Hewlett Packard
Enterprise

Inspired by the Brain:
Computing from Architecture to Devices

R. Stanley Williams
HPE Senior Fellow

Neuromorphic Computing Symposium
July 18, 2017

What's Next?

The end of Moore's law could be the best thing that has happened in computing since the beginning of Moore's law.

The Chua Lectures: A 12-Part Series at HPE Labs

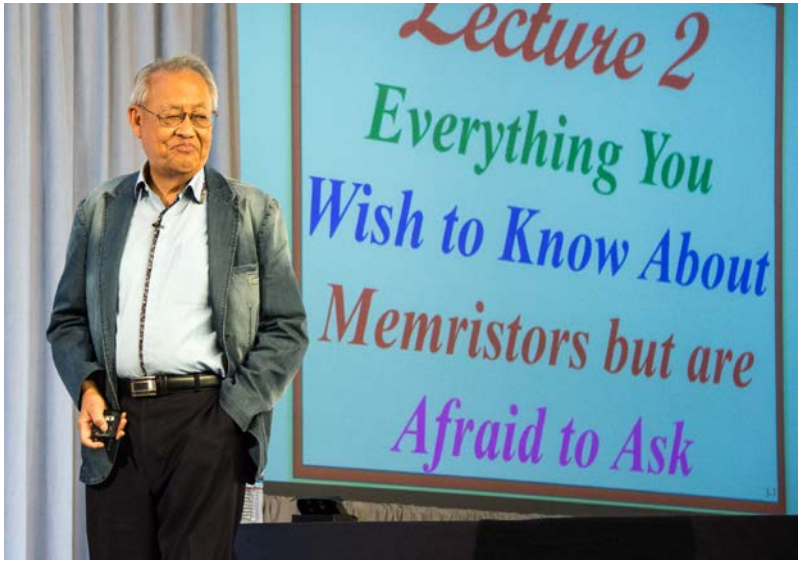
From Memristors and Cellular Nonlinear Networks to the Edge of Chaos

https://www.youtube.com/playlist?list=PLtS6YX0YOX4eAQ6IrOZSta3xjRXzpcXyi

or enter “The Chua Lectures” into your favorite browser

What’s missing?

‘Linearize then analyze’ is not valid for understanding nanodevices or neurons – a new nonlinear dynamical theory of ‘electronic’ circuits is needed, and was developed 50 years ago by Leon Chua, father of nonlinear circuit theory and Cellular NNs.



Structure of a US Neuromorphic Science Computing Program

1. Connect Theory of Computation with Neuroscience and Nonlinear Dynamics
   e.g. Boolean logic, CNN, Baysian Inference, Energy-Based Models, Markov Chain
2. Architecture of the Brain and Relation to Computing and Learning
   Theories of Mind: Albus, Eliasmith, Grossberg, Mead, many others
3. Simulation of Computational Models and Systems – need supercomputers!
4. System Software, Algorithms & Apps – Make it Programmable/Adaptable
5. Chip Design – System-on-Chip: Accelerators, Learning and Controllers
   Compatible with standard processors, memory and network fabric (Gen-Z)
6. Chip Processing and Integration – Full Service Back End of Line on CMOS
   DoE Nanoscale Science Research Centers (NSRCs) – e.g. CINT
7. Devices and Materials – *in situ* and *in operando* test and measurement
   Most likely materials will be adopted from Non-Volatile Memory

Future exponential increases in computer performance and efficiency will require multiple advances:

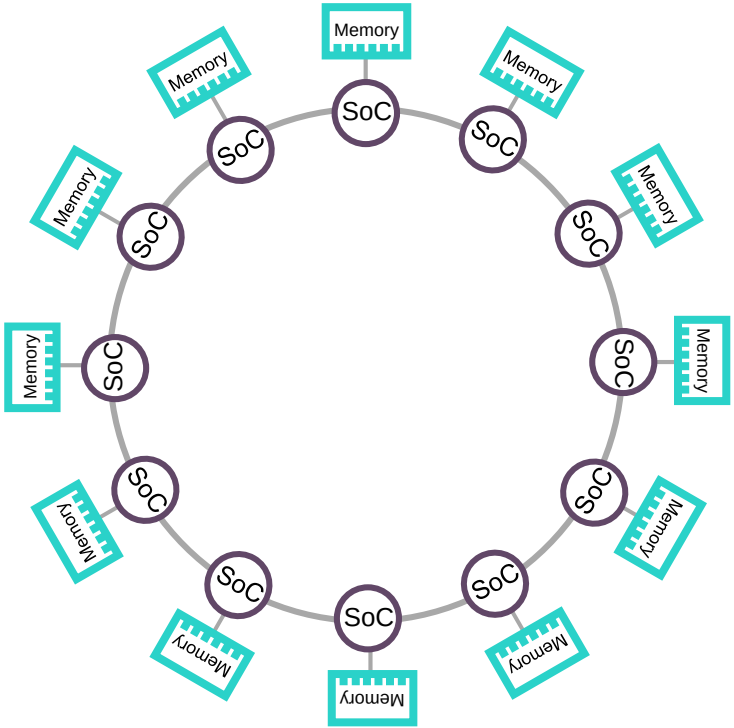
Systems and Architectures

- Memory-centric computing – no von Neumann bottleneck
- Gen-Z – high performance open fabric to democratize computing (http://genzconsortium.org/)

- Dot-product engine: memristor-based vector-matrix multiplication accelerator for neural nets and signal processing
- Mimicking Synapse and Neuron Dynamics with Memristors

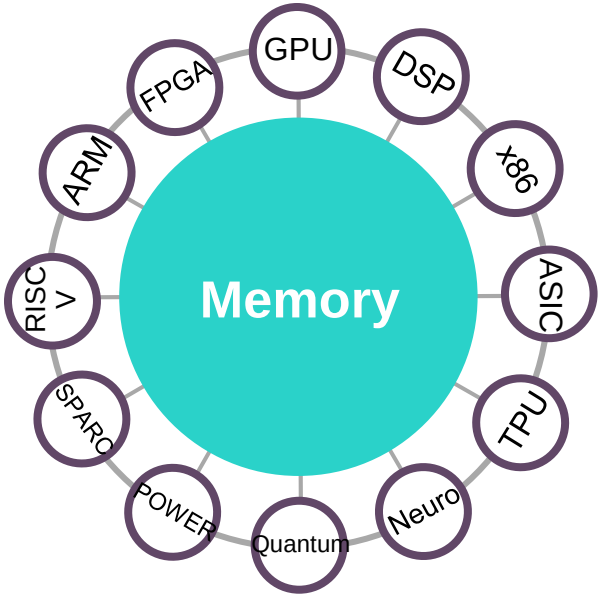
- Chaos as a computing resource for constrained optimization problem solving (Hopfield network)

Devices



From processor-centric computing…

the traditional von Neumann architecture

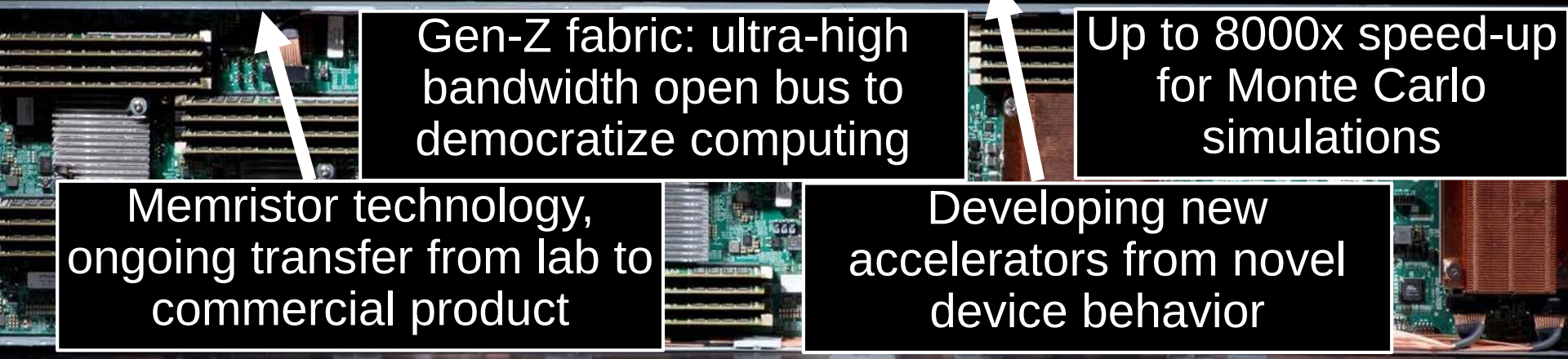


…to Memory-Driven Computing

with Gen-Z open (nonproprietary) fabric, computing is plug and play

Memory-Driven Computing (MDC) is a reality: The Machine

Fast, persistent memory	Fast memory fabric	Task-specific processing	New and Adapted software
Combining memory and storage in a stable environment to increase processing speed and improve energy efficiency	Using photonics where necessary to eliminate distance and create otherwise impossible topologies	Optimizing processing from general to specific tasks	Radically simplifying programming and enabling new applications that we can't even begin to build today



Gen-Z fabric: ultra-high bandwidth open bus to democratize computing

Up to 8000x speed-up for Monte Carlo simulations

Memristor technology, ongoing transfer from lab to commercial product

Developing new accelerators from novel device behavior

http://genzconsortium.org/en-Z

High Bandwidth Low Latency

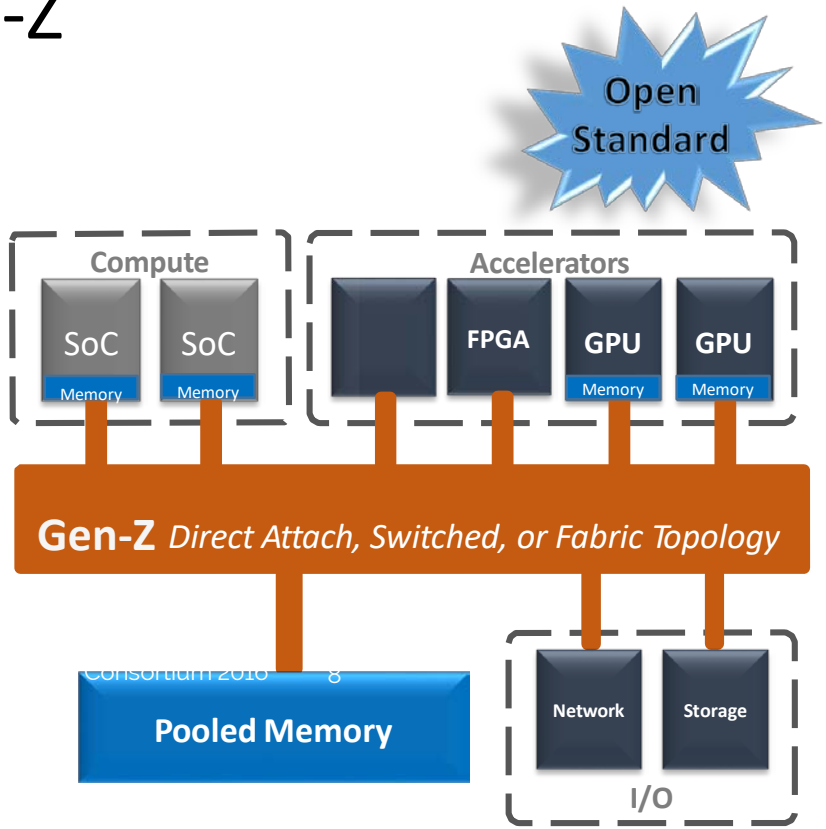
- Memory Semantics – simple Reads and Writes
- From tens to several hundred GB/s of bandwidth
- Sub-100 ns load-to-use memory latency

Advanced Workloads & Technologies

- Real time analytics
- Enables data centric and hybrid computing
- Scalable memory pools for in memory applications
- Abstracts media interface from SoC to unlock new media innovation

Secure Compatible Economical

- Provides end-to-end secure connectivity from node level to rack scale
- Supports unmodified OS for SW compatibility
- Graduated implementation from simple, low cost to highly capable and robust
- Leverages high-volume IEEE physical layers and broad, deep industry ecosystem

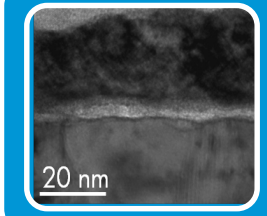


Rack Scale

Hewlett Packard Enterprise

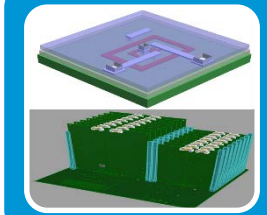
Founding Members of the Gen-Z Consortium (38 today)

Brain Inspired Accelerators: devices → circuits →applications



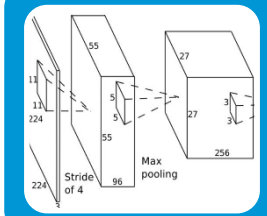
Materials engineering – device properties

- High resistance
- Large OFF/ON ratio
- Low switching current
- Linear electronic transport



Cell integration, arrays, and circuits

- Integration/fab of memristors with selectors/transistors
- Optimizing reading and writing circuits
- Construct platform to write/read device arrays
- Compact modeling of memristor characteristics



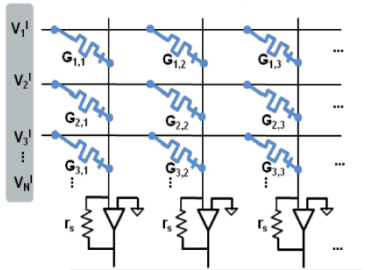
Architectures, Algorithms, Applications

- Matching capability to favorable applications
- Benchmarking and optimizing performance; GOPS/Watts
- Quantifying sources of error and bottlenecks

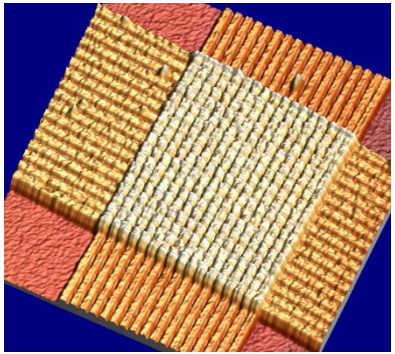
Application-aware research into materials/cell/circuit

Dot Product Engine: memristor arrays accelerate vector-matrix multiplication (C. elegans – 800 papers/yr)

Input Voltage vector



Output current $I_j^O = \sum_i G_{ij} \cdot V_i^I$



- **Parallel multiply & add through Kirchoff's and Ohm's laws**

  1961, K. Steinbuch "*Die Lernmatrix*"– suggests using "ferromagnetic toroids"

- **Memristors as highly scalable, tunable analog resistors**

  High ON/OFF ratio (~10^5), supporting multiple levels

- **Well suited for streaming workloads like neural nets**

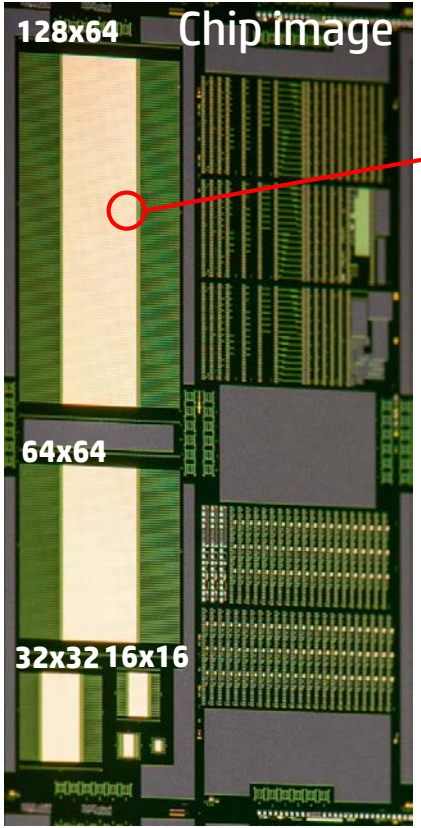
- **Many ways to scale up**

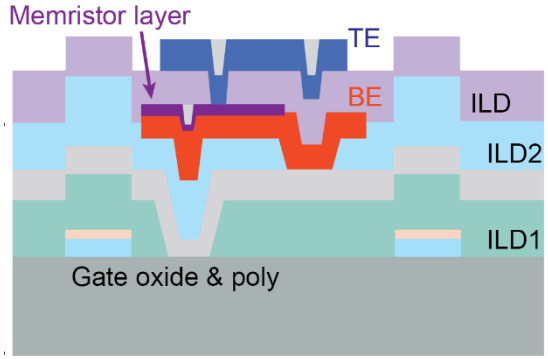
  Memristor levels, array size, wire pitch, 3D layer, DAC/ADC speed & width etc.

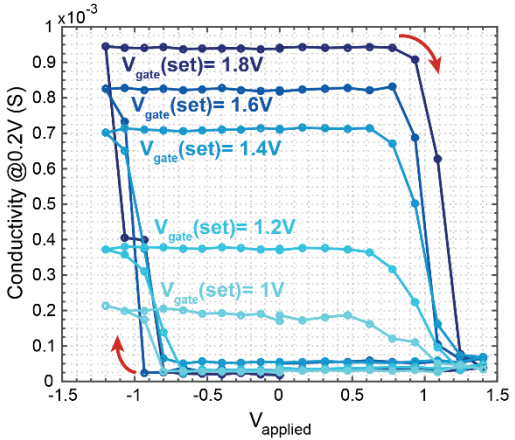
- **Performance (execution time) improvements >1000x and energy efficiency >100x over GPUs for particular applications**

The DPE: memristor-based analog computing platform

Flexible system for programming and computing on arrays of integrated Transistor-Memristors (1T1M)







Hewlett Packard Enterprise

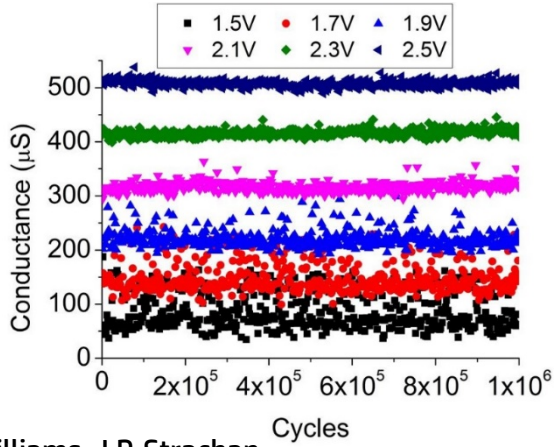
Programming memristor arrays for computation

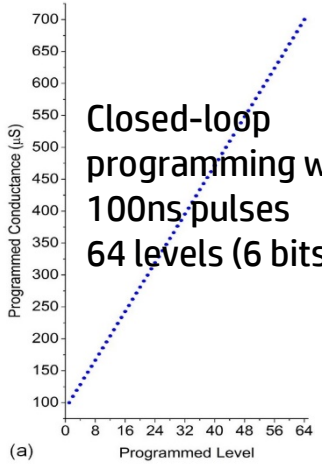
64x128 array = 8192 memristors

Each “pixel” is a continuously tunable memristor cell

Grayscale - 182 distinct conductance levels (~7-8 bits)

Each memristor can be reprogrammed >1e6 times





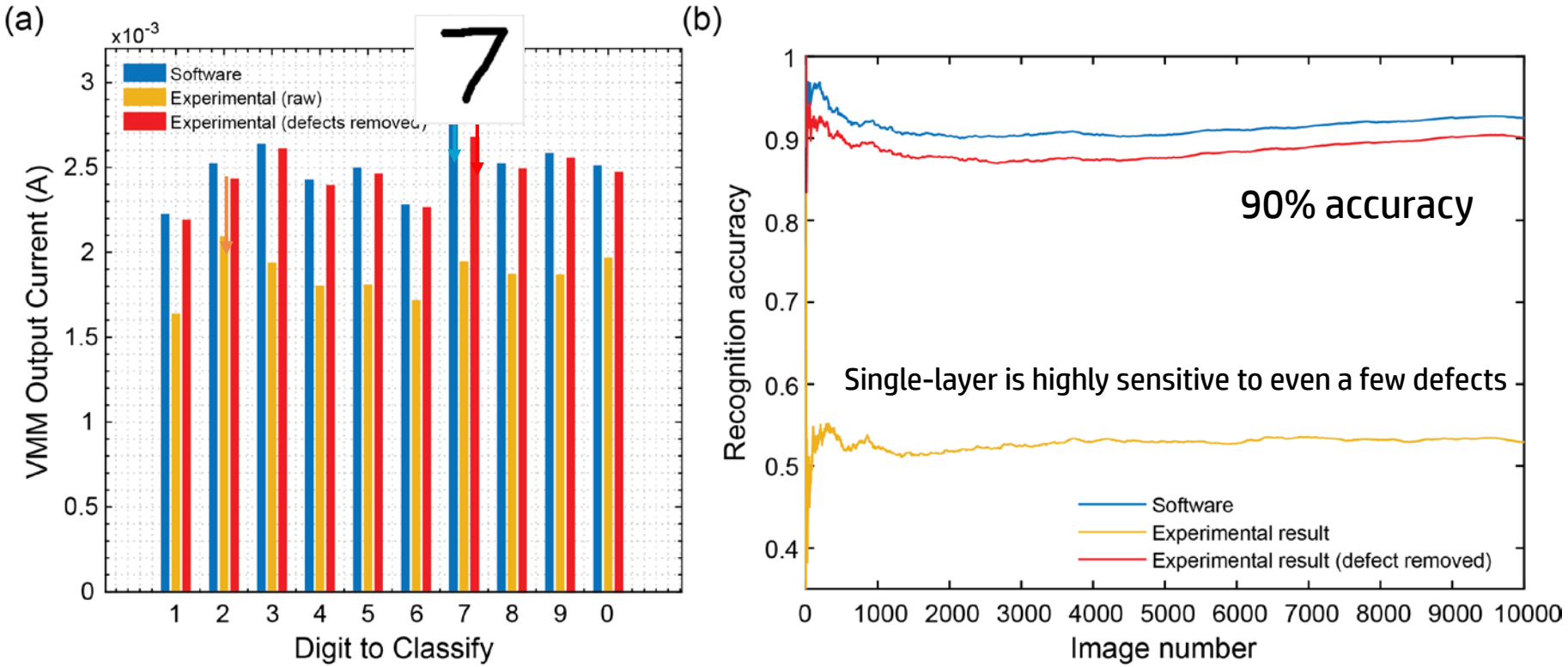
Closed-loop programming with 100ns pulses
64 levels (6 bits)

E.J. Merced-Grafals, N. Dávila, N. Ge, R.S. Williams, J.P. Strachan

Nanotechnology 27, 365202 (2016).

Hewlett Packard Enterprise

Single Layer NN for MNIST classification

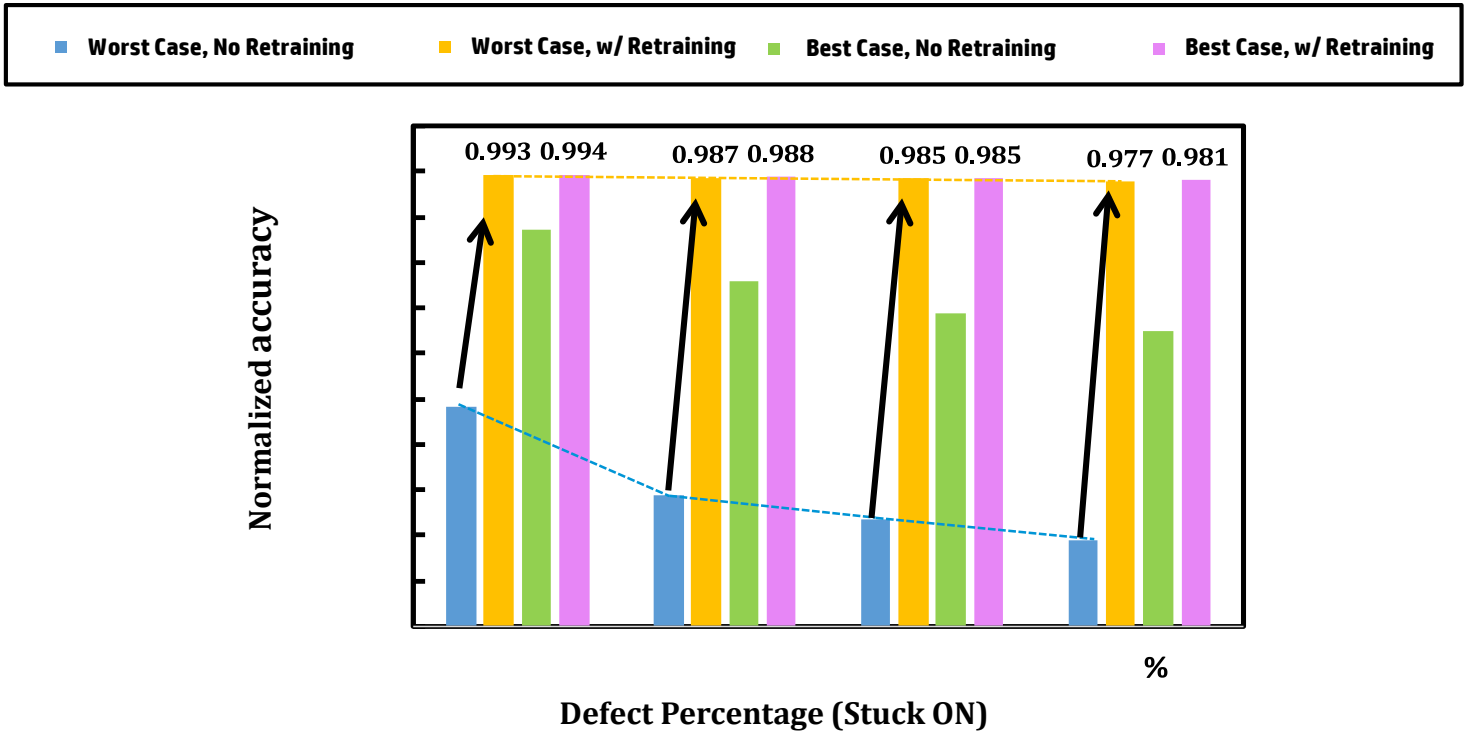


Miao Hu, et al. *in preparation*

Hewlett Packard Enterprise

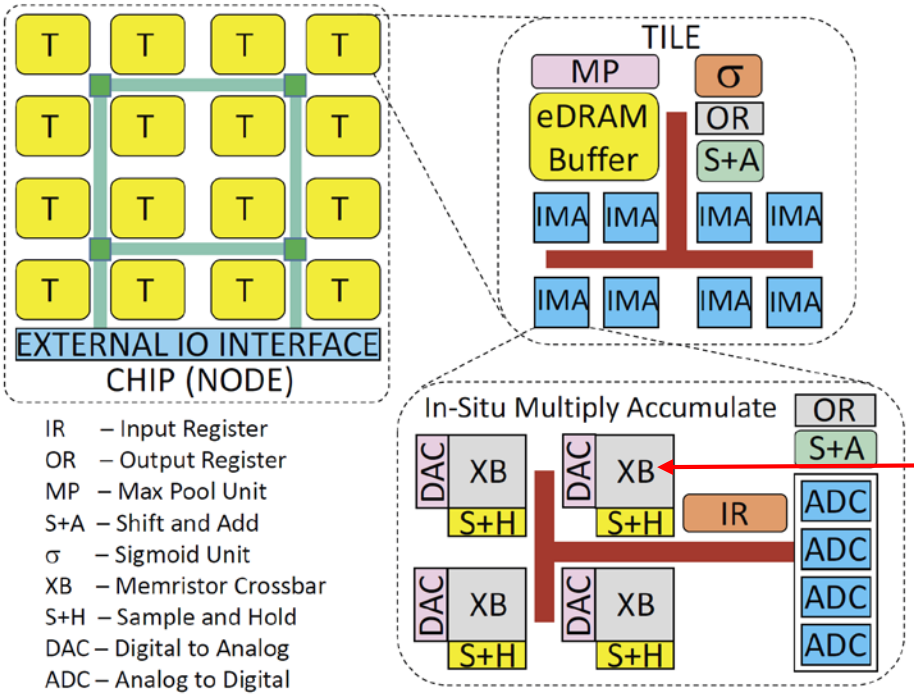
Dealing with “defects” - retraining

Retrain the network - surrounding weights can compensate for bad devices



C Liu, M Hu, JP Strachan, H Li , *DAC 2017*

“ISAAC” - architecture acceleration using DPEs



Store and re-use Kernels in non-volatile memristors – reduce data fetching

Heavy pipelining

Speedup of >5,000x over GPUs and 800x less energy

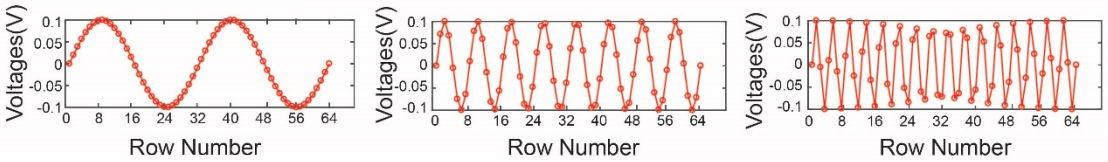
Speedup of 15x over Digital ASIC and 5.5x less energy

128x128 memristor Xbar arrays, 2 bits per cell, operating at 10 MHz

A. Shafiee, et al., “ISAAC: A Convolutional Neural Network Accelerator with In-Situ Analog Arithmetic in Crossbars”, *International Symposium on Computer Architecture (ISCA) 2016.*

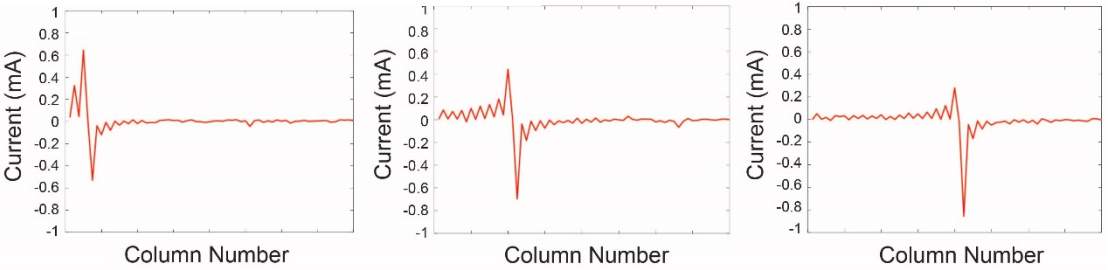
Signal processing on a DPE system
(UMass collaboration) – instantaneous cosine transform

Time-domain inputs (applied to rows)

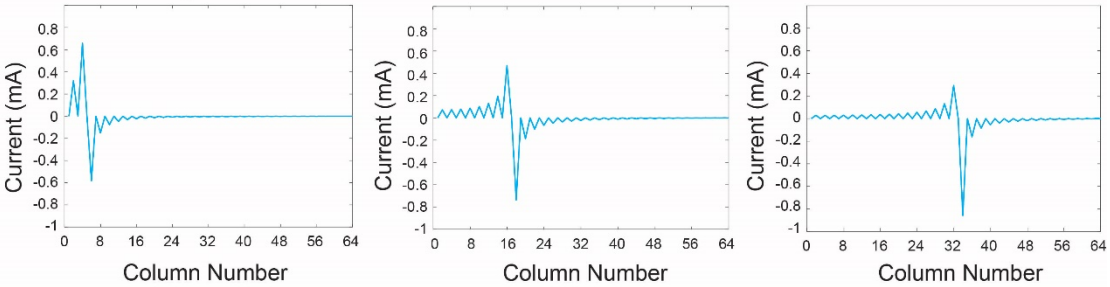


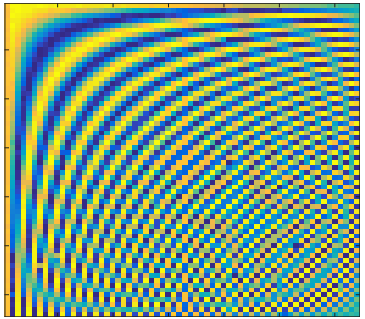
Freq-domain (DCT) outputs (read from columns)

Real-time experimental data from DPE



Software cosine transform

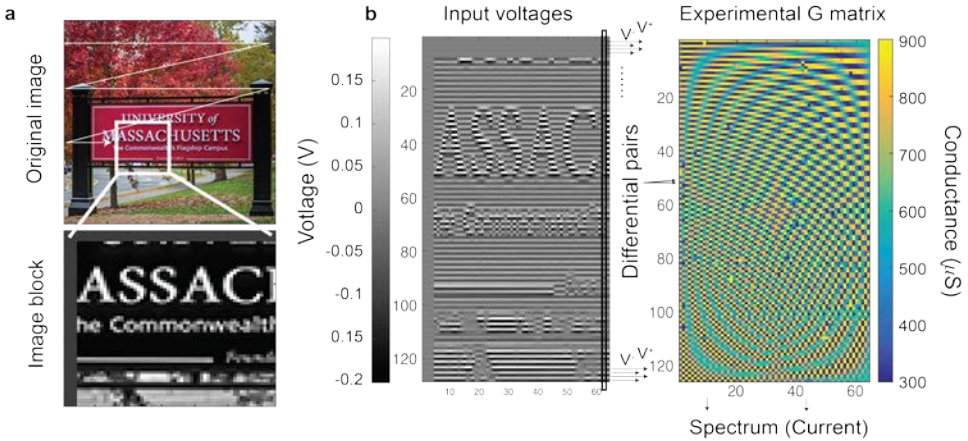




C. Li, et al., *submitted*
Umass Amherst

Hewlett Packard Enterprise

Image Compression with cosine transform on the DPE



c
7:1 compression ratio

Software encoder

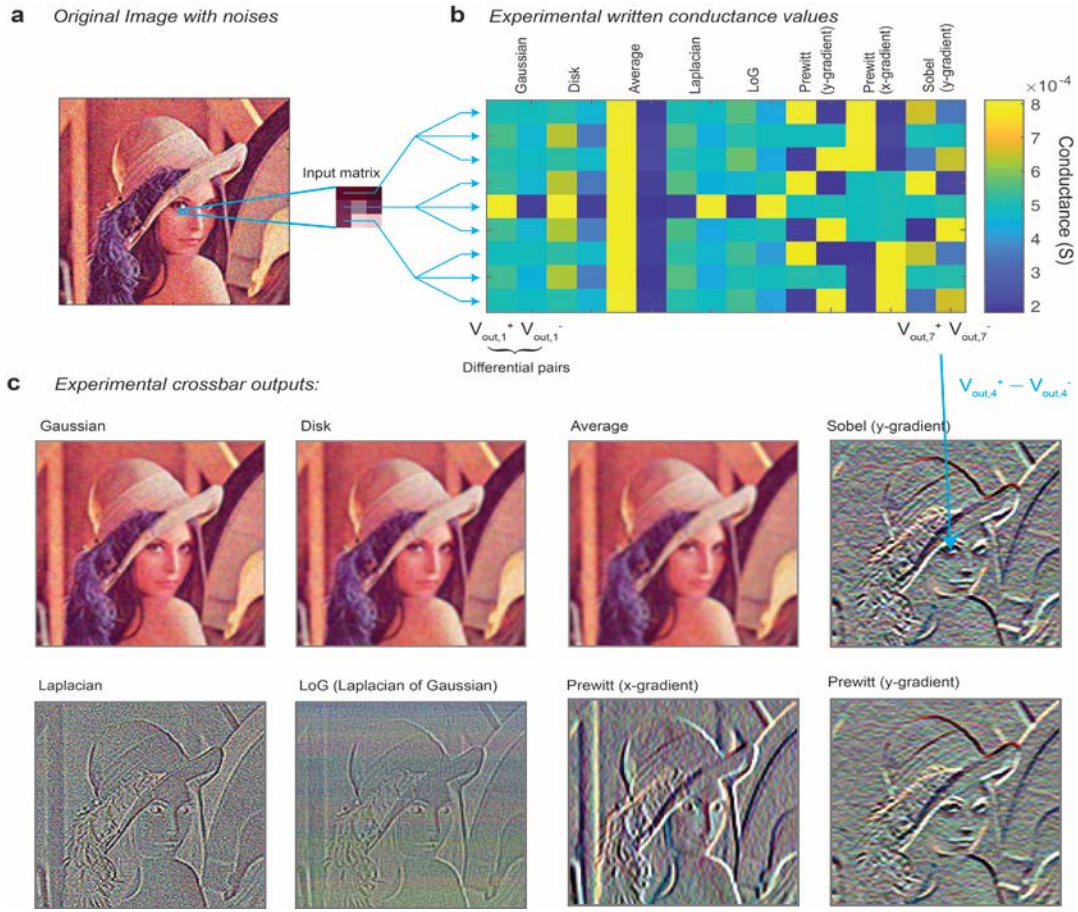
d
7:1 compression ratio

Experimental crossbar encoder

C. Li, et al., *submitted*
UMass Amherst

Hewlett Packard Enterprise

Convolutional Filters with the DPE



C. Li, et al., *submitted*

Umass Amherst

Writing software for a memristor accelerator

- Initial development of OS/Compiler/HW emulator
- Co-designed with hardware efforts
- Generalize/Broaden to other matrix-heavy applications (medical imaging, sci computing, network management)

Application Layer

- Network specification (Torch, Tensorflow, Caffe)– MLP, CNN, RNN, SNN etc.

Compiler

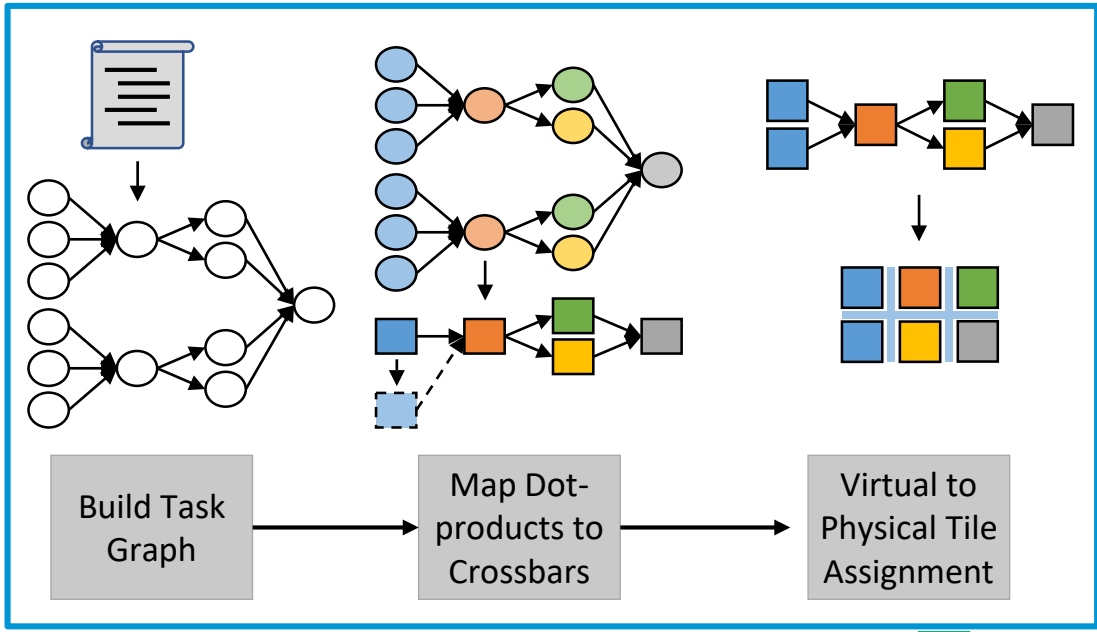
- Map specification to hardware (map table, bit precision, data routing)

Emulator

- Simulate and provide output metrics (accuracy, energy, efficiency, etc.)

DPE Hardware

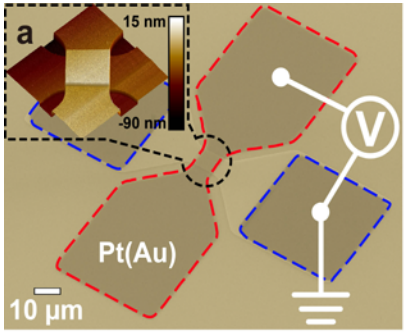
- Final hardware after the design is frozen based on design space explorations

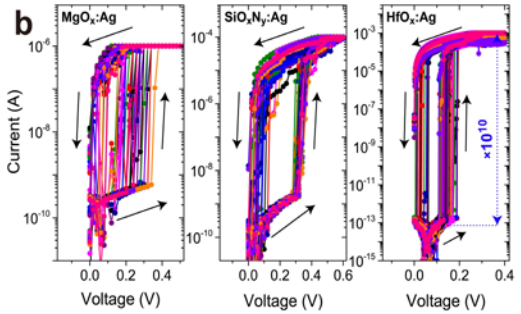


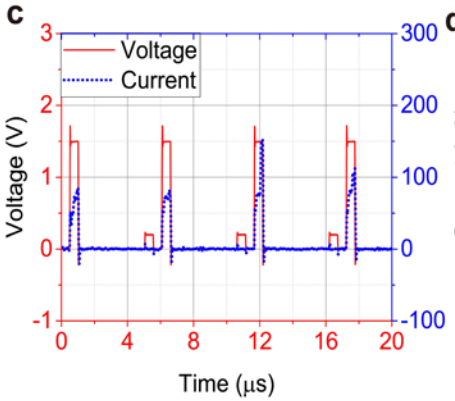
Collaboration with Dejan Milojicic, Sai Chalamalasetti, Izzat El Hajj, Pedro Bruel, Aayush Ankit

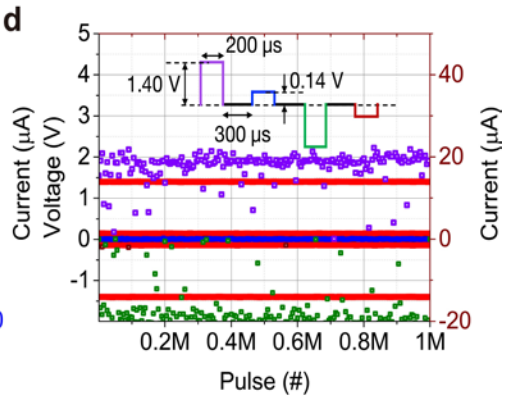
Hewlett Packard Enterprise

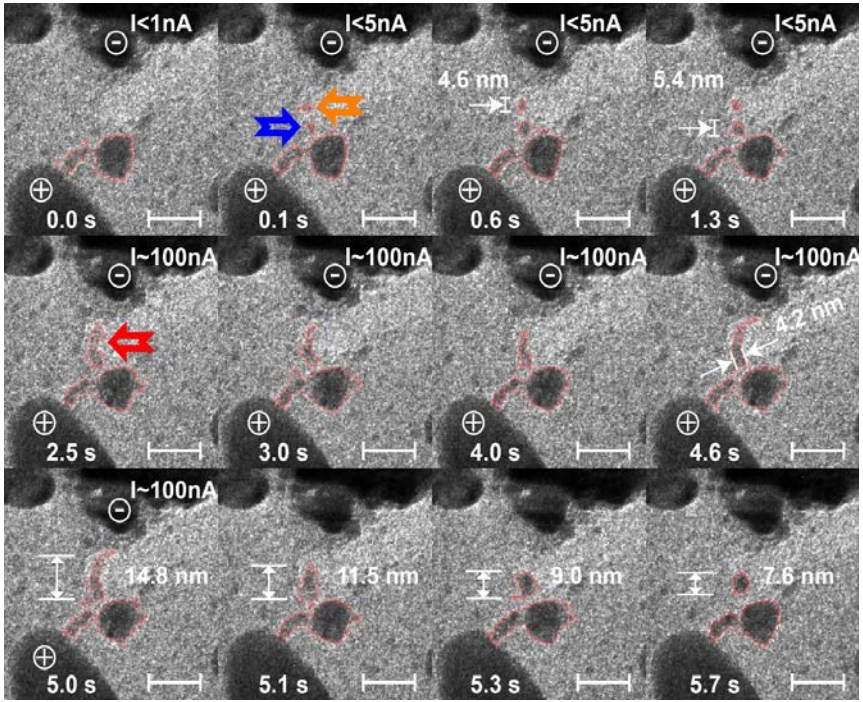
Ag in Metal Oxide Diffusive Memristor – Emulating Dynamics







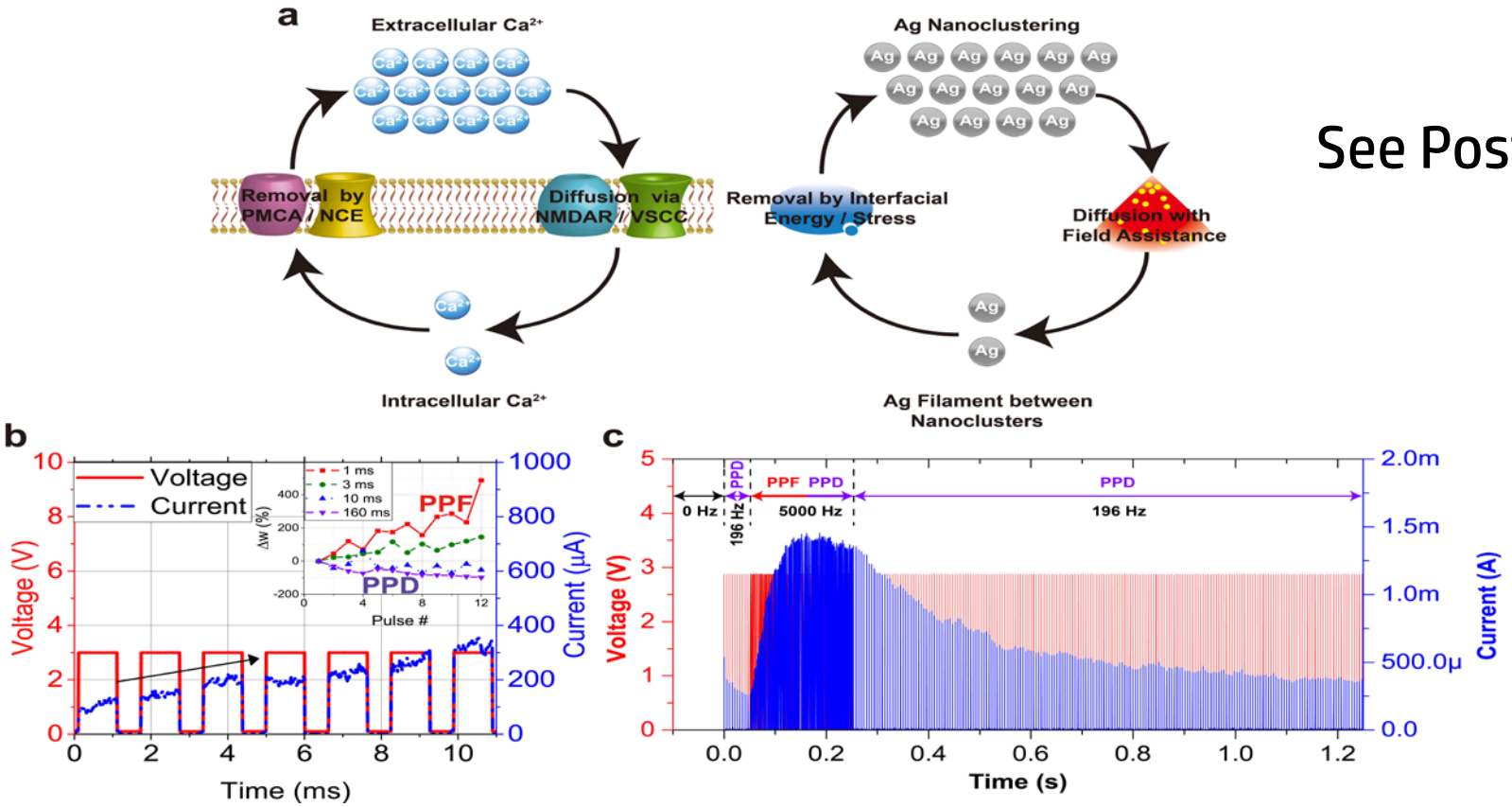




Z Wang, S Joshi, SE Savel'ev et al., Nature Materials 16 (1), 101-108 UMass Amherst

Hewlett Packard Enterprise

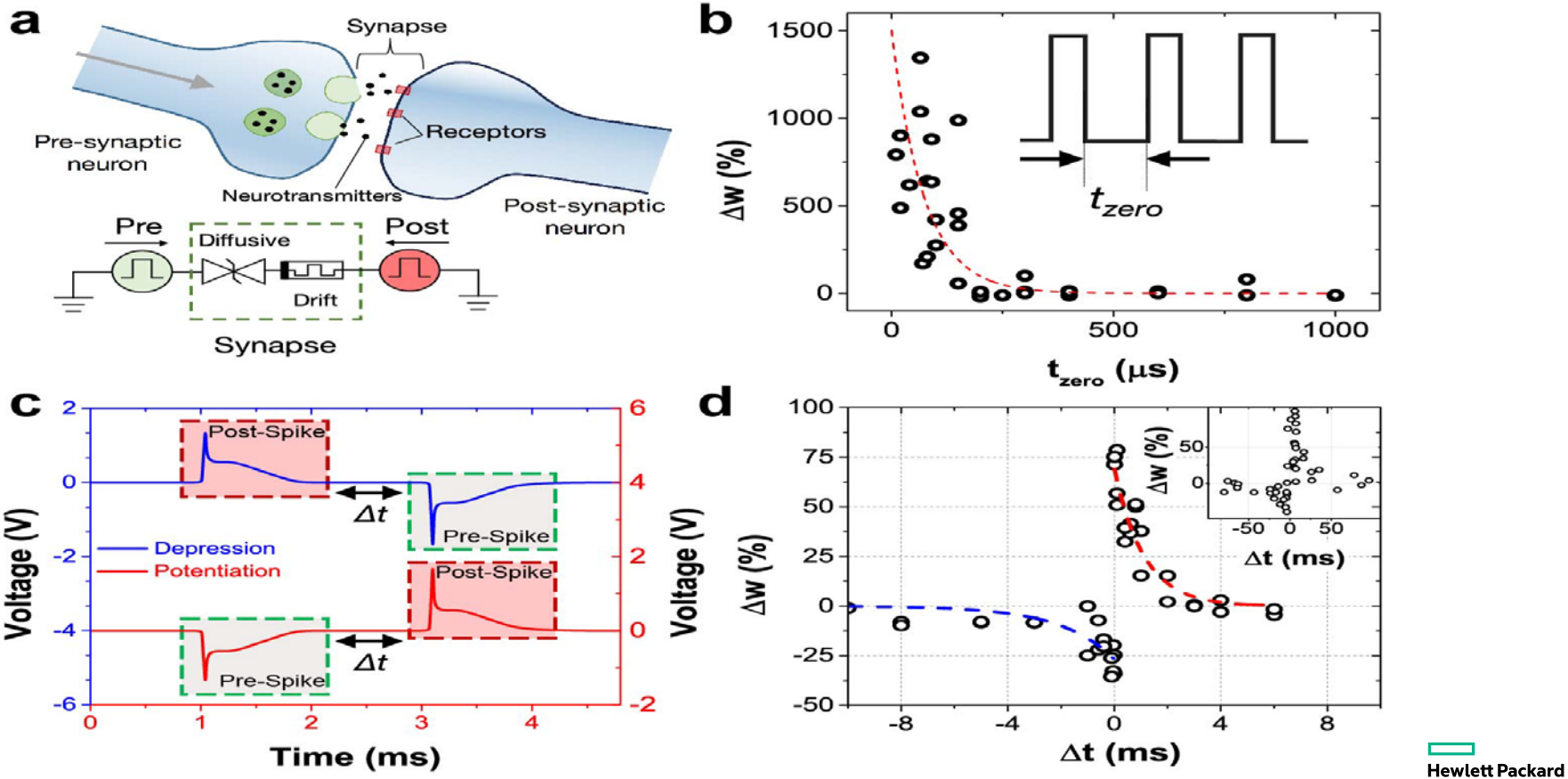
Ag Diffusive Memristor Mimics Ca^{2+} Ion Dynamics in Synaptic Gap



See Poster!

Z Wang, S Joshi, SE Savel'ev et al., Nature Materials 16 (1), 101-108 UMass Amherst

Diffusive Decay Provides an Extended Window for STDP



Z Wang, S Joshi, SE Savel'ev et al., Nature Materials 16 (1), 101-108 UMass Amherst

Modified Hopfield network for optimization problems

Turing believed that intelligence required randomness and would necessarily lead to mistakes

Chua has shown that neurons are ‘poised on the edge of chaos’, and that chaos leads to complexity and emergent behavior

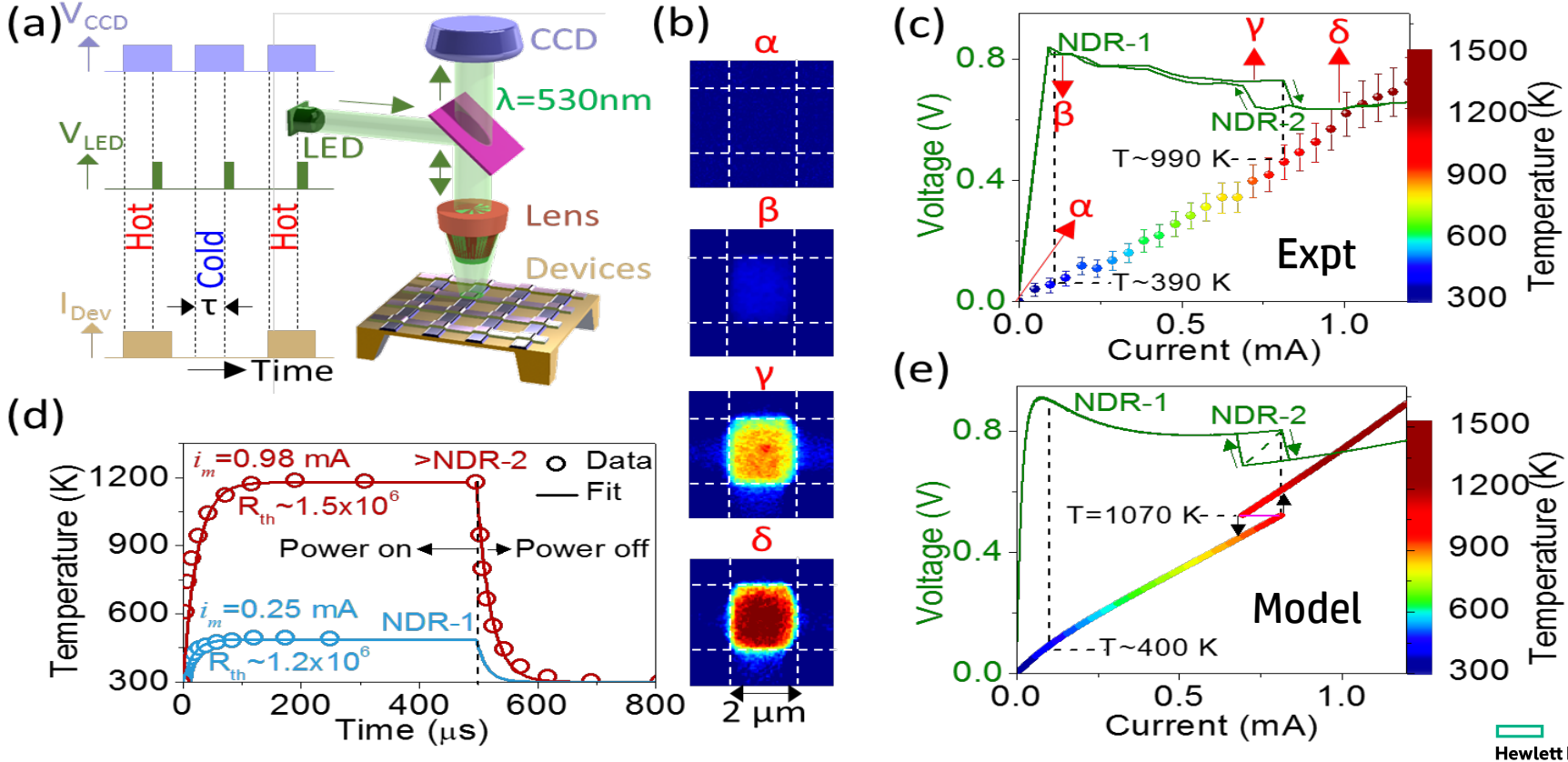
Epilepsy occurs when the neurons in the brain synchronize, i.e. lack of chaos

Chaos as a computing resource

May allow construction of a highly scalable “annealing” machine

Benchmarking to compare to other Classical and Quantum Annealing machines

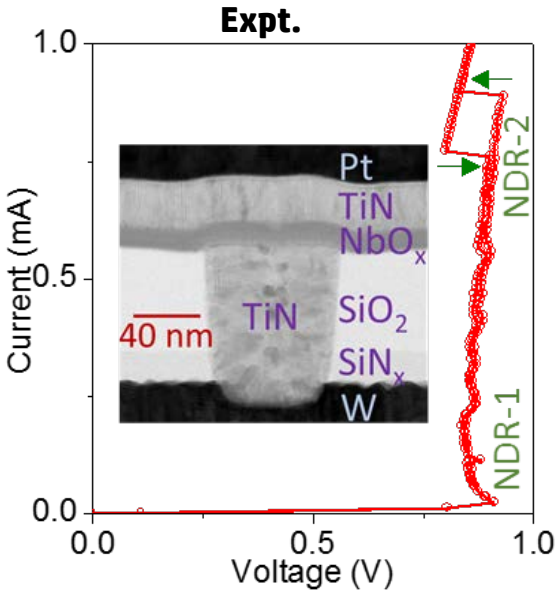
Thermoreflectance Imaging of NbO_2 current control behavior

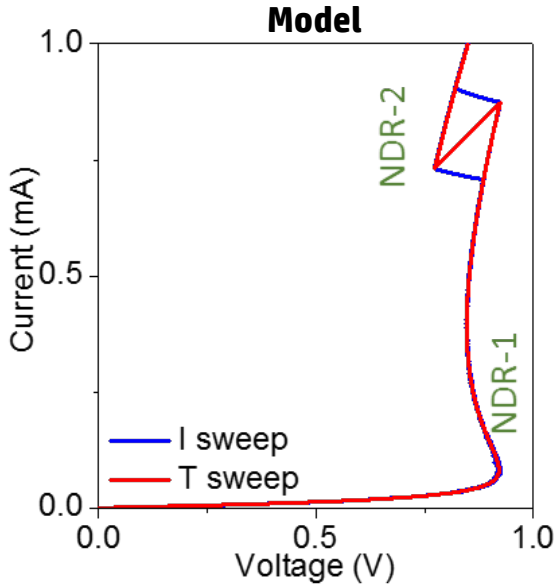


S Kumar et al., submitted for publication

Hewlett Packard Enterprise

Niobium oxide memristor: behavior and model





Model: Modified 3D Poole-Frenkel

$$i_m = \left[2A \times 10^4 e^{-\frac{0.301}{k_B T}} \left(\frac{k_B T}{\beta\sqrt{v_m}}\right)^2 \left\{1 + \left(\frac{\beta\sqrt{v_m/t}}{k_B T} - 1\right) e^{\frac{\beta\sqrt{v_m/t}}{k_B T}}\right\} + \frac{2A \times 10^4 e^{-\frac{0.301}{k_B T}}/t}{2}\right] v_m$$

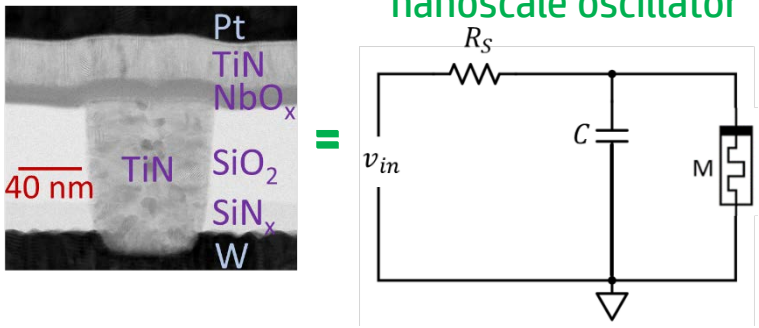
Temperature dynamics

$$\frac{dT}{dt} = \frac{i_m v_m}{C_{th}} - \frac{T - T_{amb}}{C_{th} R_{th}(T)}$$

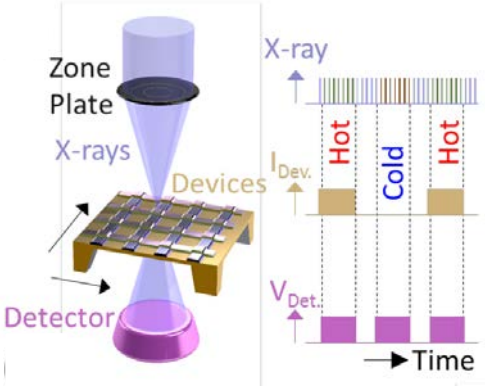
Mott transition in R_{th}

$$R_{th}(T) = 1.4 \times 10^6 \text{ (for } T \le T_C) \text{ and } 2 \times 10^6 \text{ (for } T > T_C)$$

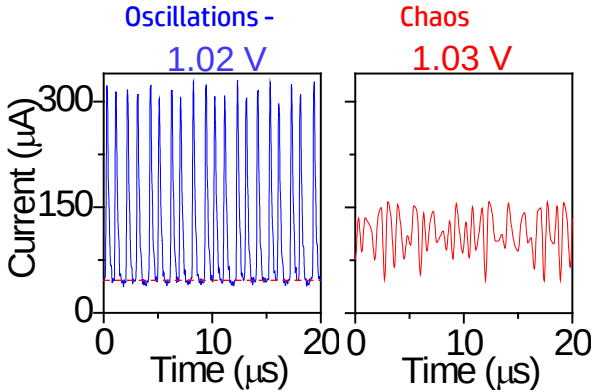
NbO_2 memristor: computing with chaos



in situ and *in operando* scanning transmission x-ray microscopy (STXM) at ALS critical for analysis of memristor operation

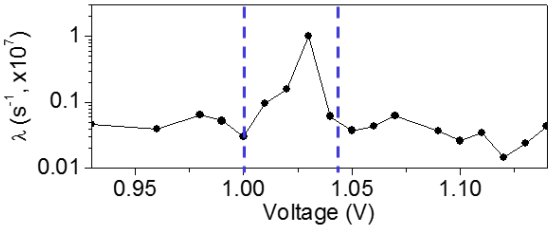


Dynamical Behavior:

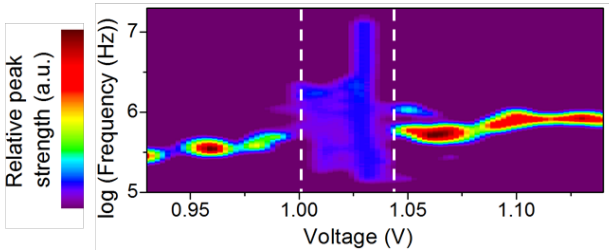


Lyapunov exponent

$$\lambda = \lim_{t\to\infty} \lim_{\delta \mathbf{Z}_0 \to 0} \frac{1}{t} \ln \frac{|\delta \mathbf{Z}(t)|}{|\delta \mathbf{Z}_0|}$$



Frequency analysis



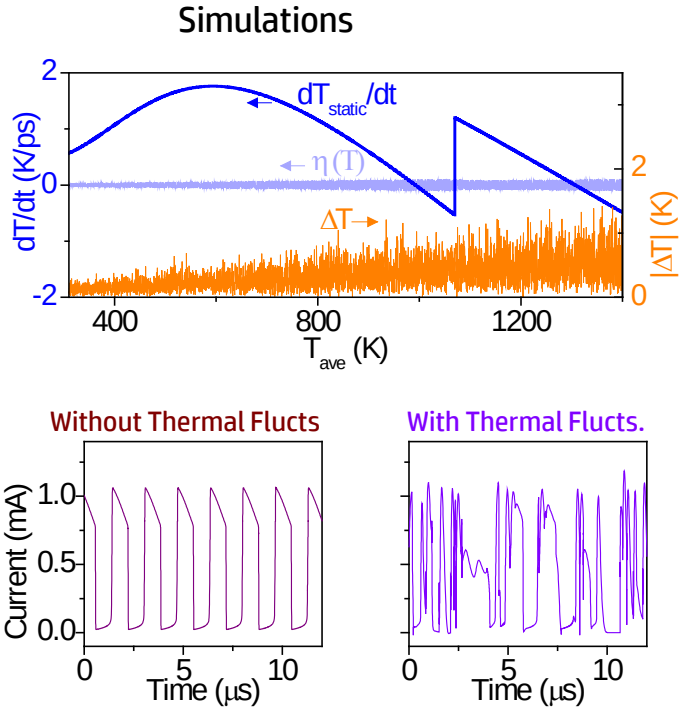
Able to describe by incorporating thermal fluctuations

$$\eta(T) = T\left(\frac{k_B}{C_{th}}\right)^{\frac{1}{2}} \frac{4\pi}{R_{th}C_{th}} rand(-1 \leftrightarrow 1)$$

Negative Differential Resistance in the system leads to positive feedback in presence of fluctuations

Thermal capacitance C_{th} shrinks with device size

→ increasing importance of thermal fluctuations

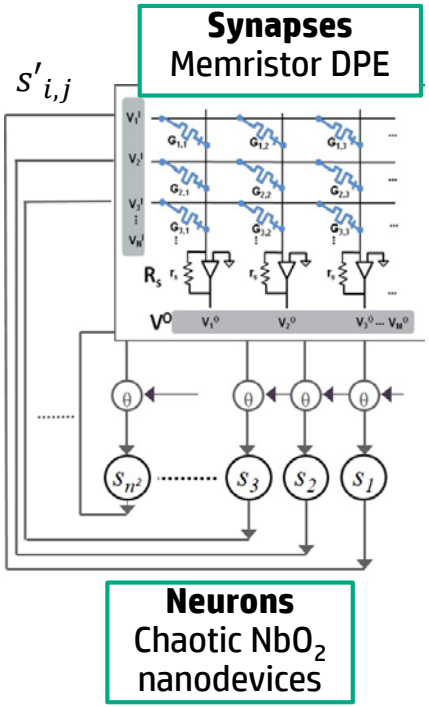


Hewlett Packard Enterprise

Modified Hopfield network for optimization problems

Traveling Salesman problem

NP hard: $O(n^3 2^n)$

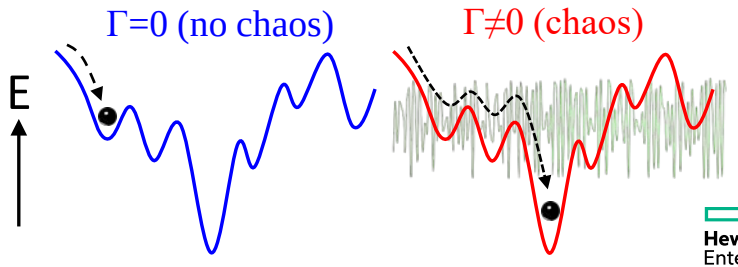


Encode any TSP instance in the DPE weight matrix

Defines an "energy" of the system to be minimized

$$E = -\frac{1}{2}\sum_i \sum_j s_{i,j} \sum_k \sum_l s_{k,l} W_{(i,j),(k,l)} + \sum_i \sum_j s_{i,j}\theta$$

Follow update rule: $s_{i,j} = \begin{cases} 1 \; if \; Ws'_{i,j} > \theta \\ -1 \; if \; Ws'_{i,j} < \theta \end{cases}$



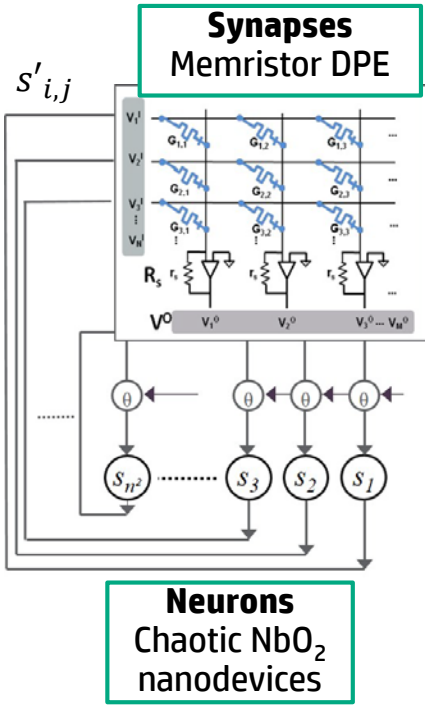
S Kumar, et al. Nature, in press – on line Aug. 9

Modified Hopfield network for optimization problems

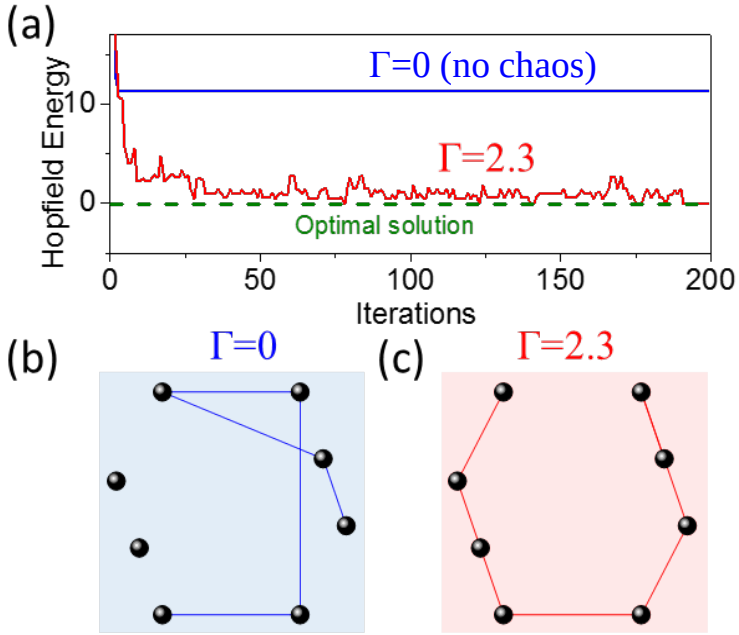
Traveling Salesman problem

$O(n2^n)$?

still NP hard



Example solutions w/ and w/o chaos



S Kumar, et al. Nature, in press

Hewlett Packard Enterprise

Acknowledgments

HPE Labs

John Paul Strachan
Miao Hu
Suhas Kumar
Catherine Graves
Emmanuelle Merced
Dejan Milojicic
Ali Shafiee
Naveen Muralimanohar
Gary Gibson
Max Zhang

UMASS AMHERST

J. Joshua Yang
Qiangfei Xia
Can Li
Yunning Li
Saumil Joshi
Zhongrui Wang

Izzat El Hajj (Univ Illinois Urbana-Champagne)
Pedro Bruel (IME, Universidade de São Paulo)
Sergey Saveliev (Loughborough U, UK)

Funding assistance

Hewlett Packard Enterprise