Future opportunities in high performance and low power computing with emerging technologies and novel architectures

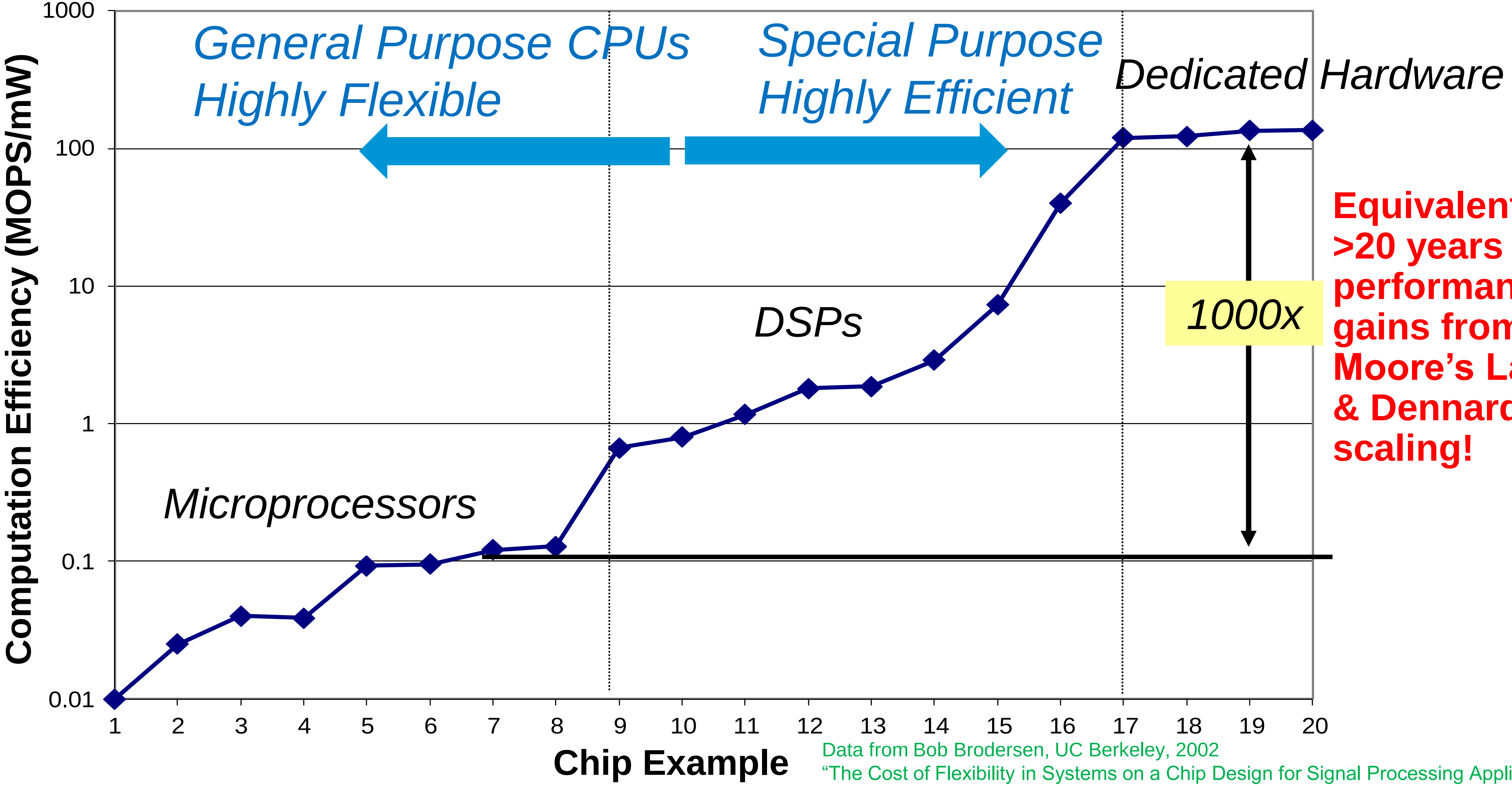
John Paul Strachan
Hewlett Packard Labs, HPE

LETI Devices Workshop – December 2, 2018

Outline

- The rise and demand for efficient accelerators
- The memristor-based accelerator for A.I./Machine Learning
- Future opportunities: brain-inspired approaches and alternatives to quantum computing

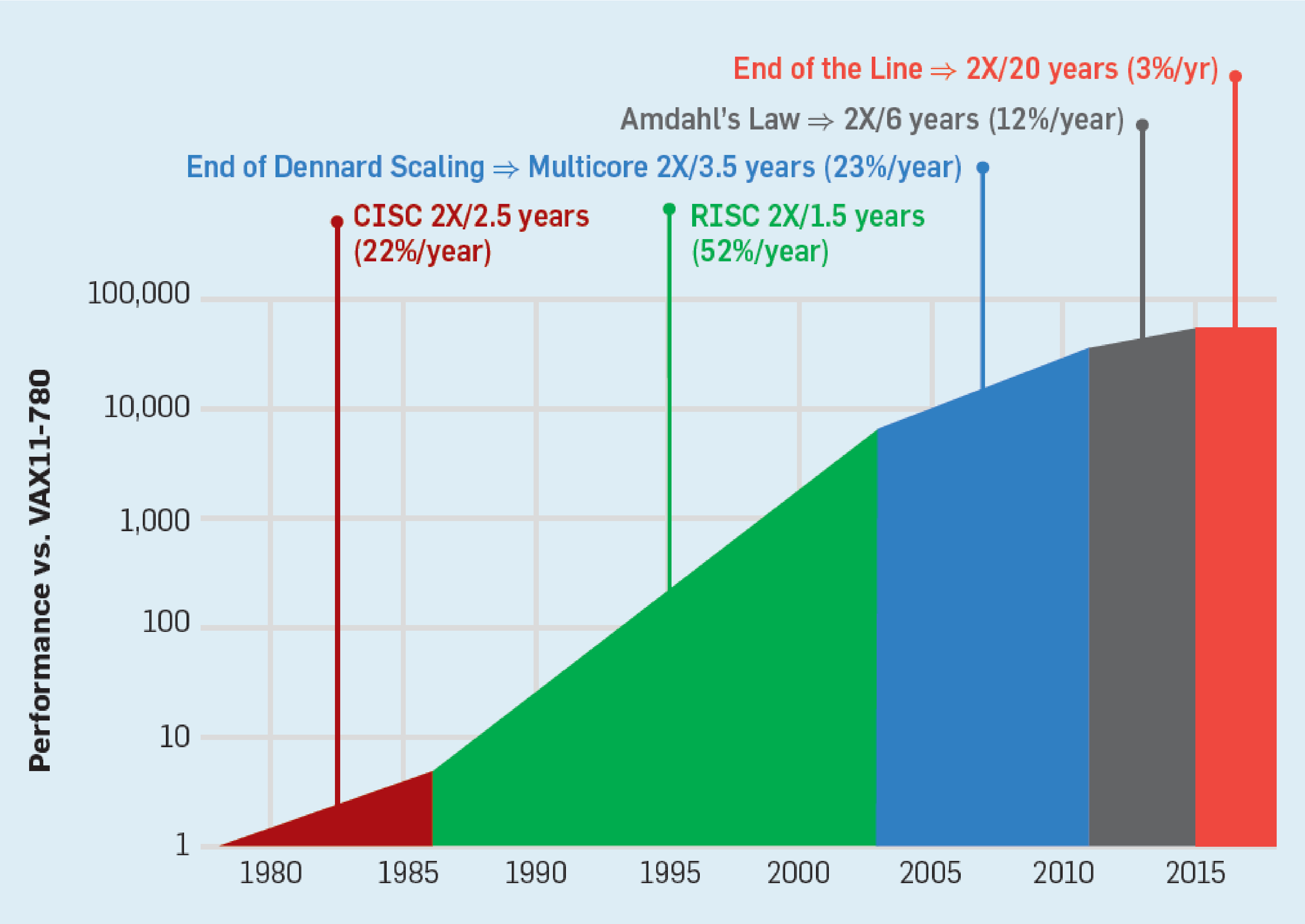
HW accelerators – increased performance for special cases



Equivalent to >20 years performance gains from Moore's Law & Dennard scaling!

Data from Bob Brodersen, UC Berkeley, 2002
"The Cost of Flexibility in Systems on a Chip Design for Signal Processing Applications"

Unlike before, we work hard for limited performance gains



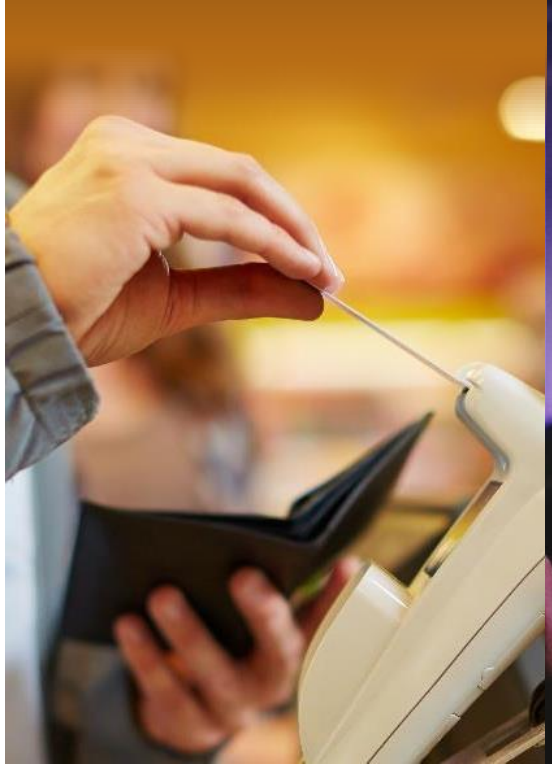
Source: Hennessy and Patterson

Some Key Drivers for Specialization: Data Explosion & AI

Structured data

40 petabytes

Walmart's transaction database (2017)



Human interaction data

4 petabytes a day

Per-day posting to Facebook across 2 billion users (2017)

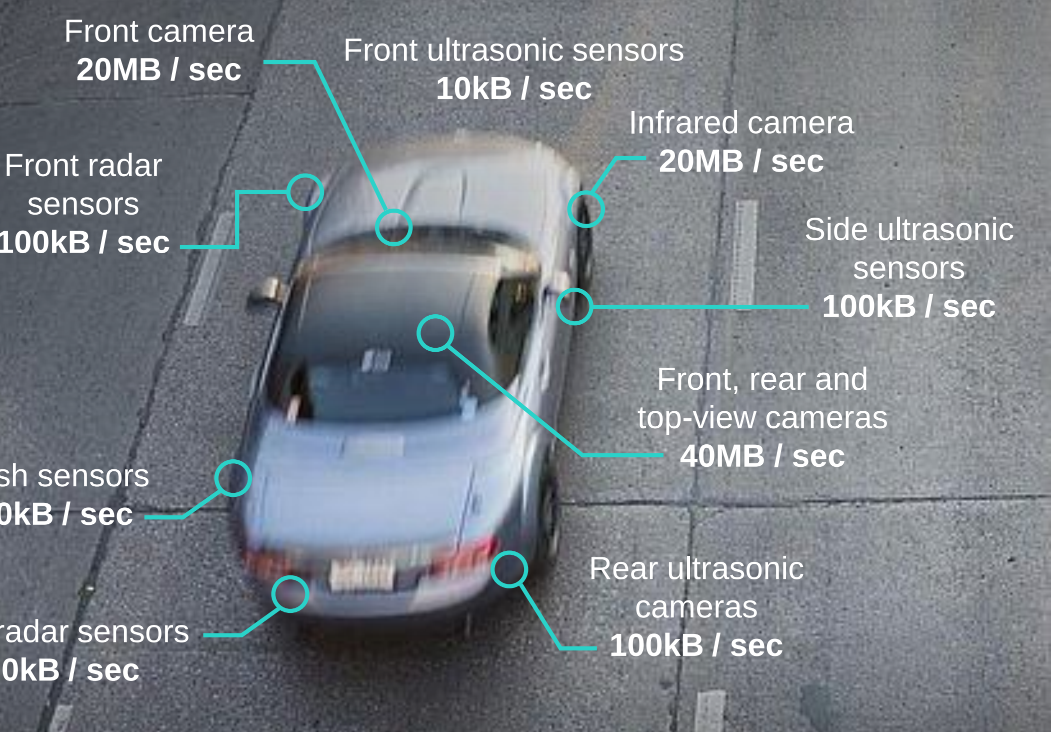
2MB per active user

The world is replacing programming with training

Digitization of analog reality

40,000 petabytes a day*

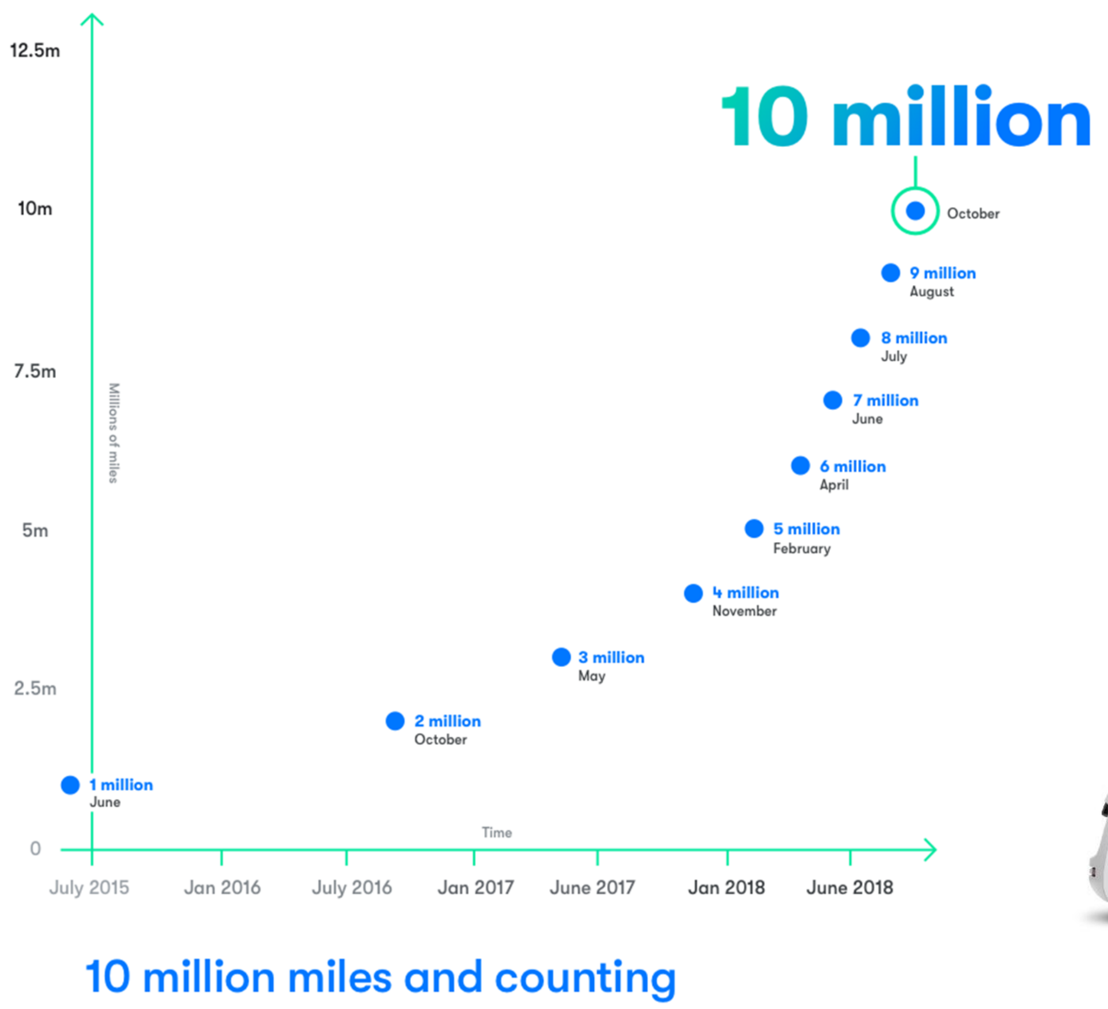
10m connected cars by 2020



* Driver assistance systems only

Hewlett Packard Enterprise

Motivating example: Autonomous/Assisted Driving



- 4 TB/day per instrumented vehicle
- 1 PB/day for a 250-car fleet
- Not practical to move all data to the Data Center

Need all sorts of accelerators at the Edge!

Hewlett Packard Enterprise

But we need Billions of miles for safety

How many miles (years[a]) would autonomous vehicles have to be driven...	(A) 1.09 fatalities per 100 million miles?
(1) without failure to demonstrate with 95% confidence that their failure rate is at most...	275 million miles (12.5 years)
(2) to demonstrate with 95% confidence their failure rate to within 20% of the true rate of...	8.8 billion miles (400 years)
(3) to demonstrate with 95% confidence and 80% power that their failure rate is 20% better than the human driver failure rate of...	11 billion miles (500 years)

Source: RAND Corp." Driving to Safety"

Source: Cognata

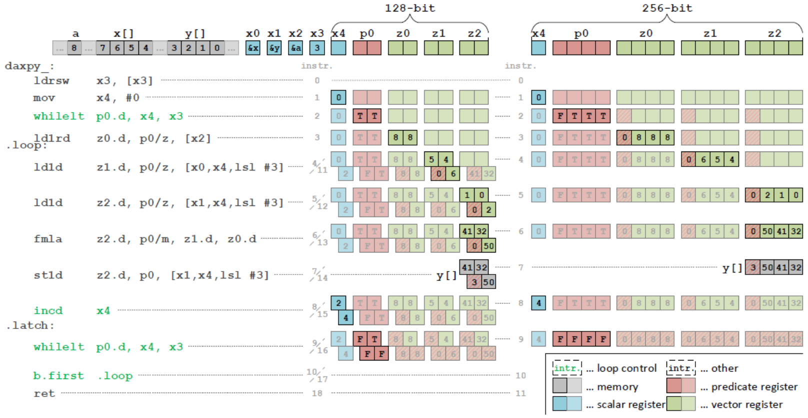
Safe, autonomous vehicles depend on billions of miles of <u>simulated</u> driving

Need for accelerators in the Data Center!

Conventional accelerators

CPU extensions
ISA-level acceleration

- Vector and matrix extensions
- Reduced precision
- Example: ARM SVE2

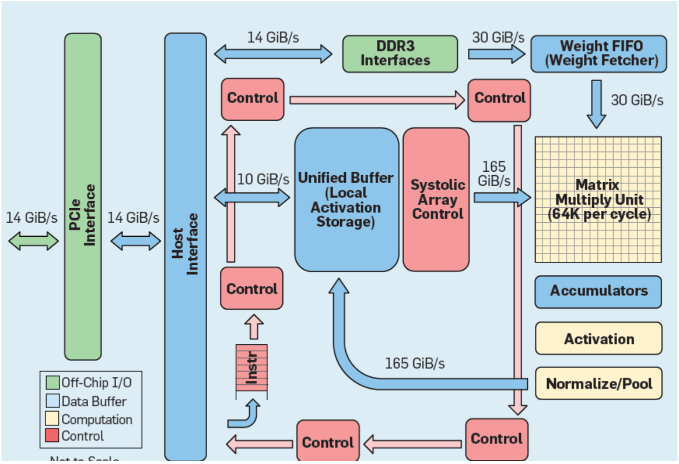


GPUs
Data parallel calculations

- Optimized for throughput
- High-bandwidth memory
- Example: Nvidia, AMD

Deep Learning Accelerators
ASIC-like flexible performance

- Data-flow inspired, systolic, spatial
- Cost optimized
- Example: Google's TPU, FPGAs

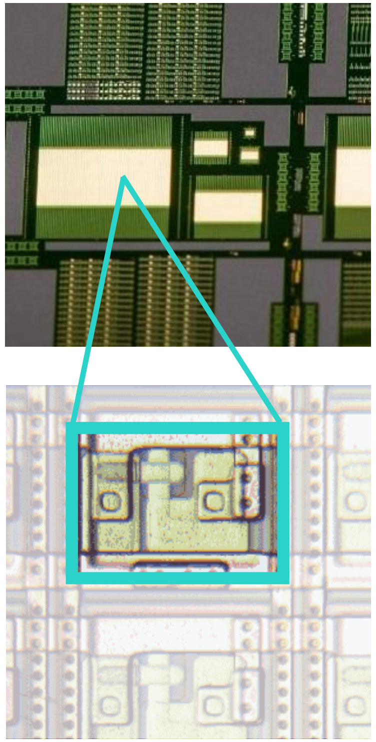


Unconventional accelerators

Analog neuromorphic computing

Massive speedup for AI training and inference

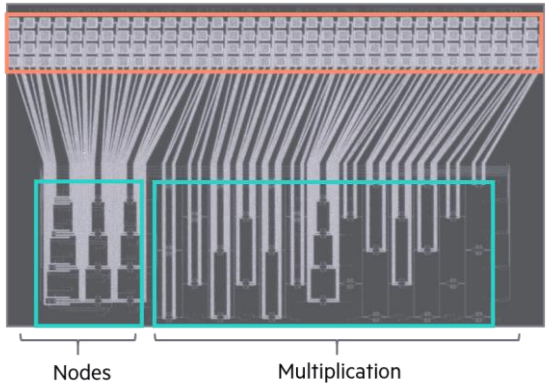
- Complex matrix calculations in one step
- 10-100x faster
- 10-1000x more energy efficient
  (Compared to GPU)



Optical Computing

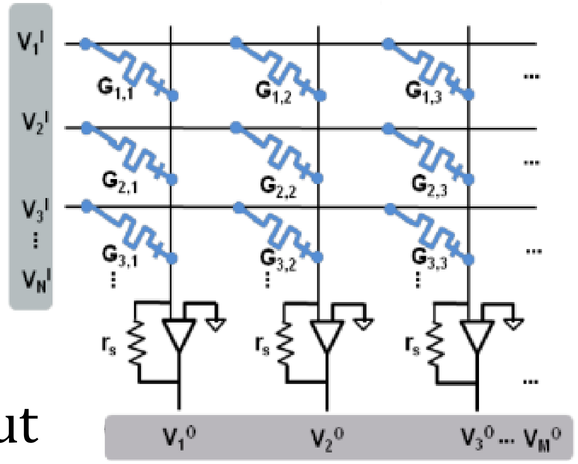
Designed for “unsolvable” optimization problems

- Harnessing the properties of light at the microscale
- Prototype has world record 1,000 optical components
- Scalable to 100,000 components



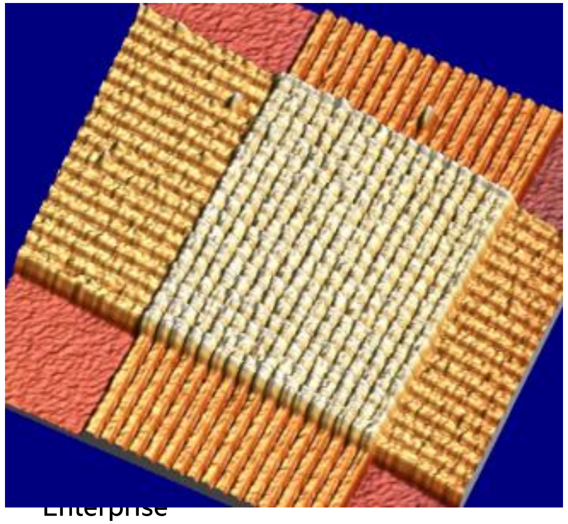
The memristor Dot Product Engine (DPE)

Input Voltage vector V_i



Output current vector

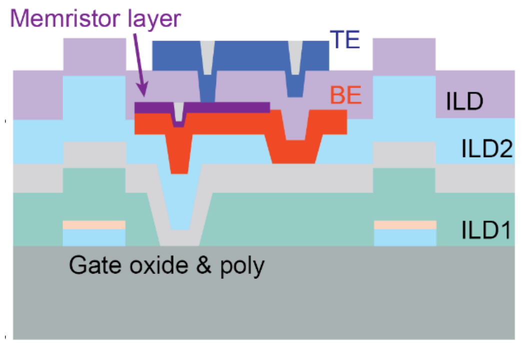
$$I_j = \sum_i G_{ij} \cdot V_i$$

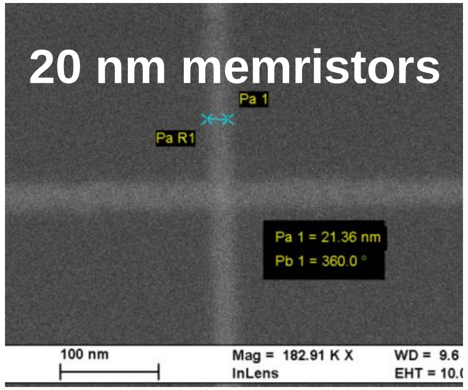


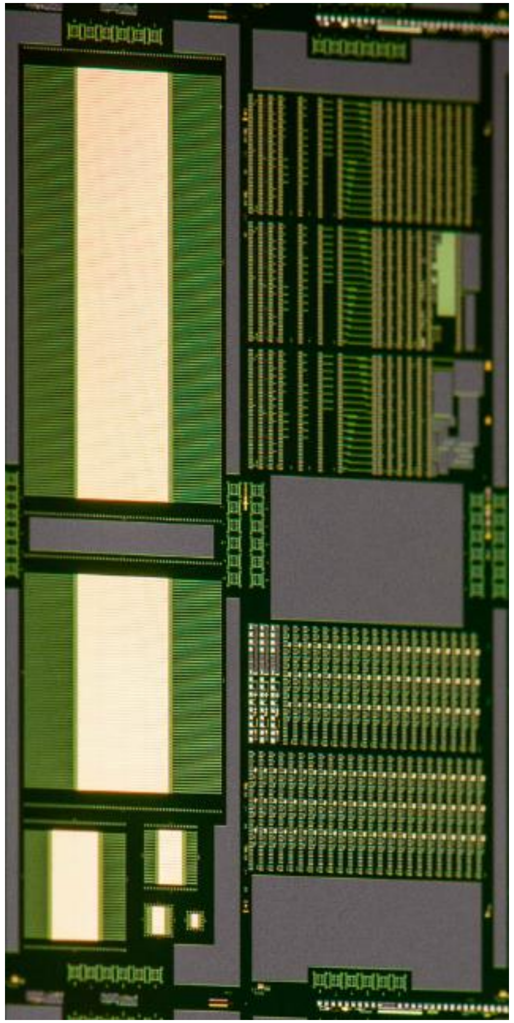
- Harness memristors in dense crossbar arrays
- Memristor = non-volatile, analog memory cell
- Parallel activation of every row and column in crossbar
- Vector-matrix multiplication (VMM) in a single cycle
- Computing = read operation
- Efficient multiply & add in <u>analog domain</u>
- Key advantage is <u>in-memory processing</u>

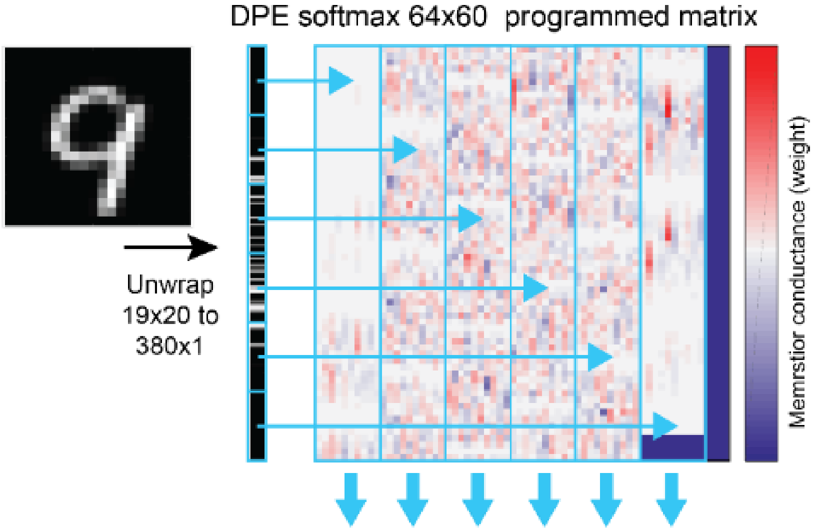
Dot Product Engine: working prototype chip

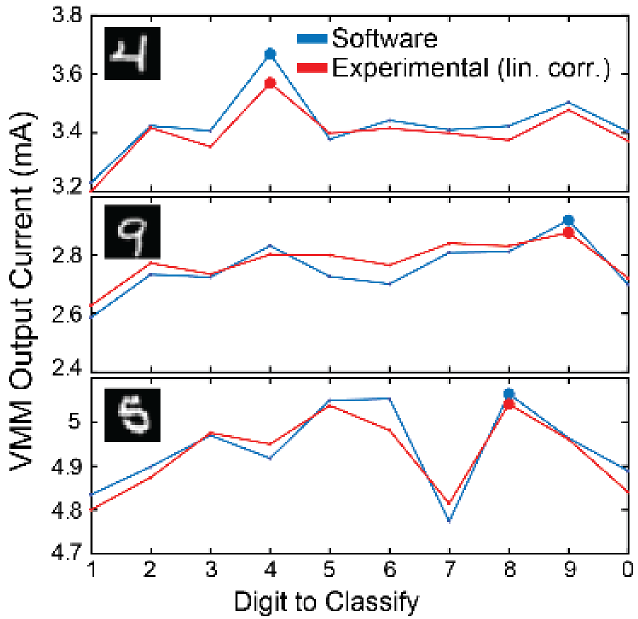
Back-end (BEOL) integration of memristors with CMOS











Successful MNIST Neural Network inference with memristor-based analog computing

M. Hu, et. al, Adv. Mater. *2018*

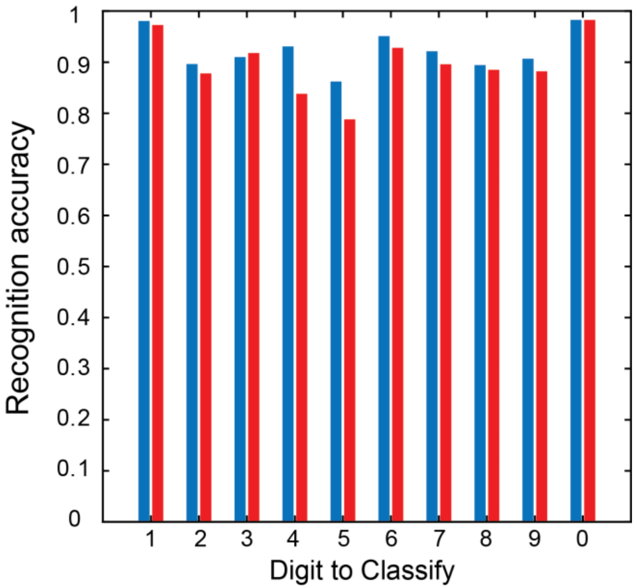
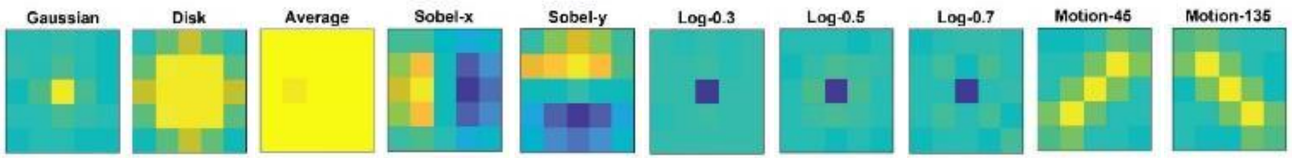
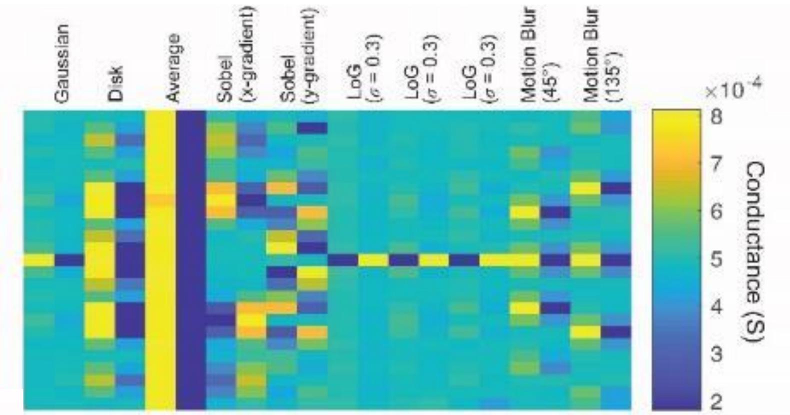


Image Processing on memristor-DPE system

10 different 5x5 convolutions

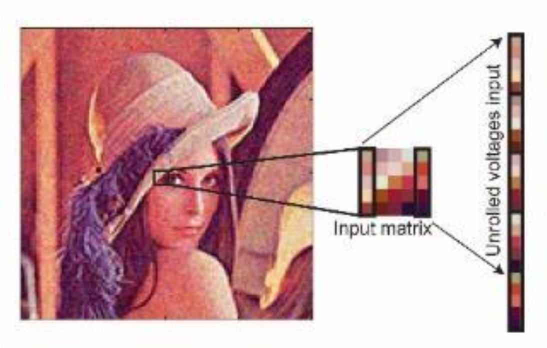


Unroll each into a 25 (=5x5) element column



Experimental conductance pattern of memristor array

Input image
5x5 portions at a time
Apply 25 Voltages to array



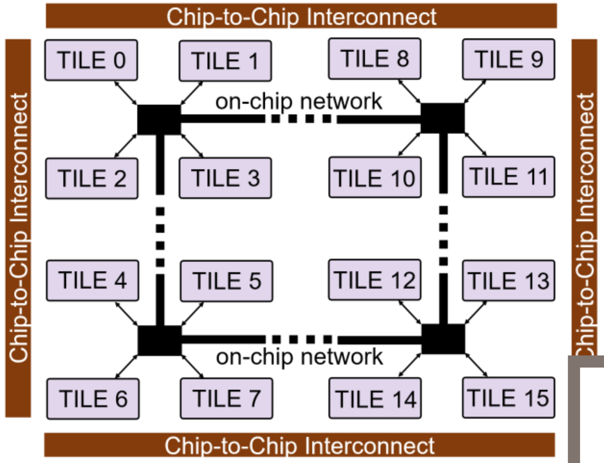
Experimental outputs – 10 filtered images output in parallel

Reduces computations from $O(Cm^2n^2)$ operations to $O(n^2)$

C. Li, et. al, *Nature Electronics*, (2018)

System Architecture, Compiler, & Software Support

- Developed Architecture supporting all state-of-the-art neural networks (CNN, LSTM, MLPs, RBMs, etc.)
- Developed an "Assembly" code (ISA) for our memristor accelerator
- Built a compiler, with support for standard ONNX format

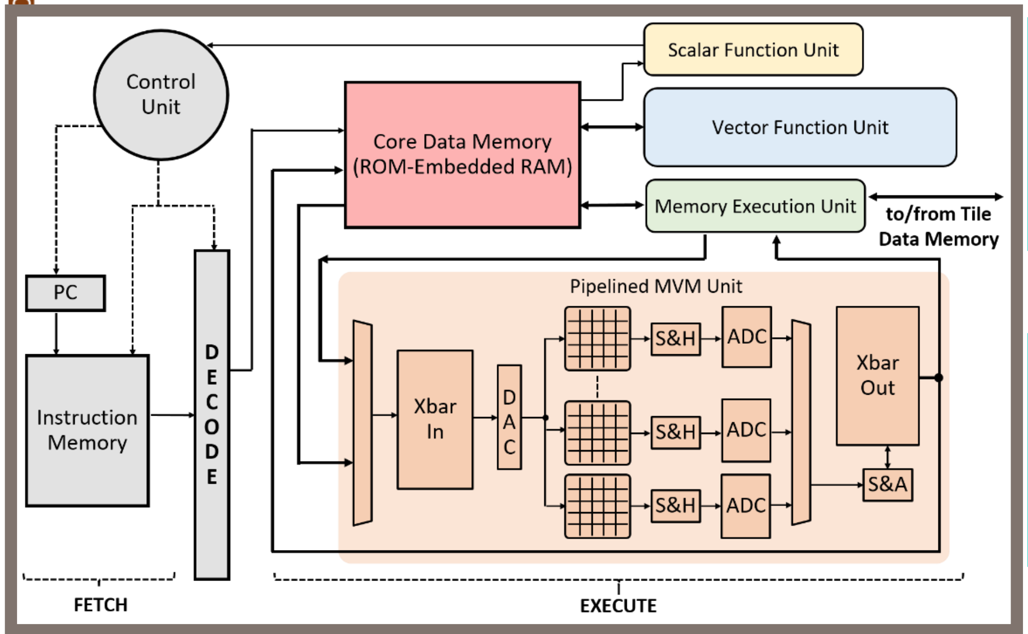


A. Ankit, et. al, *ASPLOS*, (2019)

Architecture: PUMA – Programmable Ultra-efficient memristor-based Accelerator

10-100k memristor xbars (128x128) performing matrix vector multiplications

Digital units for other operations (logic, scalar, and vector units). 3-stage pipeline, instruction decoder, and instruction memory.



Application Layer

- Neural Network specification (ONNX) – CNN, LSTM, etc

Compiler

- Convert to DPE Assembly; Map to crossbars

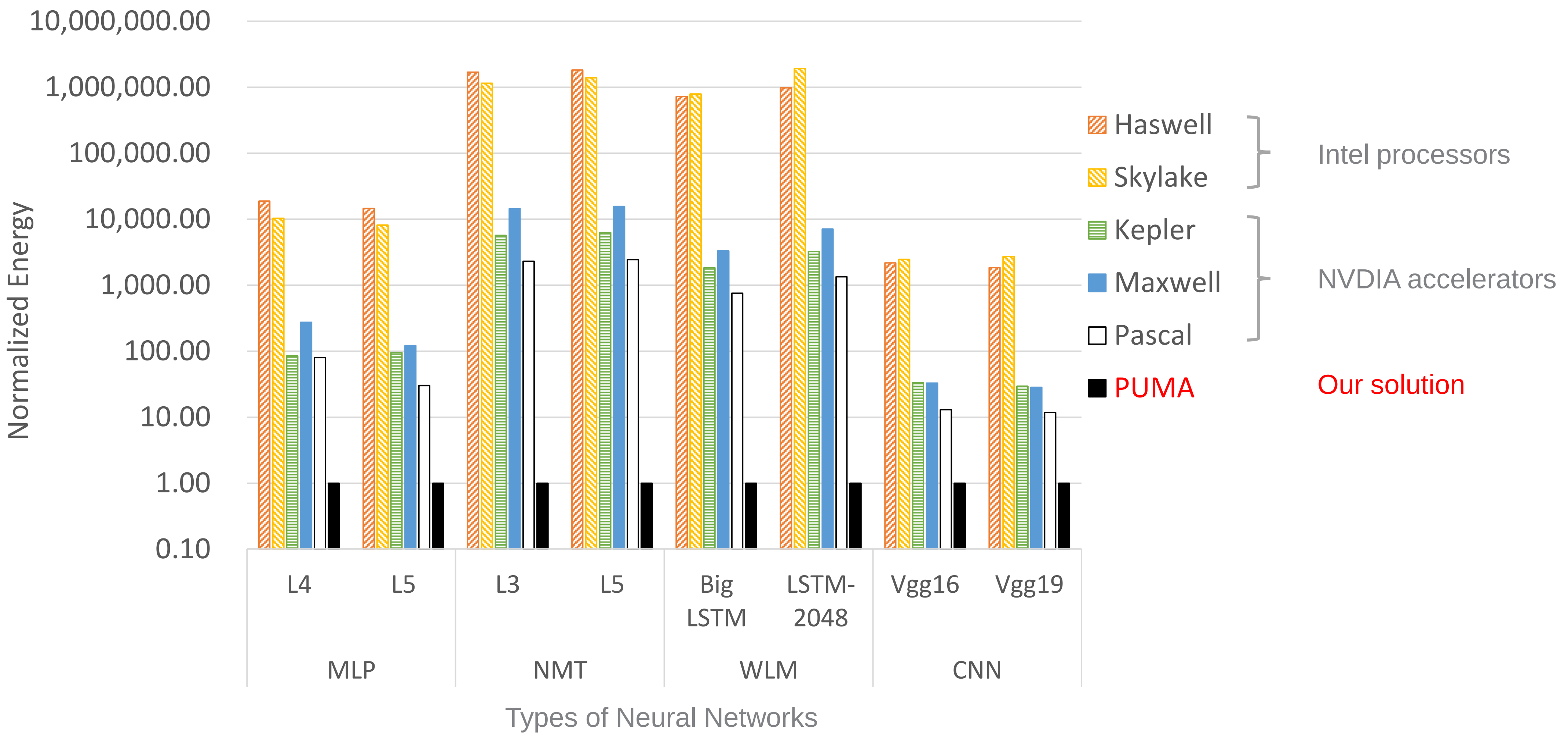
Simulator

- Provide performance metrics (accuracy, energy, latency, etc.)

Hewlett Packard Enterprise

Benchmarking

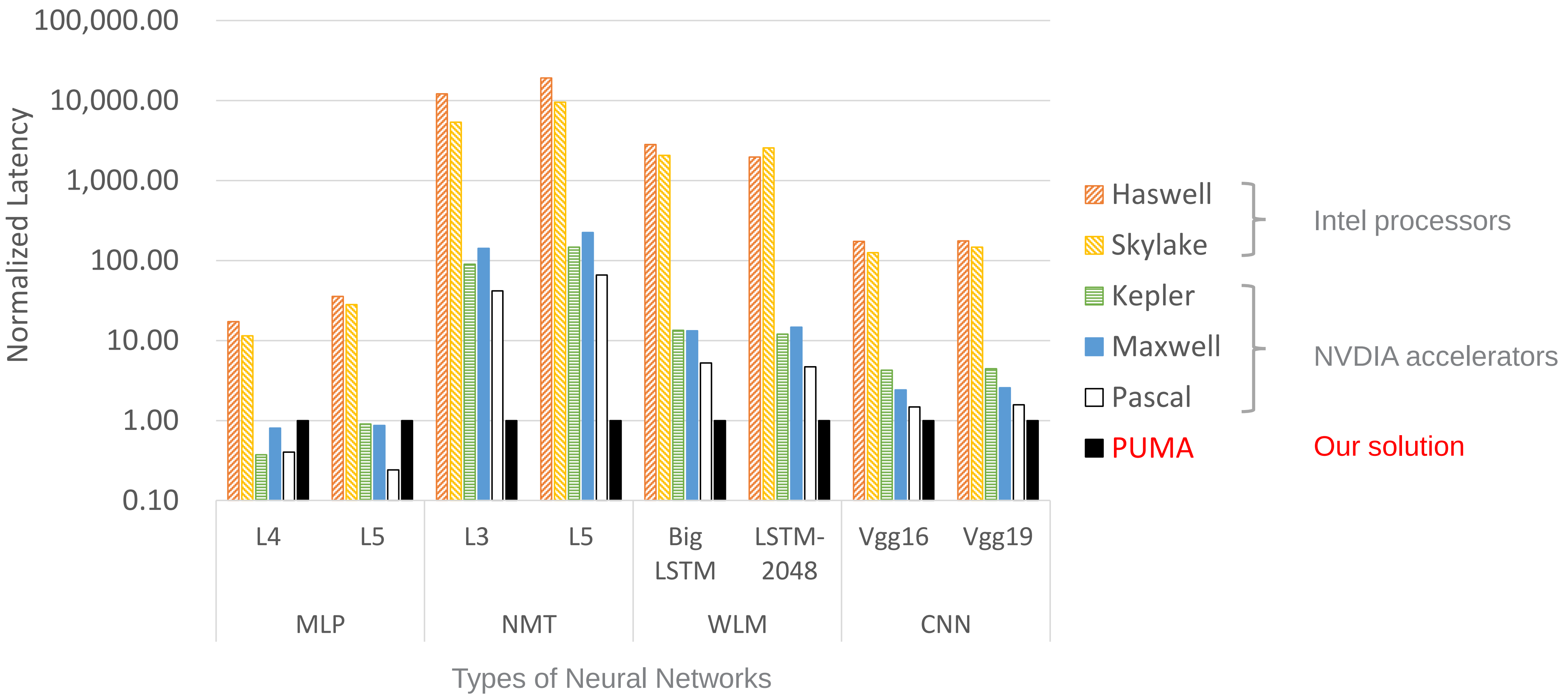
Inference energy normalized to PUMA (lower is better)



Lower energy than CPUs (1,000 - 1,000,000x) and NVIDIA GPUs (10 - 1,000x)
Larger networks (NMT, WLM) benefit the most

Benchmarking

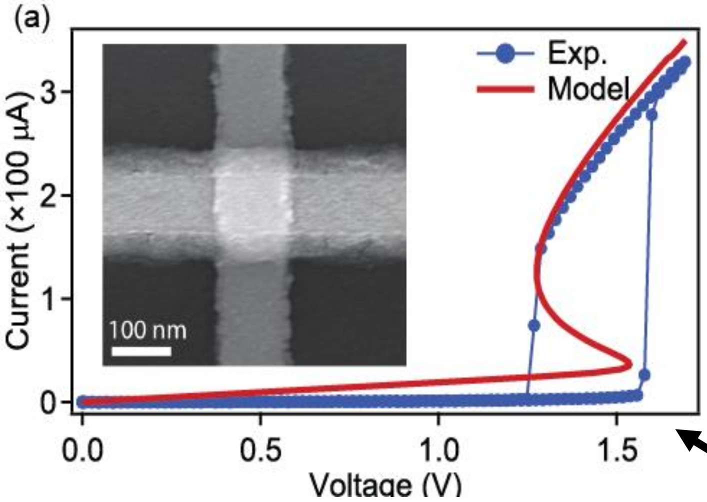
Inference latency normalized to PUMA (lower is better)



Lower latency than CPUs (10-10,000x) and NVIDIA GPUs (10-100x)
Larger networks (NMT, WLM) benefit the most

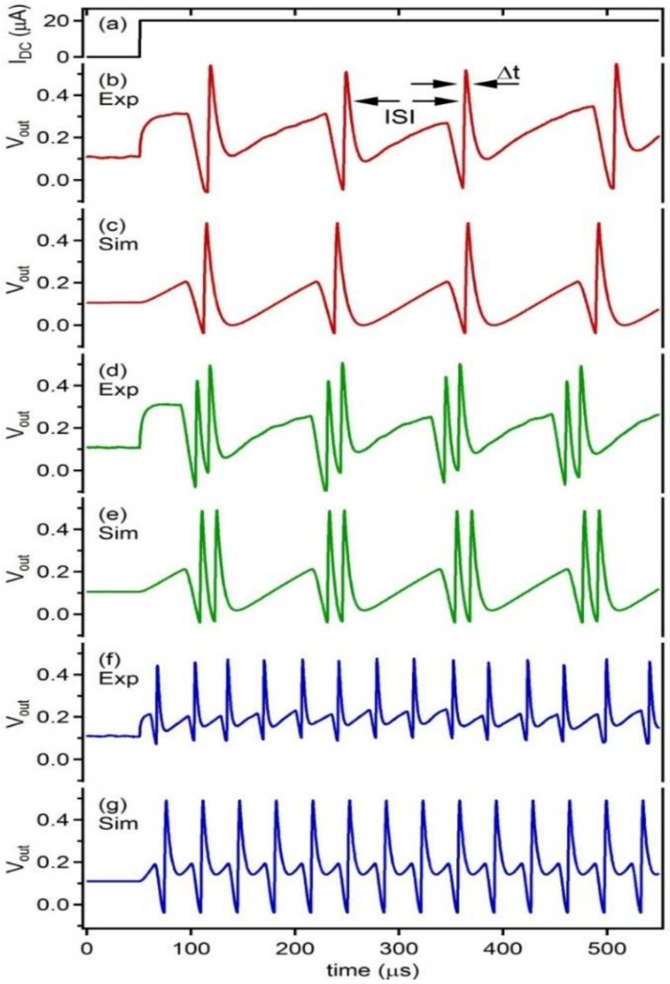
Future opportunities: brain-inspired approaches as alternative to quantum computing

Memristors also provide neuron-like behavior



Can build a “neuronic” circuit element from a memristor (NbO_2 device shown here)

Directly emulates signals seen in brains



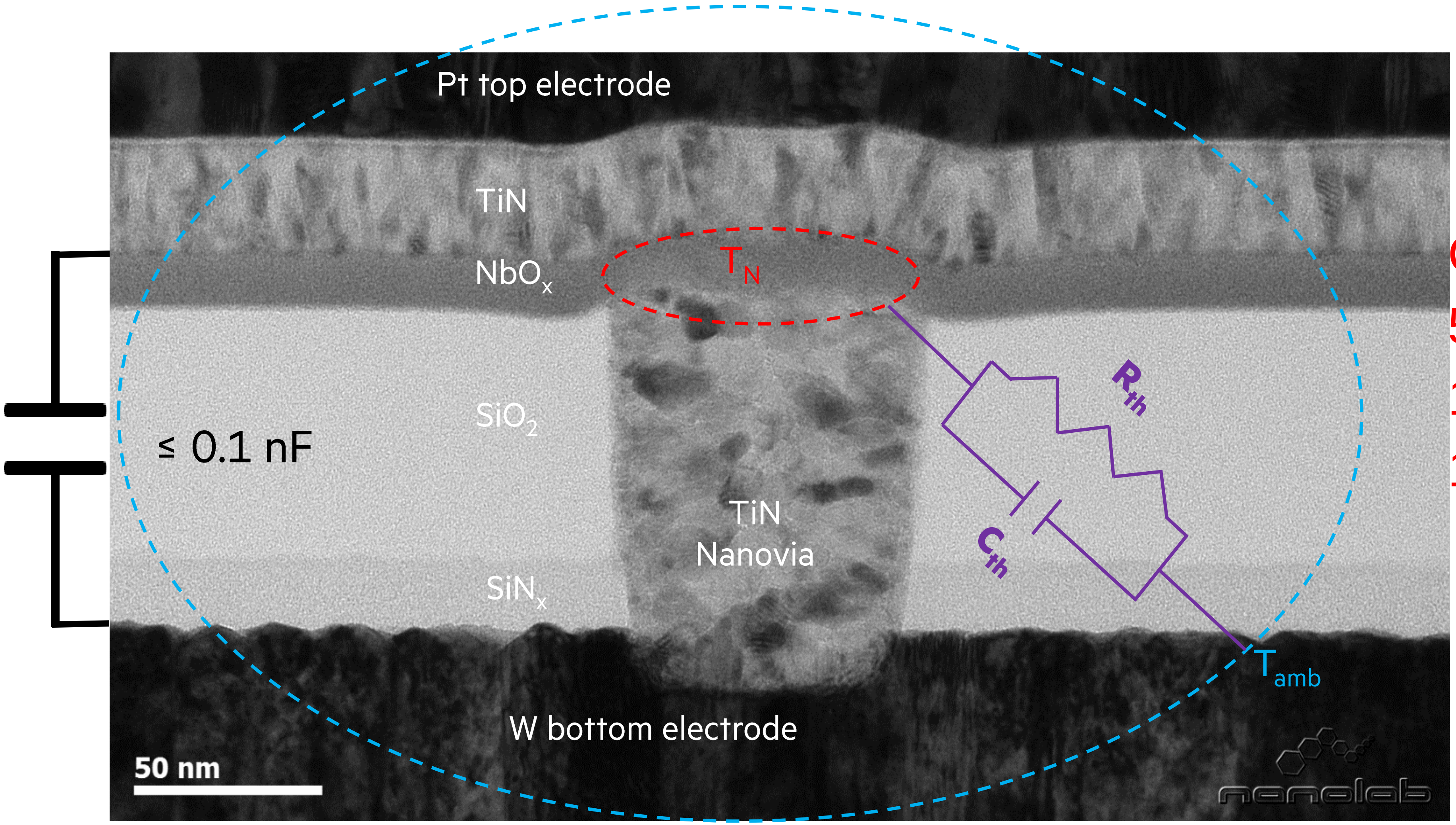
“Regular Spiking”
C_1=5.1 nF, C_2=0.75 nF

“Chattering”
C_1 = 5.1 nF, C_2 = 0.5 nF

“Fast Spiking”
C_1=1.6 nF, C_2=0.5 nF

M.D. Pickett, et al. *Nature Materials* 2013

Highly compact artificial neuron



Compared to brain:

500x frequency

100x less energy/spike

100 nm vs 100 µm

Dark field cross-sectional TEM image of NbO_x memristor $R_{th}C_{th} \leq 0.1$ ns

Apply to Important Optimization Problems

NP-hard and NP-complete problems:

For a problem of size N, running time or memory use grows >> exp(N)

Important Graph Problems:

“Set Cover” - applies to airline flight scheduling

“Traveling salesmen” – UPS, shipping

“Max-cut” – applies to VLSI layout, routing

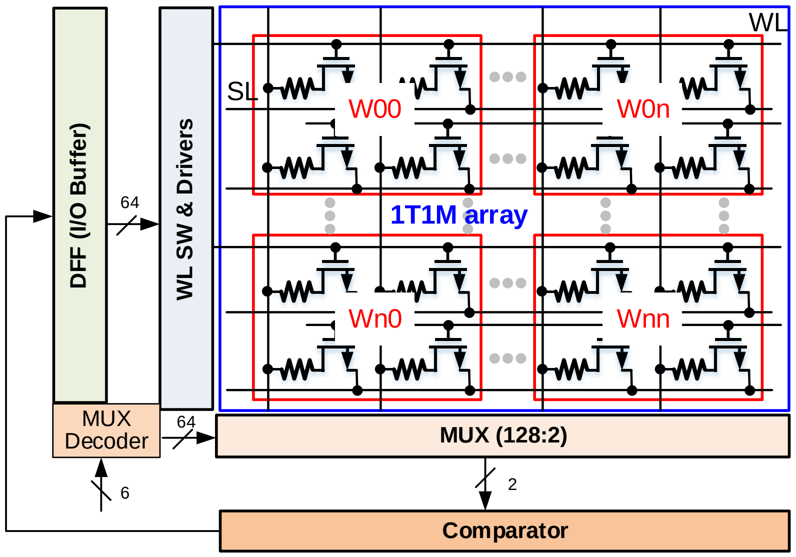
Example :

Every year, the National Football League (NFL) builds their 256-game schedule for the next season

- ➢ Have to consider team match-ups, stadium usage by other events, traffic, etc.
- ➢ Takes ~3months on a 1000-core system to solve!

**Source: Gurobi CEO Edward Rothberg*

Optimization Accelerator: memristor- Hopfield Network



Synapses w_{ij}
Memristor DPE

+

Neurons s_i
Memristors with Non-Linear threshold

Traveling Salesman problem (TSP):

Find shortest route visiting all cities

Encode any TSP instance in the DPE xbar

Defines an “energy” of the system to be minimized

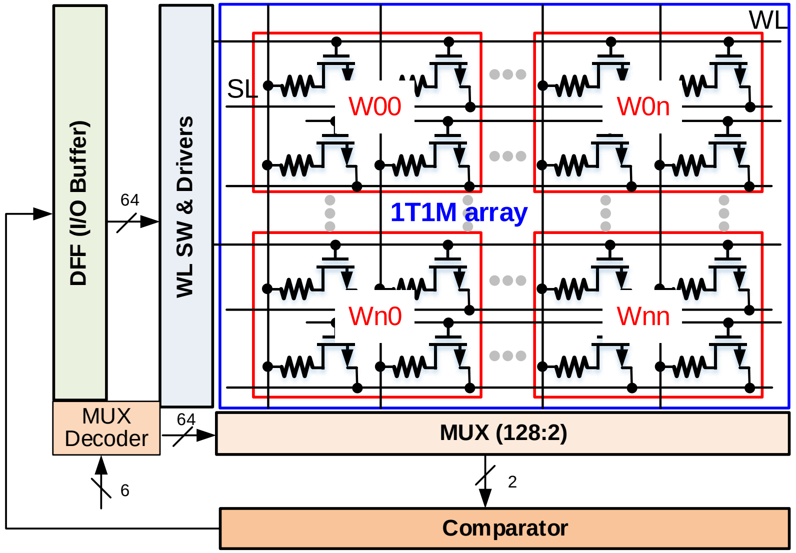
$$E = -\frac{1}{2}\sum_i \sum_j s_{i,j} \sum_k \sum_l s_{k,l} w_{(i,j),(k,l)} + \sum_i \sum_j s_{i,j}\theta$$

Follows simple 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 (2017)

Hewlett Packard Enterprise

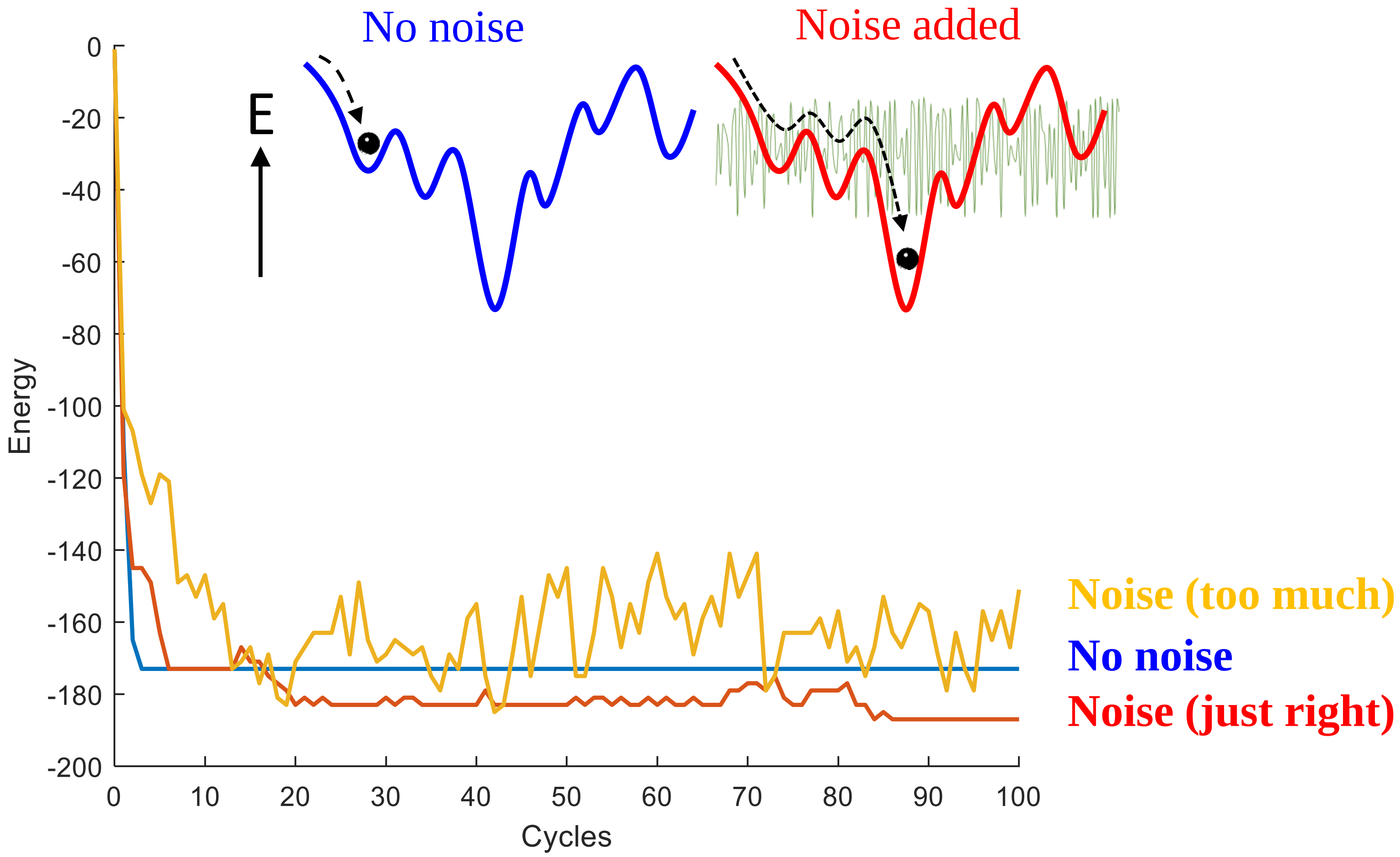
Optimization Accelerator: memristor- Hopfield Network



Synapses
Memristor DPE

+

Neurons
Memristors with Non-Linear threshold



S Kumar, et al. Nature (2017)
F. Cai, et al., manuscript in preparation

Hewlett Packard Enterprise

Summary

- –The computing world has become **heterogeneous**, there is no turning back
- –Big opportunities to speed up applications with significant markets
- –You can jump >20 years into the tech future with a special purpose accelerator
- –Harness emerging devices to build new architectures
- –But we also **need software to rise to the challenge**
  - • Can't depend on hardware to keep up performance growth
- –We must consider **system balance** (compute, memory bandwidth, cooling)
- –We are kicking off a new Cambrian explosion, with plenty of extinctions coming – an exciting time to be designing computing systems!

Thank you

labs.hpe.com

Hewlett Packard
Enterprise

Acknowledgments

HPE Labs

Catherine Graves
Suhas Kumar
Miao Hu
Xia Sheng
Xuema Li
Martin Foltin
Dejan Milojicic
Amit Sharma
Fuxi Cai
Rui Liu

University Collaborators

Jianhua Yang (UMass Amherst)
Qiangfei Xia
Can Li
Aayush Ankit (Purdue)
Kaushik Roy
Izzat El Hajj (UIUC)
Wen-Mei Hwu
Wei Liu (U Michigan)
Shimeng Yu (GeorgiaTech)

Program support

Karl Roenigk
Jeffrey Weinschenk
Richart Slusher
Chad Meiners (Lincoln Lab)
Chris Algire (NGA)