SAND2019-7110C

Artificial synapses based on electrochemical lithium insertion into metal oxides for neuromorphic computing

PRESENTED BY

Yiyang Li, Elliot Fuller, Sapan Agarwal, Alec Talin

Sandia National Laboratories

Shiva Asapu, Joshua Yang

University of Massachusetts, Amherst

Scott Keene, Armantas Melianas, Alberto Salleo

Stanford University

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and 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.

Artificial neural networks and deep learning

Computers are fast and efficient at executing instructions

$$\frac{dx}{dt} = x^2$$

```
x=1;
dt = 0.01;
for i = 1:1000 {
   dxdt = x.^2;
   xnew = dt*dxdt;
   x = xnew; }
end
```

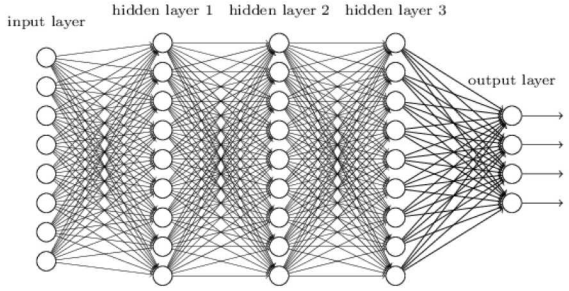
Computers struggle when there are no clear instructions for the task

Which one of these images is a cat?

Image recognition
Autonomous driving
Natural language processing

Artificial neural networks: use training examples and error backpropagation to find the matrix weights that correctly maps the input **x** onto the desired output **y**

$$\begin{bmatrix} y_1 \\ \vdots \\ y_m \end{bmatrix} = \begin{bmatrix} w_{1,1} & \cdots & w_{1,n} \\ \vdots & \ddots & \vdots \\ w_{m,1} & \cdots & w_{m,n} \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$$



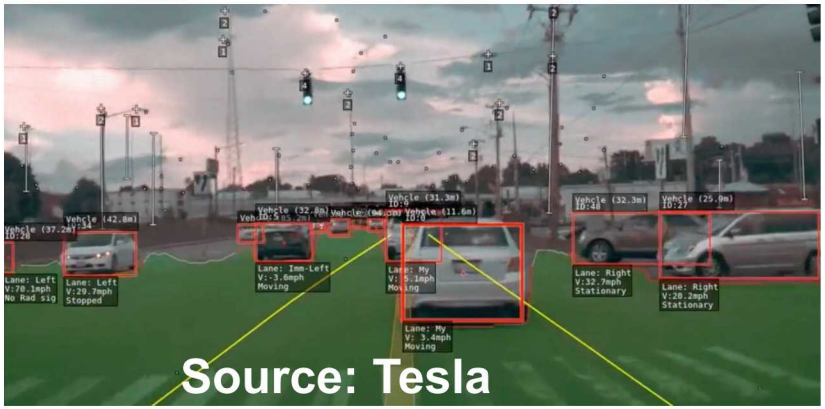
Inference and training are very energy expensive

Andrew Ng, Coursera
Nawrocki et al. *IEEE Elec. Dev.* 2016

n, m > 1000

Energy in autonomous driving

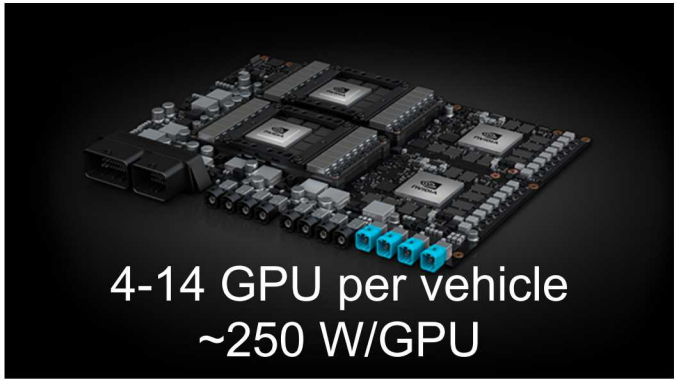
Chevy Bolt



~100 GB/sec

Human brain: 10 Watt

Computing: 1-4 kW

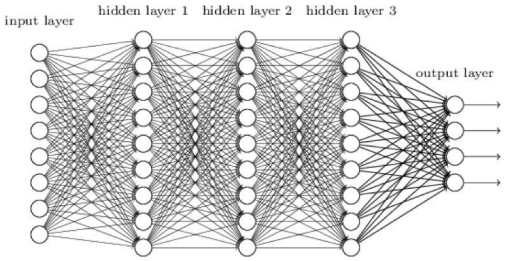


Nvidia

Digital and analog implementations of neural networks

$$\begin{bmatrix} y_1 \\ \vdots \\ y_m \end{bmatrix} = \begin{bmatrix} w_{1,1} & \cdots & w_{1,n} \\ \vdots & \ddots & \vdots \\ w_{m,1} & \cdots & w_{m,n} \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$$

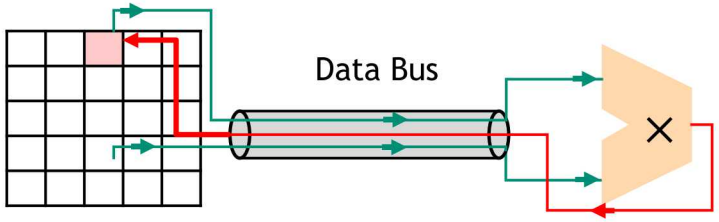
n, m > 1000



Von Neumann Digital
Separate logic and memory structures

SRAM to store the weights

Arithmetic logic unit for multiplication

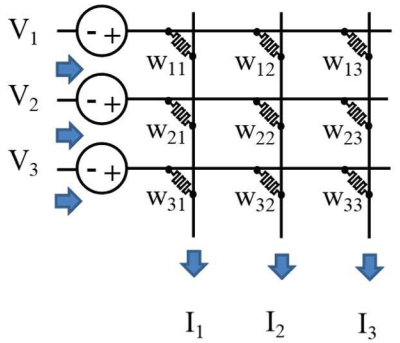


Uses established CMOS technology
Data bus results in latency and power

In-memory Parallel Analog
Use non-volatile memory

Crossbar for matrix multiplication

Conductance of each element can be changed in a predictable manner



$I_1 = V_1W_{11} + V_2W_{21} + V_3W_{31}$

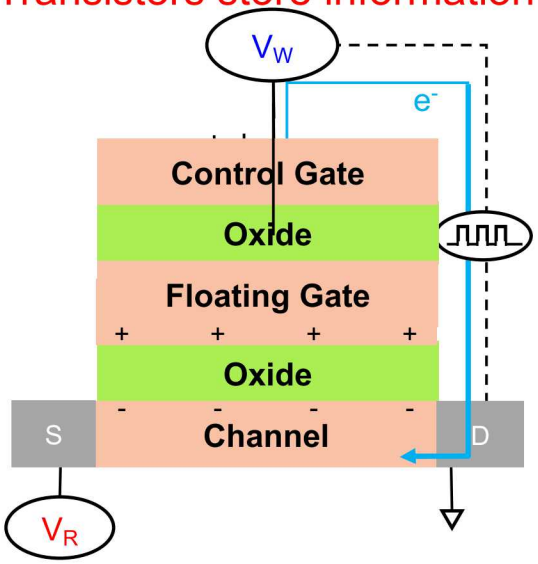
Simultaneous logic and memory
3 orders of magnitude less power

M. Marinella, *IEEE Circuits and Systems*, 8, 86-101, 2018

Challenge: find an appropriate low-energy, highly predictable nonvolatile memory

Three-terminal memory

Transistors store information by moving charge between the gate and the channel



$$\tau_{retention} = C_{gate}\, R_{OFF}$$

$$C_{gate} = \epsilon_0 \epsilon_r \frac{A}{t_{ox}} \approx 10^{-6}\ \mathrm{F\ cm^{-2}} = 10^{-16}\ \mathrm{F}\ \ (A = 100\ \mathrm{nm^2})$$

Option 1: transistor switch: $R_{OFF} \sim 10^{15}\ \Omega$

$\tau_{retention} < 1$ s (Resistive Processing Unit, dynamic RAM)

S. Ambrogio, G. Burr, et al. Nature, **558**, 60 (2018)

Option 2: floating gate oxide switch, $R_{OFF} \sim \infty$

$\tau_{retention} > 10$ yr (Flash memory)

10V, high power, limited endurance (10^5)

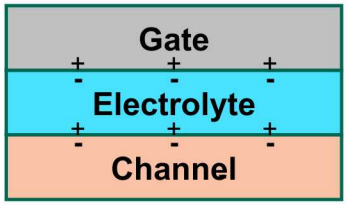
Improving retention by increasing C_{gate}? $C = \frac{\Delta Q}{\Delta V}$

Electrochemical double-layer transistors

Electrons stored at the **interface**

$C_{gate} \sim 10^{-4}\ \mathrm{F\ cm^{-2}}$

$t_{"ox"} < 1\ nm$



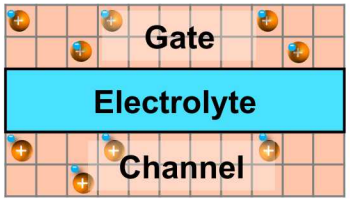
P. Gkoupidenis, G. Malliaras, Adv Mater, 27, 7176 (2015)

Redox transistors

Electrons (and ions) are stored in the **bulk** of the material (battery)

Electron/ion neutral ambipolar pair enables ΔQ without electrostatic charging

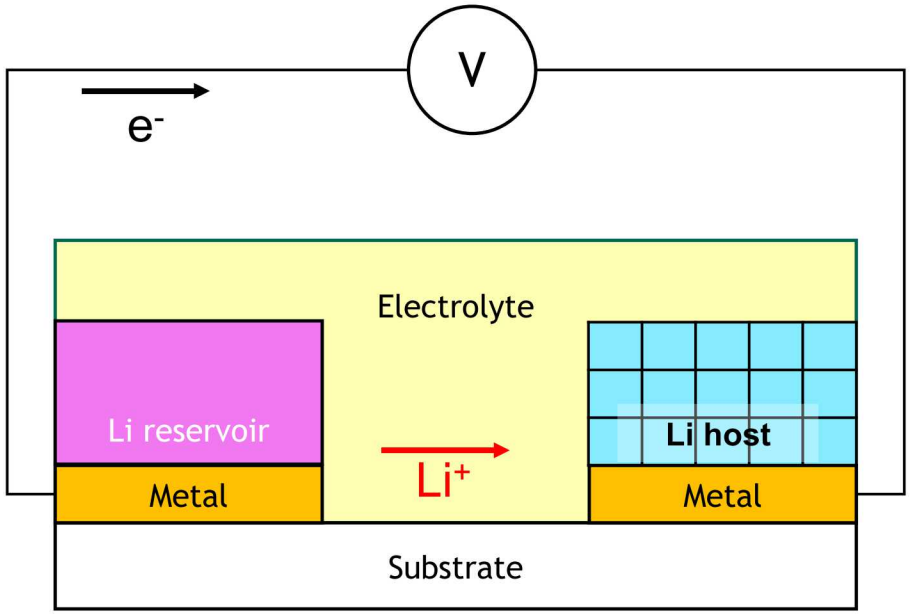
$C_{gate} \sim 10^{-1}\ \mathrm{F\ cm^{-2}}$ (100-nm film)



E. Fuller, Adv Mater. 29, 1604310 (2017)

Y. van de Burgt, Nat. Mater., 16, 414 (2017)

Electrochemical ion insertion

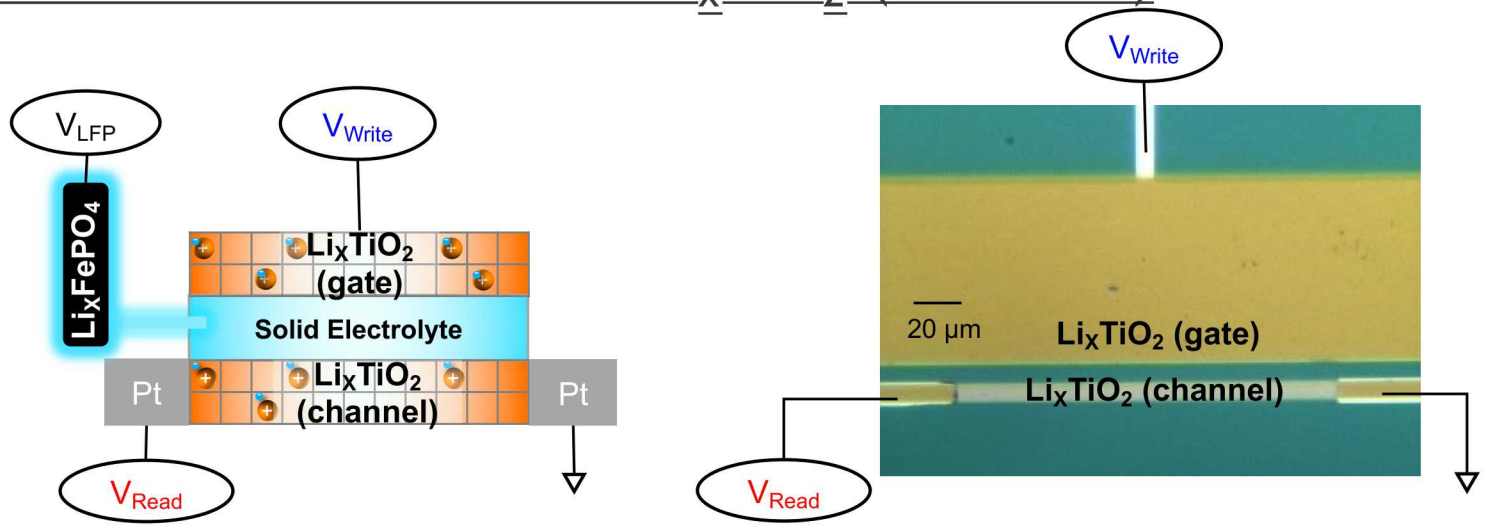


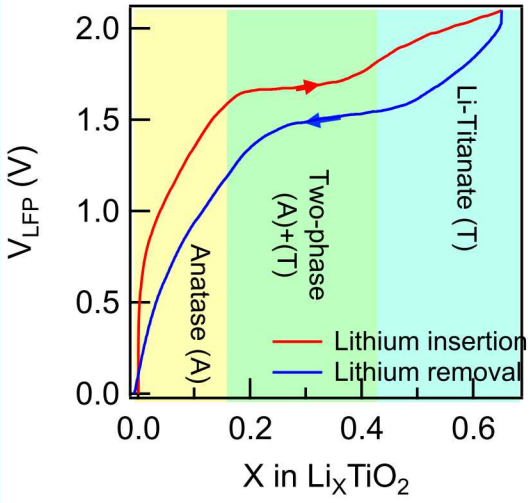
Charge storage with *minimal* electrostatic charging

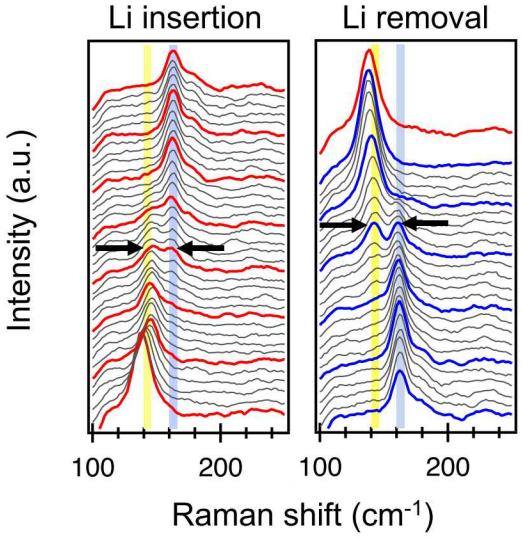
In batteries, intercalation provides a high density of energy storage;
In redox transistors, intercalation provides a high density of information storage

Y. Li, W. Chueh. Ann. Rev. Mater. Res. 48, 137 (2018)

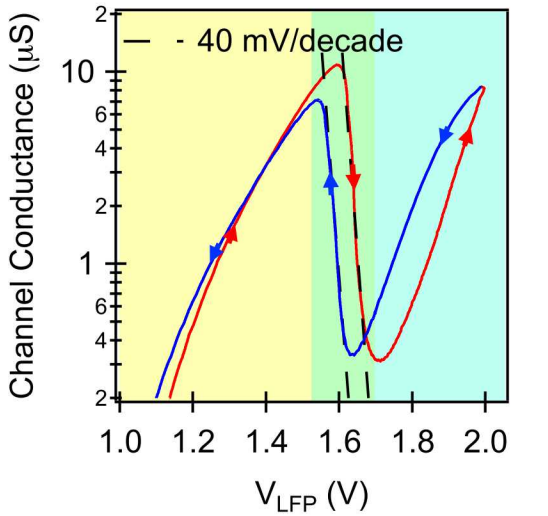
Lithium insertion into Li_xTiO_2 (anatase)



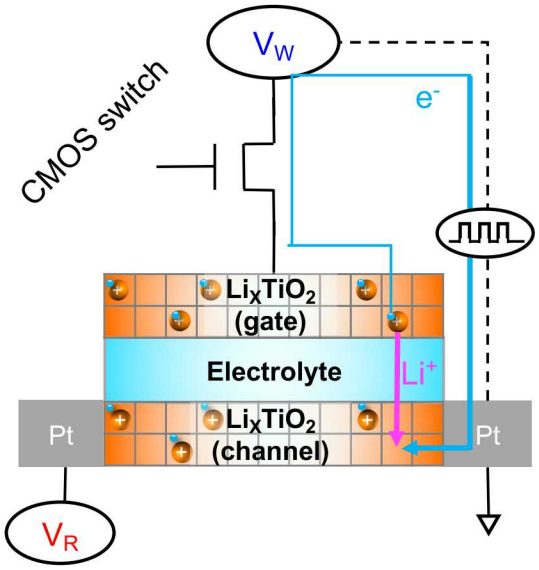


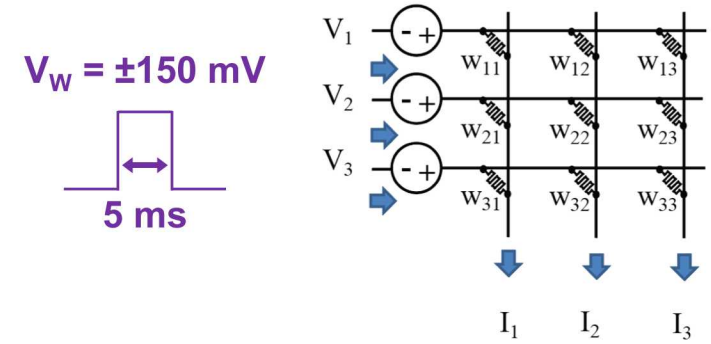


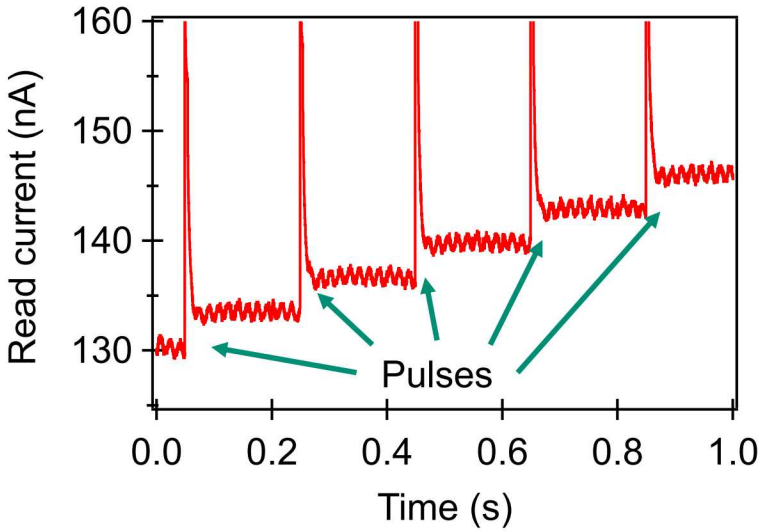
$$Li^+ + e^- + Ti_{Ti}^X \rightarrow Li_i^\circ + \boldsymbol{Ti'_{Ti}}$$



Redox transistor using Li_XTiO_2 (anatase)

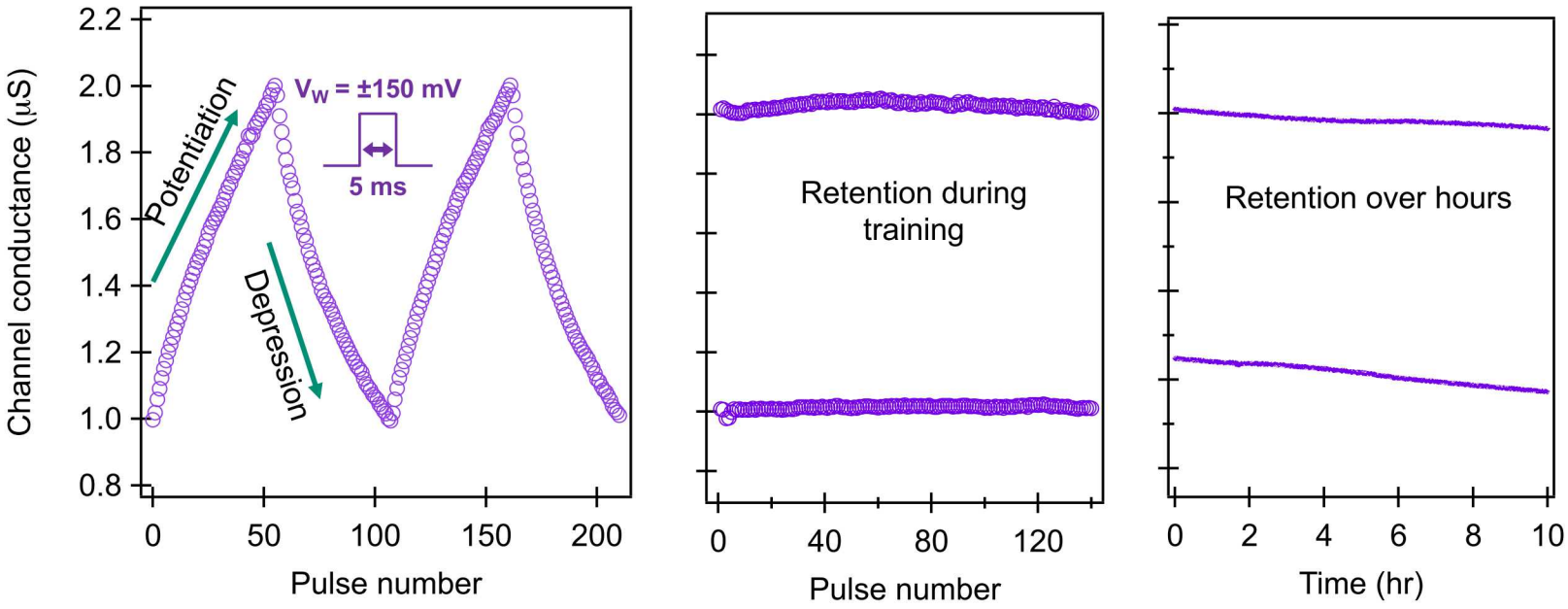






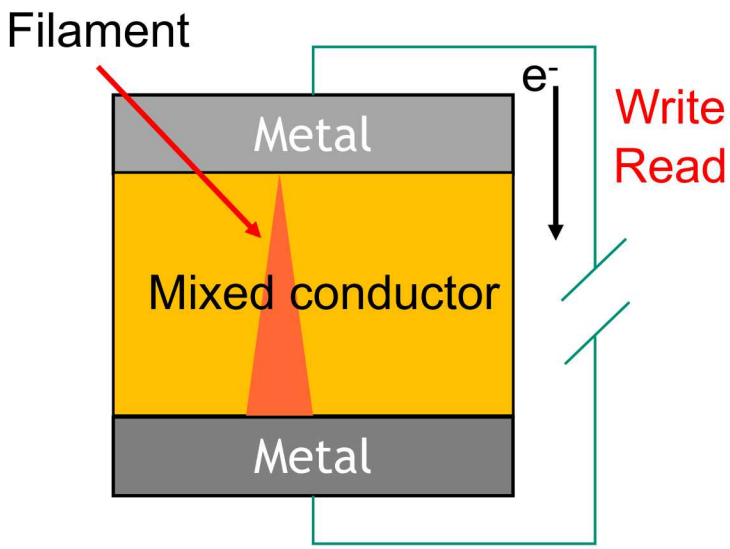
- Linear and symmetric programming
  - 150-mV write pulses
  - Analog, non-volatile states

Redox transistor using Li_XTiO_2 (anatase)



Two-terminal vs three-terminal synaptic memory

Memristor

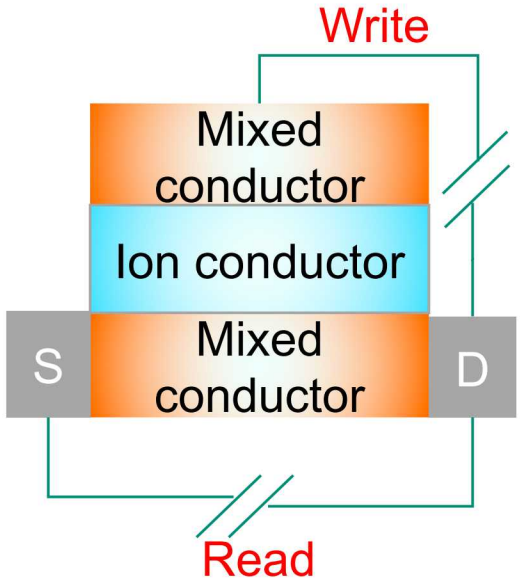


Electron current dominates write

$$I_{electron}^{write} \gg I_{ion}^{write}$$

Large joule heating
Unpredictable weight updates

Redox transistor



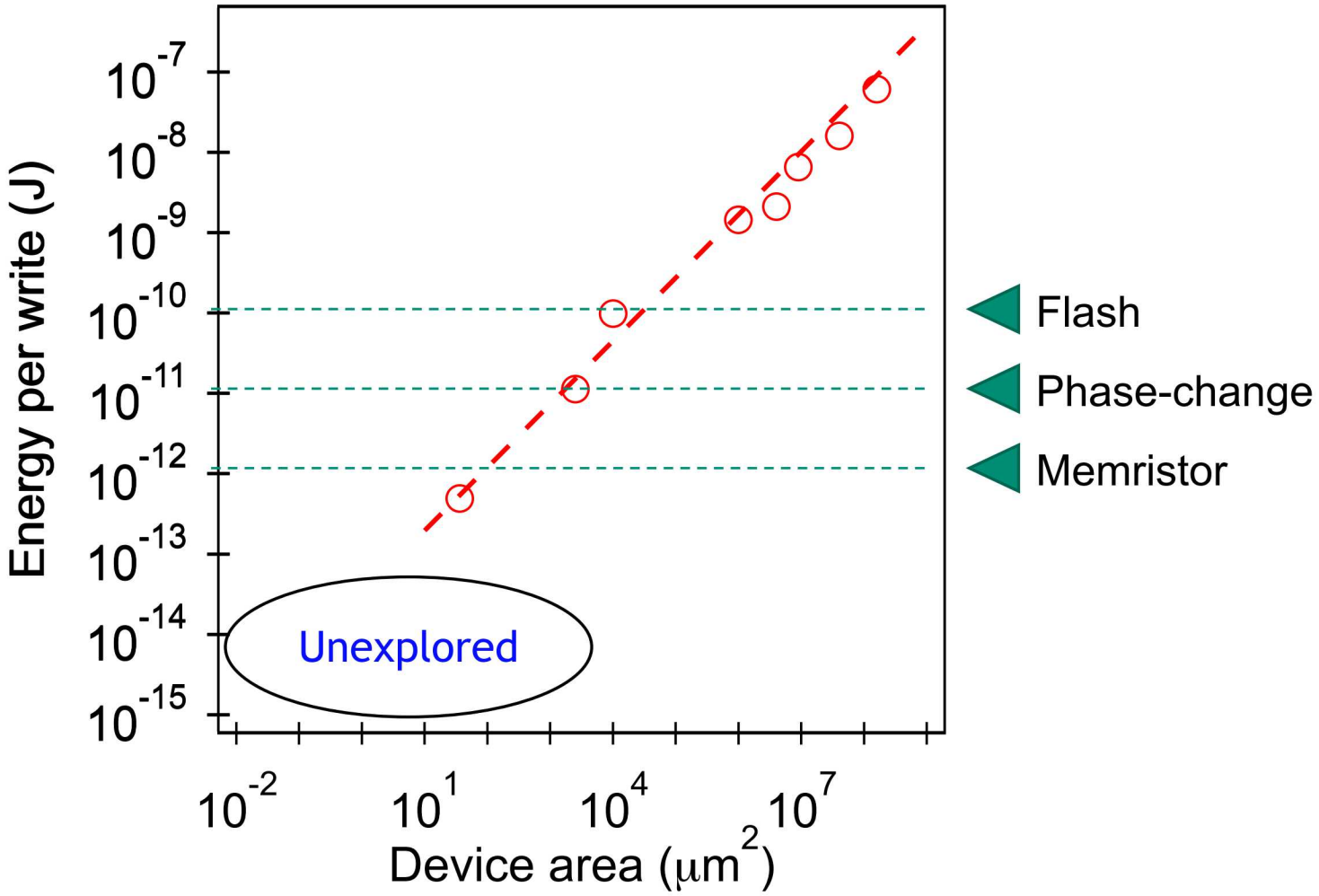
Electrolyte blocks electrons

$$I_{electron}^{write} = I_{ion}^{write}$$

Low write energy
Predictable /linear weight updates

Q. Xia and J. Yang, *Nat. Mater.*, 18, 309 (2018)

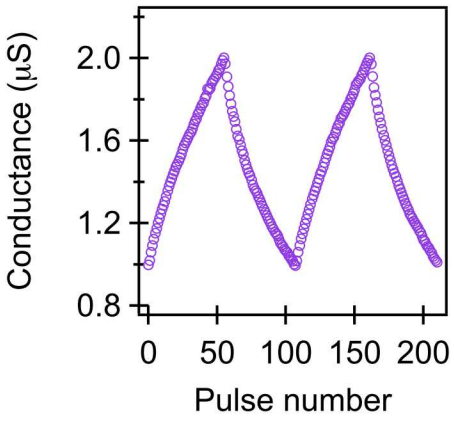
Exceptional low-energy consumption



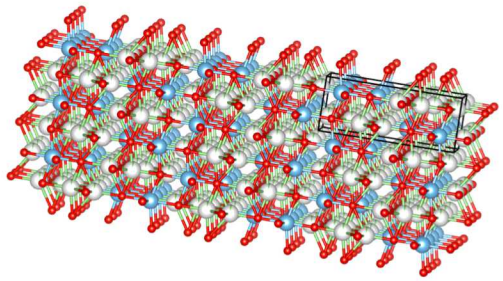
Y. Van de Burgt, *Nat Mater*, 16, 414(2017)
M. Sharbati *Adv. Mater.* 30, 1802353 (2018)
J. Yang et al. *Nat. Nano.* 8, 13(2012)

Linear programming and high accuracy

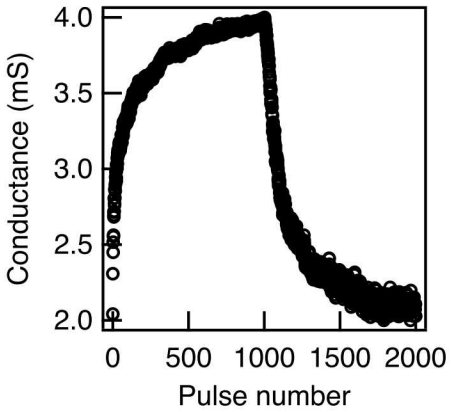
Li_XTiO_2 transistor
Voltage: ± 0.15 V



Redox transistors store information continuously as dopants in a crystal

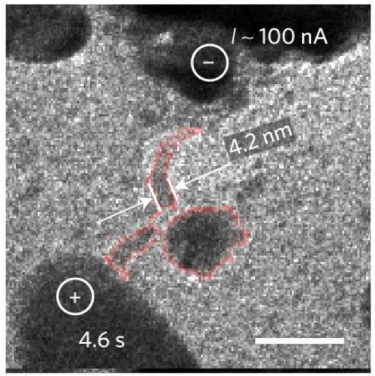


TaO_x Memristor
Voltage: ±1 V



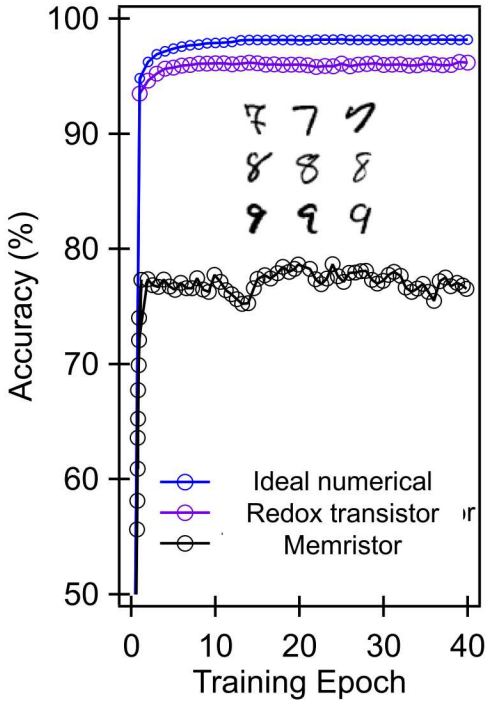
R. Jacobs-Gedrim et al. *ICRC*, 2017

Memristors store information at filaments



Wang et al. *Nat Mater.* 2017

Linearity is essential to train an accurate neural network

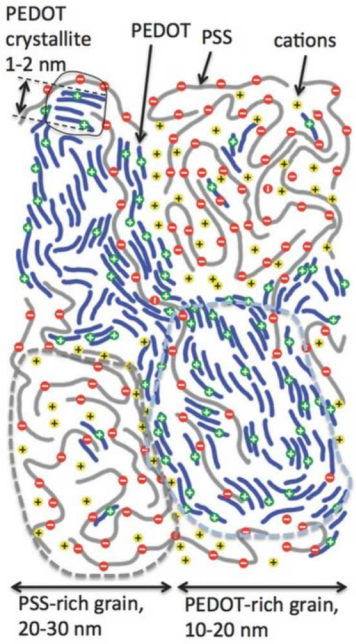


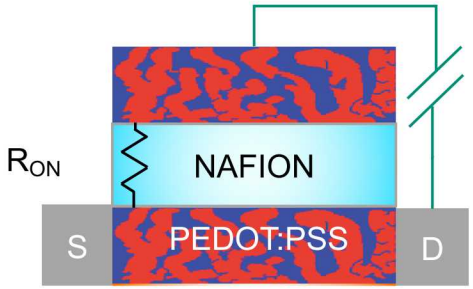
Crossbar Simulator

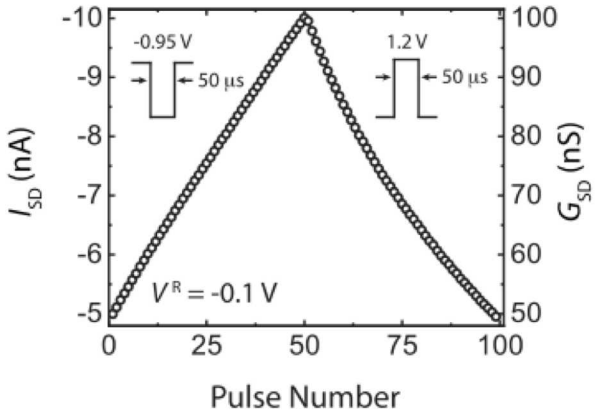
Polymer-based “super-capacitive” redox transistors

Alberto Salleo, Stanford University

PEDOT (hole conductor)
PSS (proton conductor)

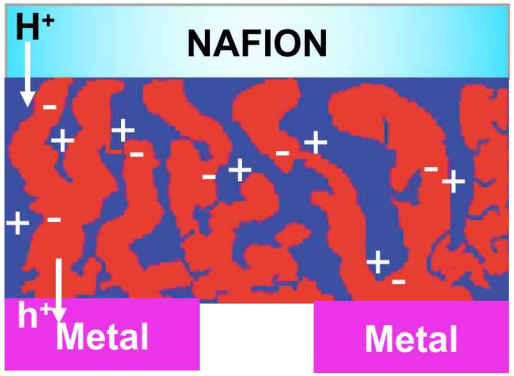






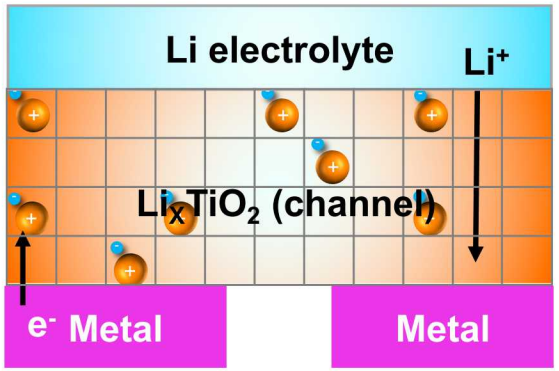
A. Volkov, et al. *Adv. Funct. Mater.* 27, 1700329 (2017); Y. Van de Burgt, et al. *Nat Mater*, 16, 414(2017); E. Fuller, et al. *Science*, 364, 570 (2019)

Super-capacitative vs intercalation redox transistor

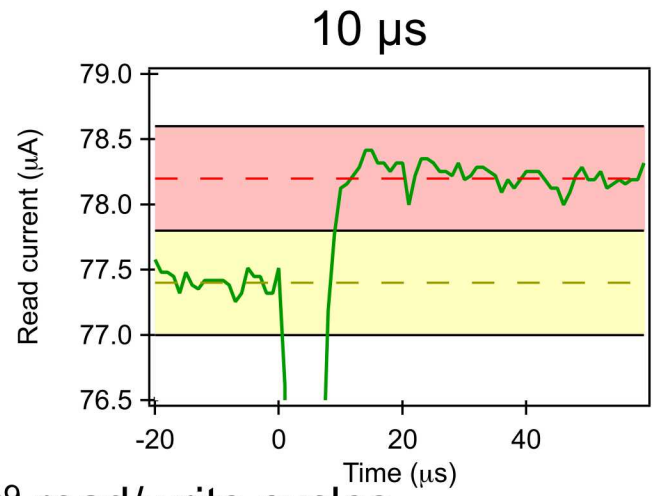


H⁺: proton; h⁺: hole
PEDOT:PSS

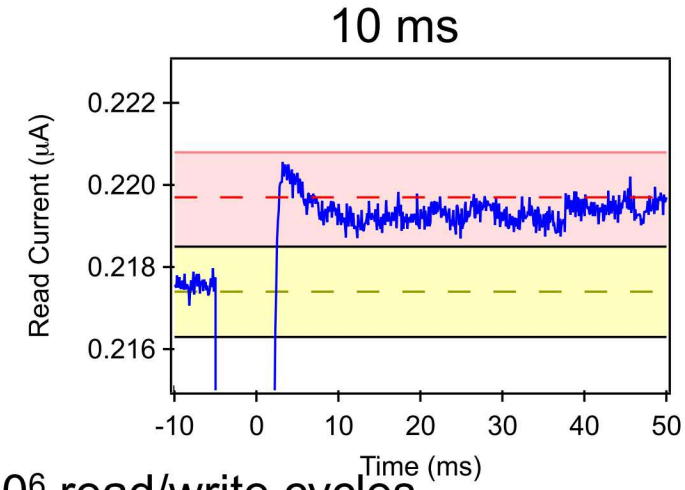
No charge transfer;
Fast single-specie diffusion
Low capacitance: 40 F cm^{-3}



Interfacial charge transfer
Slow chemical diffusion
High capacitance: 5000 F cm^{-3}

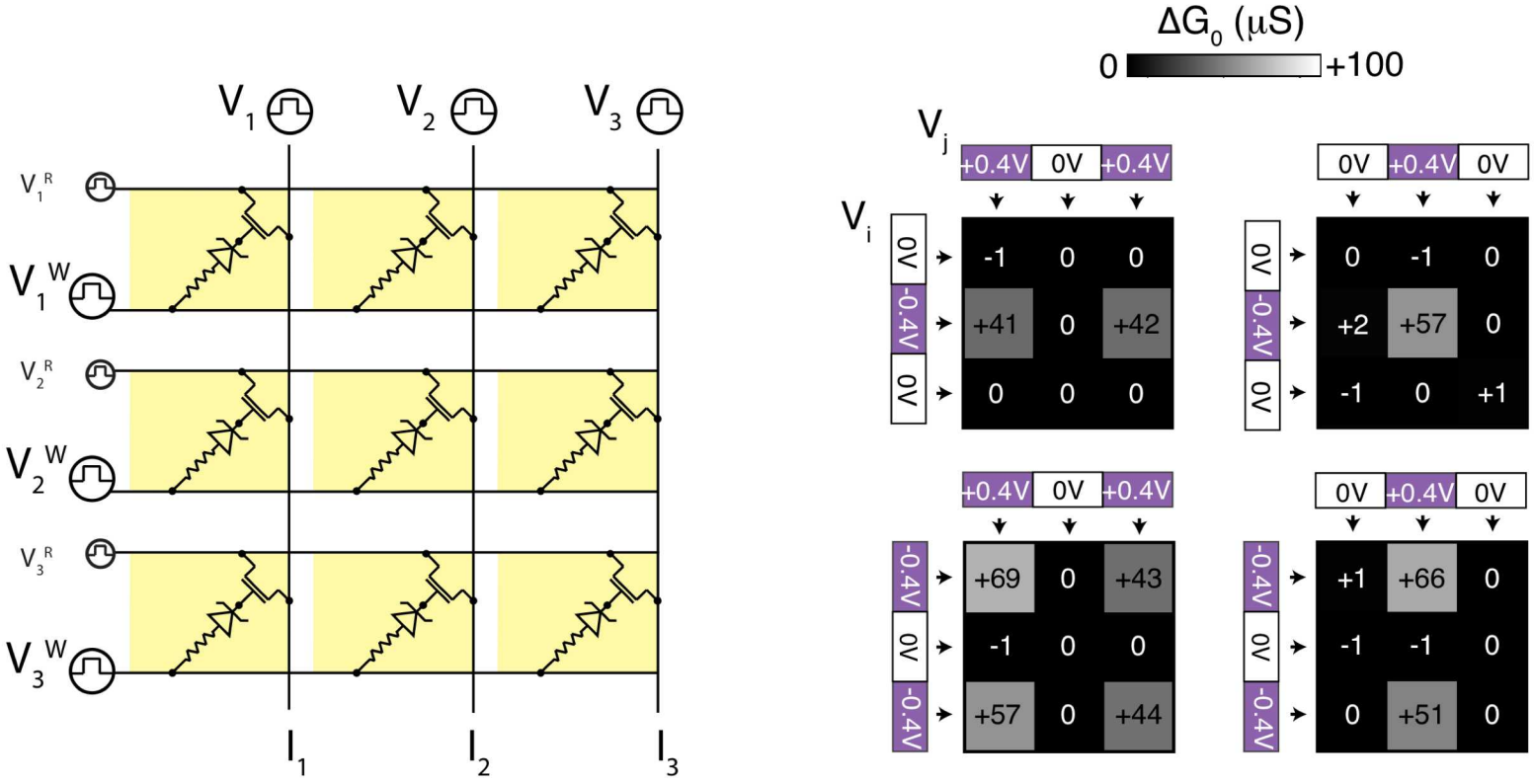


10^9 read/write cycles



10^6 read/write cycles

Parallel weight updates in crossbar arrays

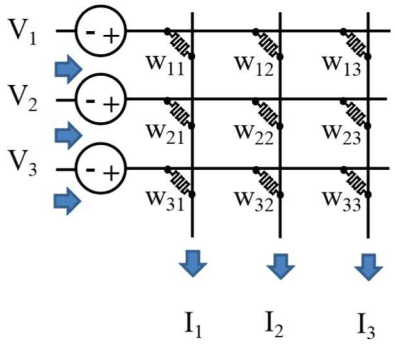


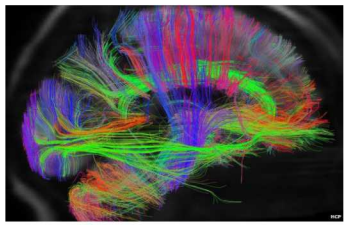
We can select the rows and columns in this crossbar to update the weights of many synapses in a highly parallel fashion in accordance with an outer product update

E. Fuller, et al. *Science*, 364, 570 (2019)

Redox transistors for neuromorphic computing

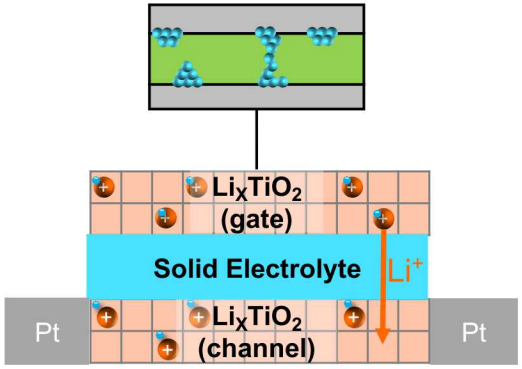
Analog neuromorphic computing provides lower energy and more parallelism compared to digital computing.

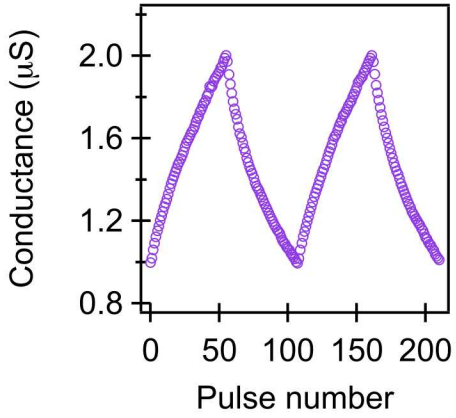




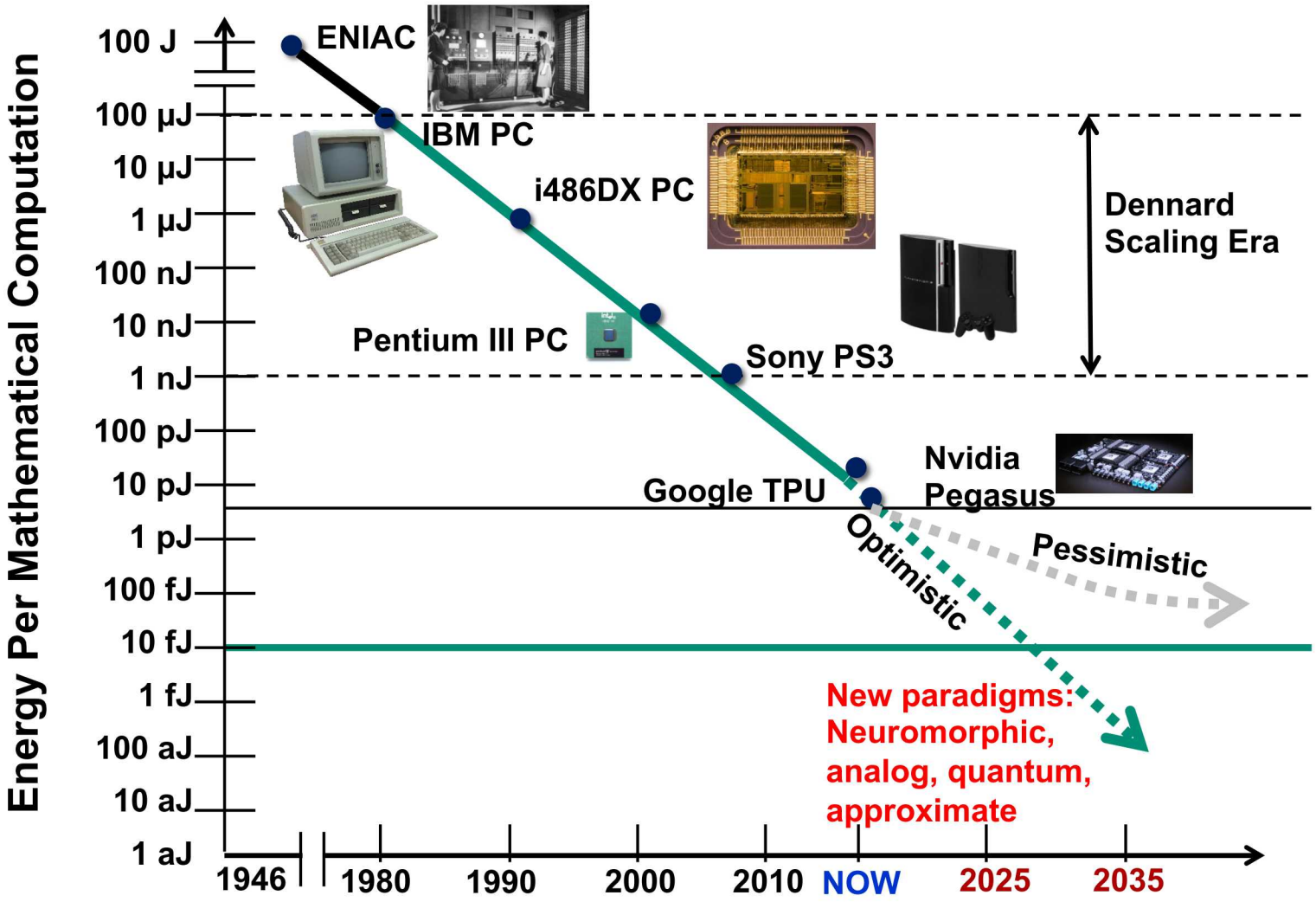
100 billion neurons
100 trillion synapses

Electrochemical ion insertion can be used to create highly linear, low voltage non-volatile analog transistors





Evolution of Computing Machinery



17

Acknowledgements

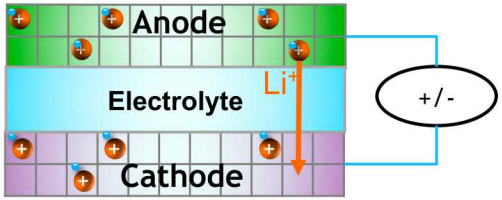
Sandia National Laboratories

Zhiyong Li, Francois Leonard, David Heredia, Maria Pilar Prieto, Tomochika Kurita, Erik Isele

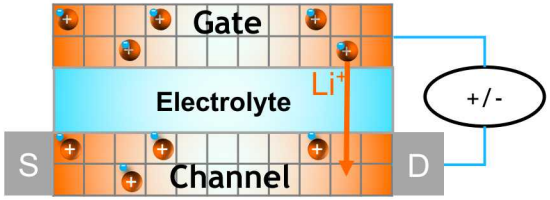
This work is supported by the Sandia National Laboratories Truman Fellowship Program, which is funded by the Laboratory Directed Research and Development (LDRD) Program. Sandia National Laboratories is a multi-mission laboratory managed and operated by National Technology and Engineering Solutions of 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-NA-0003525.

From batteries to redox transistors

Batteries: Maximize open-circuit voltage to store energy

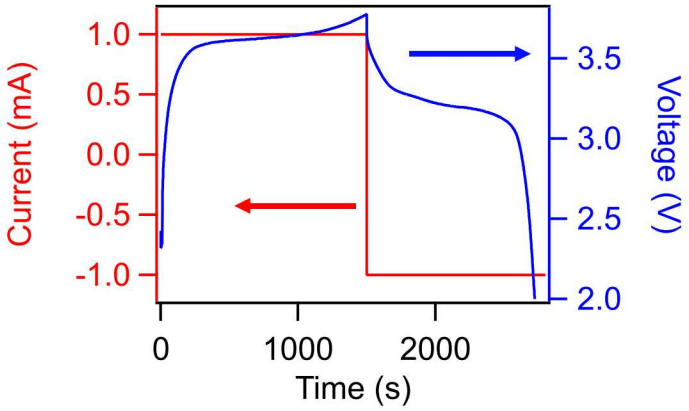


Redox transistors: Minimize open-circuit voltage for array integration

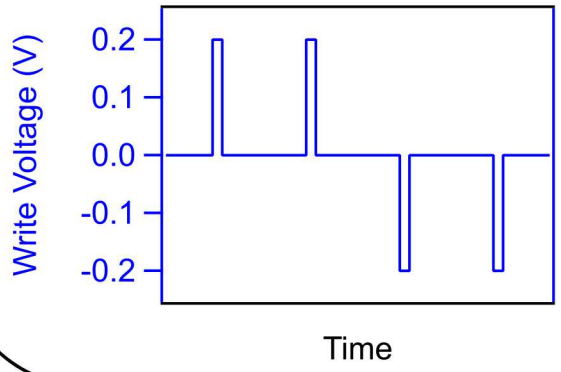


Materials should have a significant change in conductance with a minimal change in write voltage

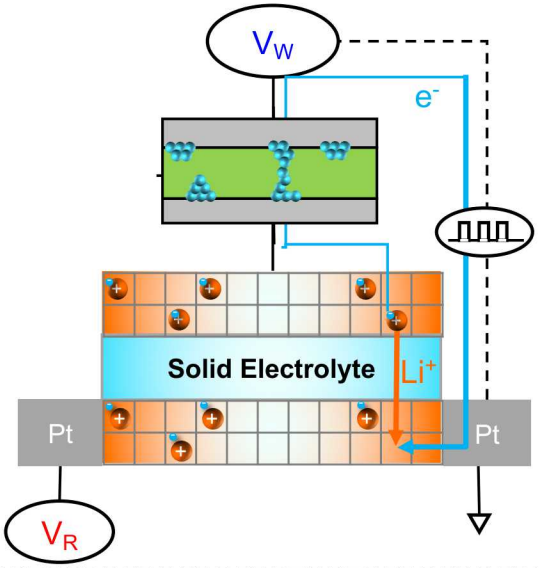
Batteries: Current control



Redox transistors: Voltage control



Low-voltage, Si-free electrochemical memory



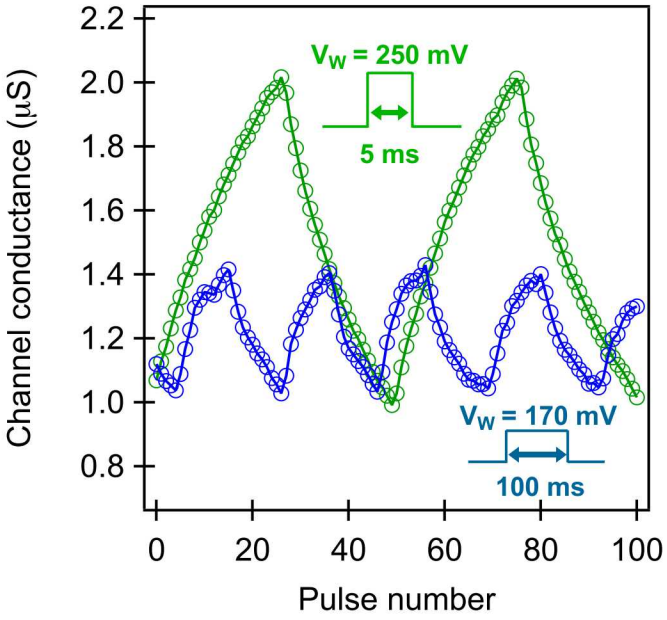
Diffusive memristor (Ag in SiO_X): high ON/OFF ratio

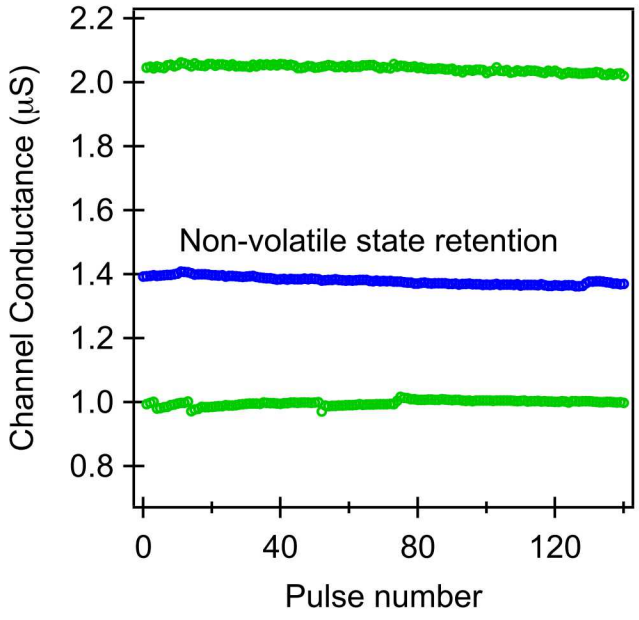
Redox transistor: high charge density via bulk storage

Both: Low switching voltages

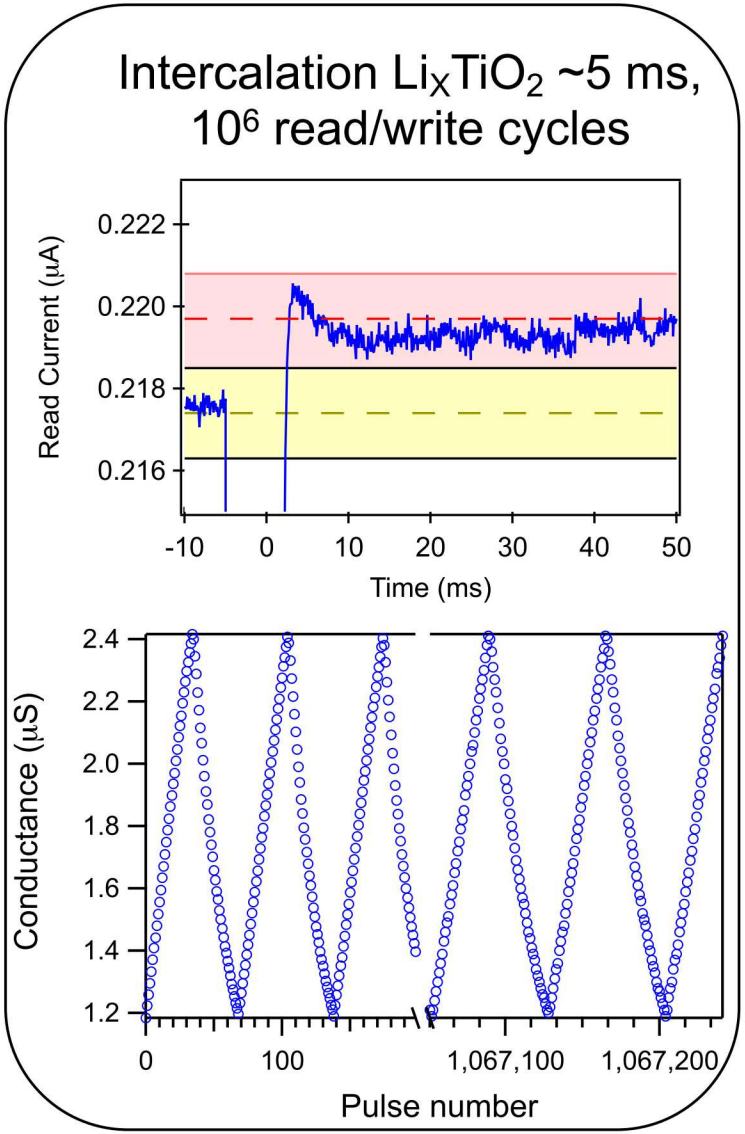
Wang et al. Nat. Mater. 2017

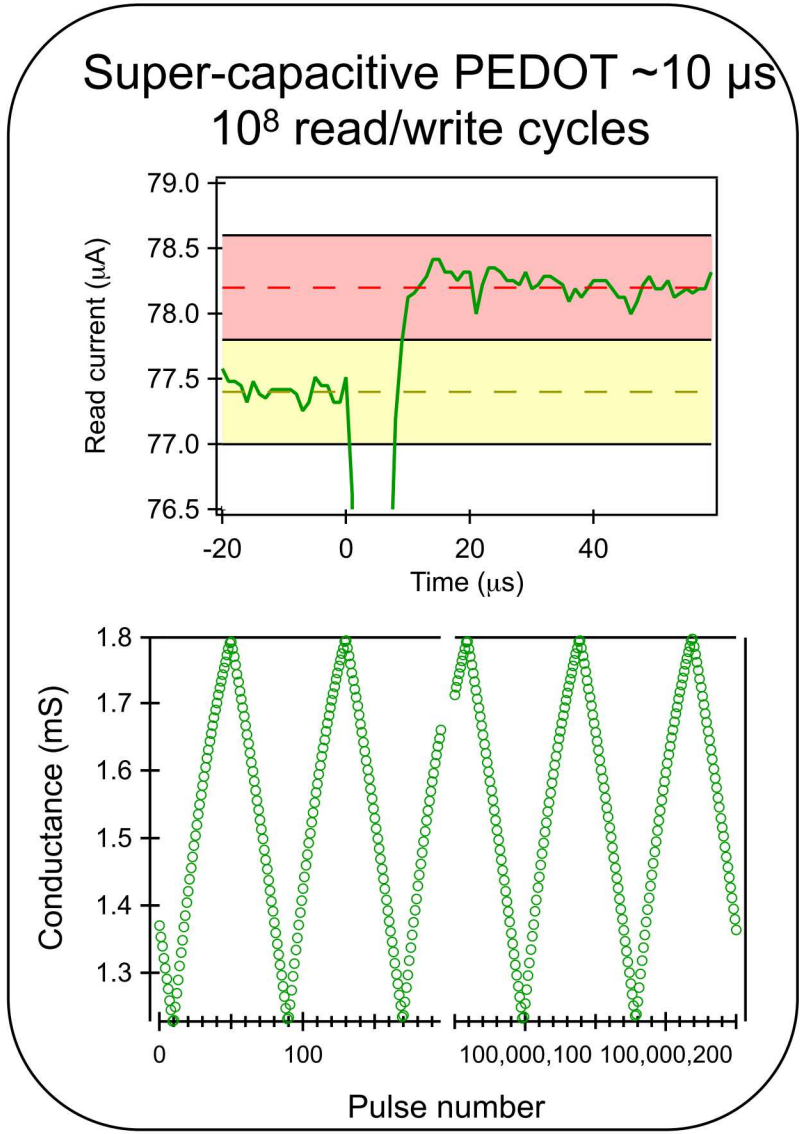
Non-volatile memory that switches at just 6 times the thermal voltage





Switching speed and endurance



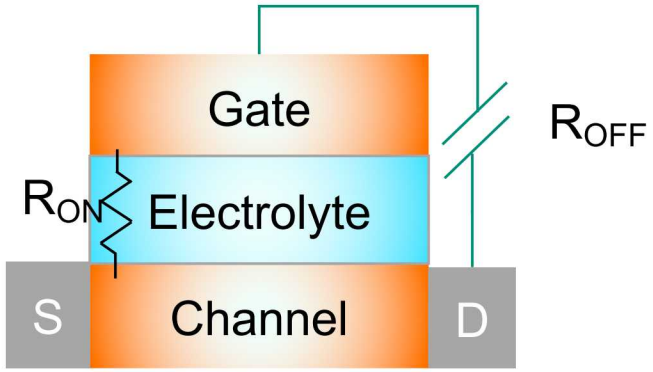


E. Fuller, et al. *Science*, 364, 570 (2019)

Scaling of energy, speed, and retention

ΔQ: amount of charge moved per weight update

$\Delta Q \sim Volume \sim Area$



$Energy = V_{write}\,\Delta Q$

$Energy \sim Area$

$$\tau_{write} = \frac{\Delta Q}{I} = \frac{\Delta Q}{V_{write}/R_{ON}} = \frac{R_{ON}\Delta Q}{V_{write}}$$

$R_{ON} \sim 1/Area$

Write time is independent of size in this geometry

$$\tau_{retain} = \frac{\Delta Q}{I_{leak}} = \frac{\Delta Q}{V_{gate-channel}/R_{OFF}} \sim C_{volumetric} R_{off}$$

Retention time is inversely proportional to the capacitance and the device area