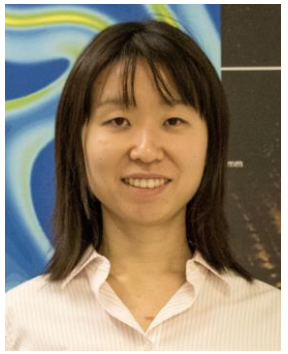
ISCAS 2020
2020 IEEE International Symposium on Circuits and Systems
Virtual, October 10-21, 2020
CAS

A Mott Insulator-Based Oscillator Circuit for Reservoir Computing



Wen Ma[1], Tyler Hennen[2], Martin Lueker-Boden[1], Rick Galbraith[3], Jonas Goode[3], Won Ho Choi[1], Pi-Feng Chiu[1], Jonathan A. J. Rupp[2], Dirk J. Wouters[2], Rainer Waser[2,4], and Daniel Bedau[1]

[1] Western Digital Research, San Jose, CA 95119, USA [2] IWE II, RWTH Aachen University, 52074 Aachen, Germany [3] Western Digital HDD R&D, Rochester, MN 55901, USA [4] Peter Grünberg Institute, Forschungszentrum Jülich GmbH, 52428 Jülich, Germany

2020 IEEE International Symposium on Circuits and Systems
Virtual, October 10-21, 2020

Western Digital

RWTH AACHEN UNIVERSITY

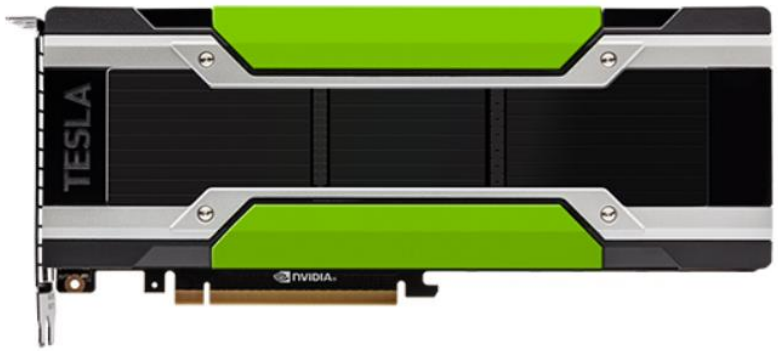
daniel.bedau@wdc.com

Presentation Outline

- Introduction
- Device modeling and oscillator circuit
- Reservoir computing system
  - Spoken digit recognition
  - Handwritten digit recognition (MNIST stroke)
  - HDD channel decoding
- Comparison with other works
- Conclusion

ML has Throughput and Energy Issues

- Inference: fast object detection with YOLO v2, **2 fps** on 8K video with Tesla P100! (Franchetti et al. 2018)
- Training ImageNet takes 1 hour on 256 GPUs (Goyal et al. 2017)
  - On-line training for many applications are not realistic



10k$ GPU
250W

Training a single AI model can emit as much carbon as five cars in their lifetimes

Deep learning has a terrible carbon footprint.

by Karen Hao Jun 6, 2019

The artificial-intelligence industry is often compared to the oil industry: once mined and refined, data, like oil, can be a highly lucrative commodity. Now it seems the metaphor may extend even further. Like its fossil-fuel counterpart, the process of deep learning has an outsize environmental impact.

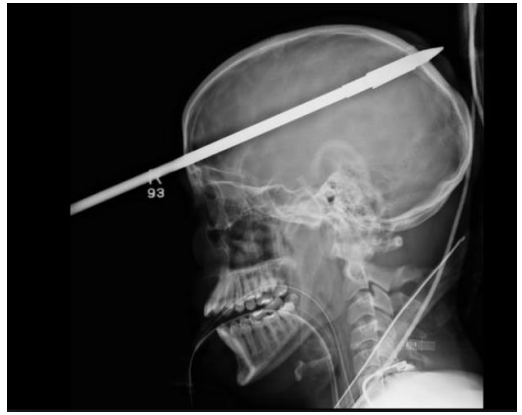
In a new paper, researchers at the University of Massachusetts, Amherst, performed a life cycle assessment for training several common large AI models. They found that the process can emit more than 626,000 pounds of carbon dioxide equivalent—nearly five times the lifetime emissions of the average American car (and that includes manufacture of the car itself).

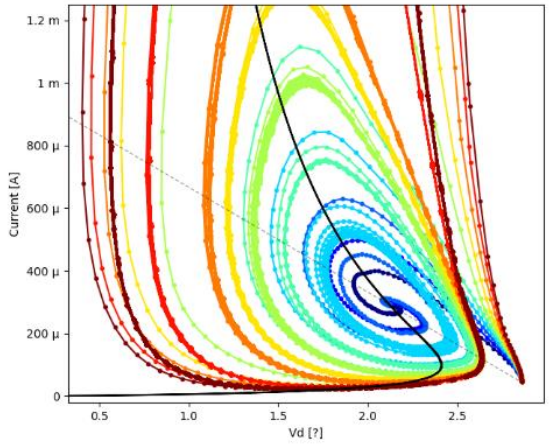
Energy and Policy Considerations for Deep Learning in NLP

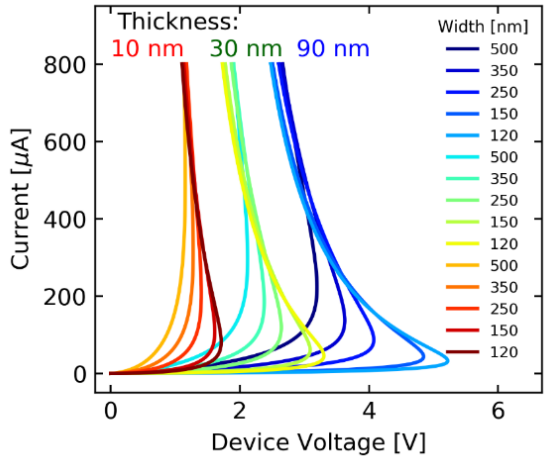
Emma Strubell **Ananya Ganesh** **Andrew McCallum**
College of Information and Computer Sciences
University of Massachusetts Amherst
{strubell, aganesh, mccallum}@cs.umass.edu

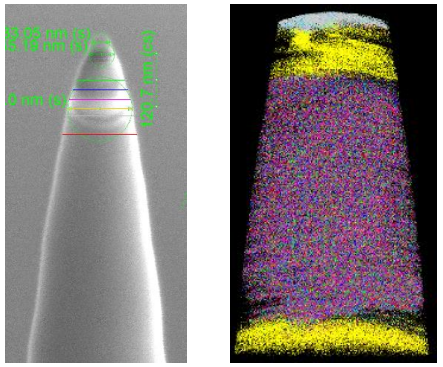
Neuromorphic and Non-CMOS Computation

- Human brain uses 20 W, extremely error tolerant
- Substantial progress in neuromorphic computing (IBM TrueNorth, Intel Loihi)
- New materials are expected to provide far better scaling and lower power consumption than CMOS chips

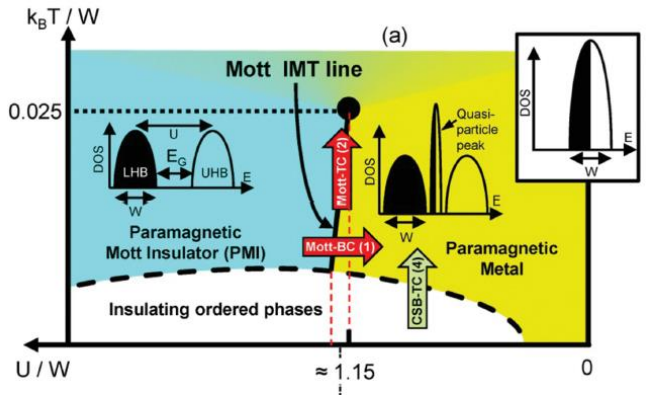




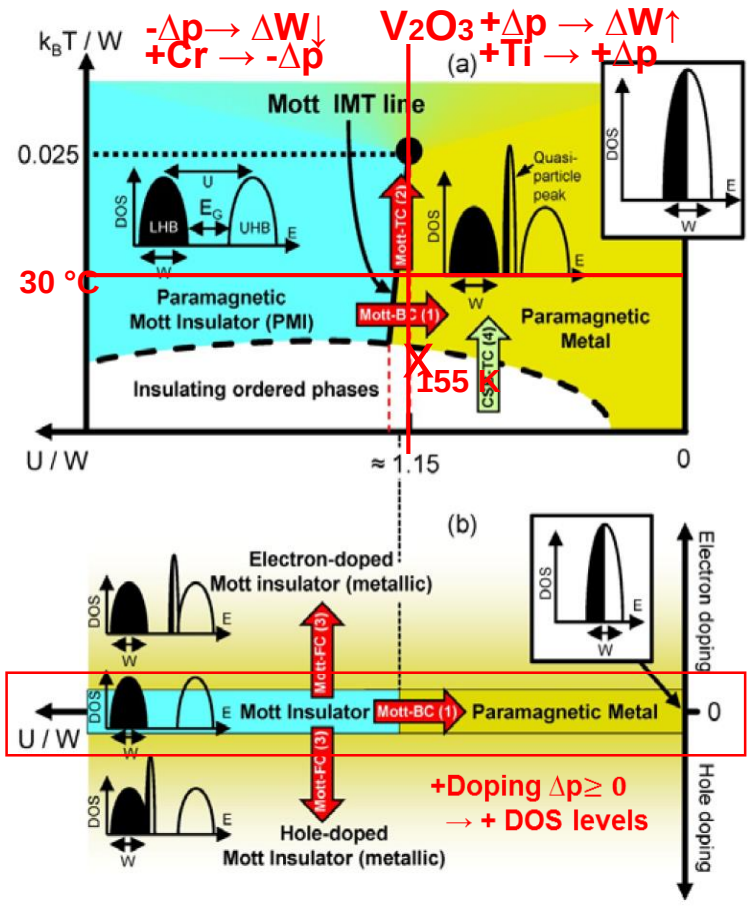




V_2O_3 device atom probe

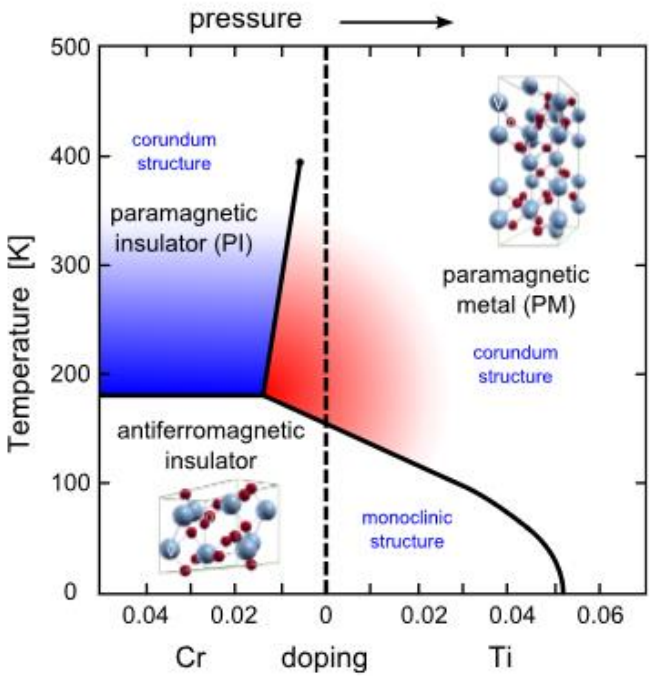


Mott Materials



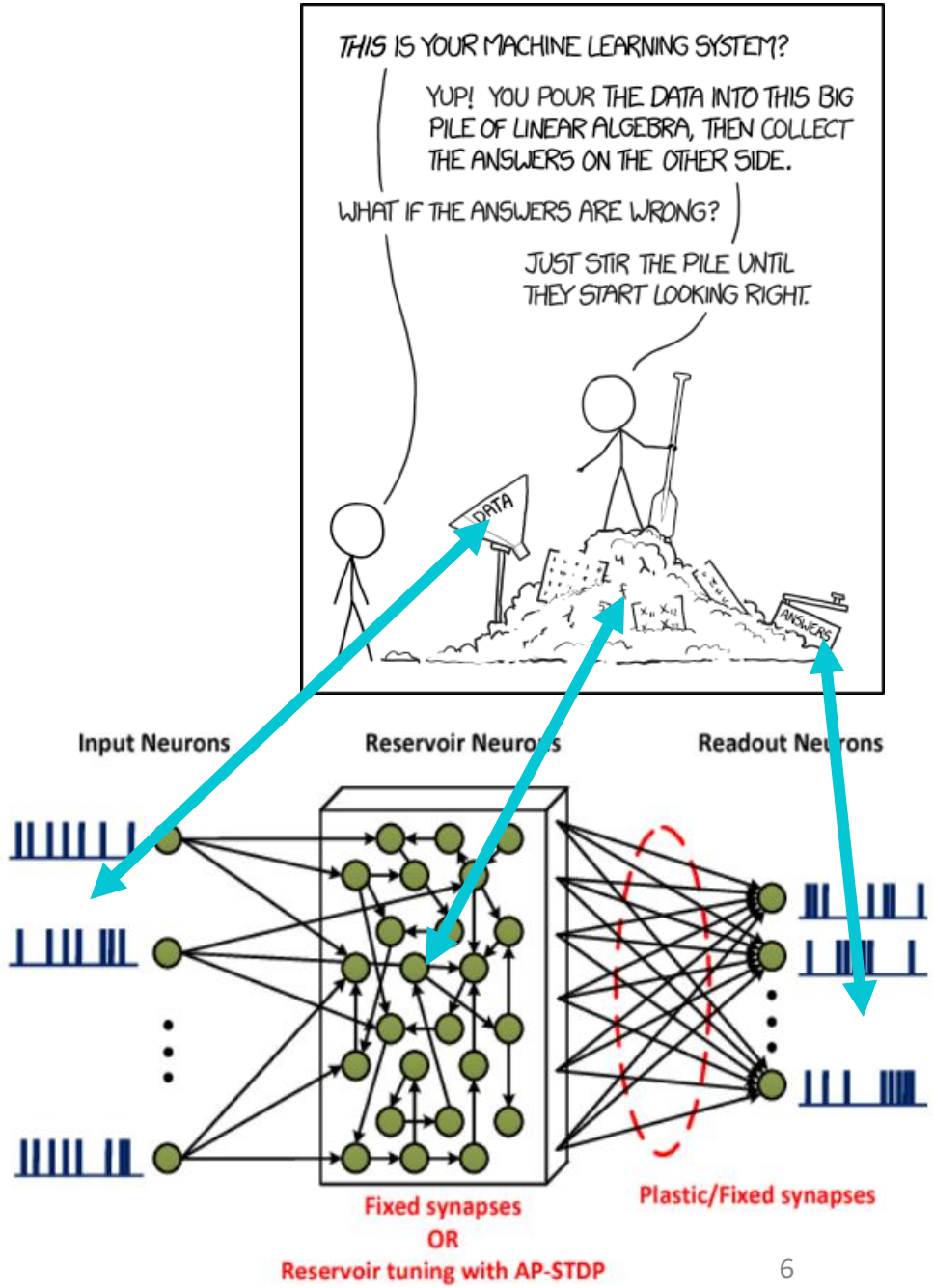
[2015, 2016 Janod]

- Need a small scalable switch
- Switching material – Mott insulators
- **Cr:VO_x**
  - Switch between metal and insulator with large memory window
  - High endurance (T. Hennen 2018)
  - Electronic mechanism, no forming, no snapback
  - Volatile/non-volatile switching
  - Threshold switching
  - Leaky integrate-and-fire neuron (Cario 2017, Corraze 2018)
  - Integration with transistors possible
- Excellent choice to replicate dynamical systems



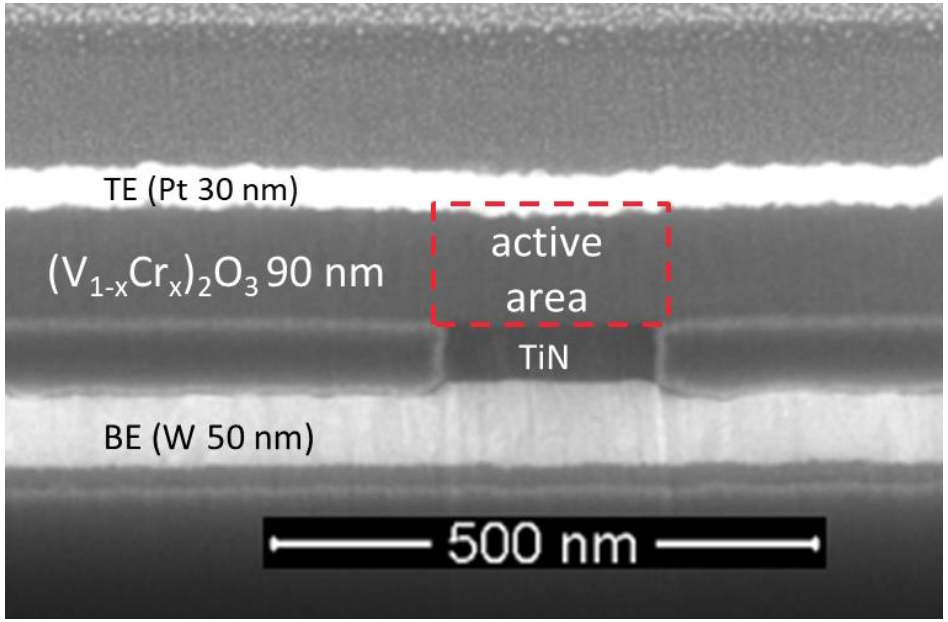
Reservoir Computing

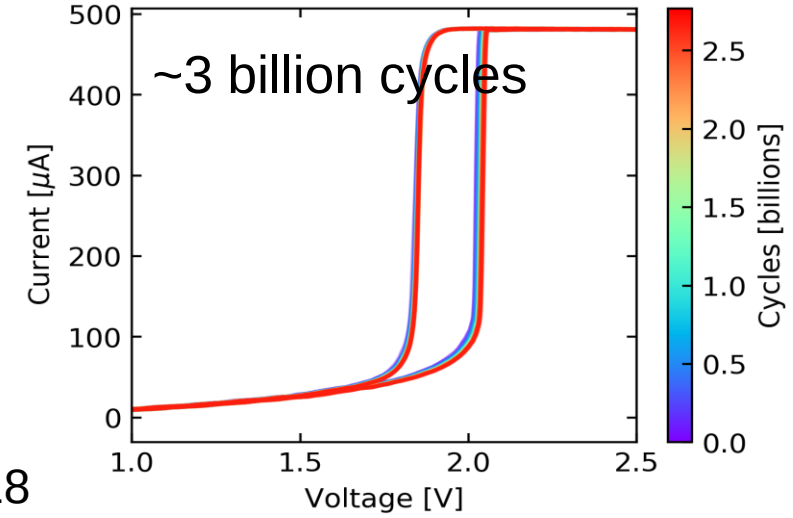
- Bio-inspired RNN
- Fundamental idea:
  - Build a dynamical system that maps the input stream into a linearly separable space
  - Fixed random weights from input to reservoir and within reservoir
  - Only train the classification layer
- Advantages:
  - Scales well
  - Good theory for dynamical systems
  - Similar to higher order brain functions
  - **No backpropagation**
  - **Easy on-line learning**
- Can be built with Mott insulators



Device Structure

- MIM structure
- Demonstrated film thickness down to 5 nm
- Device size 120 nm – 500 nm
- Very stable, exercised cell for over **10^{12}** cycles with no degradation
- **5%** device variation
- Great for integration and scaling up

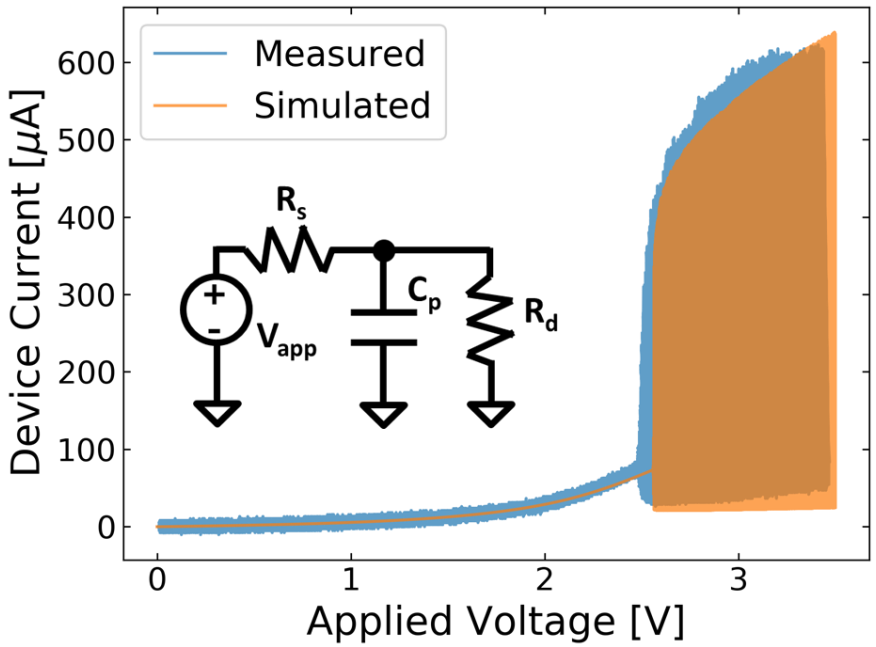




T. Hennen *et al.*, IEDM 2018

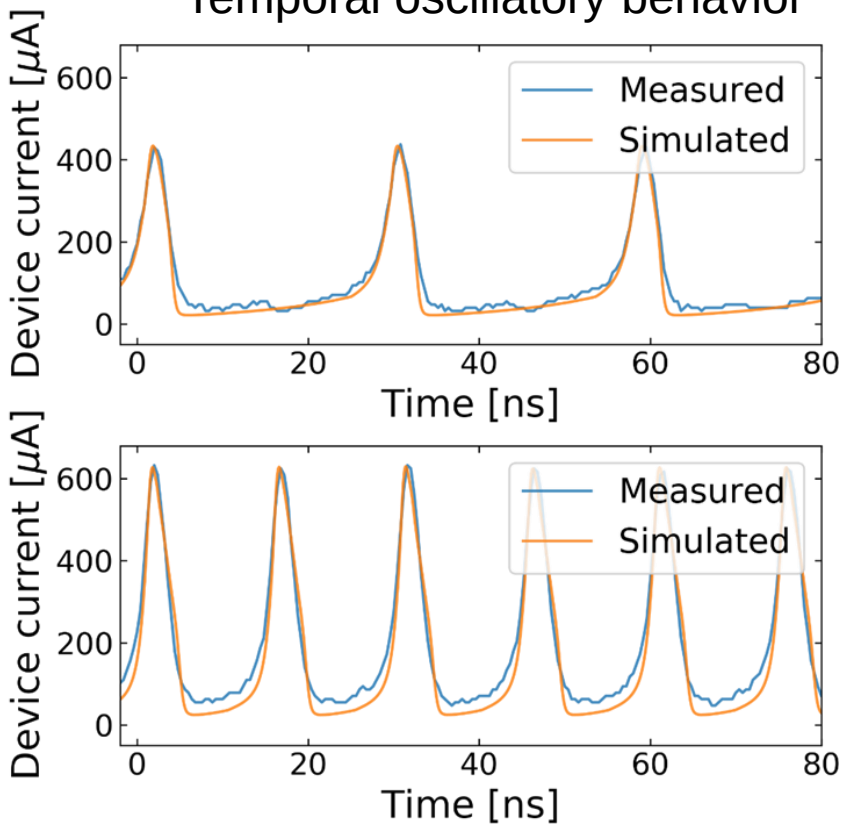
Oscillatory Behavior of RC circuit

- Mott insulator device in an RC circuit emulates neuron spiking



150nm width, 30nm thickness, formed device

Temporal oscillatory behavior



T. Hennen *et al, IMW* 2019

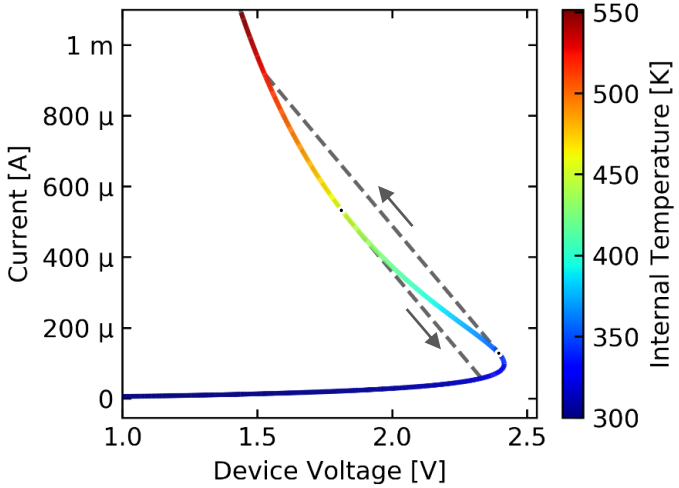
Dynamic V_2O_3 Model

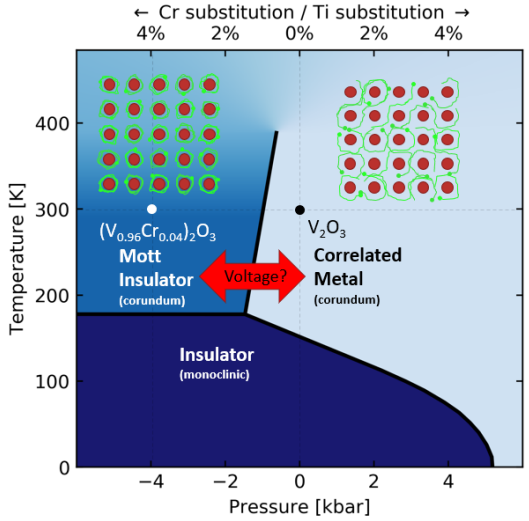
- Physics-based SPICE model
- Combining electrical and heat transport
- 1D Equations, good enough for RC
  - Empirical temperature dependent conduction equation
  - Newton's law of cooling
- Co-simulation with python

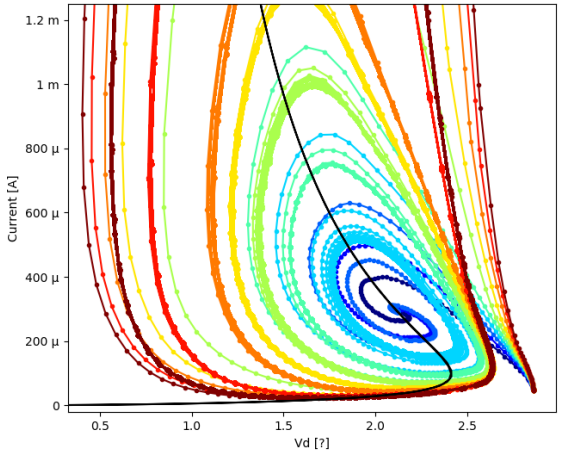
$$j = \alpha E \exp\left(\frac{-\beta}{k_B T}\right) \exp\left(\gamma\sqrt{E}\right)$$

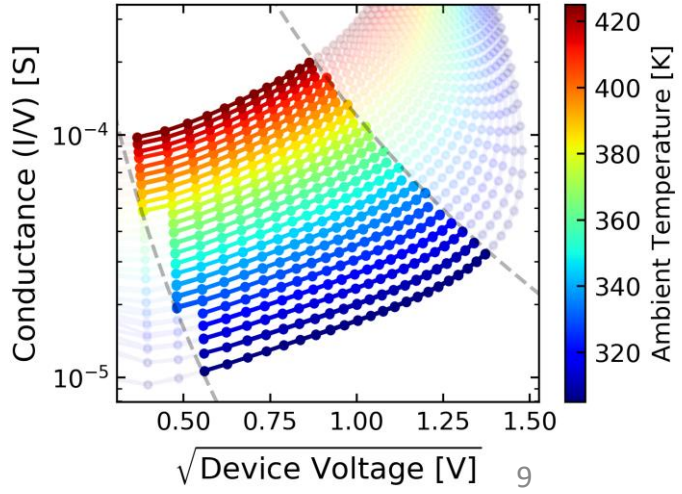
$$C_{th}\frac{dT}{dt} = \frac{T - T_0}{R_{th}} + jE$$

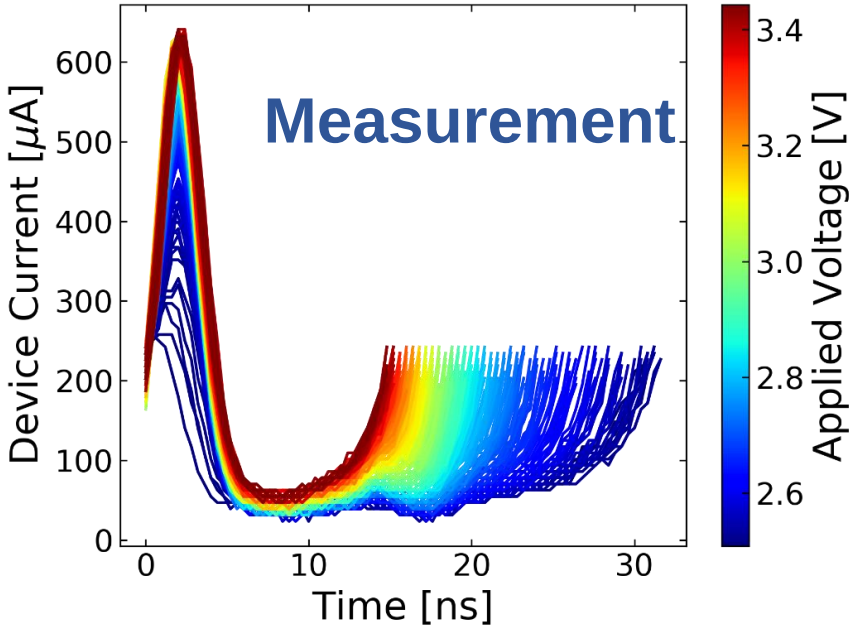
α, β, γ: conduction fitting parameters
C_{th}: thermal capacitance
R_{th}: thermal resistance

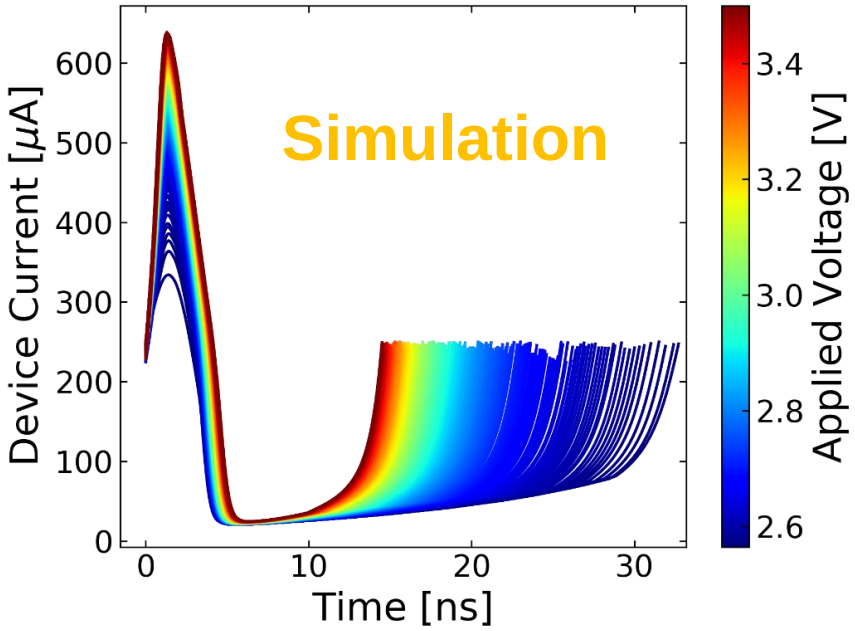


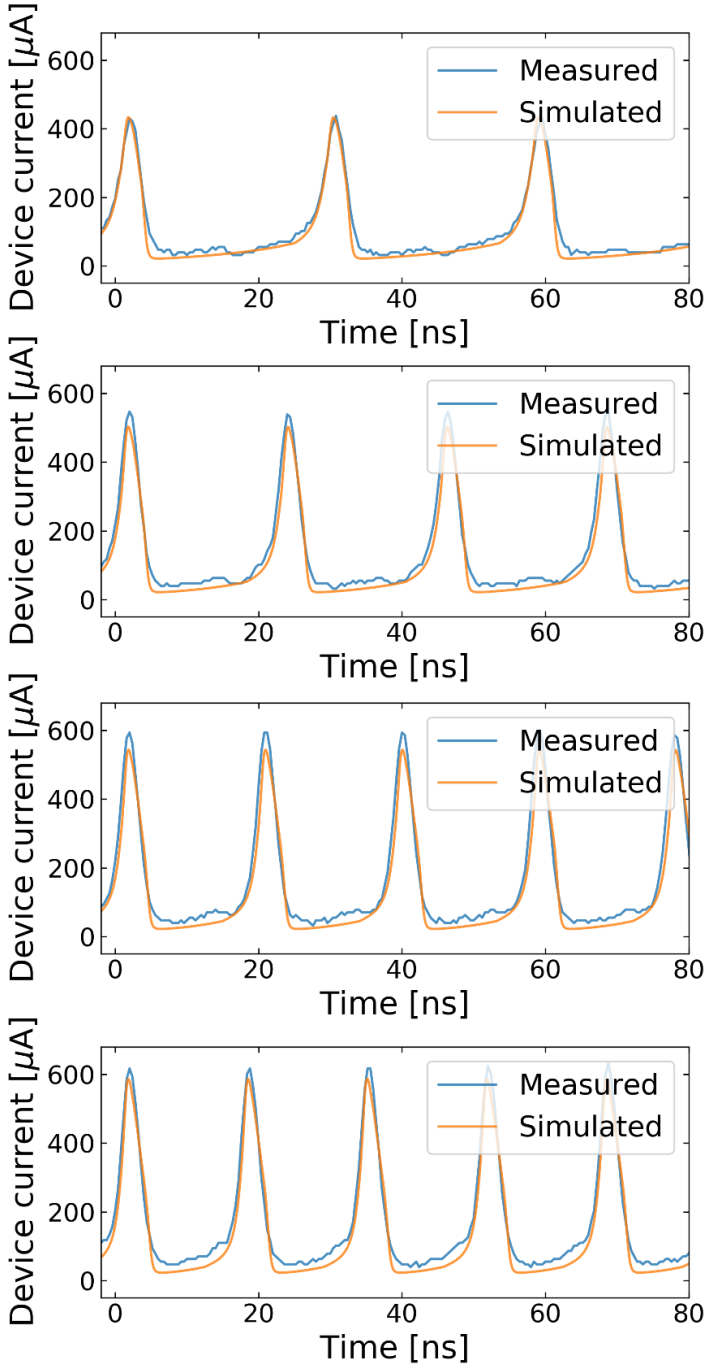


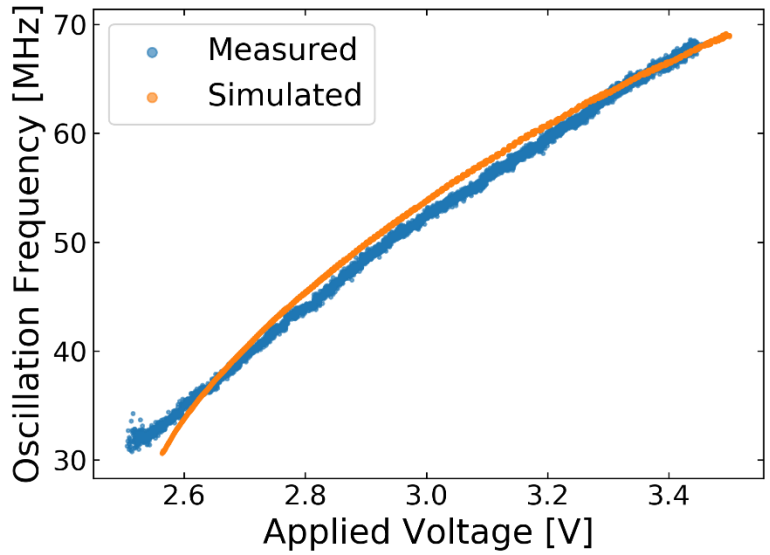


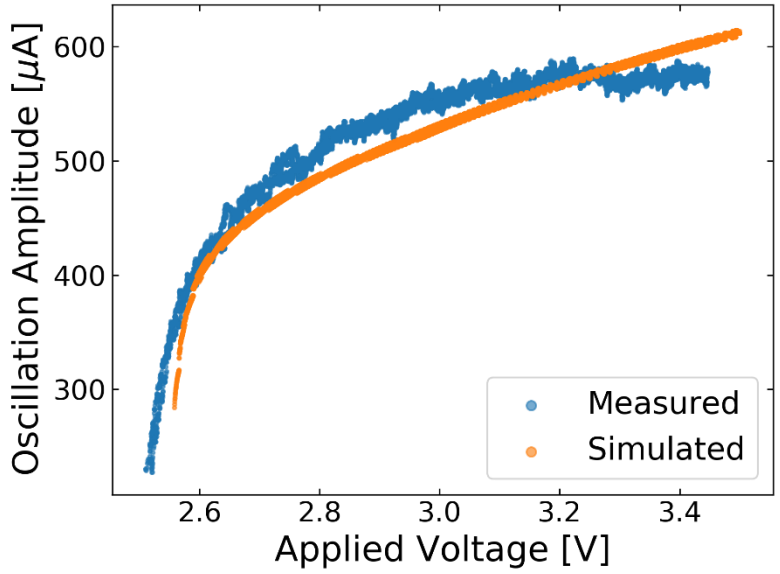



Measurement
Device Current [μA]
Time [ns]
Applied Voltage [V]


Simulation
Device Current [μA]
Time [ns]
Applied Voltage [V]

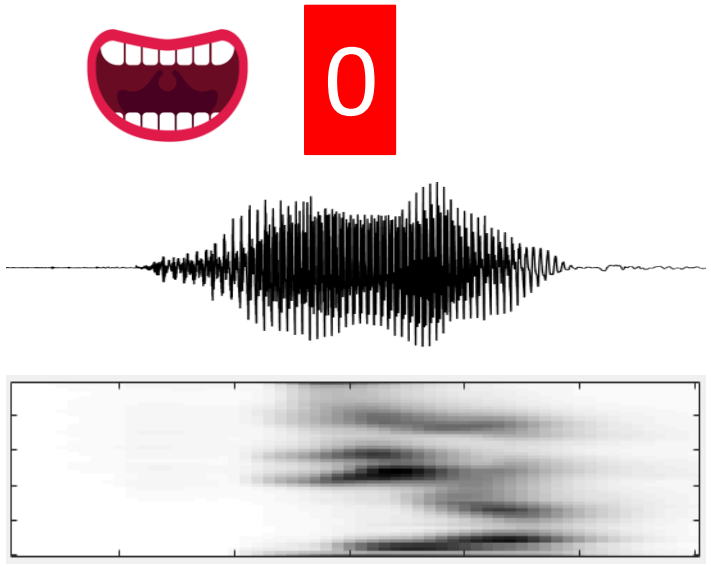

Measured
Simulated
Device current [μA]
Time [ns]

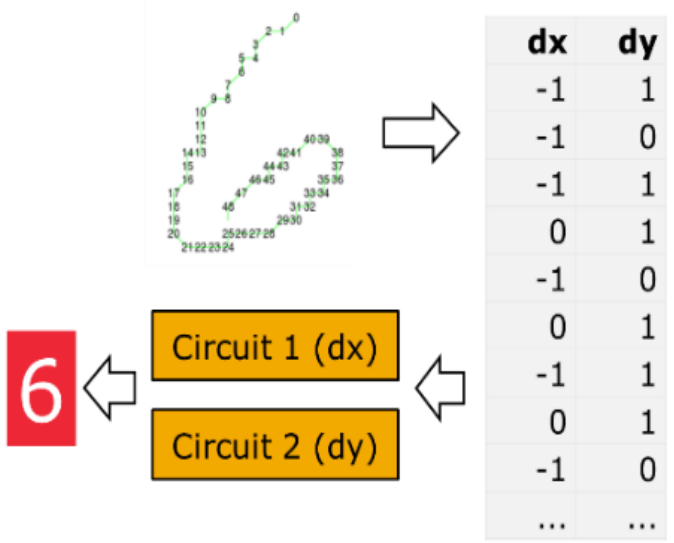

Measured
Simulated
Oscillation Frequency [MHz]
Applied Voltage [V]

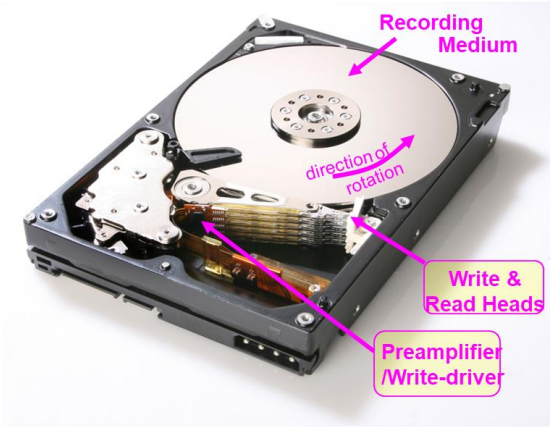

Measured
Simulated
Oscillation Amplitude [μA]
Applied Voltage [V]

Model Systems for RC Evaluation

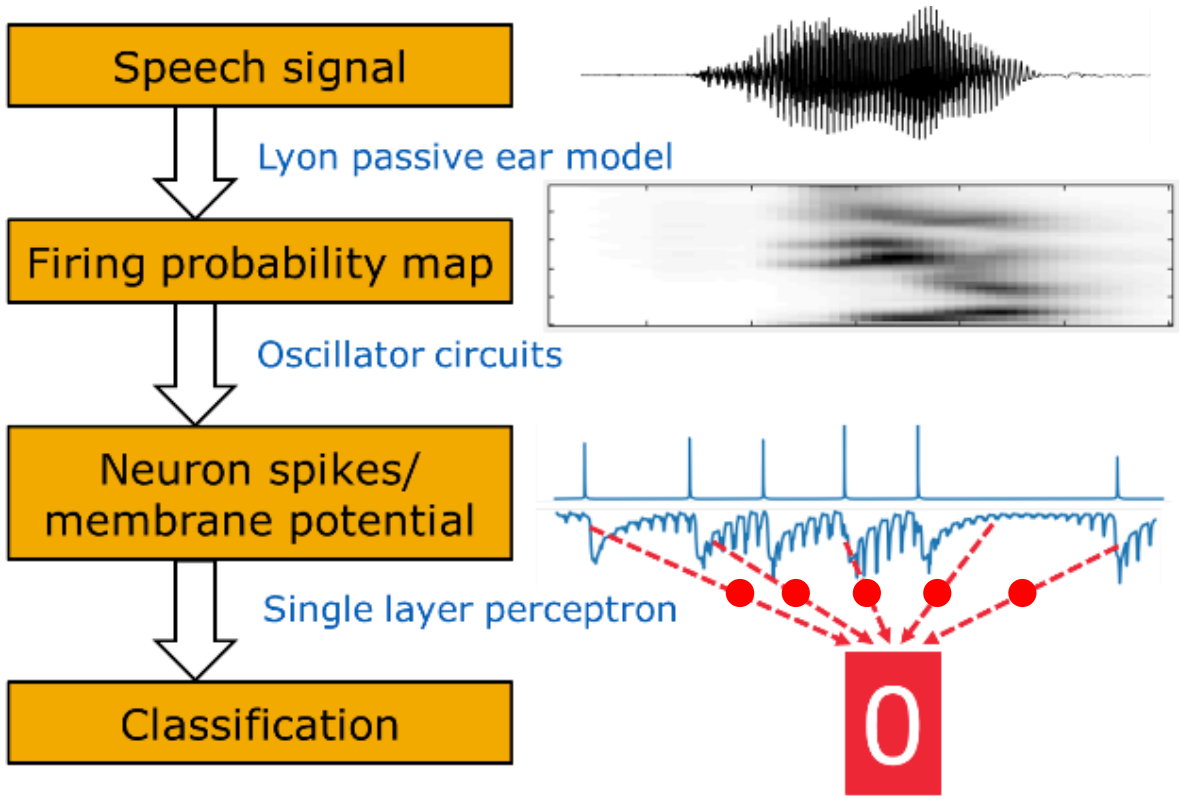
- Studied 3 model systems
  - Spoken digit recognition
  - Handwritten digit recognition (MNIST stroke)
  - HDD channel decoding



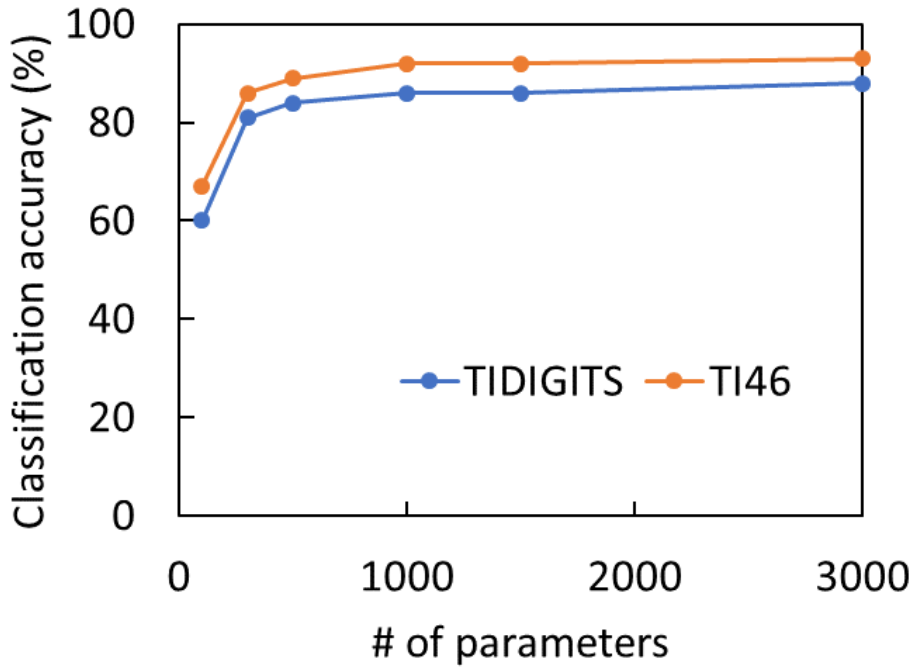




Speech Recognition with RC

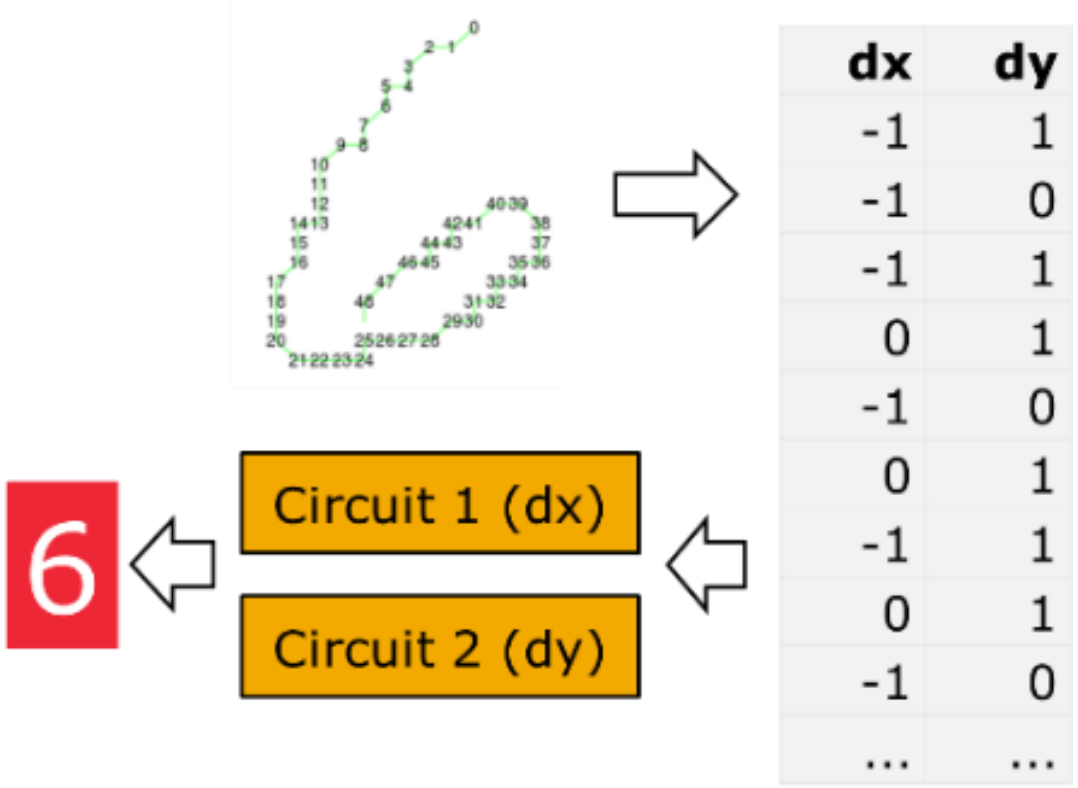


● Trainable weights

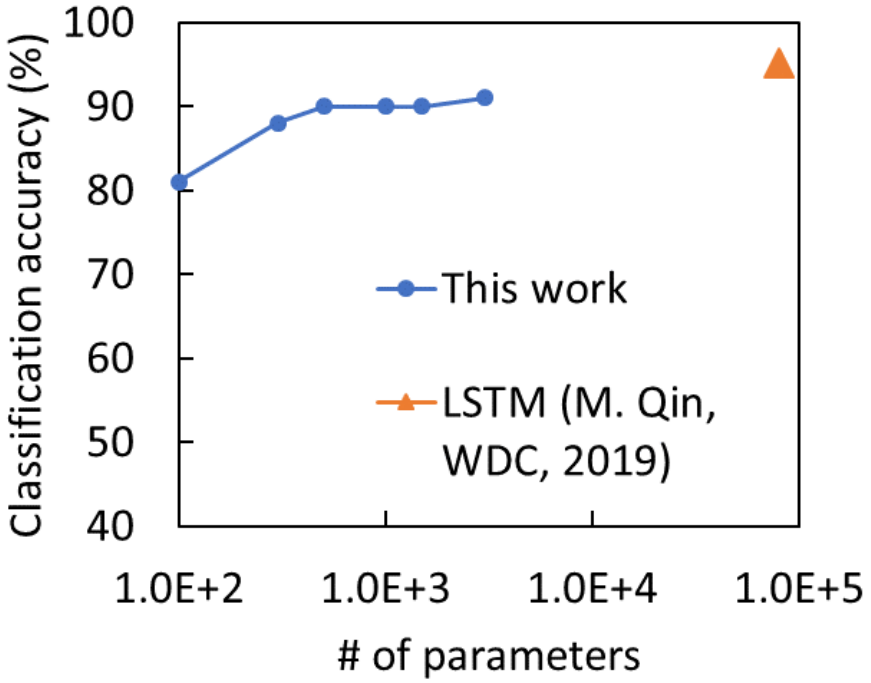


- 93% for TI46 and 88% for TIDIGITS with limited amount of trainable weights

MNIST Stroke Sequence Recognition with RC

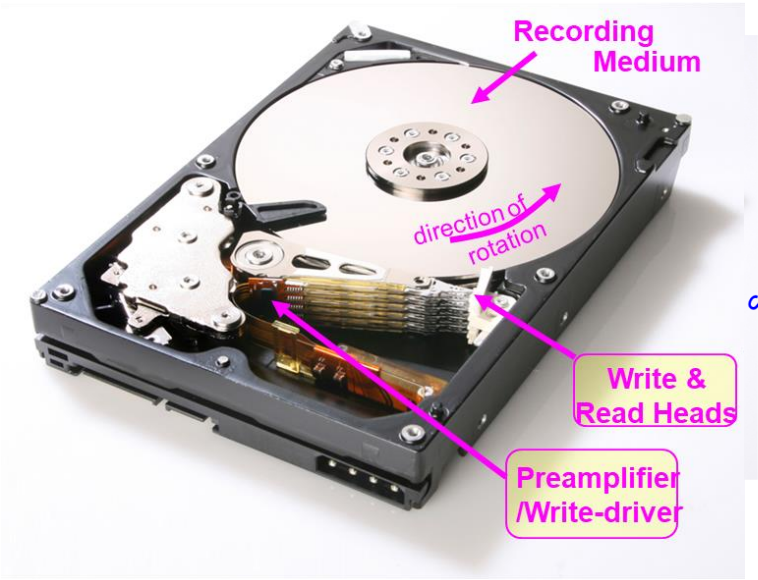


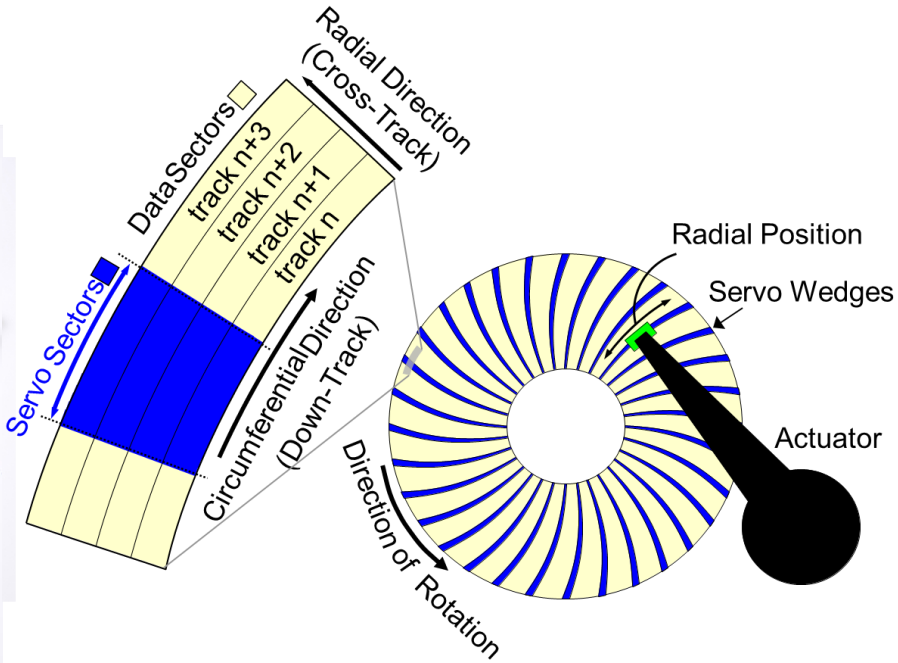
dx	dy
-1	1
-1	0
-1	1
0	1
-1	0
0	1
-1	1
0	1
-1	0
…	…

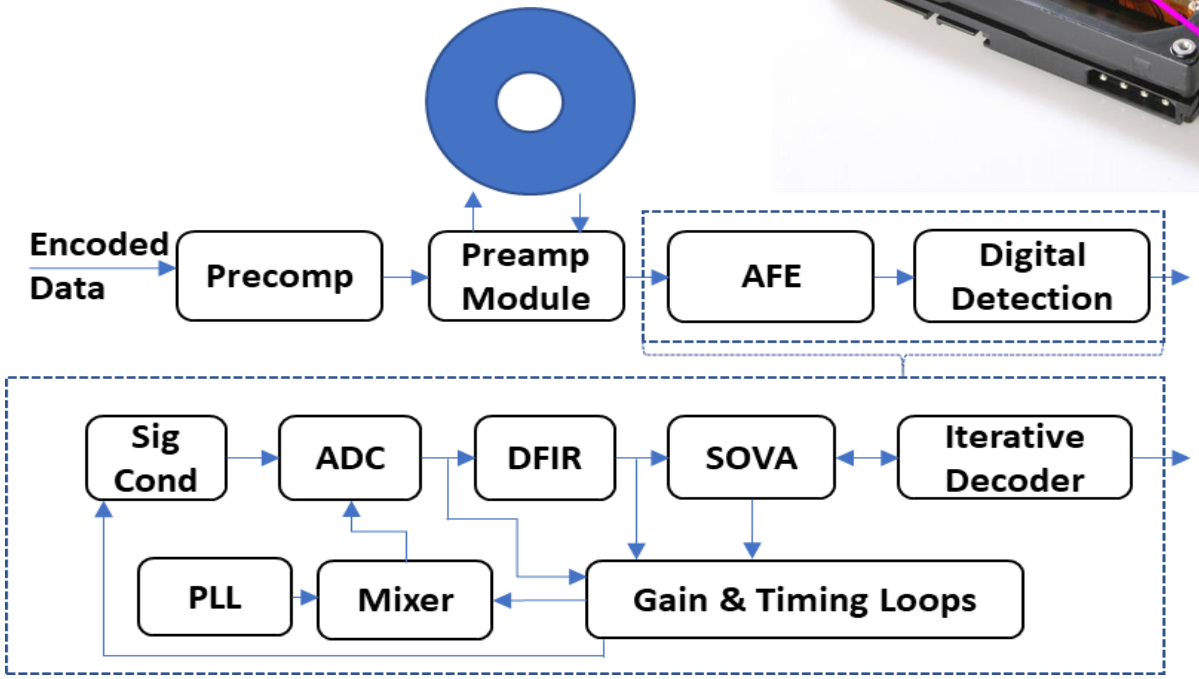


- 91% achieved with RC, compared to 95% achieved with LSTM

HDD Channels

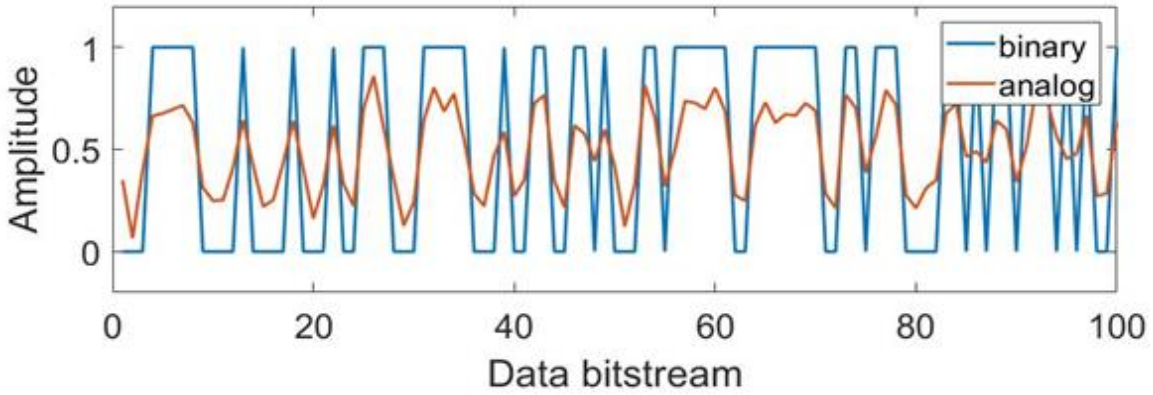


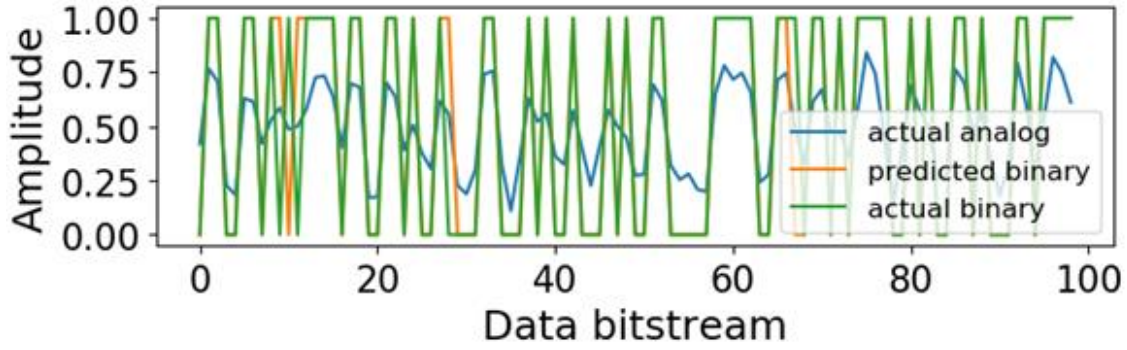




- An HDD recording system contains a very complex channel decoder

HDD channel decoding with RC



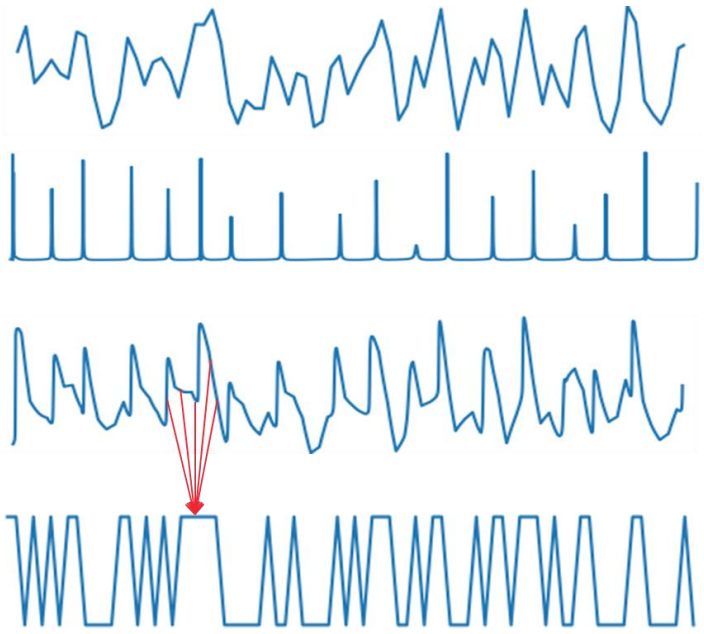


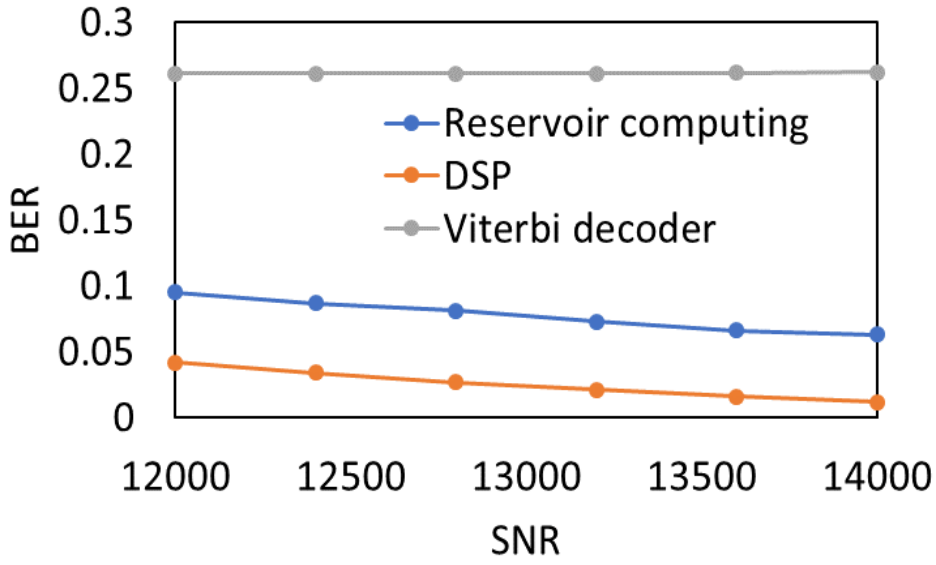
Input voltage:
Analog sequence

Rd current:

Rs current:

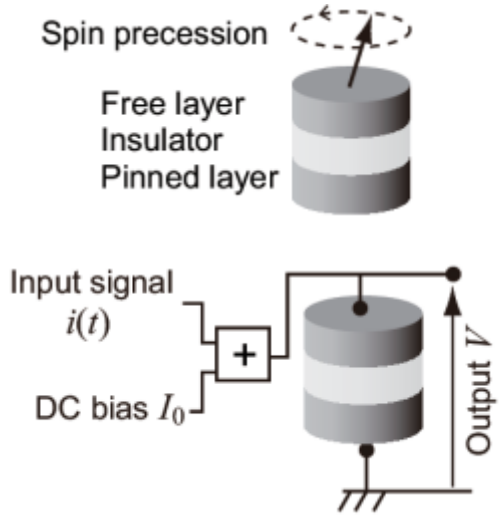
Output:
Binary sequence





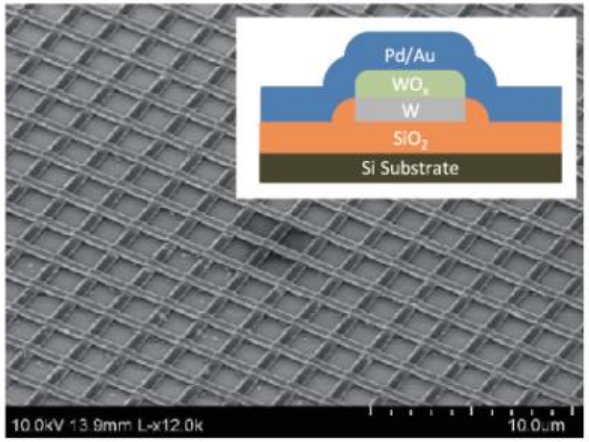
Comparison with Other Works

	CPU (S. Han, 2016)	**GPU** (S. Han, 2016)	**FPGA** (S. Han, 2016)	**This work**
Latency (us)	6017	240	82	80
Power (W)	111	202	42	0.001-0.05



Torrejon et al, 2017

	Spintronic reservoir computing (reviewed by G. Tanaka, 2019)	**RRAM reservoir computing** (reviewed by G. Tanaka, 2019)	**This work**
# of devices used	1	88	2-50
Spoken digit recognition	98% measurement	N/A	93% simulation
Handwritten digit recognition	N/A	88% measurement	91% simulation



Du *et al*, 2017

Conclusion

- ML is getting limited by available power
- Reservoir computing (RC) is a promising candidate with high BW, low latency, low power
- RC is very hardware friendly, uses very small number of weights
- On-line learning becomes feasible - Simple LSQ calculation
- Next steps:
  - Experimental demonstration
  - Power estimation on the whole system with CMOS peripherals

Thanks for listening