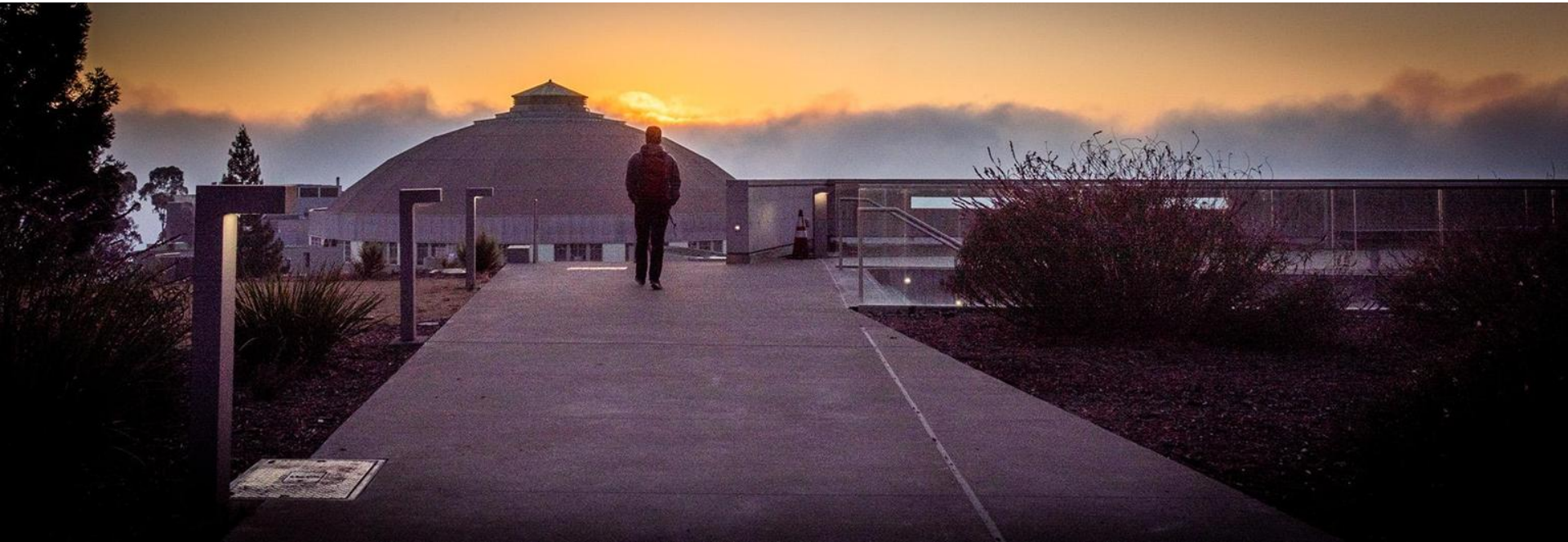
BERKELEY LAB
Bringing Science Solutions to the World

U.S. DEPARTMENT OF ENERGY
Office of Science

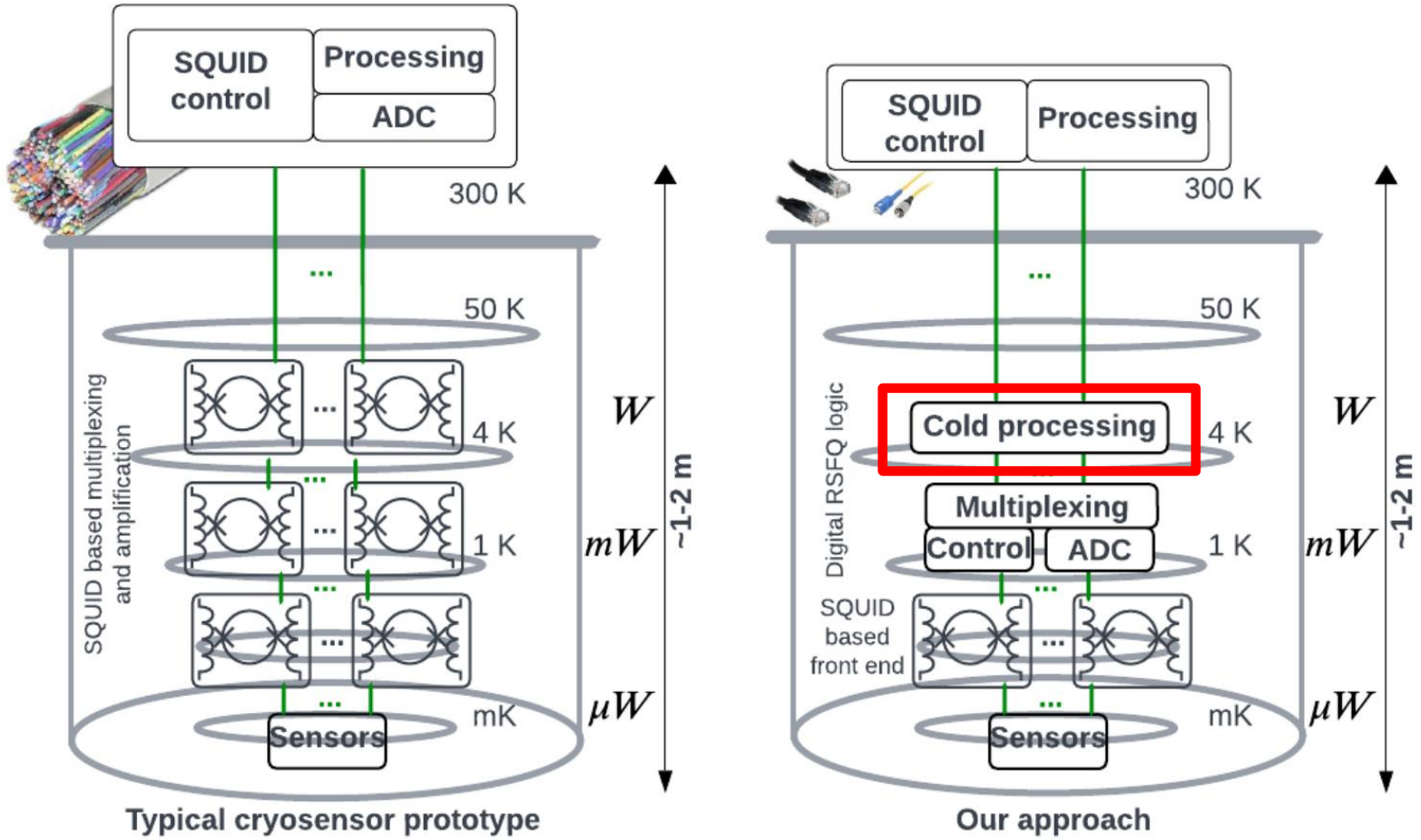
Temporal and Pulse-Train Computing for Reliable and Efficient Computing

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



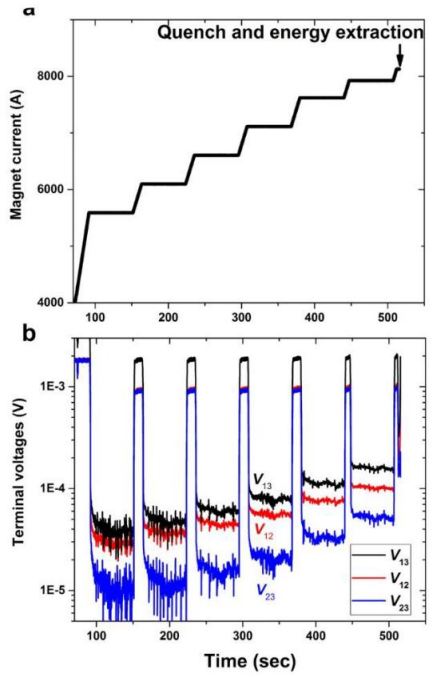
Our Current Focus: Cryogenic Sensors/Qubits

- Expensive and noise-prone cables to move data to room temperature
- Lower-temperature environments are more power constrained and are more noisy



Superconducting Magnet Quench Detection

A new quench avoidance paradigm for HTS magnets



Current (top) and voltage (bottom) plots for a subscale Bi-2212 HTS coil. From: T. Shen, *et al*. 2019 *Sci Rep* **9**, 10170.

- Quench propagation in HTS conductors is one to two orders of magnitude slower than in conventional superconductors
- HTS magnet protection is difficult, as voltage rise associated with the normal zone may be too low to detect it reliably
- Experiments show that HTS conductors can operate in a stable dissipative regime before entering thermal runaway

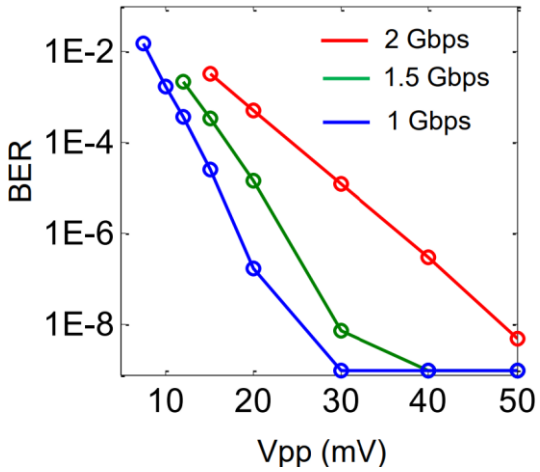
➢ Therefore, a new protection paradigm for HTS magnets has emerged, aiming at **avoiding quenching** altogether

➢ We will **detect the dissipative regime** using advanced non-voltage diagnostics and **estimate proximity** to the runaway

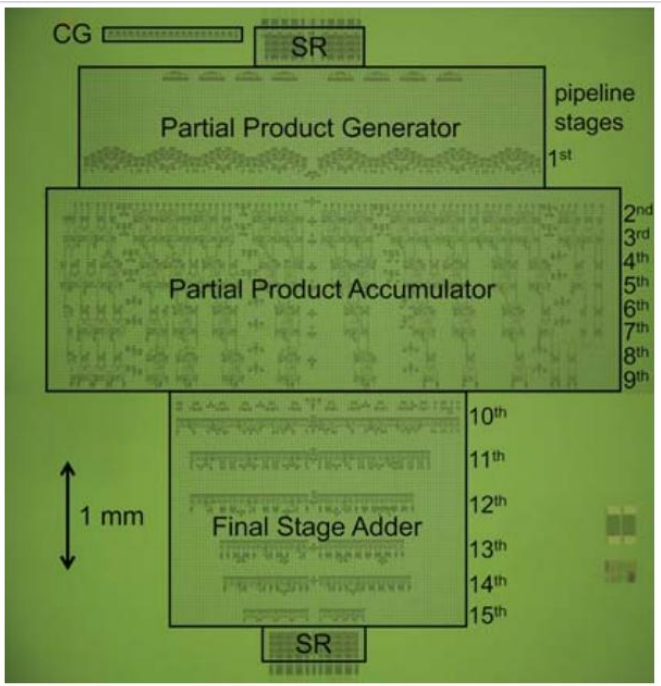
Slide from Maxim Marchevsky

Motivating Challenges

- Device density
  - Makes area a primary constraint
  - And memory capacity
- Cables to room temperature
  - Mostly reliability
- Circuit reliability under harsh environments
  - Or with thermal noise, cooling variation



P Pintus et al., "Ultralow voltage, high-speed, and energy-efficient cryogenic electro-optic modulator", Optica Vol. 9, Issue 10, pp. 1176-1182 (2022)

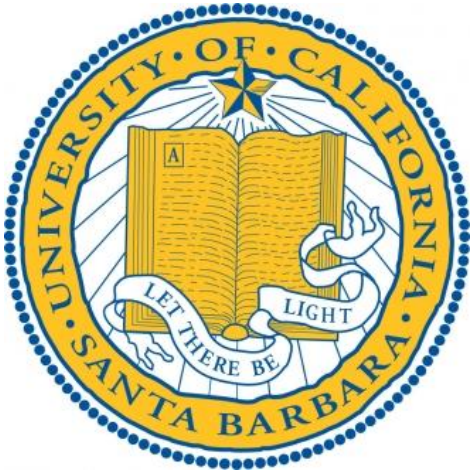


I Nagaoka et al., "A 48GHz 5.6mW Gate-Level-Pipelined Multiplier Using Single-Flux Quantum Logic", ISSCC 2019

Temporal Computing

The Team For The Early Work

LBNL and UCSB team. Funded by ARO 2019 - 2022







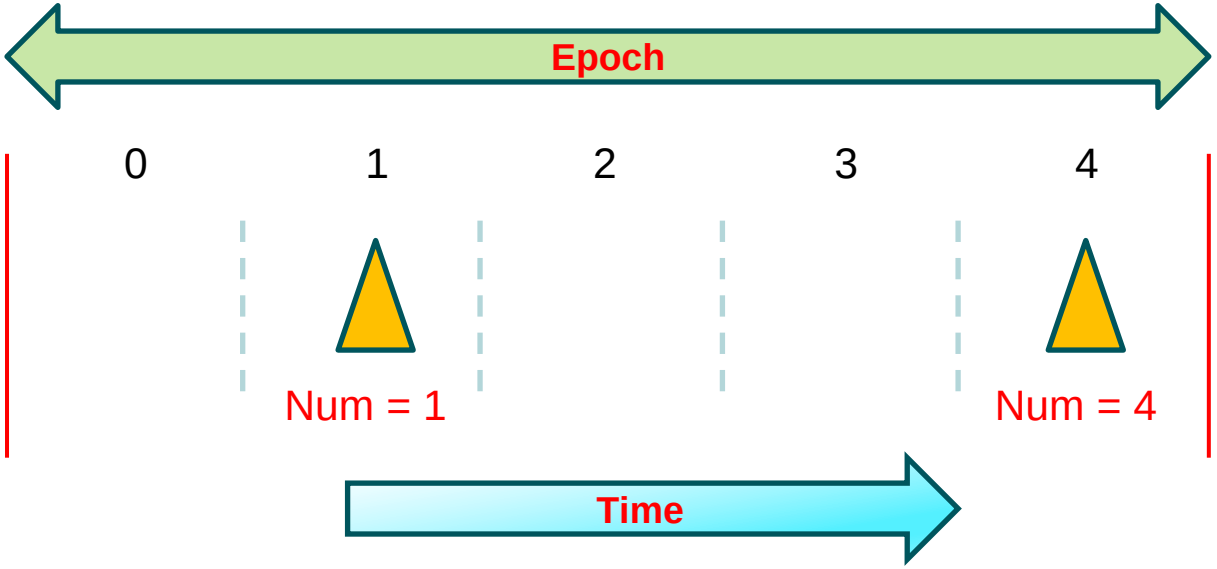





Data Encoding in Race Logic

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

- Epochs repeat
- Epoch duration = TimeSlotDuration × NumTimeSlots
- Each pulse represents an equivalent 2^N binary number
  - (N = NumTimeSlots)
- Can efficiently represent non power of two number ranges



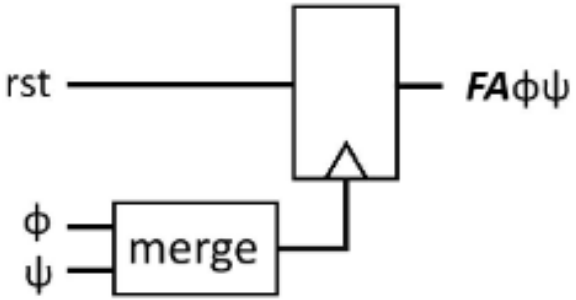
G Tzimpragos et al., “A computational temporal logic for superconducting accelerators”, ASPLOS 2020

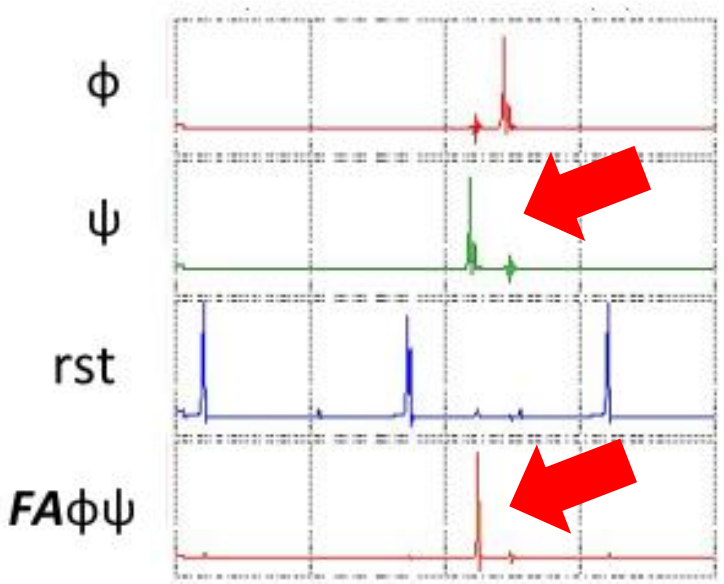
First Arrival – The MIN Function

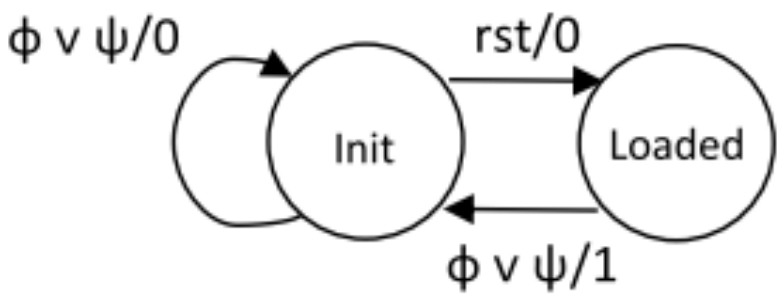
First incoming pulse causes an output pulse. Has a reset

MIN(φ,ψ)

DFF's clock input repurposed





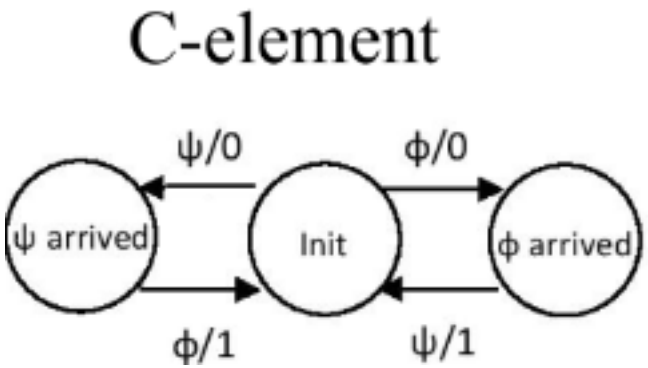


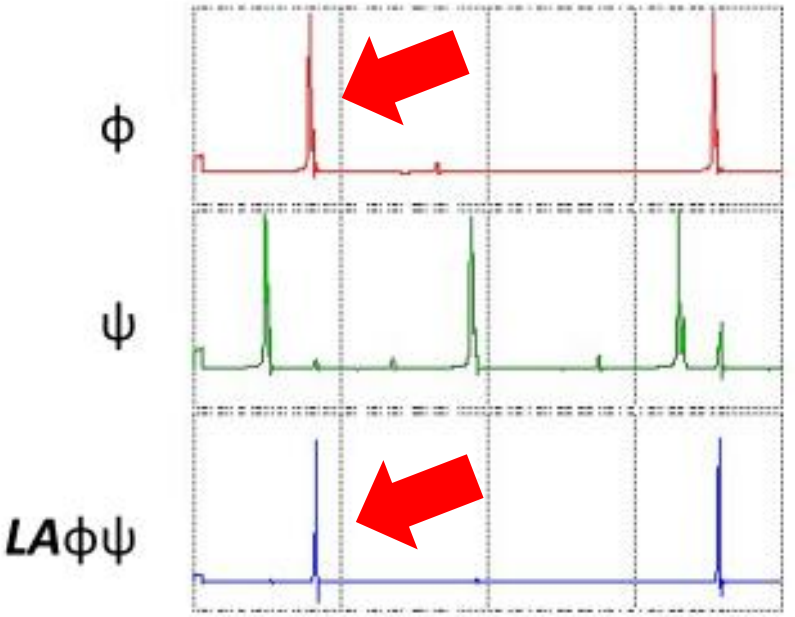
G Tzimpragos et al., "A computational temporal logic for superconducting accelerators", ASPLOS 2020

Last Arrival – The MAX Function

Last incoming pulse causes an output pulse

MAX(φ,ψ)



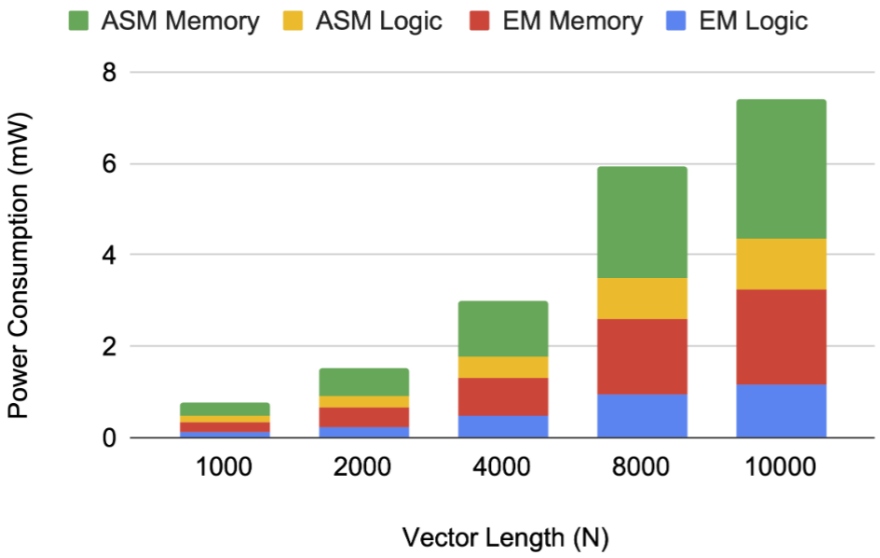


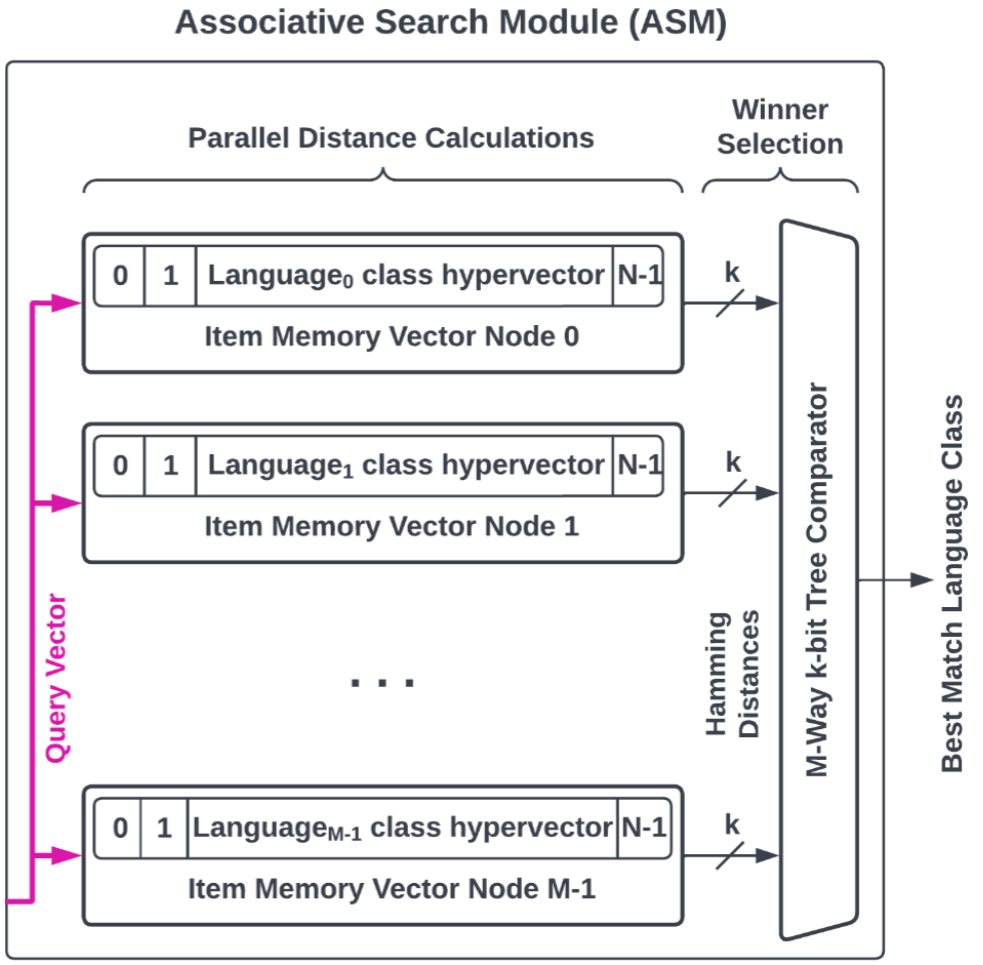
G Tzimpragos et al., “A computational temporal logic for superconducting accelerators”, ASPLOS 2020

Race Logic Makes Comparison Easy

Which is one of the scaling bottlenecks of hyperdimensional computing (HDC)

- Area still large (due to other HDC modules), but much smaller than binary

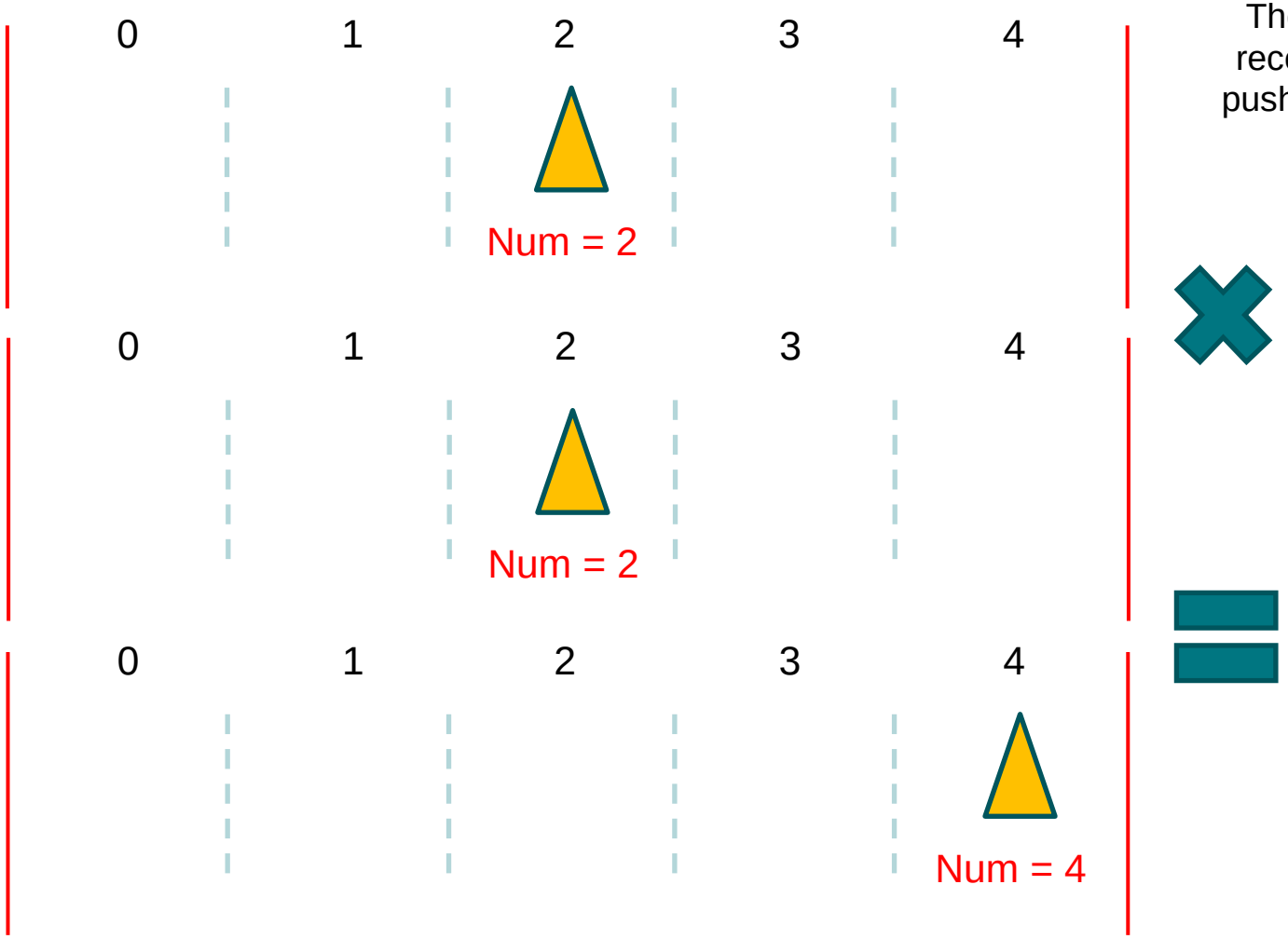




Kylie Huch et al., “Superconducting Hyperdimensional Associative Memory Circuit for Scalable Machine Learning” IEEE Transactions on applied superconductivity 2023

Arithmetic in RL is Expensive

For instance, multiplying two race logic pulses



There are some recent works that push this boundary

U-SFQ: Temporal and Pulse Streams Encoding

Instead: Unipolar and Bipolar Race Logic

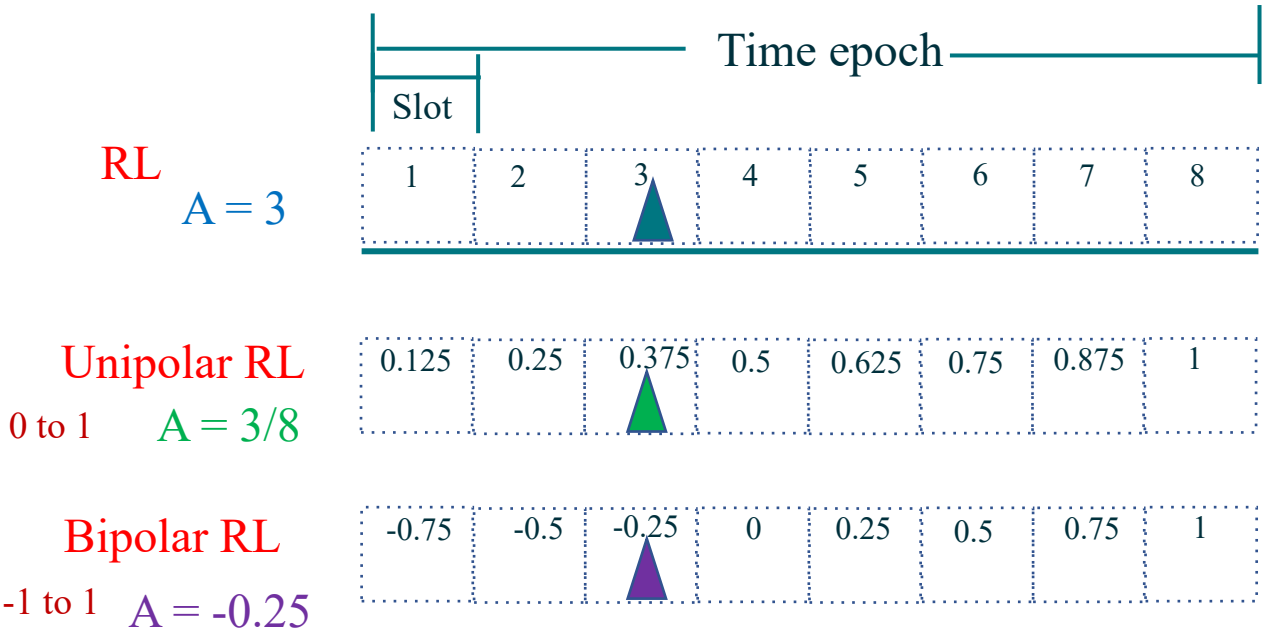
Changing the range of representation to [0,1] (unipolar)

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

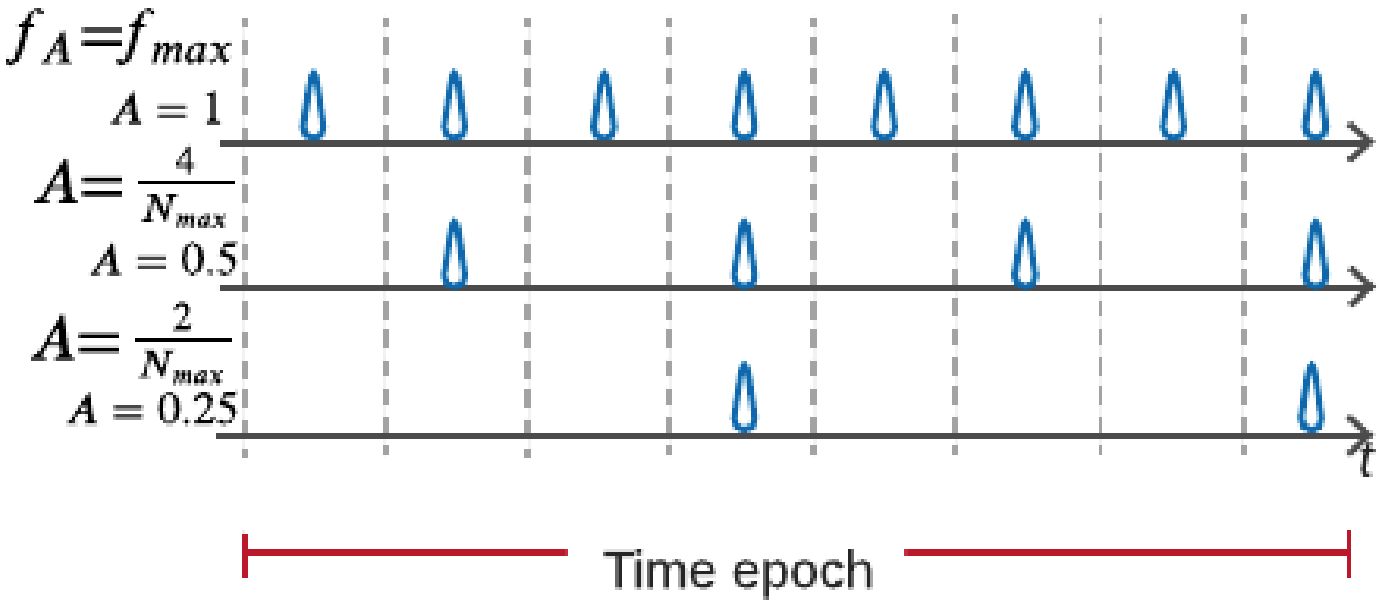
Maps a value to the number of pulses. “1” is for the maximum number of pulses

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

$A = n/N_{max}$

To obtain bipolar representation (not shown)

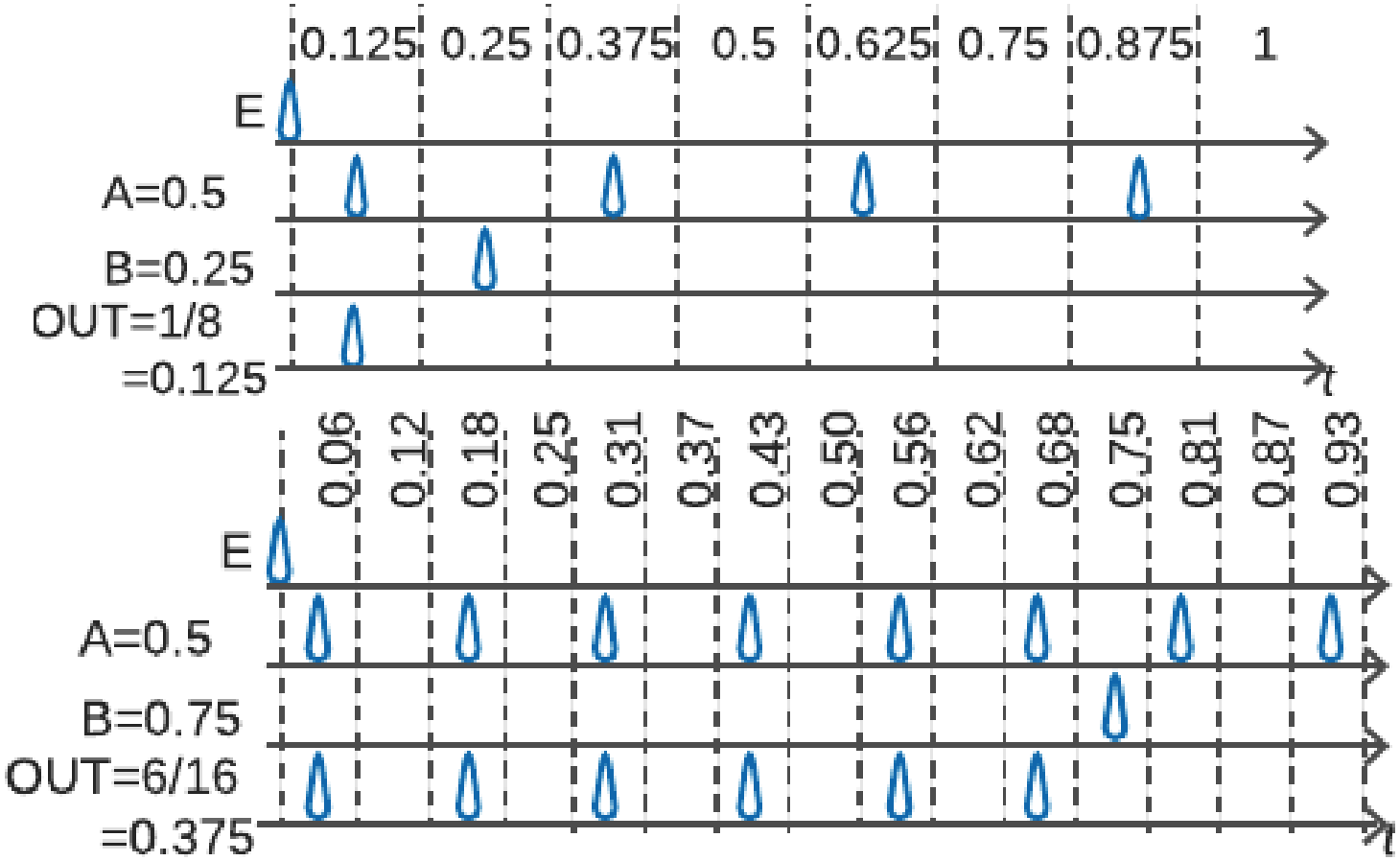
$A_b = 2A_u - 1$



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



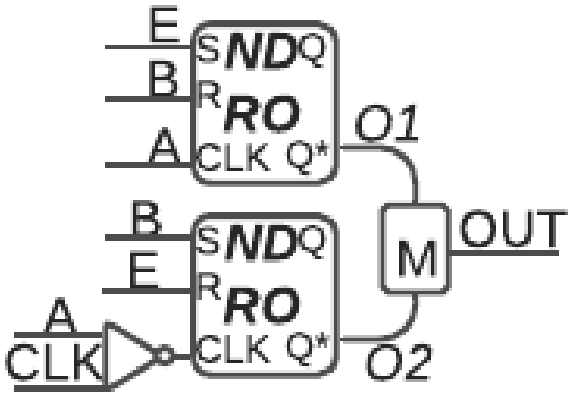
P Gonzalez-Guerrero et al., “Temporal and SFQ pulse-streams encoding for area-efficient superconducting accelerators”, ASPLOS 2022

Multiplication With Just One or Four 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

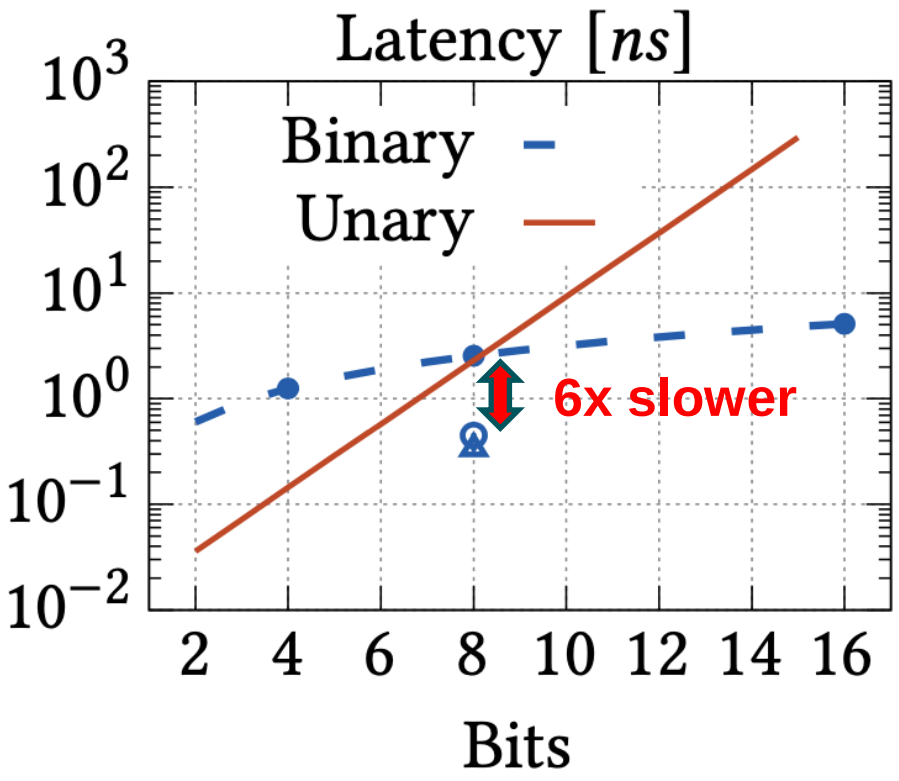
“Clk” denotes time slot boundaries

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

U-SFQ Multiplier Exposes an Area-Latency Tradeoff

A fundamental tradeoff in race logic compute circuits.

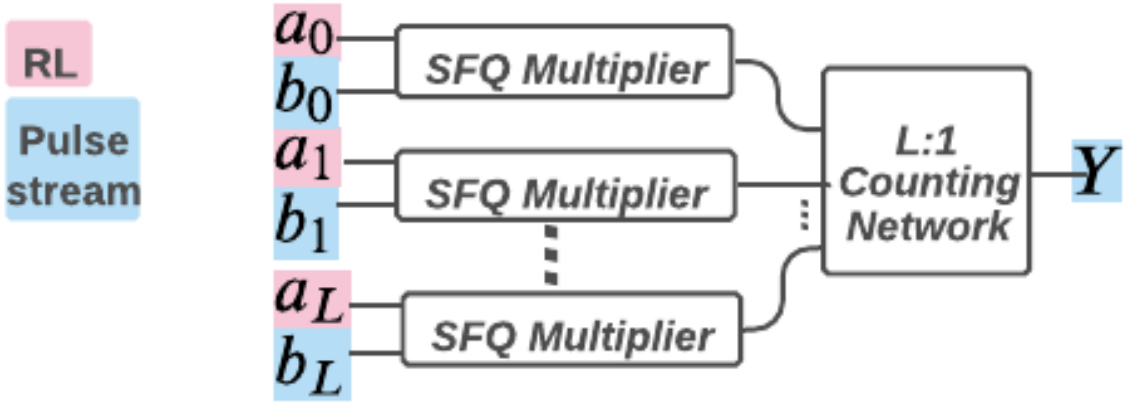
U-SFQ provides higher performance over area

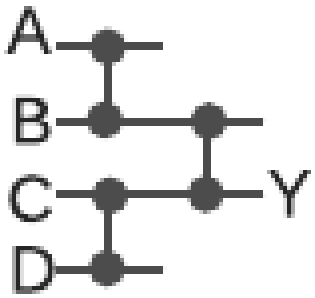


P Gonzalez-Guerrero et al., “Temporal and SFQ pulse-streams encoding for area-efficient superconducting accelerators”, ASPLOS 2022

Multiply-Accumulate Unit

Final result is a pulse stream





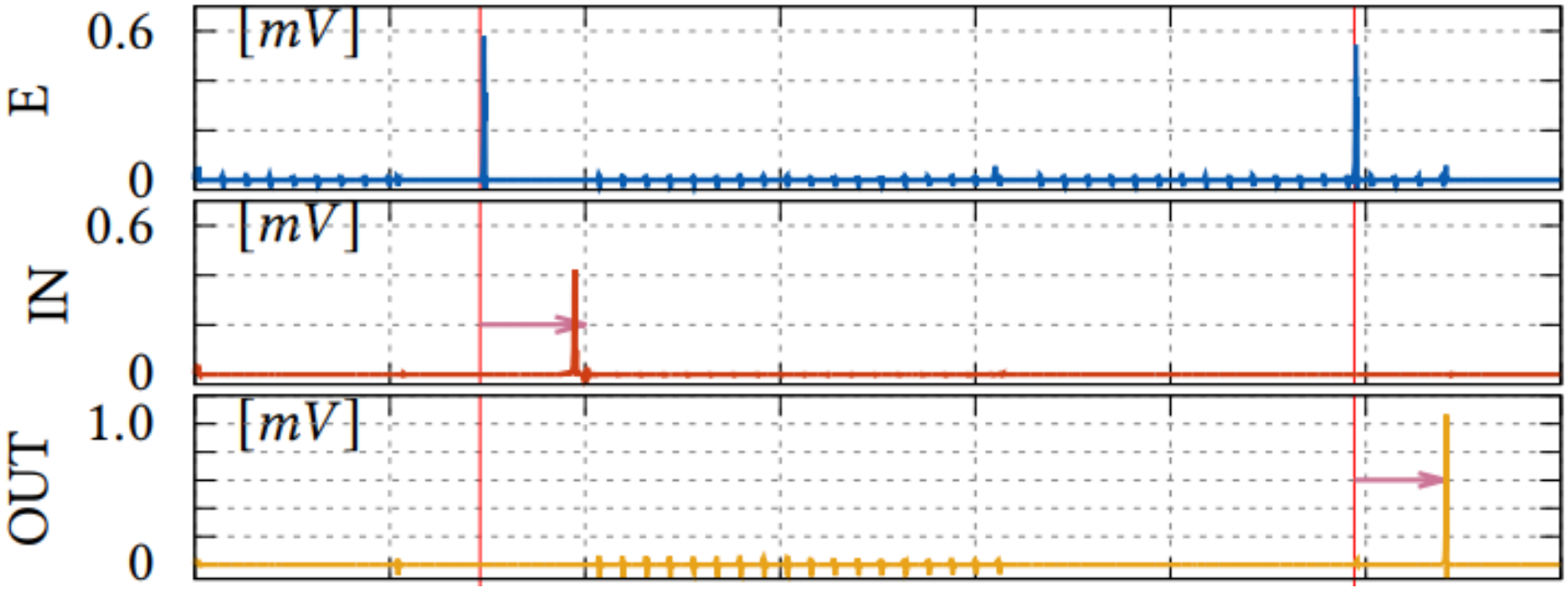
11x – 200x less area

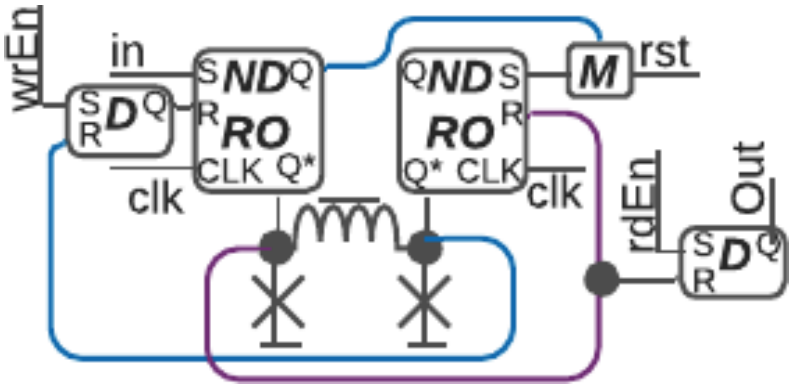
$$y = a \circ b = \sum_{i=0}^{L-1} a[i]b[i]$$

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

Memory Cell for Race Logic

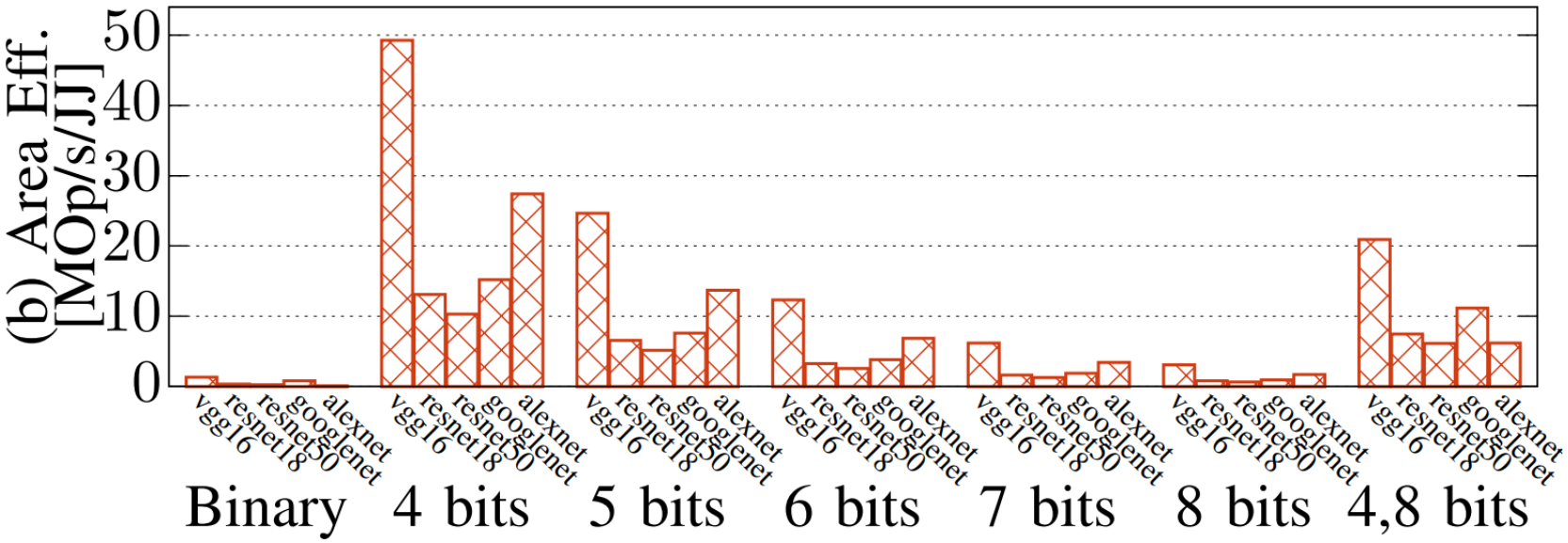
Inductor-based. Releases an incoming pulse in the same time slot of a future epoch





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

CNN Accelerator



Binary uses 8 bits

(4,8) means inner layers 4 bits, outer 8 bits

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

On-Chip Data Movement and Networks

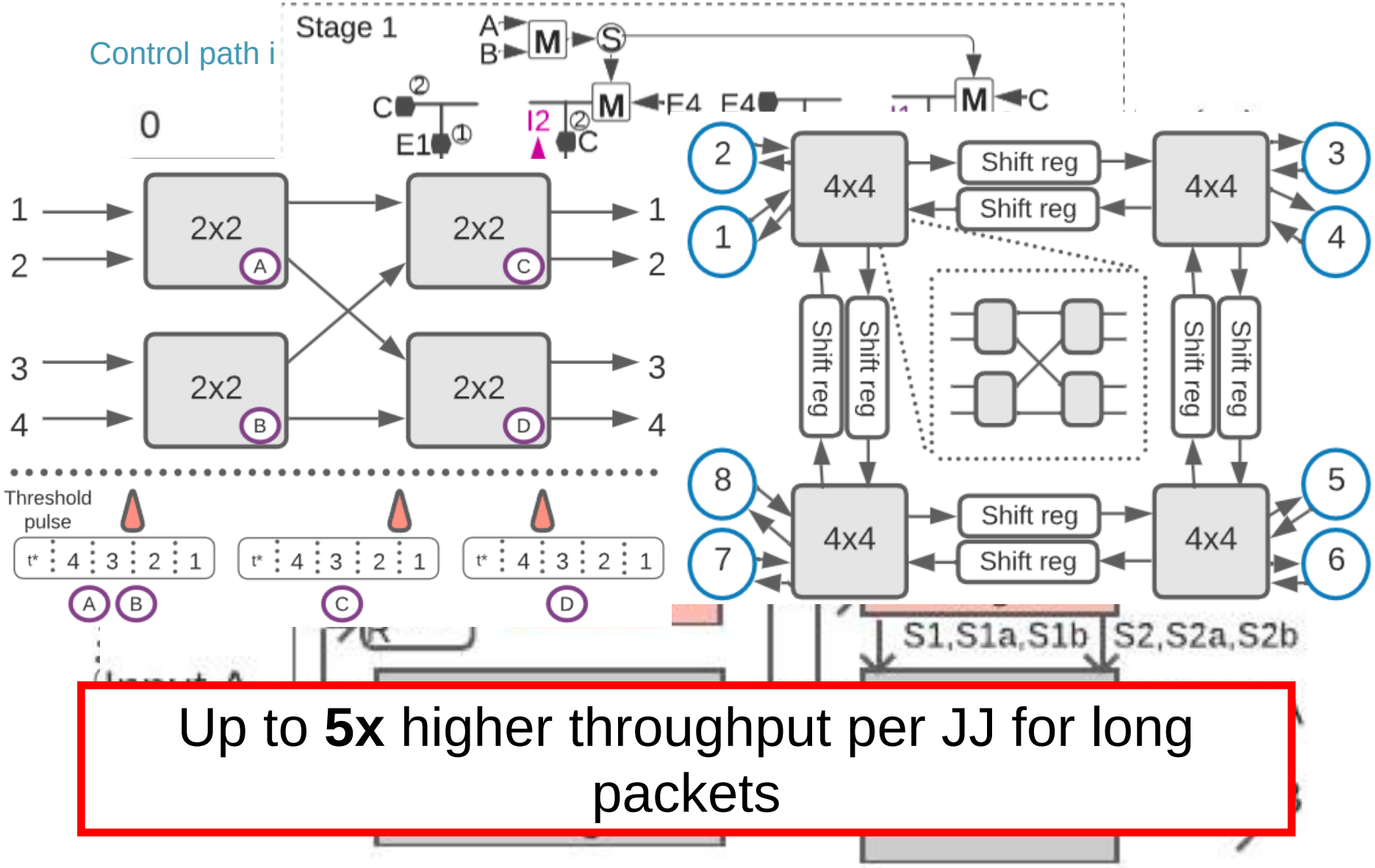
Dynamic energy to traverse wires small

Buffers are expensive (in area and power)

On-chip networks should adapt to novel compute models and be area efficient

D Lyles et al., “PaST-NoC: A Packet-Switched Superconducting Temporal NoC”, IEEE Tran Appl Sup 2023

PaST-NoC: A RL Packet-Switched On-Chip Network

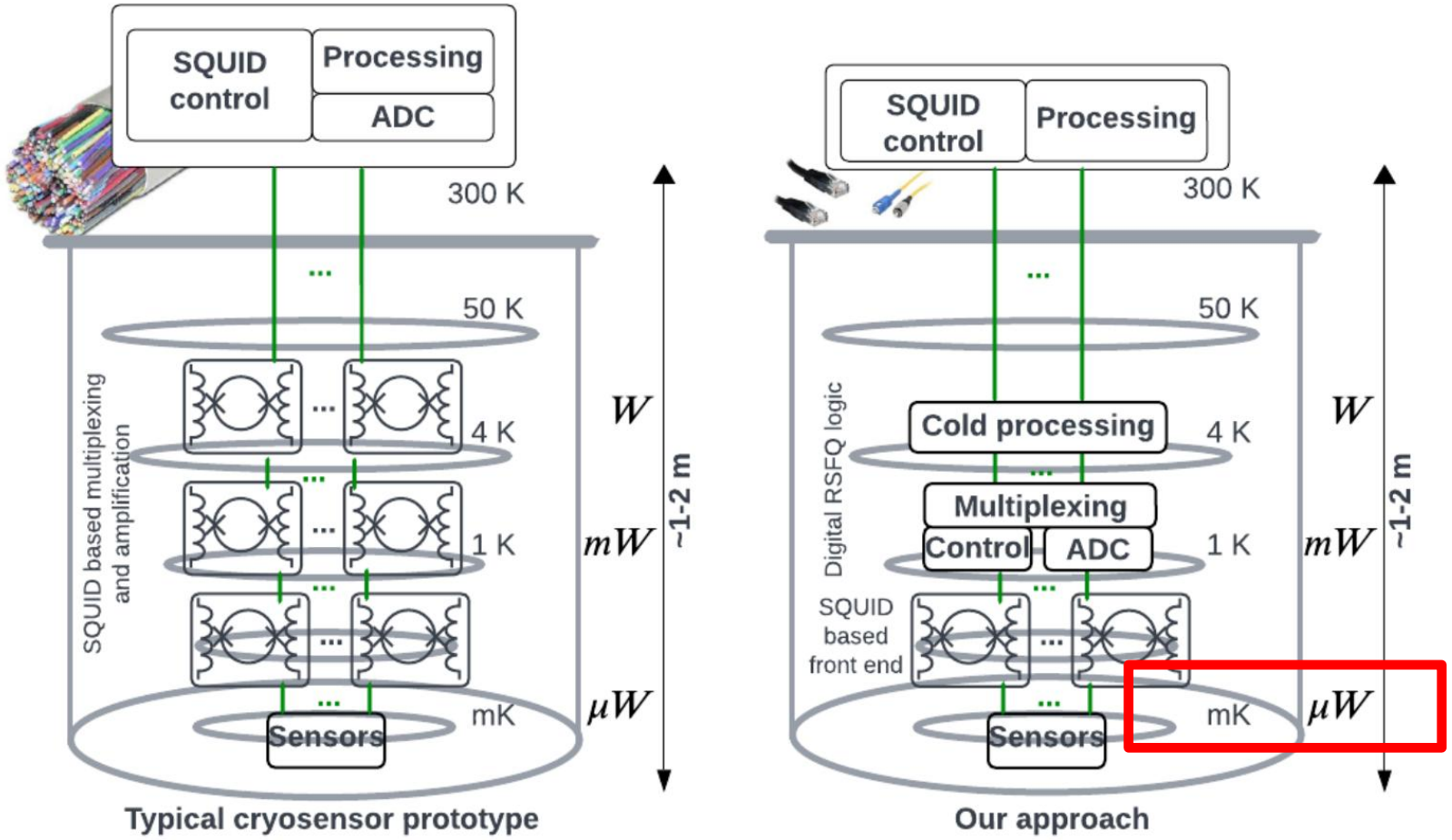


Up to **5x** higher throughput per JJ for long packets

What's Next?

Challenges at Lower Temperatures

At a sizeable application scale



https://github.com/andreamurillomtz/decimation_filter.git

Power Limits

Simple 16-bit adder. Results with qPalace. 50% activity factor in every input

10mW active power

Unit:uW	power consumption	percentage
Total power consumption:	8188.75	100
Static power consumption:	5240.64	64.00
Register cell power consumption:	1575.13	19.24
Combinational cell power consumption:	335.65	4.10
Clock tree power consumption:	1037.33	12.67

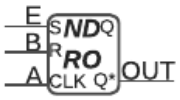
STA (ignore paths about PI & PO)	Corner	Nominal
Minimum workable clock period (ps)	40.1	40.1
Maximum workable clock frequency (GHz)	24.9	24.9

~1000x gap with µW target

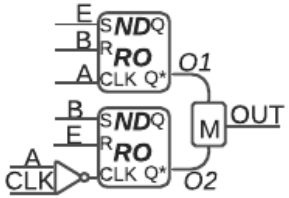
How to reduce power?

- ERSFQ/eSFQ (static power)
- Fewer gates (smaller circuit)
- Lower activity factor
- Lower clock frequency

U-SFQ's Power Requirements

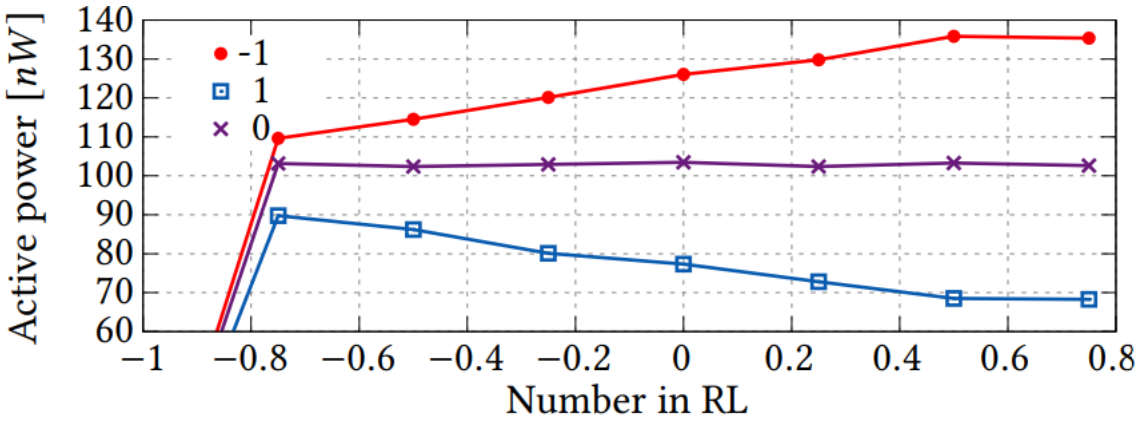


Unipolar SFQ multiplier



Bipolar SFQ multiplier

Power depends on numerical value of inputs. Higher numbers -> more pulses



Active power consumption for the bipolar multiplier, using three different pulse stream frequencies representing the numbers −1, 1, and 0. We vary the RL input from -1 to 1.

32-input 8-bit U-SFQ dot-product unit

Component	Active $[mW]$	Passive $[mW]$
Multiplier	9×10^{-5}	0.05
Balancer	17×10^{-5}	0.1
DPU w/o cooling	84×10^{-4}	4.8

RSFQ. RL and pulse train inputs set to half the maximum value

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

Noise / Variability

From external factors, thermal noise, device variability, etc

- SFQ devices are usually biased around 70% of their switching voltage
- If noise levels are high, fluctuations in voltage can cause JJs to produce erroneous pulses
- Noisy grounds and power supplies can cause issues with device performance
- Worse at 300mK compared to 4K

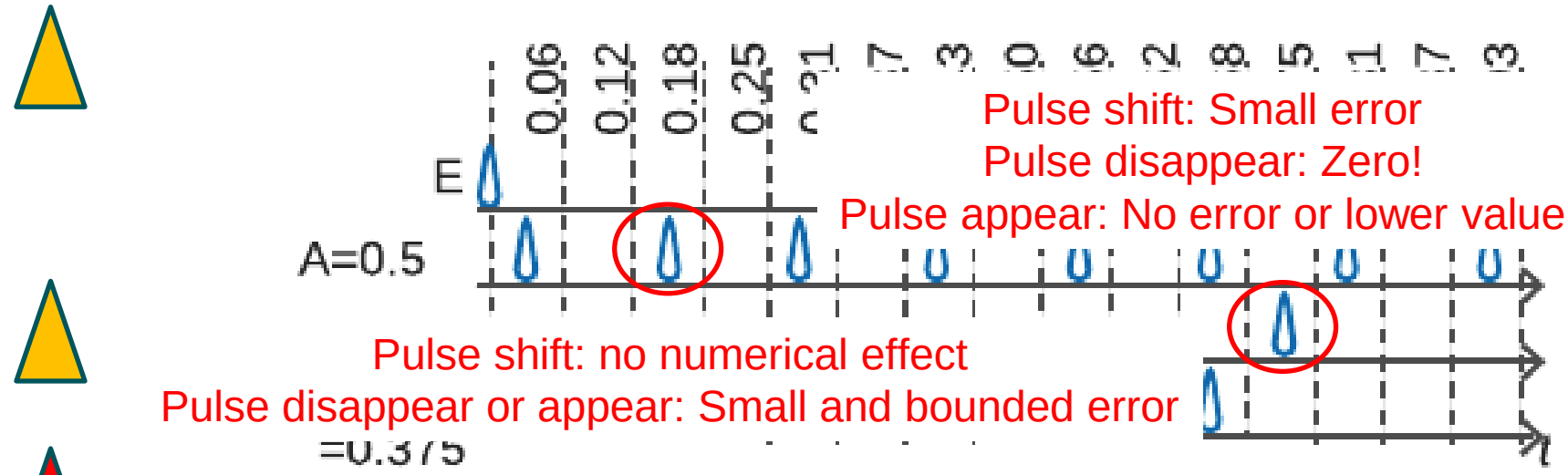
- **Can compute models accept device errors but bound the numerical error?**
- We can use a solid noise model and resulting simulation models

How Do Compute Models Affect The Impact of Errors?

Predictability of errors matters too, not just its numerical impact

- For binary, pulse trains, and race logic:
  - How does an erroneous pulse appearing affect the represented value?
  - How about a pulse disappearing?
  - How about a pulse shifting?

Binary



Erroneous pulse

Conclusion

- Compute models affect efficiency and reliability
- Lets support tools and methods to keep exploring until we settle on a logic family
  - Or we may never settle on a single winner
- What we are designing for matters a lot:
  - HPC applications, quantum, cryosensors, etc.


?
A