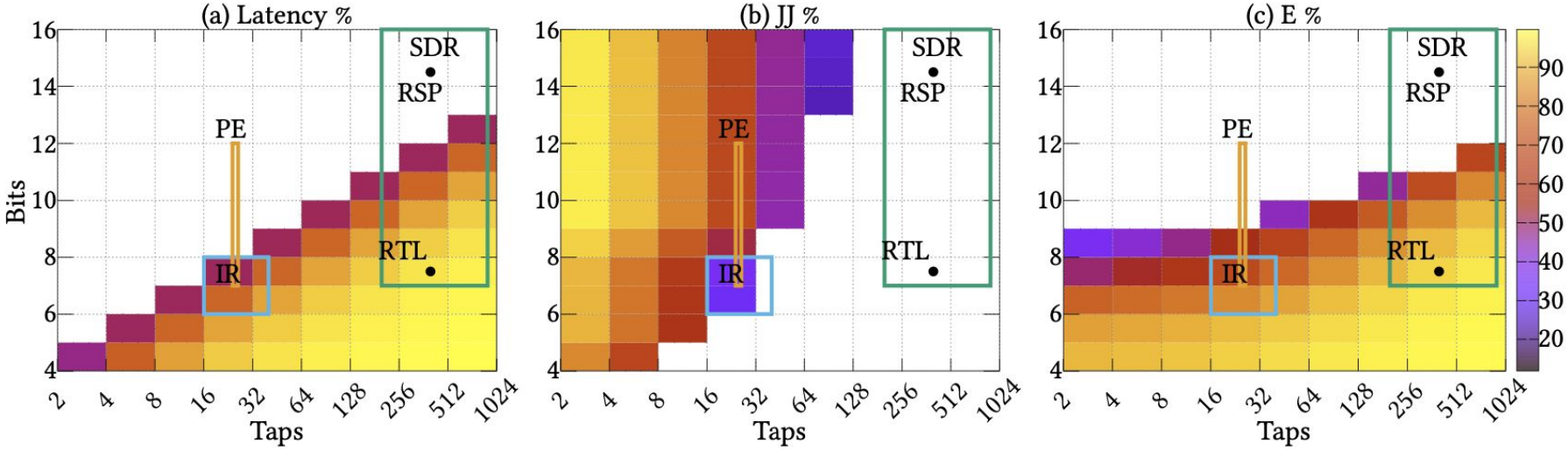
The FIR Filter accelerator shows the strengths and challenges of the proposed unary architecture





$$y[n] = \sum_{k=0}^{N-1} \boxed{h[k]} x[n-k]$$

FIR shows a design space where superconducting Unary yields significant advantages over superconducting binary



Color shows percentage savings

Thank you for your attention

Contact information:

lg4er@lbl.gov, mihelog@lbl.gov