This paper describes objective technical results and analysis. Any subjective views or opinions that might be expressed in the paper do not necessarily represent the views of the U.S. Department of Energy or the United States Government.

SAND2022-16340C

COINFLIPS: Co-designed Improved Neural Foundations Leveraging Inherent Physics Stochasticity

Suma George Cardwell

Principal Member of Technical Staff

Sandia National Laboratories

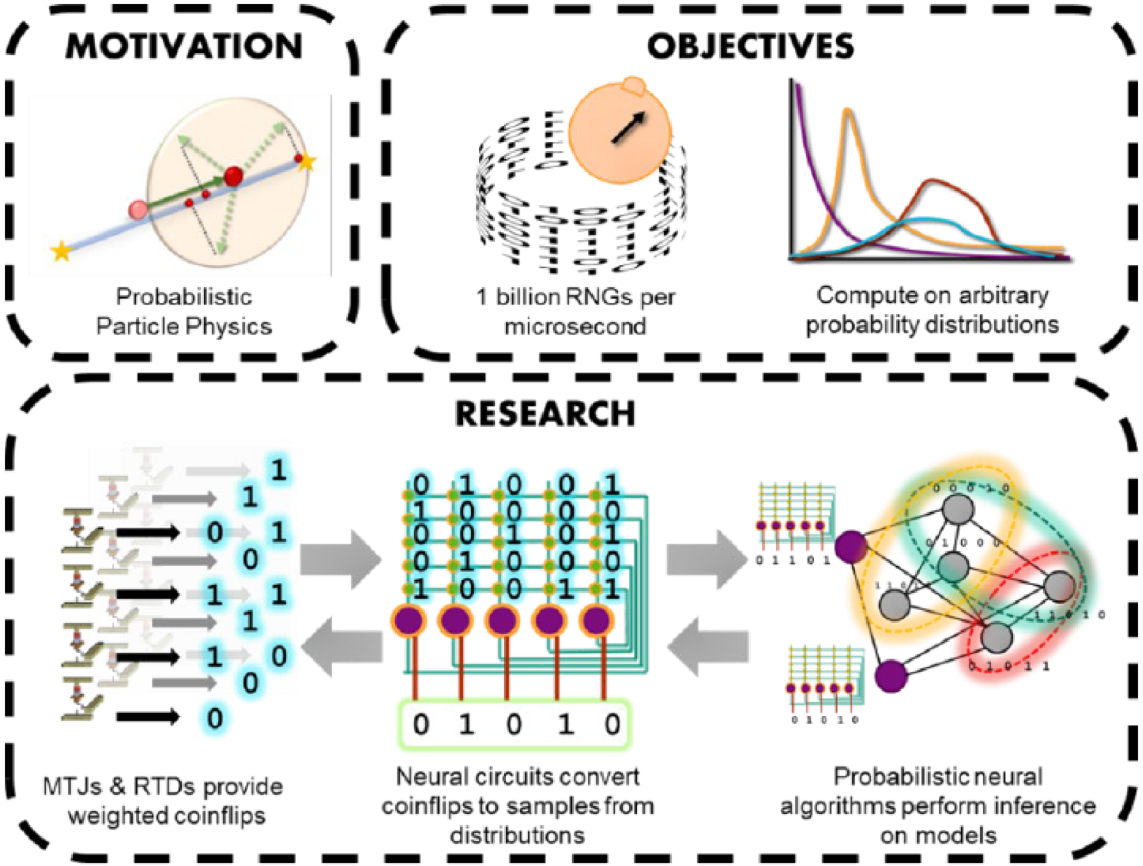
COINFLIPS

MEMRISYS 2022

November 30th, 2022

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

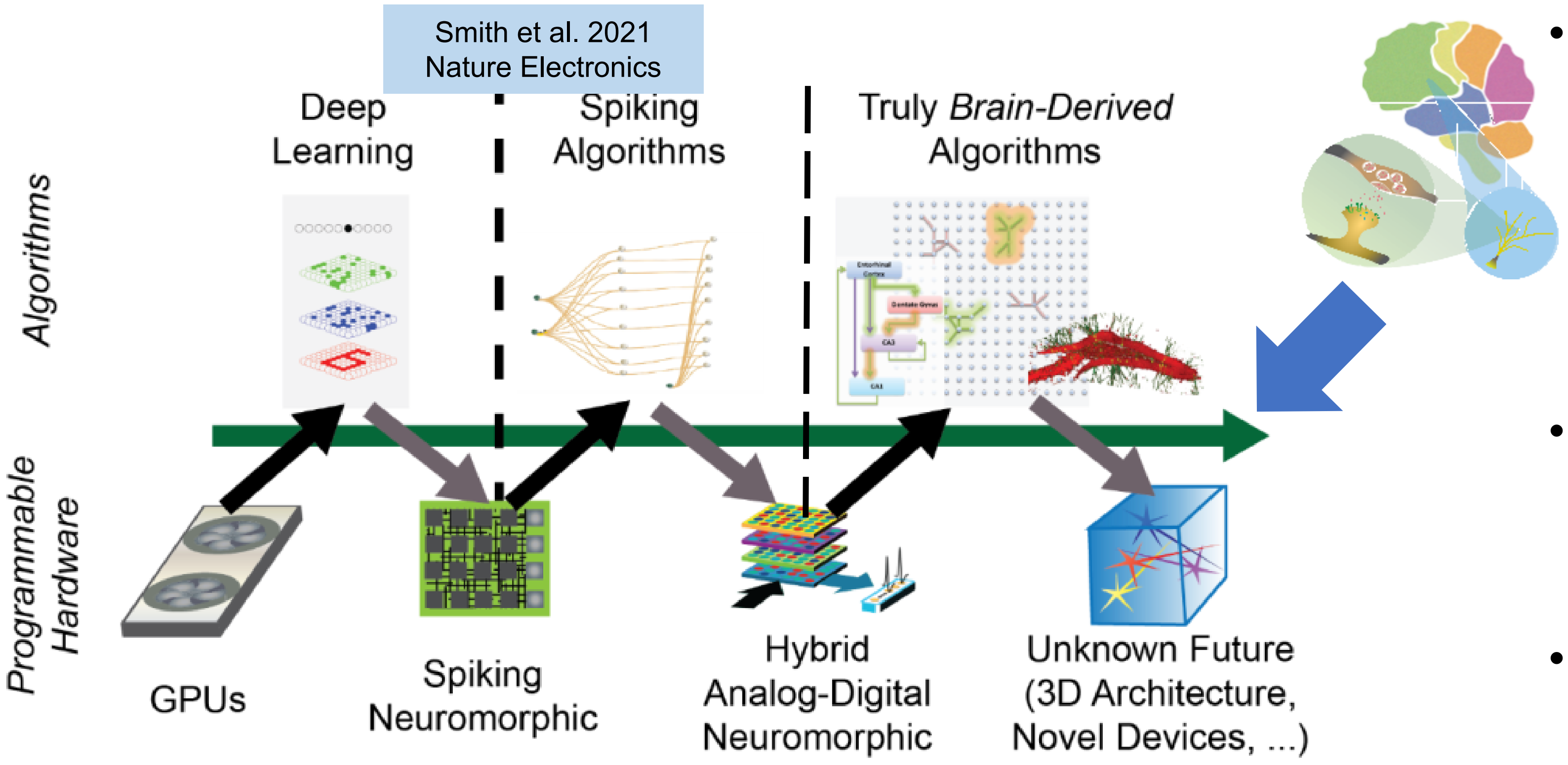
COINFLIPS: Co-designed Improved Neural Foundations Leveraging Inherent Physics Stochasticity


MOTIVATION
Probabilistic Particle Physics
OBJECTIVES
1 billion RNGs per microsecond
Compute on arbitrary probability distributions
RESEARCH
MTJs & RTDs provide weighted coinflips
Neural circuits convert coinflips to samples from distributions
Probabilistic neural algorithms perform inference on models

- Current approaches in microelectronics strive to eliminate any unpredictable behavior across the stack.
- As demand for more compute increases, relying on classical computing to scale while meeting energy constraints is untenable.
- **The key scientific goal of COINFLIPS is to fully leverage stochasticity in computing by exploiting the underlying physics of emerging random number generator (RNG) devices to build probabilistic neural architectures.**

Probabilistic Neural Computing

Inject ubiquitous stochasticity into existing neuromorphic technologies



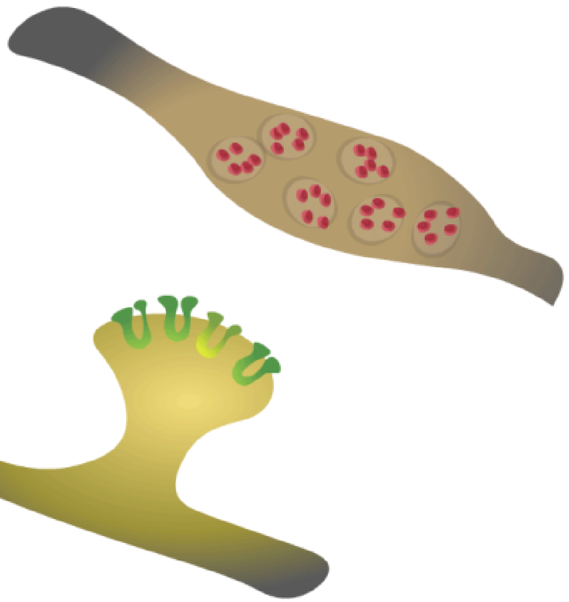
Smith et al. 2021
Nature Electronics

Algorithms
Deep Learning
Spiking Algorithms
Truly *Brain-Derived* Algorithms
Programmable Hardware
GPUs
Spiking Neuromorphic
Hybrid Analog-Digital Neuromorphic
Unknown Future (3D Architecture, Novel Devices, ...)

- Neuromorphic computing is an emerging paradigm that promises to alleviate the challenges faced by current classical computing approaches by emulating key computational principles from the brain.
- Probabilistic neural computation (a novel combination of probabilistic computing and neuromorphic computing.
- COINFLIPS will leverage stochasticity in computing, by making stochasticity ubiquitous and make it useful.

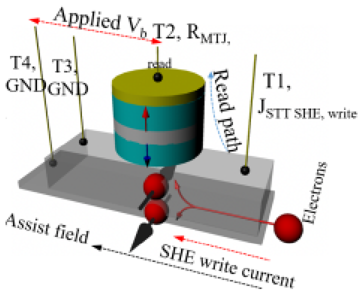
Neural algorithms can leverage RNGs in parallel to provide added capabilities to probabilistic algorithms while leveraging the energy and time advantages of neuromorphic parallelism.

STOCHASTICITY AS A FEATURE NOT A BUG

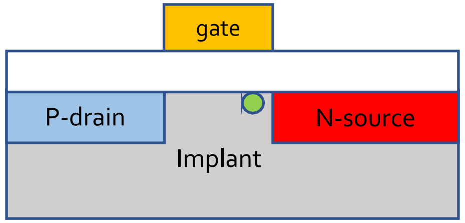


Neuroscientists have observed that stochasticity at the synapse and circuit scales allows for both synaptic development and circuit functional dynamics, phenomena that are crucial for higher-level cognitive functions

Magnetic Tunnel Junction



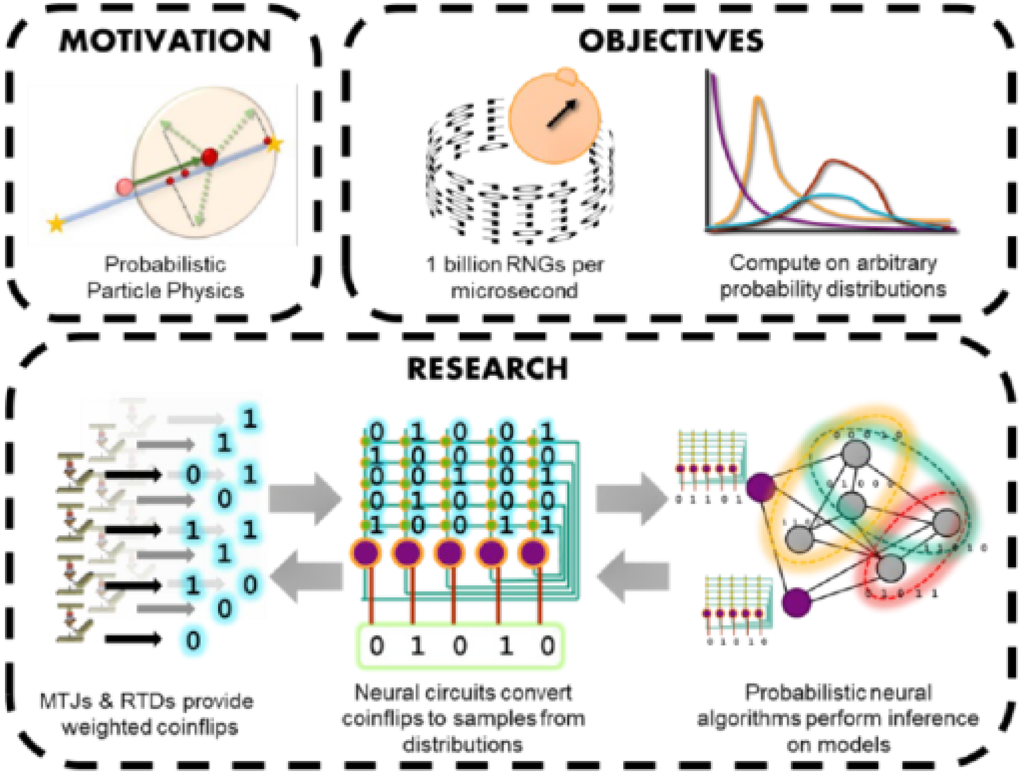
Tunnel Diode



What probabilistic logic elements that leverage the intrinsic physical properties of devices can be developed to provide more sophisticated probabilistic behaviors?

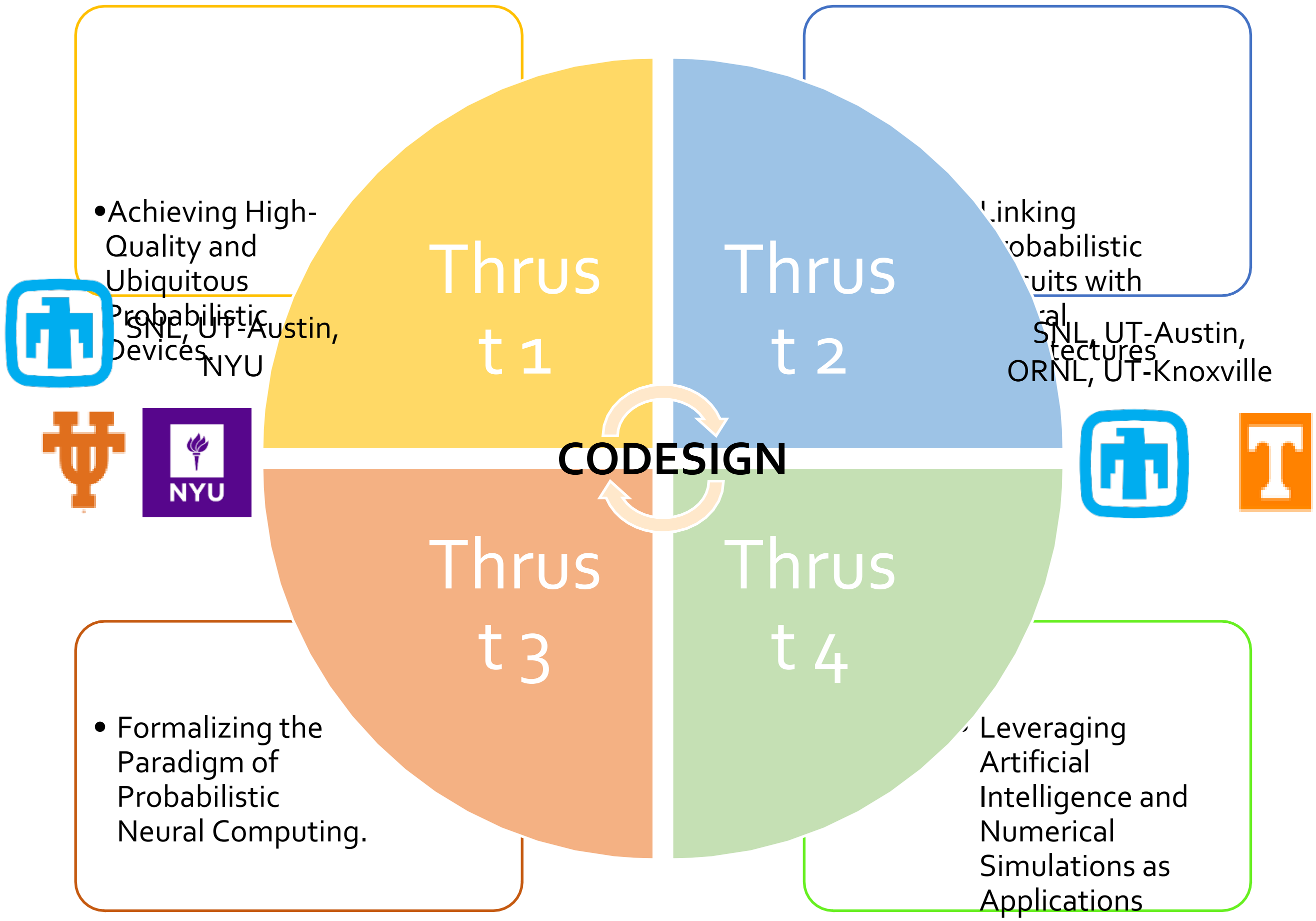
COINFLIPS Overview

Lead PI: Brad Aimone (SNL)


MOTIVATION
Probabilistic Particle Physics
OBJECTIVES
1 billion RNGs per microsecond
Compute on arbitrary probability distributions
RESEARCH
MTJs & RTDs provide weighted coinflips
Neural circuits convert coinflips to samples from distributions
Probabilistic neural algorithms perform inference on models

Collaborators: NYU, ORNL, Temple University, UT-Austin and UT-Knoxville

COINFLIPS



DEVICES

- Achieving High-Quality and Ubiquitous Probabilistic Devices

SNL, UT-Austin, NYU

CIRCUIT & ARCHITECTURE

.inking
·obabilistic
·uits with
·al
·tectures

SNL, UT-Austin, ORNL, UT-Knoxville

- Formalizing the Paradigm of Probabilistic Neural Computing.

- Leveraging Artificial Intelligence and Numerical Simulations as Applications

ALGORITHMS & THEORY

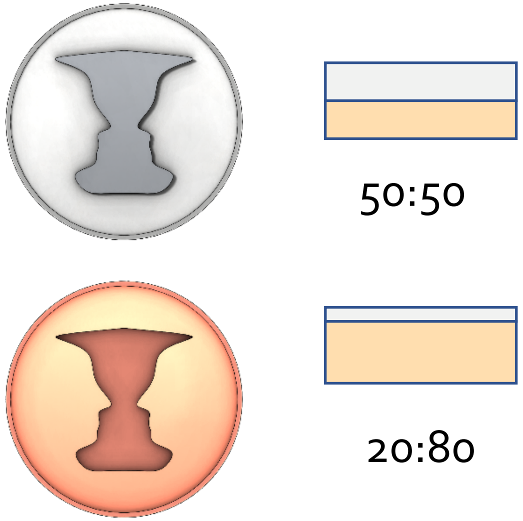
SNL, UT-Austin, ORNL

APPLICATIONS

SNL, Temple University

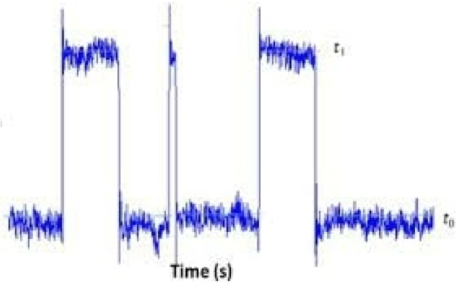
DEVICES: Tunable RNG Magnetic Tunnel Junctions & Tunnel Diodes

Tunable random number generator

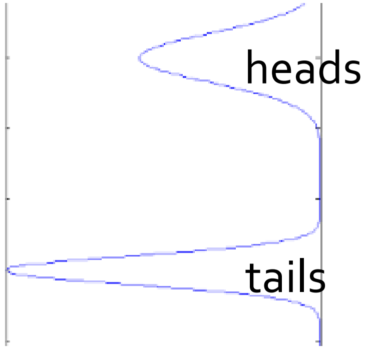


Why did we pick the devices we picked?

Large signals

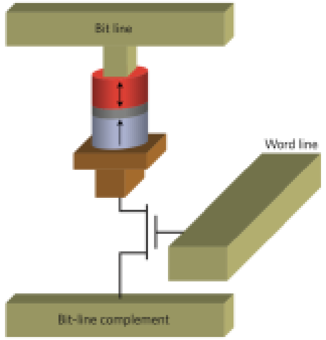


Tunable

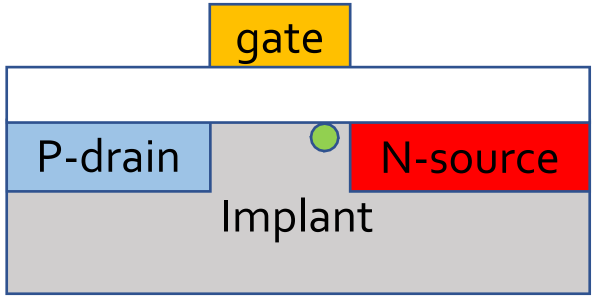


Integration

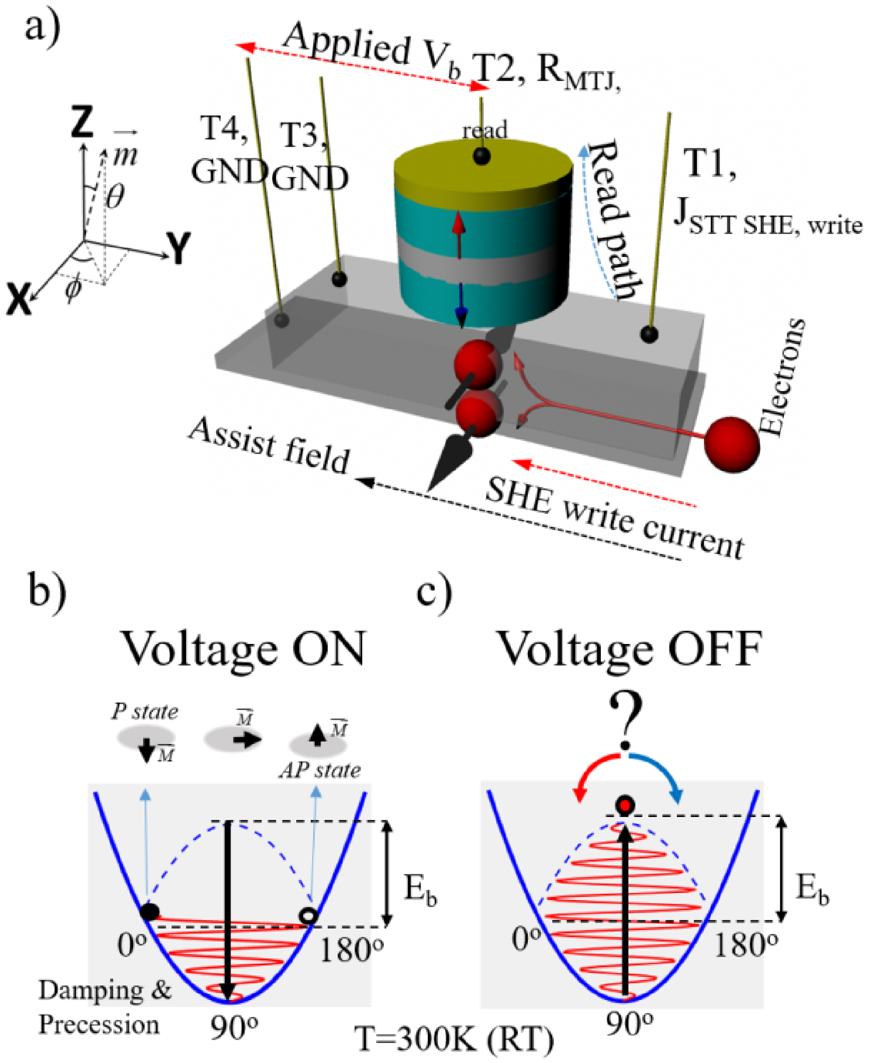
I. Magnetic Tunnel Junction



II. Tunnel diode



MAGNETIC TUNNEL JUNCTIONS



- MTJ consists of an insulating tunnel barrier between two thin ferromagnetic layers.

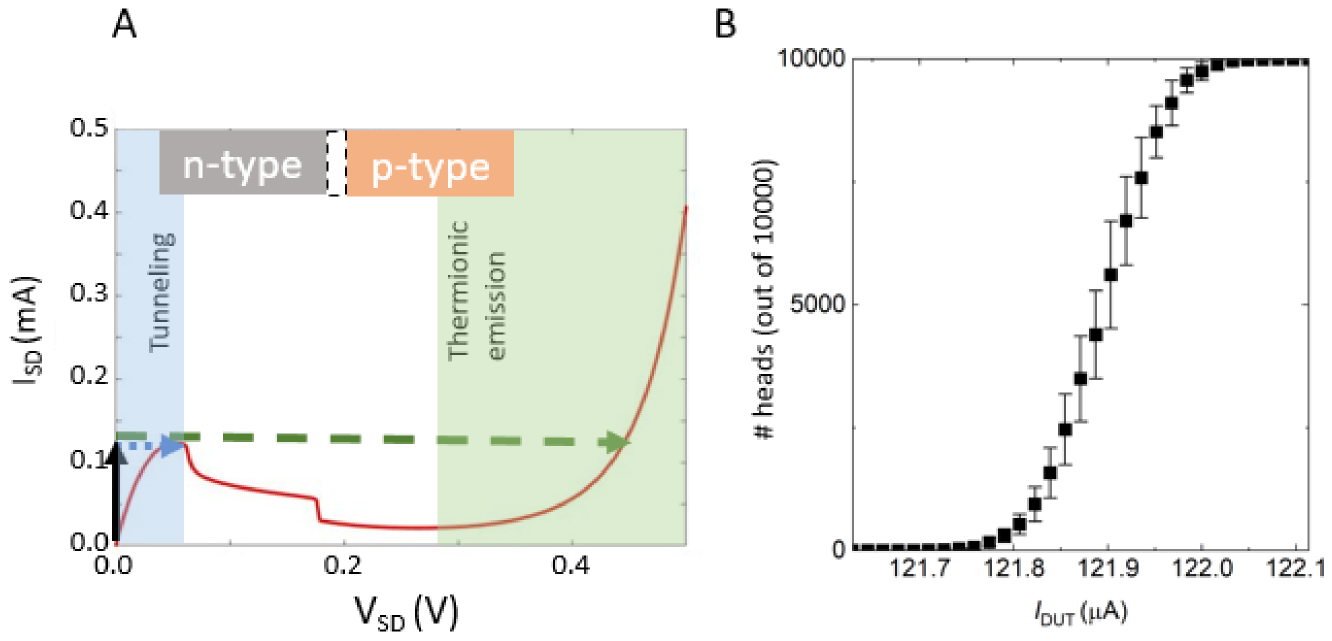
- High or low resistant state depending on orientation of magnetization of ferromagnetic layers, P (parallel) and anti-parallel.

- Devices Tested
  - MTJ-SHE (Spin Hall Effect)
  - MTJ-VCMA (Voltage Controlled anisotropy)

- Applications in memory, probability-bit device applications [Camsari et al. 2019] etc.

Cardwell et al., ICRC 2022
Kwon et al., Submitted JxCDC 2022

TUNNEL DIODES

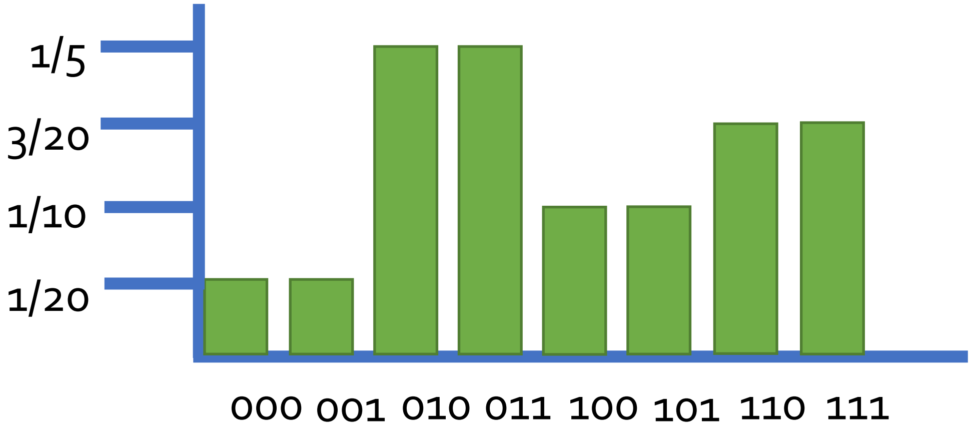


Tunnel Diode Characteristics

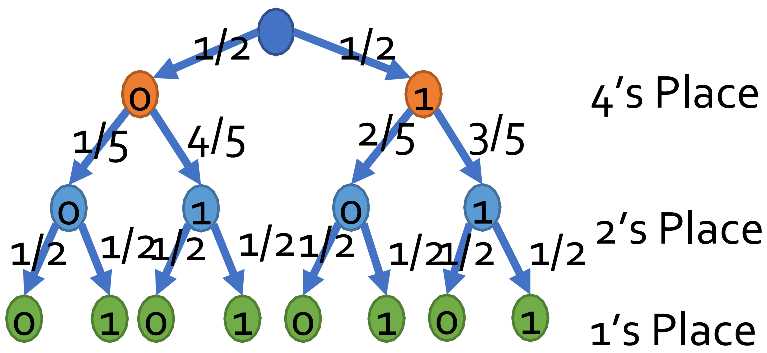
- The tunnel diode (TD) has historically been used in high-speed analog applications, and is a great candidate for a practical nanoscale random number generator.
- TD consists of a strongly n-doped and p-doped junction,
and conducts either by tunneling through or by thermionic emission over the narrow depletion region.
- Ease of CMOS integration, both scaled and high-speed devices have been demonstrated in literature.

COINFLIPS: RNG CIRCUITS

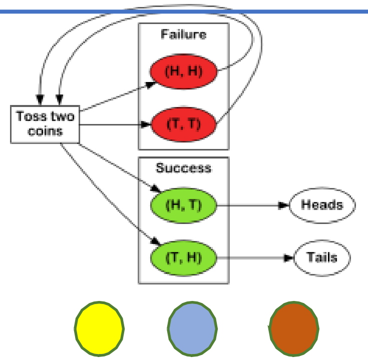
Say we have a given distribution and want a sample from it. There are many ways to approach the same problem



Tree-approach	Sampling -approach	P-bits -approach



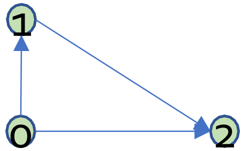
- Tune weight on coinflips
- More devices, but need not have precise tunability. Only 5 coinflips for a sample.
- (2^n-1) devices. E.g.: (31 bits for 5 bit-precision)



Von-Neumann approach to get a fair coin

Then use several 'fair' coinflips to produce weighted coinflip.

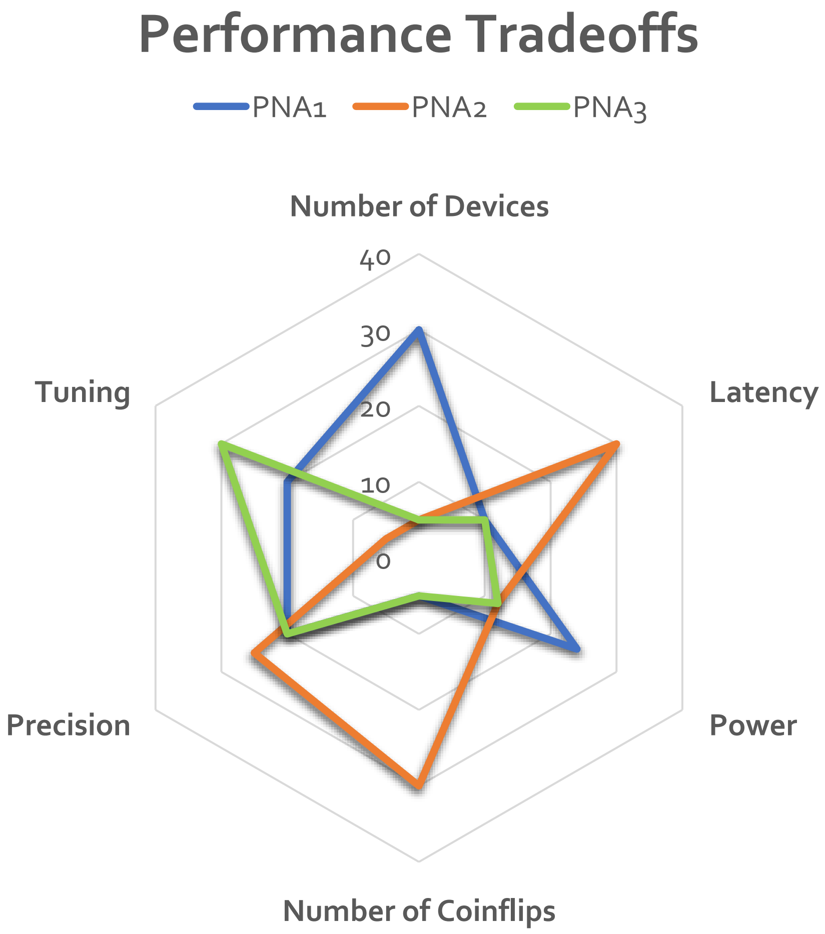
- Fewest number of bits (3 bits for 3-bit precision)
- Need to sample coin many times for fair outcome.
- Coin has unknown weight. Tunability not an issue.
- Worst case could require infinitely many coinflips.



AI-enhanced approach ?

- Fewest bits
- Costly Tunability

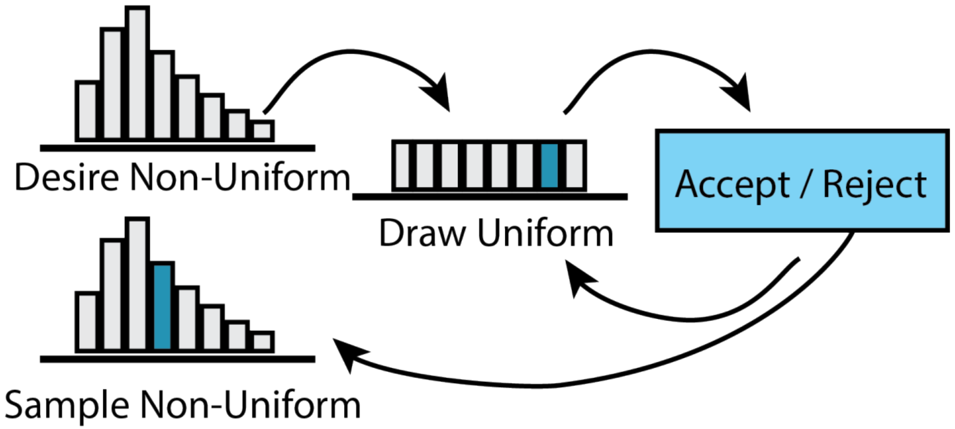
PERFORMANCE TRADEFOFFS

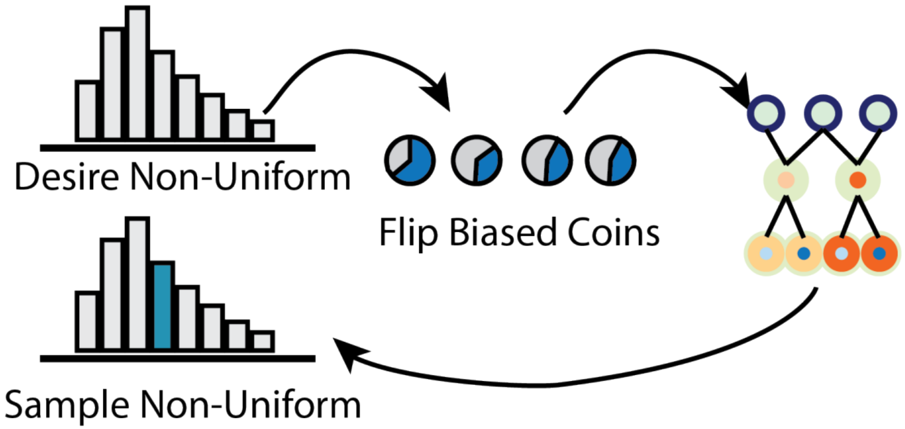


Different tradeoffs for different thrusts

- Tunability/precision/latency/energy
- Different applications will impact design choices
- Different tasks will prioritize different domains within the algorithm landscape, thus different devices, circuits and architectures.

RANDOM NUMBER GENERATION



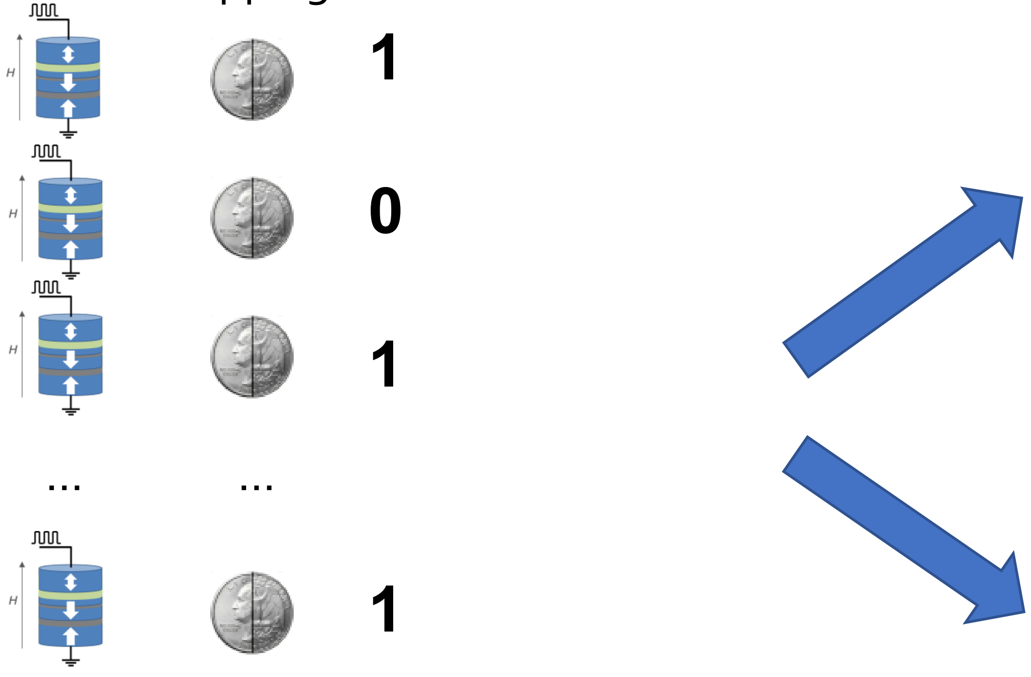


- In scientific computing applications today, uniform random numbers need to be converted to distributions of interest through expensive rejection sampling or related techniques.

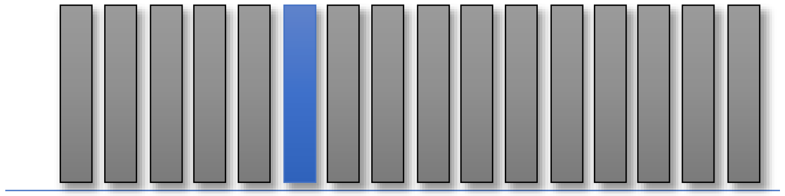
- By using device stochasticity and codesigned circuits, we can directly convert biased coin flips to our distributions of interest, avoiding repeated sampling loops.

AI-Guided Design of Neuromorphic Circuits: Arbitrary Distribution

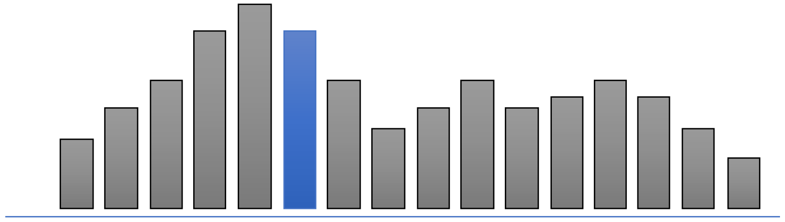
Many devices
flipping at one time

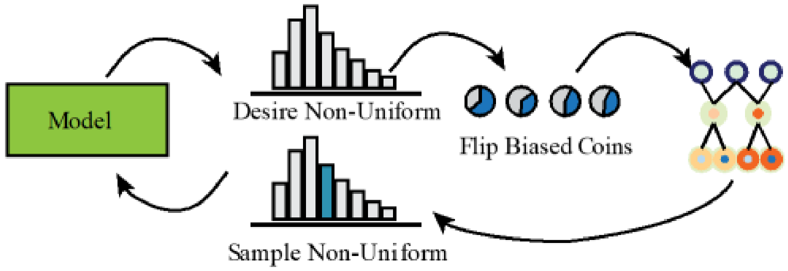


Fair coins for *uniform* distribution

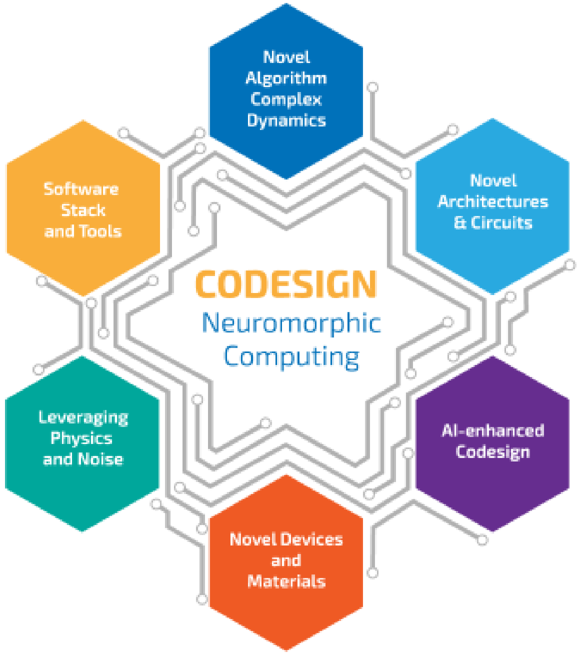


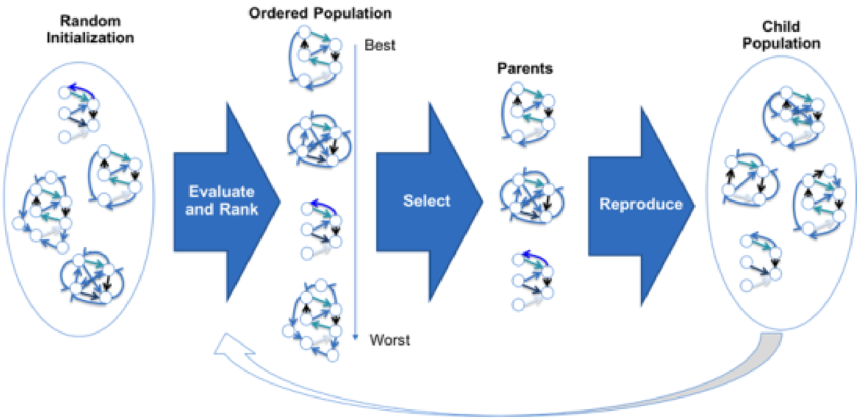
Biased coins for *non-uniform* distribution?





AI-GUIDED CODESIGN OF PROBABILSITIC CIRCUITS

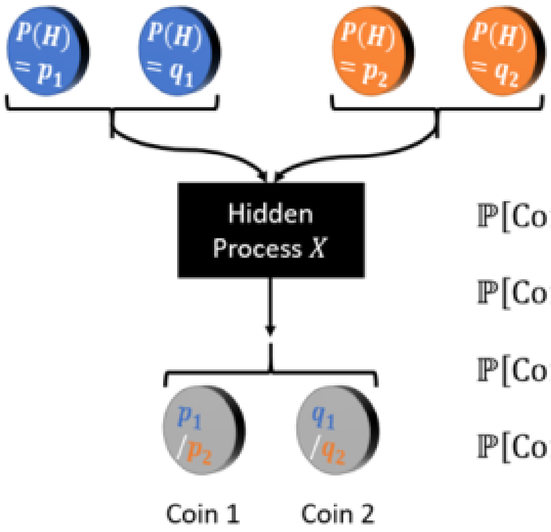

Novel Algorithm Complex Dynamics
Novel Architectures & Circuits
AI-enhanced Codesign
Novel Devices and Materials
Leveraging Physics and Noise
Software Stack and Tools
CODESIGN
Neuromorphic Computing



Evolutionary Optimization for Neuromorphic Systems (EONS)
Schuman et al. , 2020

- We leveraged evolutionary algorithms for circuit design and optimization
  - **LEAP** (Library of Evolutionary Algorithms in Python)
  - **EONS** (Evolutionary Optimization for Neuromorphic System)
- We used abstracted device models for TD and MTJ to capture functionality and energy usage.

PROBABILISTIC MIXING: HIDDEN DEPENDENCE MODEL



$$\mathbb{P}[\text{Coin } 1 = H \text{ and Coin } 2 = H] = \frac{1}{2}$$

$$\mathbb{P}[\text{Coin } 1 = H \text{ and Coin } 2 = T] = \frac{1}{6}$$

$$\mathbb{P}[\text{Coin } 1 = T \text{ and Coin } 2 = H] = \frac{1}{6}$$

$$\mathbb{P}[\text{Coin } 1 = T \text{ and Coin } 2 = H] = \frac{1}{6}$$

$$w p_1 q_1 + (1-w) p_2 q_2 = \frac{1}{2}$$

$$w p_1 (1-q_1) + (1-w) p_2 (1-q_2) = \frac{1}{6}$$

$$w (1-p_1) q_1 + (1-w)(1-p_2) q_2 = \frac{1}{6}$$

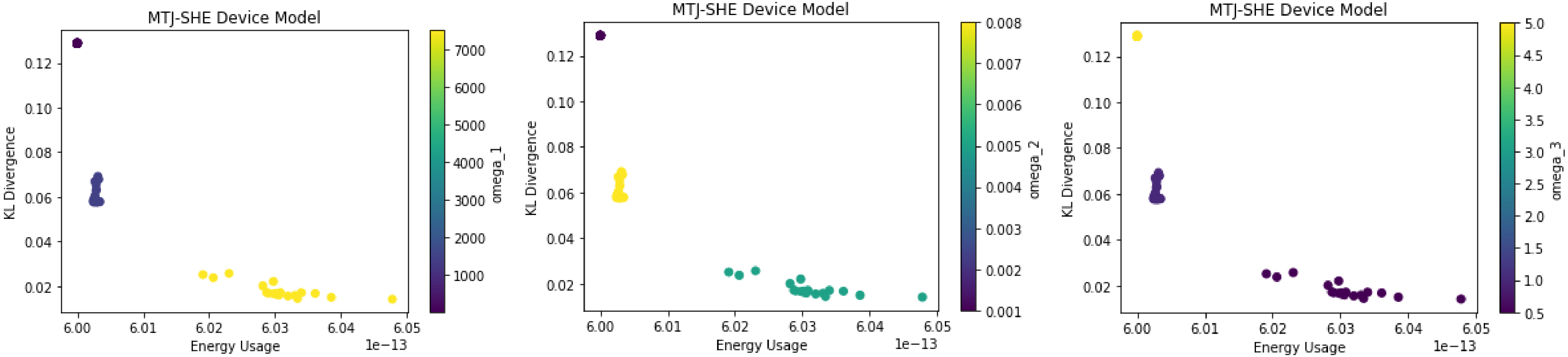
$$w (1-p_1)(1-q_1) + (1-w)(1-p_2)(1-q_2) = \frac{1}{6}$$

- In the “Hidden Dependence” model, there is a hidden process, stochastic or deterministic, that controls the probability of heads among a collection of coins.
- The hidden process chooses which set of coins is flipped. The observer only sees a single set, the effective flipping set.

Multi-Objective Optimization

FITNESS FUNCTION

$$f(w,p1,p2,q1,q2) \quad = \omega_1 \underbrace{KL(p_1,p_2,q_1,q_2)}_{\text{Kullback-Leibler Divergence}} + \omega_2\Big(\sum_{i=1}^{2}|p_i \underbrace{-0.5| + \sum_{i=1}^{2}|q_i \quad -0.5|}_{\text{Difference of weight from a fair coin}}\Big) + \omega_3 EN \underbrace{(p_1,p_2,q_1,q_2)}_{\text{Energy}}$$



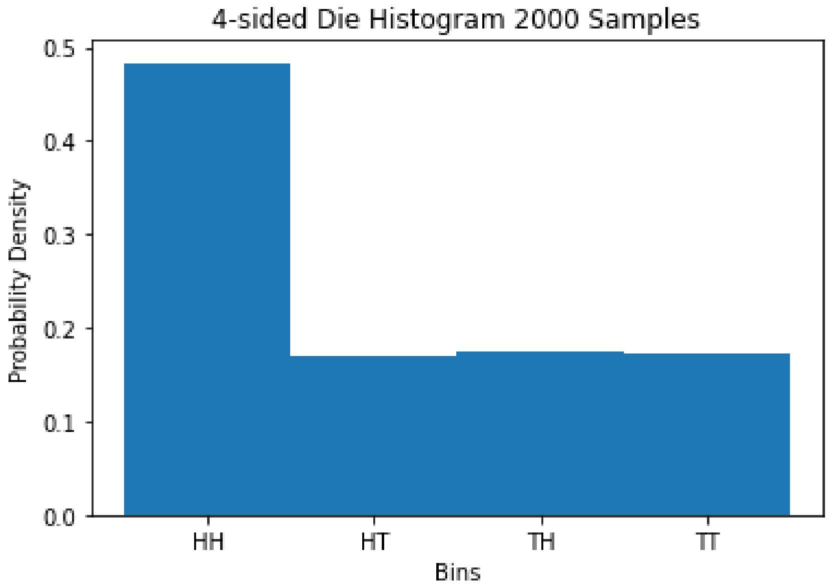
Cardwell et al.,
ICRC 2022

Multi-objective optimization of weights ω1, ω2, ω3 for optimal KL divergence and energy usage of MTJ-SHE devices

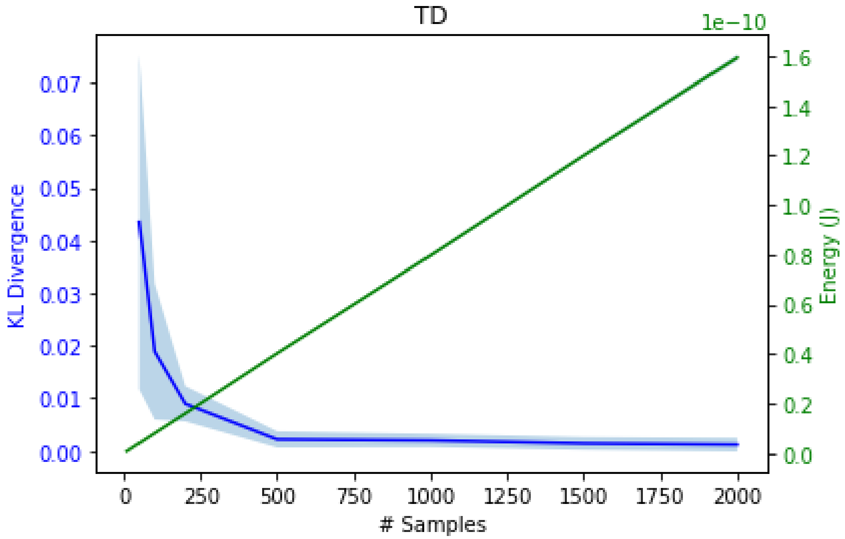
TD RNG CIRCUIT

- We selected the device configuration for lowest KL-divergence value through optimization over 1000 generations in LEAP.
- Increasing the number of samples lowers the KL-divergence from the desired probability distribution.
- However, more samples come at the cost of increased energy consumption

Empirical distribution using TD for 2000 samples in a single run



KL divergence and energy usage vs. number of samples for the given distribution



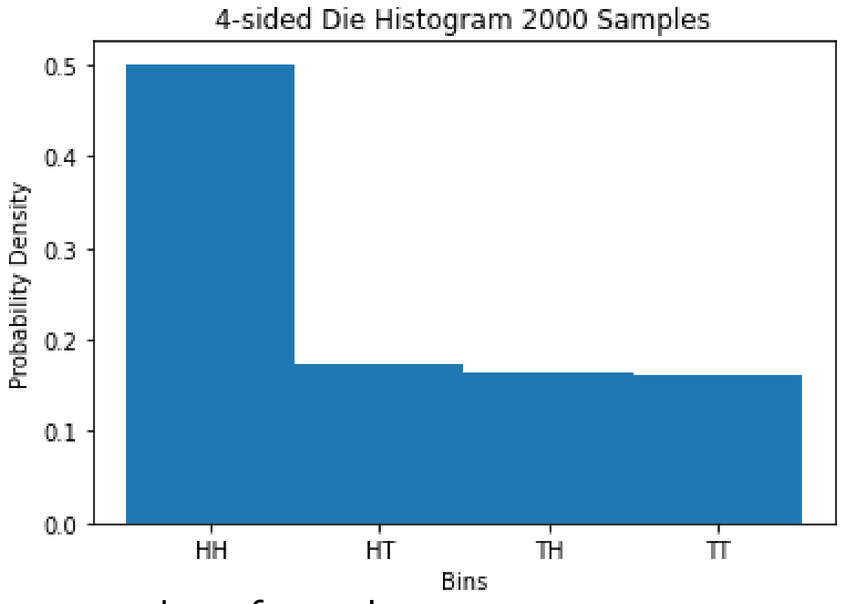
Cardwell et al., ICRC 2022

MTJ RNG CIRCUIT

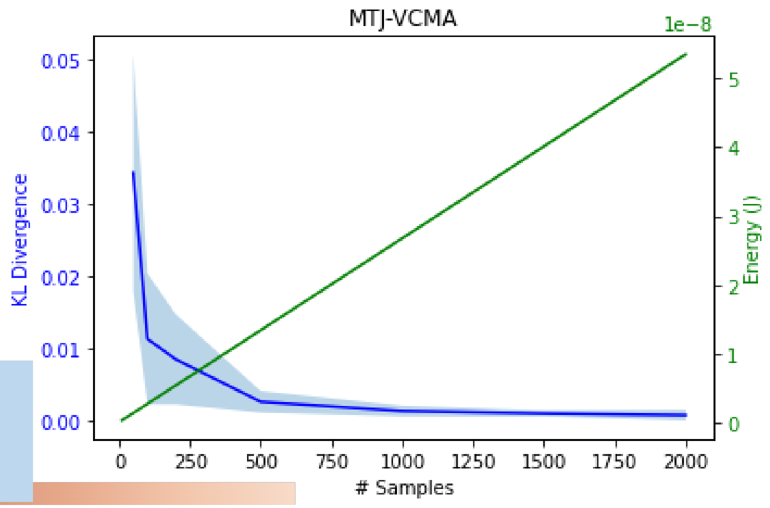
- We selected the device configuration for lowest KL-divergence value through optimization over 1000 generations in LEAP.

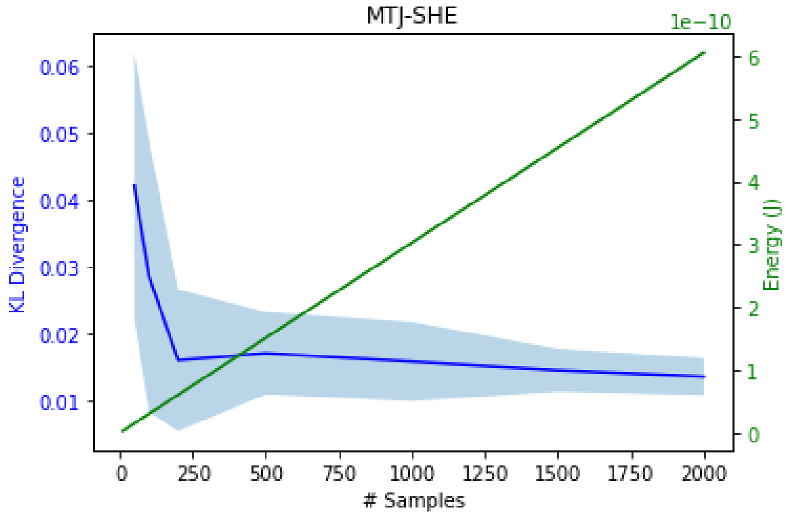
COINFLIPS

Empirical distribution using MTJ-VCMA for 2000 samples in a single run



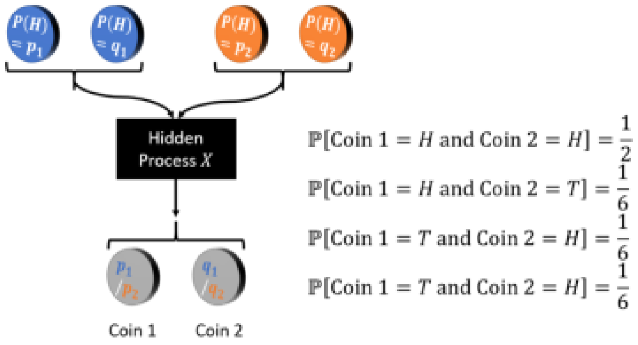
KL divergence and energy usage vs. number of samples for the given distribution



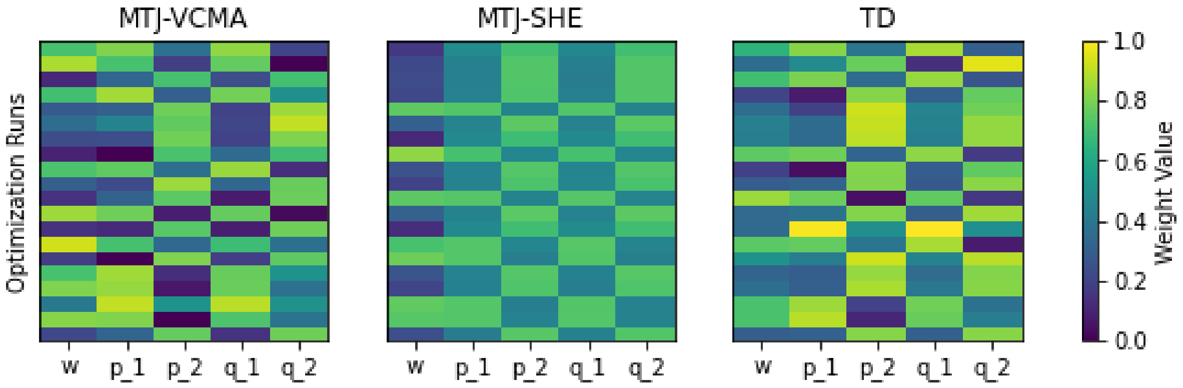


Cardwell et al., ICRC 2022

AI-GUIDED CODESIGN OF PROBABILSITIC CIRCUITS



$P(H) = p_1$
$P(H) = q_1$
$P(H) = p_2$
$P(H) = q_2$
Hidden Process X
p_1 / p_2
q_1 / q_2
Coin 1
Coin 2
$\mathbb{P}[\text{Coin } 1 = H \text{ and Coin } 2 = H] = \frac{1}{2}$
$\mathbb{P}[\text{Coin } 1 = H \text{ and Coin } 2 = T] = \frac{1}{6}$
$\mathbb{P}[\text{Coin } 1 = T \text{ and Coin } 2 = H] = \frac{1}{6}$
$\mathbb{P}[\text{Coin } 1 = T \text{ and Coin } 2 = H] = \frac{1}{6}$

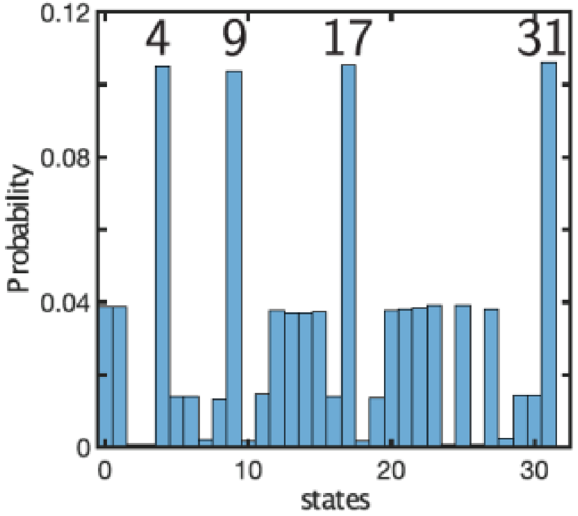


Optimized weight values for each device over twenty optimization runs using LEAP.

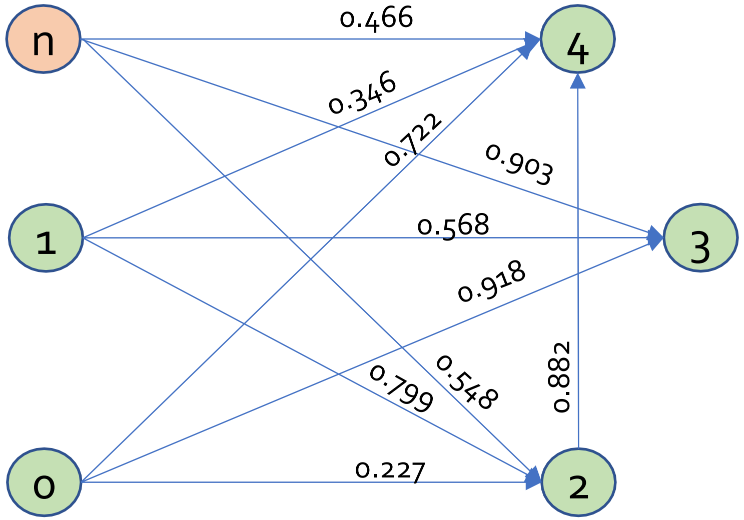
- 20 different sets of weight values that were optimized for each device type for ω1 = 7500, ω2 = 0.005, and ω3 = 0.5.
- Here, we see that the weights are customized for the device's behavior to target the best performance in terms of KL divergence and energy usage.
- One of the challenges in optimizing for both algorithms and devices was appropriately abstracting the device models and algorithmic constraints.
- The functional models developed will also be evolved in time as new device data and research emerges.
- Our framework can accommodate any emerging device type.

COINFLIPS: p-bit RNG CIRCUIT EXAMPLE

- Probability-based netlist building: Initial framework developed and testing is in progress.



Camsari et al., 2019



Network using p-bits developed leveraging EONS

Evolutionary Optimization for Neuromorphic Systems (EONS)
Schuman et al. , 2020

Each node is a p-bit, represented by a MTJ device,
“n”: control bit

MTJ: Magnetic Tunnel Junction
P-bit: Probability-bit

AI-ENHANCED CODESIGN ACROSS SCALES

COINFLIPS

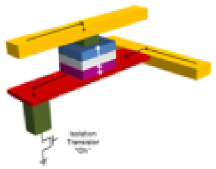
Device Design

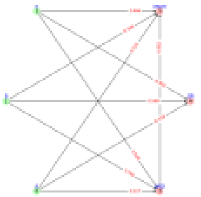
Circuit Design

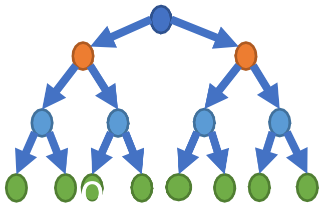
System Design

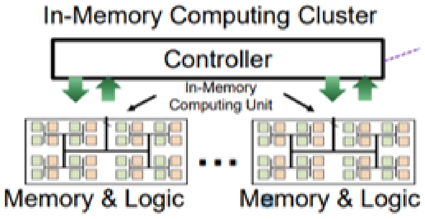
Architecture Design

Algorithm Design

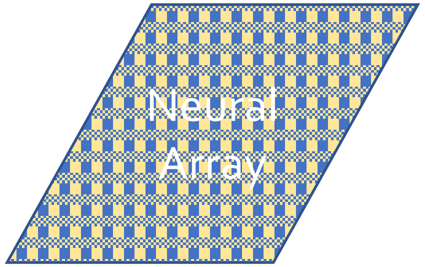


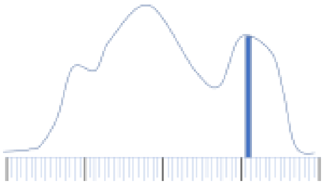


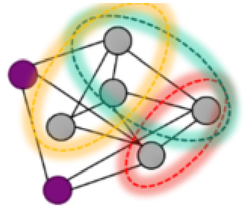




Fan (UCF), 2018

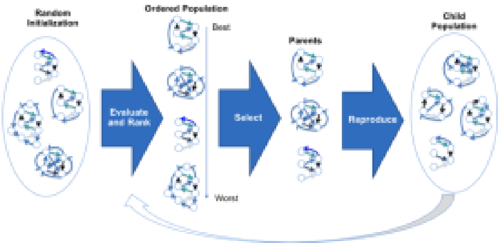




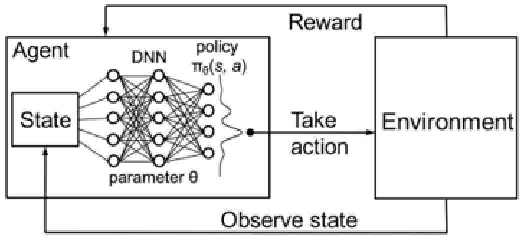


Approach

Can we leverage AI to generate specifications for novel devices?



Evolutionary/RL approaches

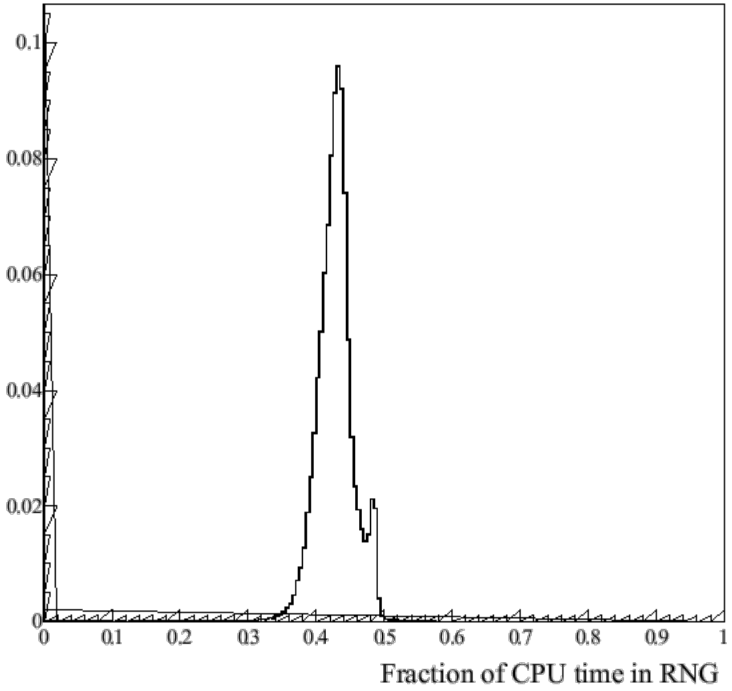


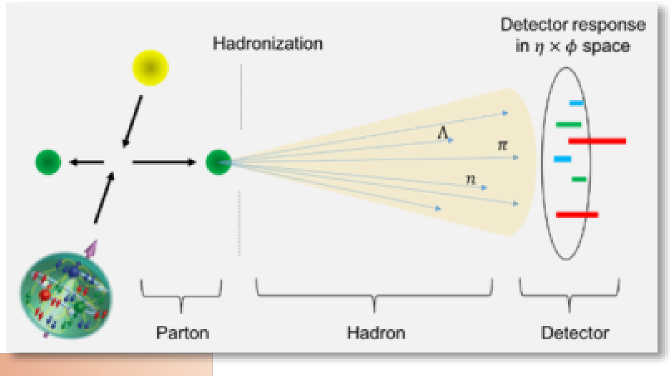
RL approaches

Analytical and cycle-accurate tools, network simulation tools

RL approaches

APPLICATIONS: NUCLEAR PHYSICS SIMULATIONS



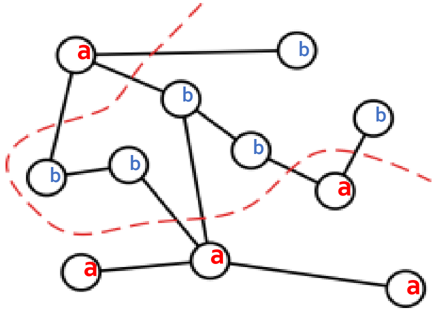


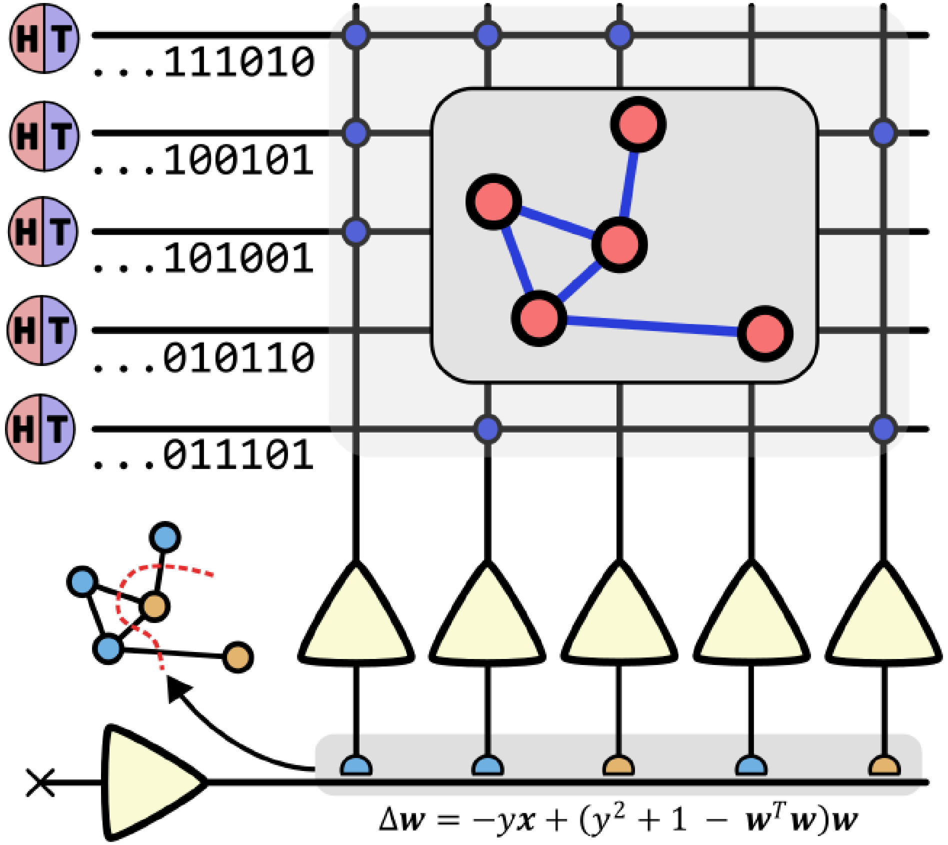
- For a particular collider physics simulation [Pierog et al., Phy Rev. 2022], ~ 270K pseudo- random numbers needed for a single event, with billions of events needing to be simulated.
- CPU time is ~ 40-50% of the total compute time
- Direct random number generation leveraging stochastic devices can promise significant energy savings for such applications

Misra et al., *Advanced Materials 2022*

Random numbers are a limiting computational cost for some nuclear physics applications

APPLICATIONS: MAXCUT

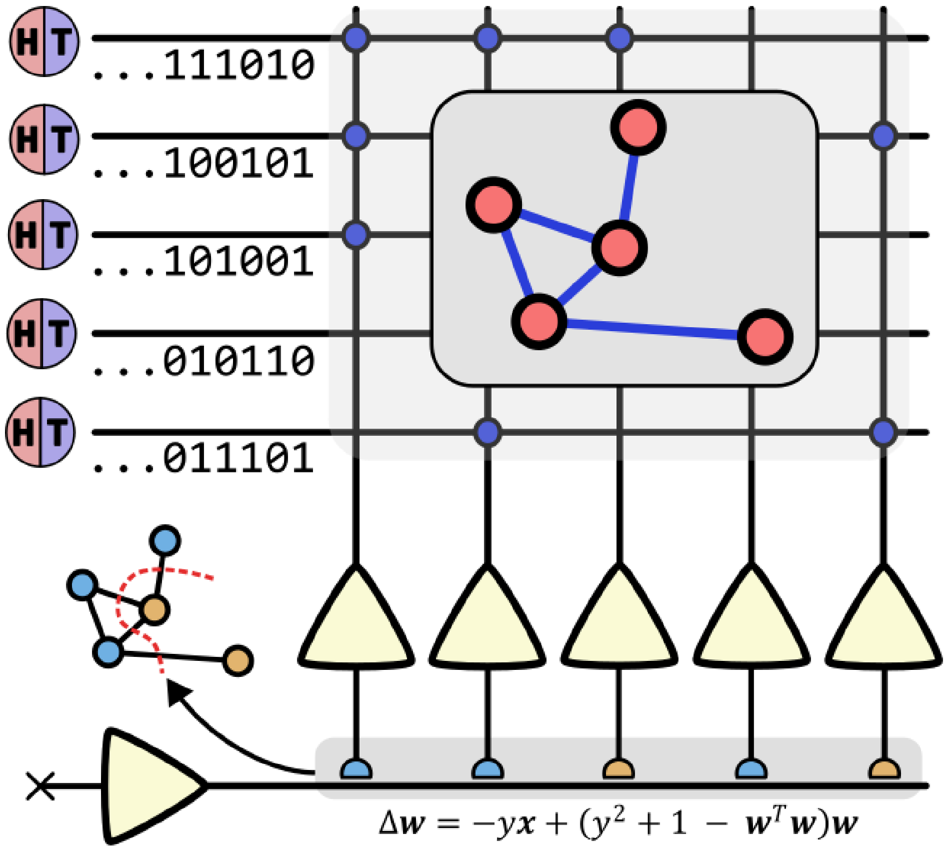


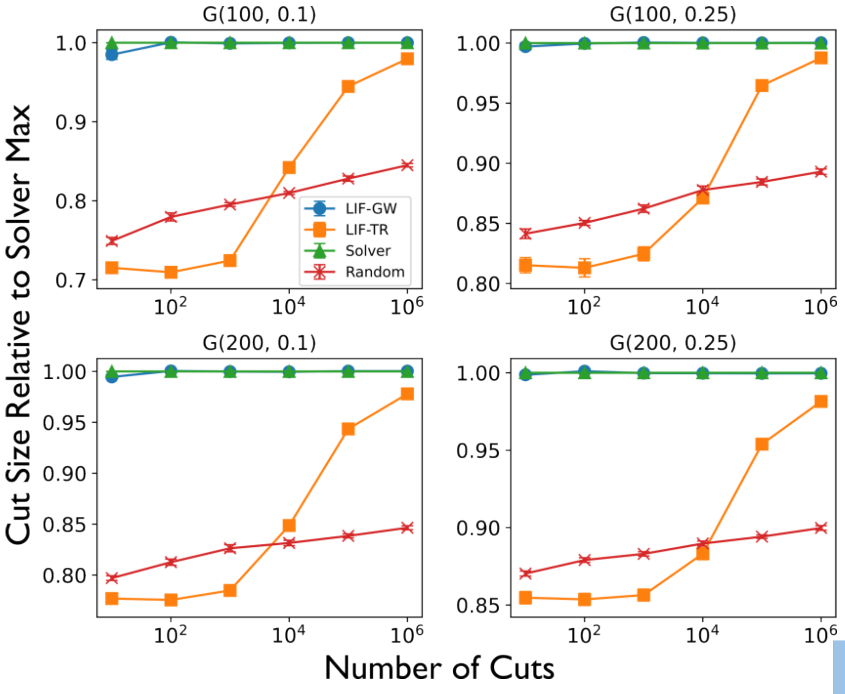


$$\Delta w = -yx + (y^2 + 1 - w^T w)w$$

- MAXCUT Applications to Ising models, VLSI circuit layout design, network design, data analysis, etc.
- Neuromorphic Implementation of Simplified Trevisan and Goemans-Williamson Sampling
- The weight vector evolves with Oja's antihebbian plasticity rule and converges to the minimum eigenvector of the LIF covariance matrix

APPLICATIONS: MAXCUT



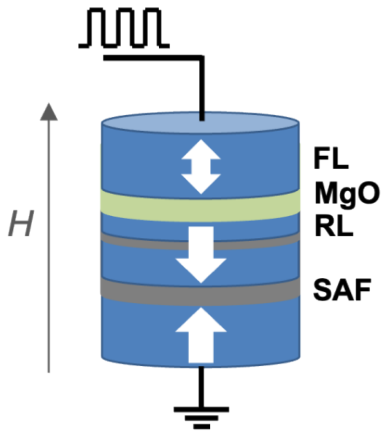


Theilman et al.; Arxiv 2210.02588

- A population of n COINFLIPS devices produces random bits.
- The weight matrix COINFLIPS > LIF is proportional to the graph adjacency matrix. By the central limit theorem, the LIF membrane potentials approximate a gaussian process with covariance determined by the weights
- The weight vector evolves with Oja's antihebbian plasticity rule and converges to the minimum eigenvector of the LIF covariance matrix
- Thresholding the output weight vector generates a graph cut.
- **Circuit generated cuts (orange curve) approach classical solver solutions (green curve)**

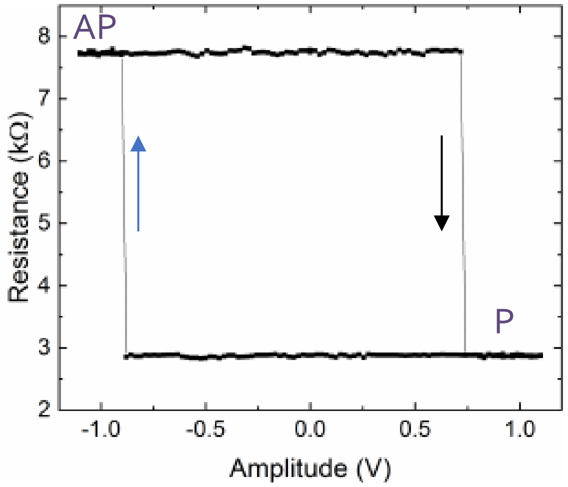
Fair coinflip device example – Magnetic Tunnel Junction (MTJ)

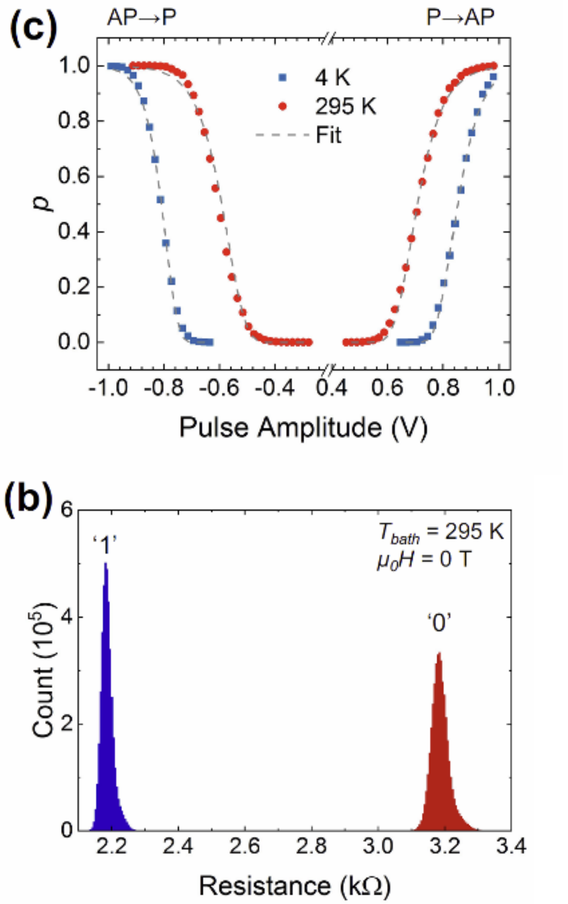
MTJ Coinflip device



40 nm circular pMTJ with CoFeB/W/CoFeB composite free layer

Reset – set metastable state – read

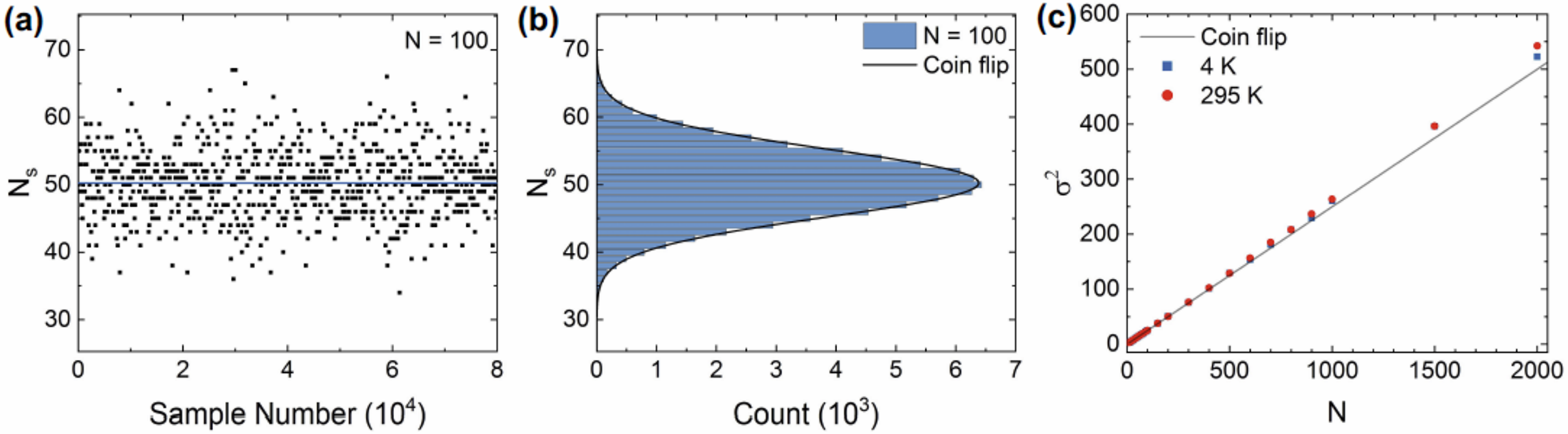




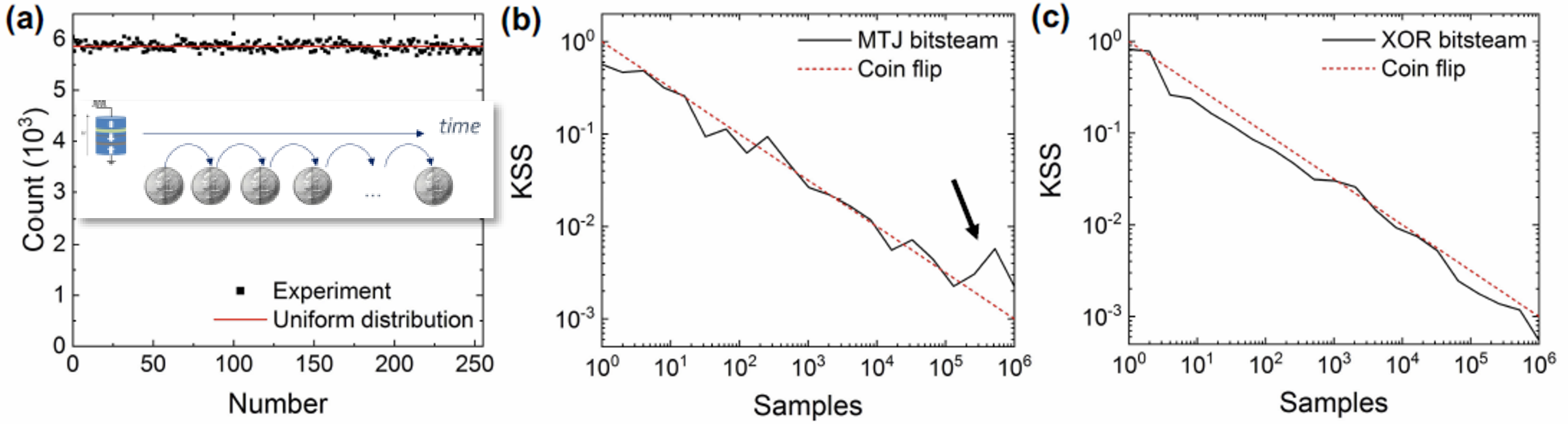
Reim et al., *Submitted arXiv 2209.01480*

Quality of coinflip directly tied to quality of sample

Blocks of 100 random coinflips show expected distribution of random samples



Generating 8-bit (integers from 0 – 255) from coinflips produces good random samples



Reim et al., *Submitted arXiv 2209.01480*

COINFLIPS TEAM

COINFLIPS

- Office of Science Co-Design in Microelectronics program
  - Co-funded through ASCR and BES, participation by NP, HEP, and FES
- COINFLIPS is partnering with a growing number of organizations
  - Sandia National Laboratories: Shashank Misra, Conrad James, Darby Smith, Suma Cardwell, Brad Theilman, Ojas Parekh, Yipu Wang, Chris Allemang, William Severa
  - Andy Kent @ New York University
  - Jean Anne Incorvia @ University of Texas Austin
  - Katie Schuman @ University of Tennessee
  - Prasanna Date @ Oak Ridge National Laboratory
  - Les Bland @ Temple University

U.S. DEPARTMENT OF ENERGY | Office of Science

Sandia National Laboratories

NYU

THE UNIVERSITY OF TEXAS AT AUSTIN

UT

OAK RIDGE National Laboratory

T

References

- "Probabilistic Neural Computing with Stochastic Devices" , Misra, Shashank, Leslie C. Bland, Suma G. Cardwell, Jean Anne C. Incorvia, Conrad D. James, Andrew D. Kent, Catherine D. Schuman, J. Darby Smith, and James B. Aimone. *Advanced Materials* (2022): 2204569.
- "p-bits for probabilistic spin logic", K. Y. Camsari, B. M. Sutton, and S. Datta, Applied Physics Reviews, vol. 6, no. 011305, pp. 1931–9401, 2019
- "Probabilistic Neural Circuits leveraging AI-Enhanced Codesign for Random Number Generation", Suma G. Cardwell, Catherine D. Schuman J. Darby Smith, Karan Patel, Jaesuk Kwon , Samuel Liu , Christopher Allemang, Shashank Misra, Jean Anne Incorvia and James B. Aimone, ICRC 2022
- "Stochastic Neuromorphic Circuits for Solving MAXCUT", Theilman, Bradley H., Yipu Wang, Ojas D. Parekh, William Severa, J. Darby Smith, and James B. Aimone. arXiv preprint arXiv:2210.02588 (2022).
- "Spin hall effect magnetic tunnel junction coinflips" Reim et al., Submitted arXiv 2209.01480
- "Random Bitstream Generation using Voltage-Controlled Magnetic Anisotropy and Spin Orbit Torque Magnetic Tunnel Junctions" Kwon et al., IEEE Journal of Exploratory Solid-State Computational Devices and Circuits, In Review
- "EPOS LHC: Test of collective hadronization with data measured at the CERN Large Hadron Collider." , Pierog, T., Iu Karpenko, Judith Maria Katzy, E. Yatsenko, and Klaus Werner. Physical Review C 92, no. 3 (2015): 034906.