Spintronic devices for artificial neural networks

Saima A Siddiqui

Materials Science and Engineering, University of Illinois Urbana Champaign

September 2, 2020

In-memory computing

Shuttled back and forth at high speeds between processor and memory is inefficient.

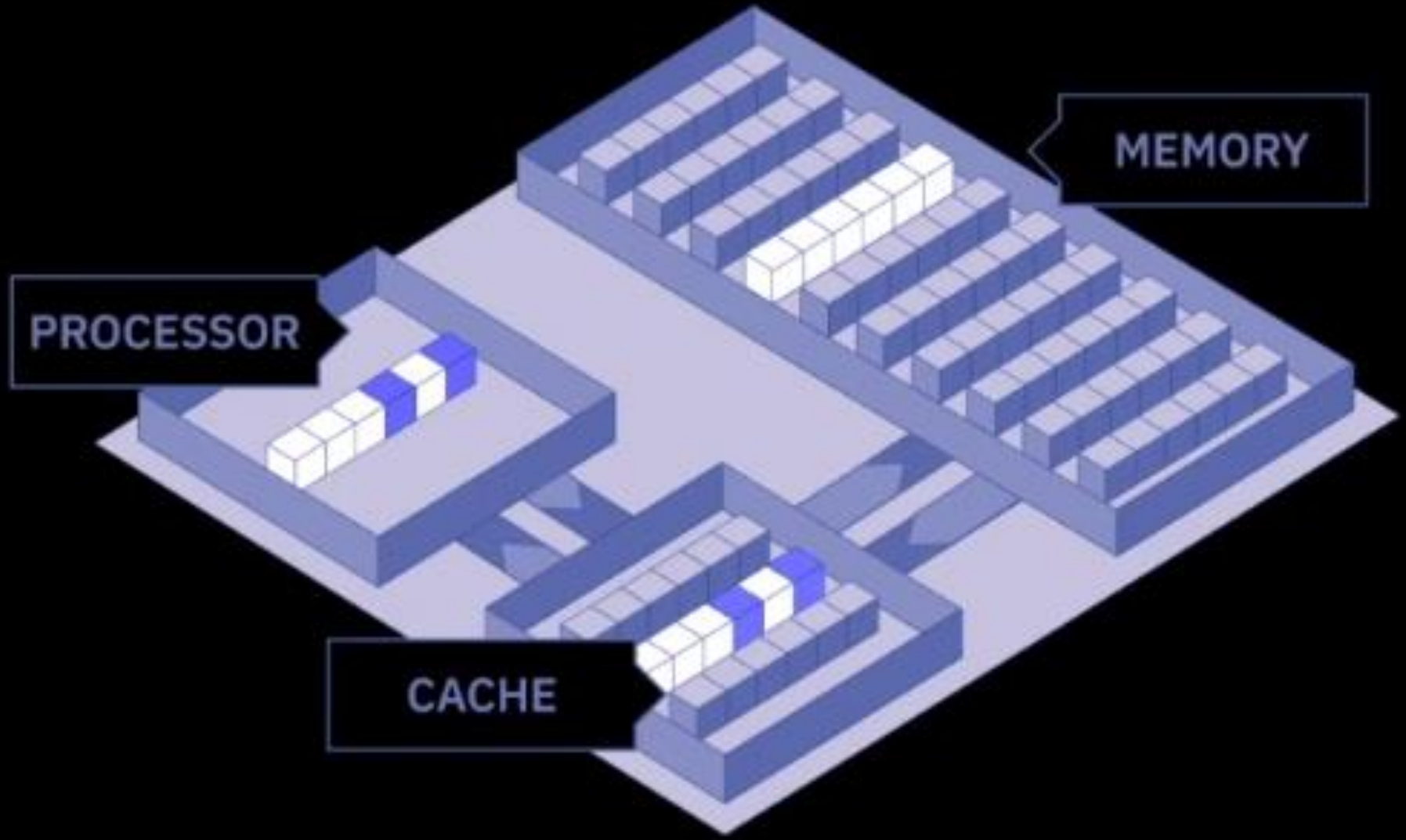


Image courtesy: IBM

Computation in cognitive era is more data centric

Traditional image processing

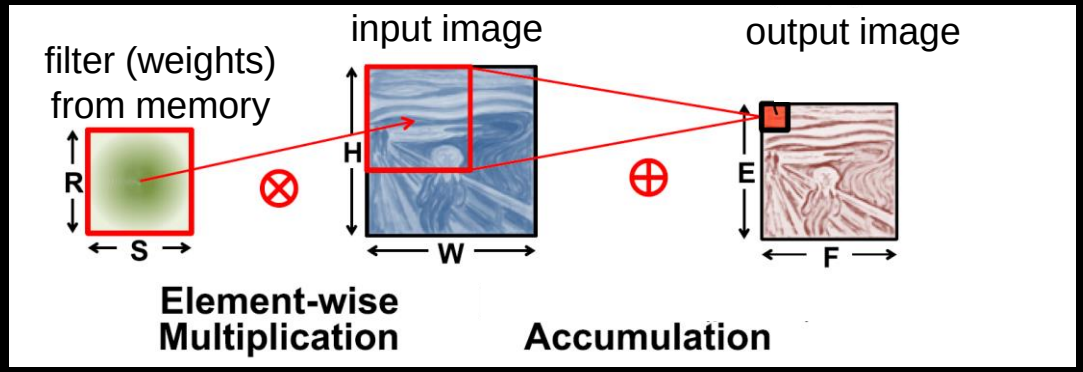
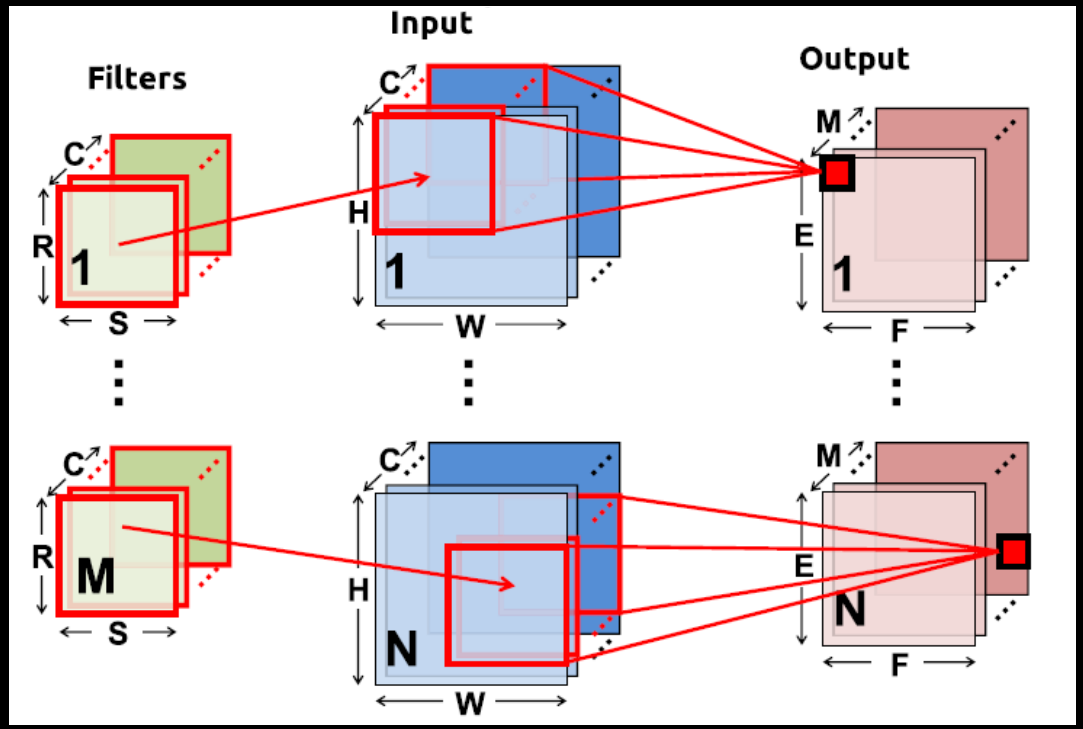
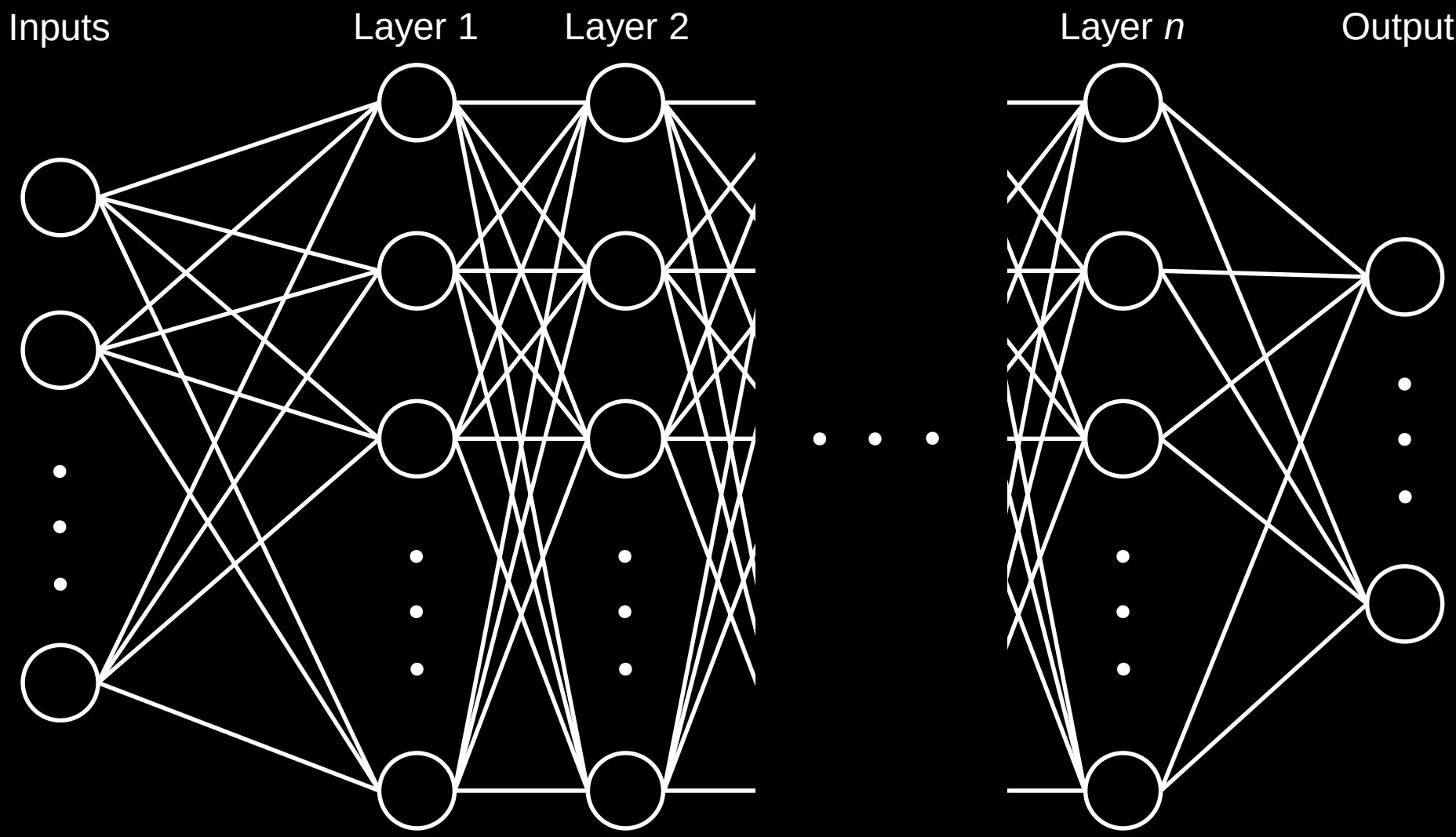


Image processing
for classification



Sze et al., Proc. of the IEEE, 105, 12 (2017)

Layer by layer operation of **neural network** for classification task

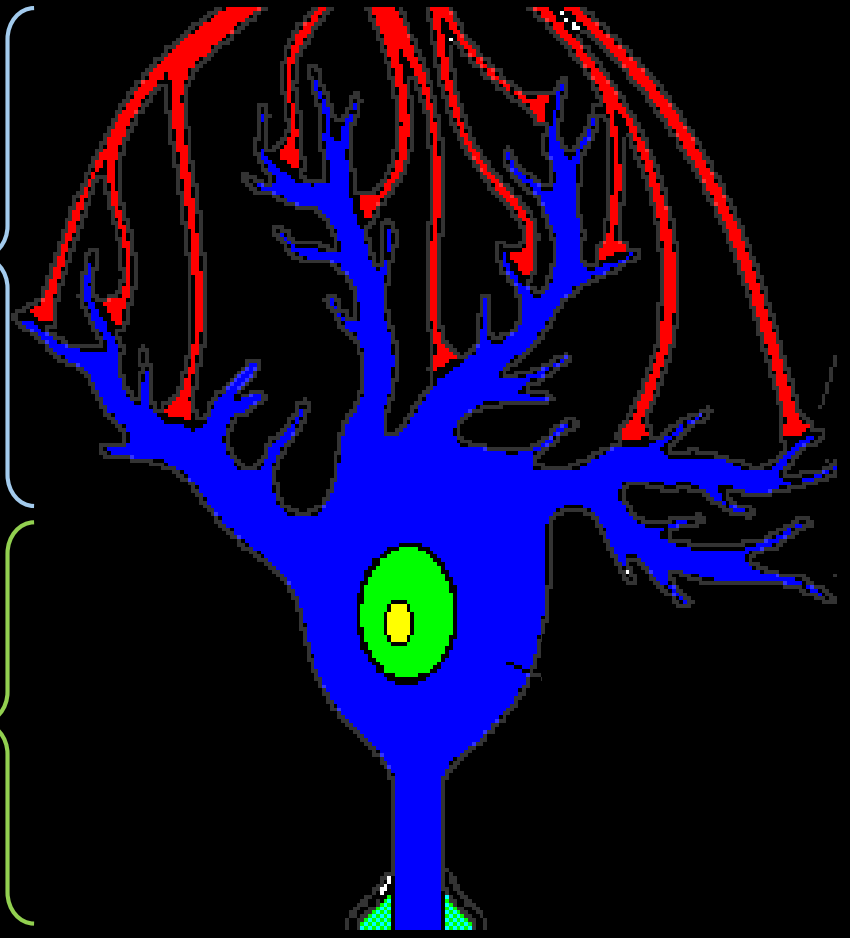


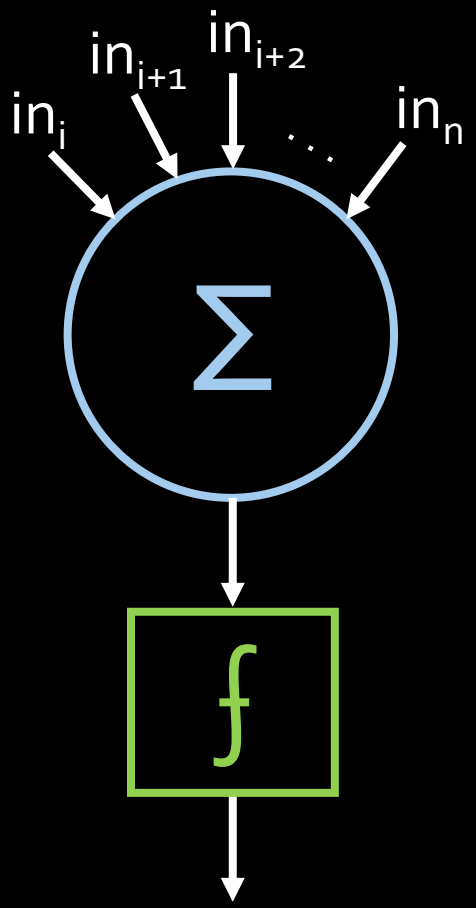
What a unit cell of Neural Networks do

A typical operation in neural networks:

Synapse

Neuron/
Activation





Multiply-accumulate:

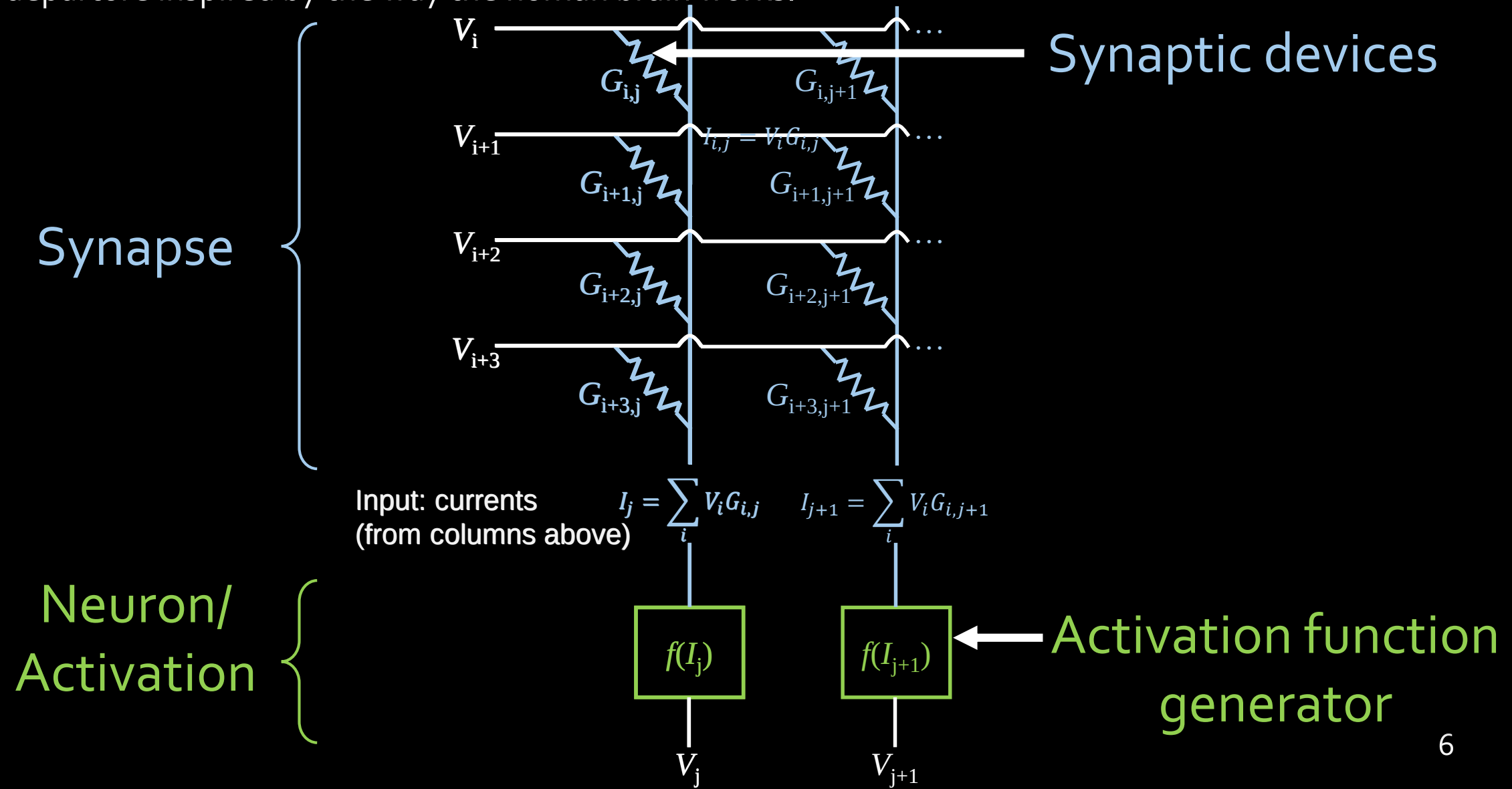
$$\boldsymbol{w} \cdot \boldsymbol{x} = \sum_{i=0}^{m} w_i x_i$$

Nonlinear activation function:

$f(\boldsymbol{w} \cdot \boldsymbol{x})$

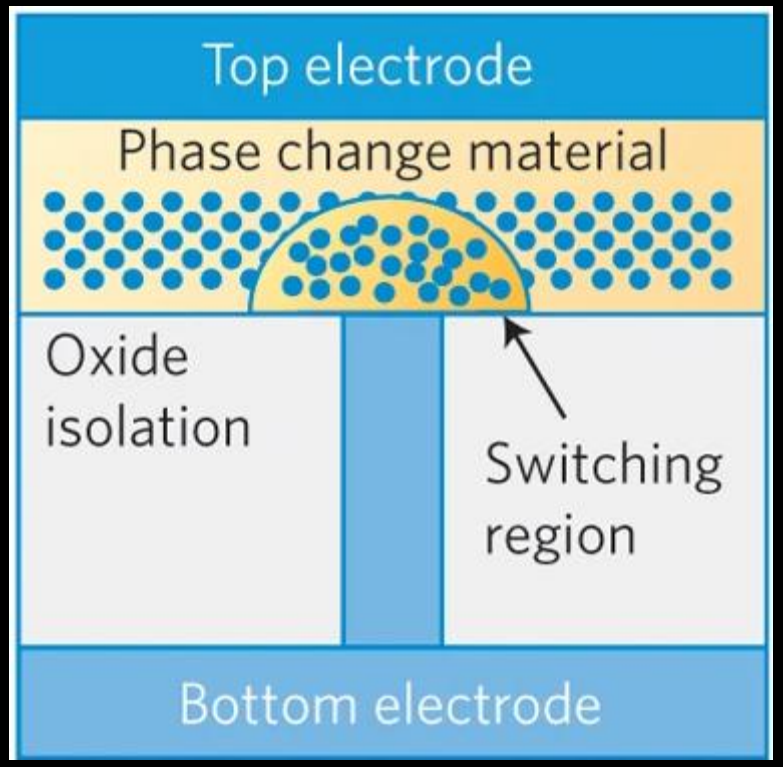
Architecture for in-memory computing: Neural network

Non-von Neumann architecture in which memory and processing coexist in some form – a radical departure inspired by the way the human brain works.

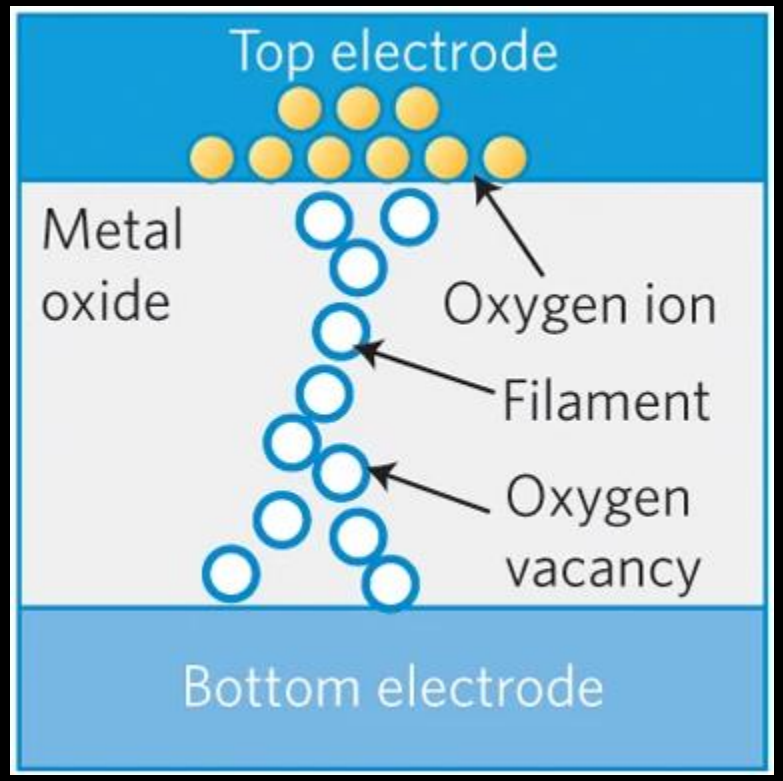


Non-magnetic non-volatile memories (NVMs)

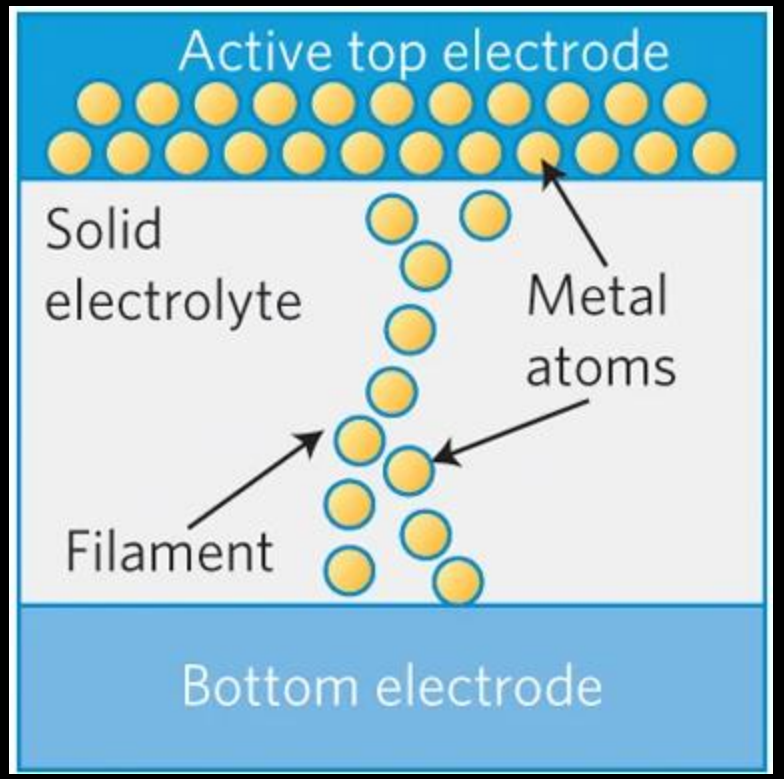
Phase change memory



Resistive random access memory



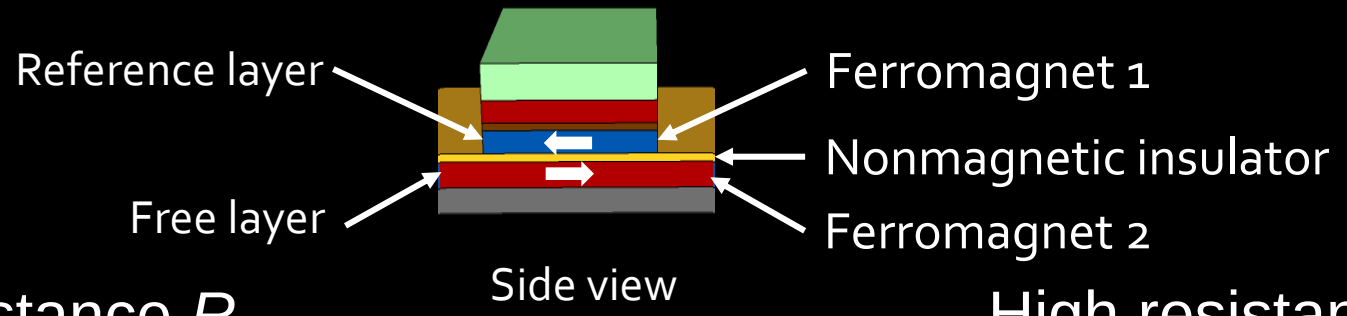
Conductive bridge random access memory



Wong et. al., **Nature Nano**, 191-194, 2015

Magnetic NVM: Magnetic random access memory

Magnetic tunnel junction (MTJ) converts spin into electrical resistance

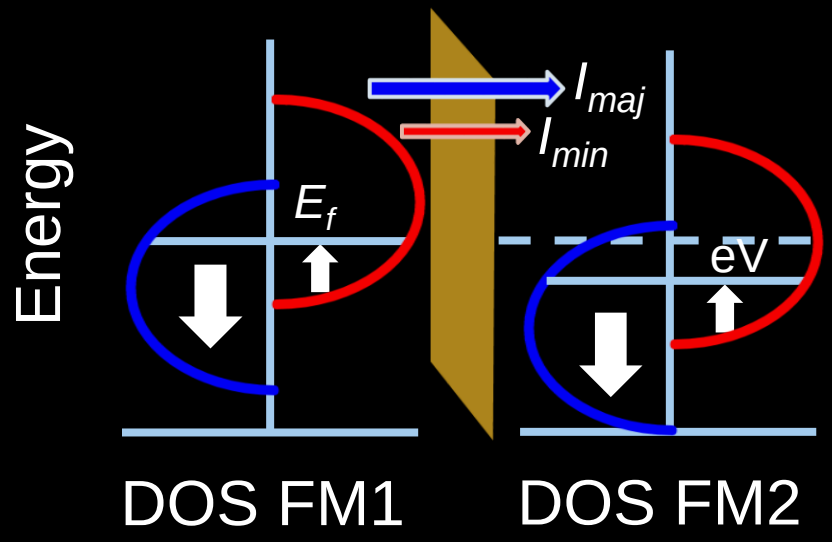


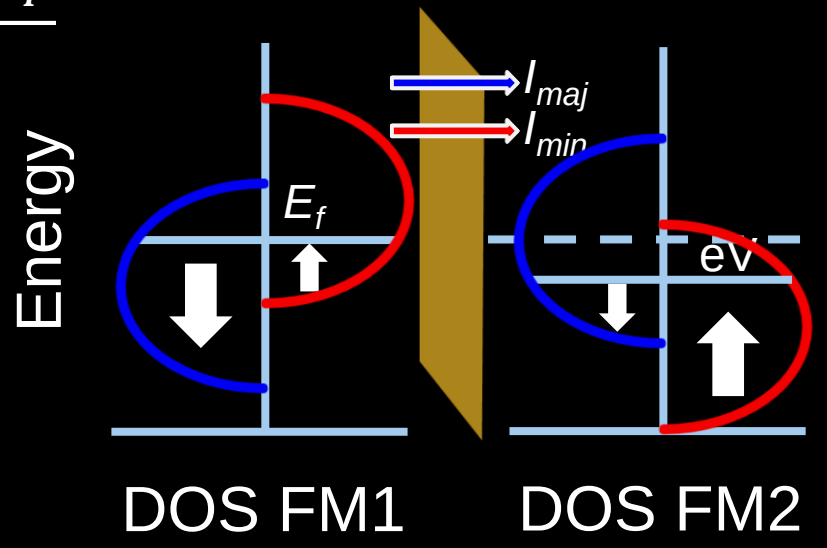
Low resistance R_P

High resistance R_{AP}

Tunnel magnetoresistance

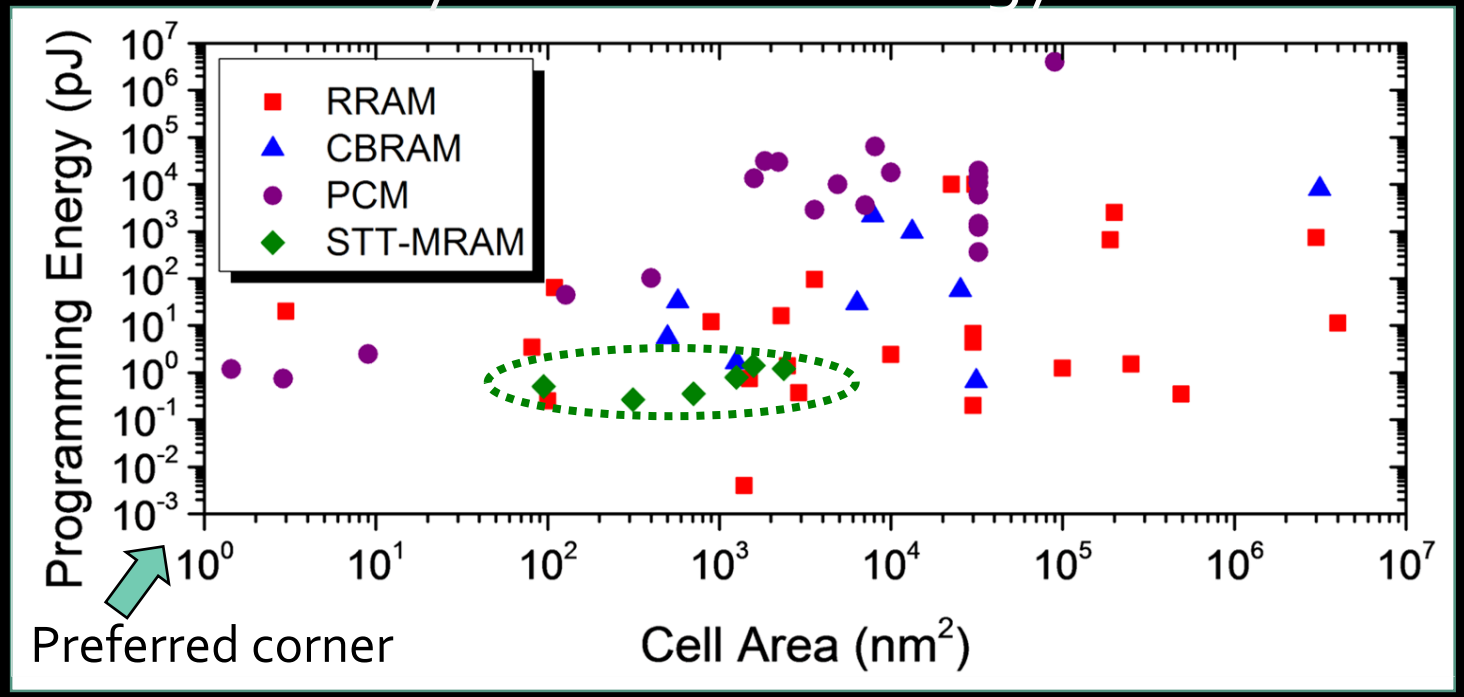
$$TMR = \frac{R_{AP} - R_P}{R_P}$$

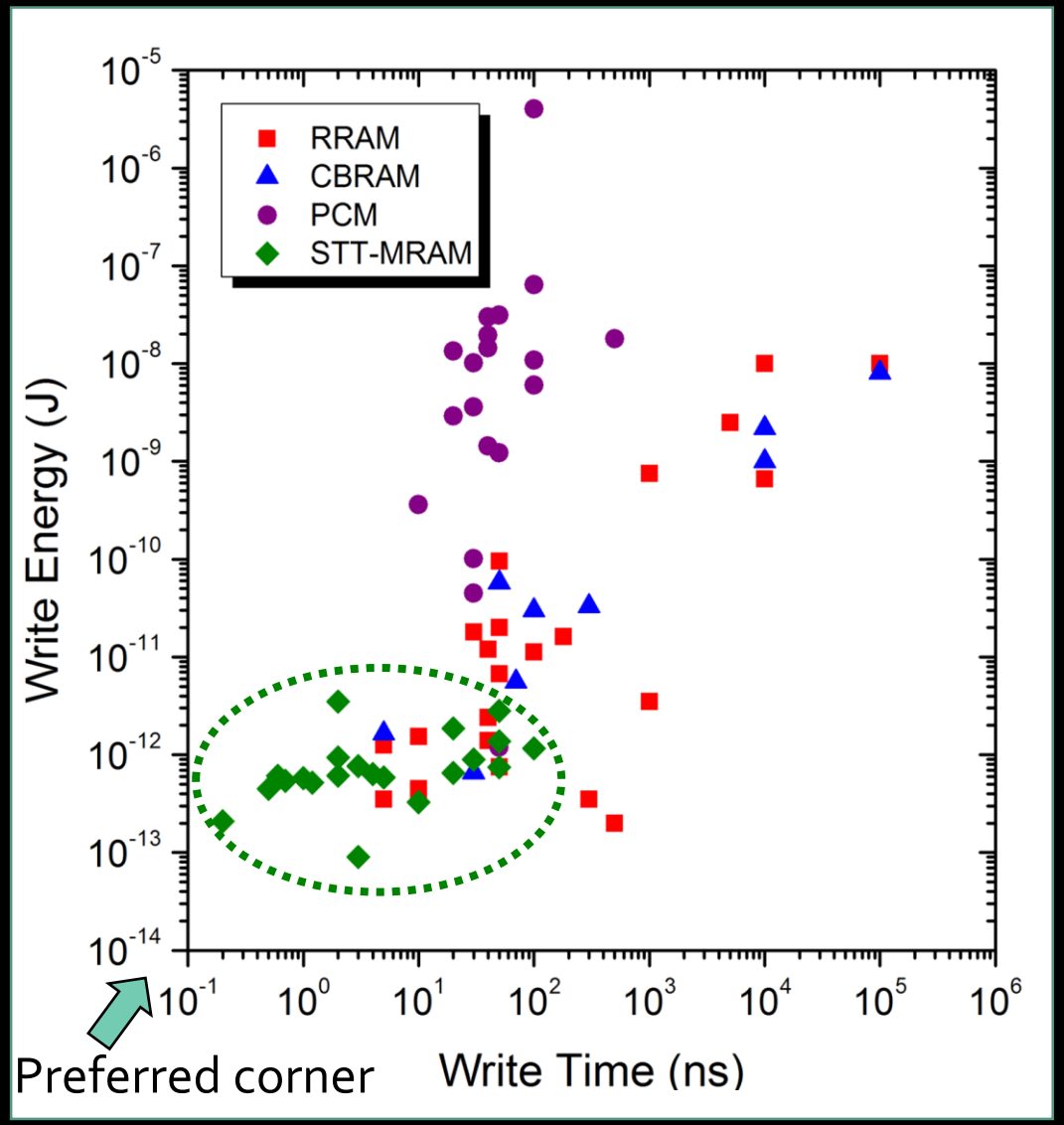




Why spintronic?

- Non-volatility
- Low switching time, switching energy
- Small area and high bit resolution
- Relatively mature technology





RRAM = Resistive Random Access Memory
CBRAM = Conductive Bridge Random Access Memory
PCM = Phase Change Memory
STT-MRAM = Spin Transfer Torque-Magnetic Random Access Memory

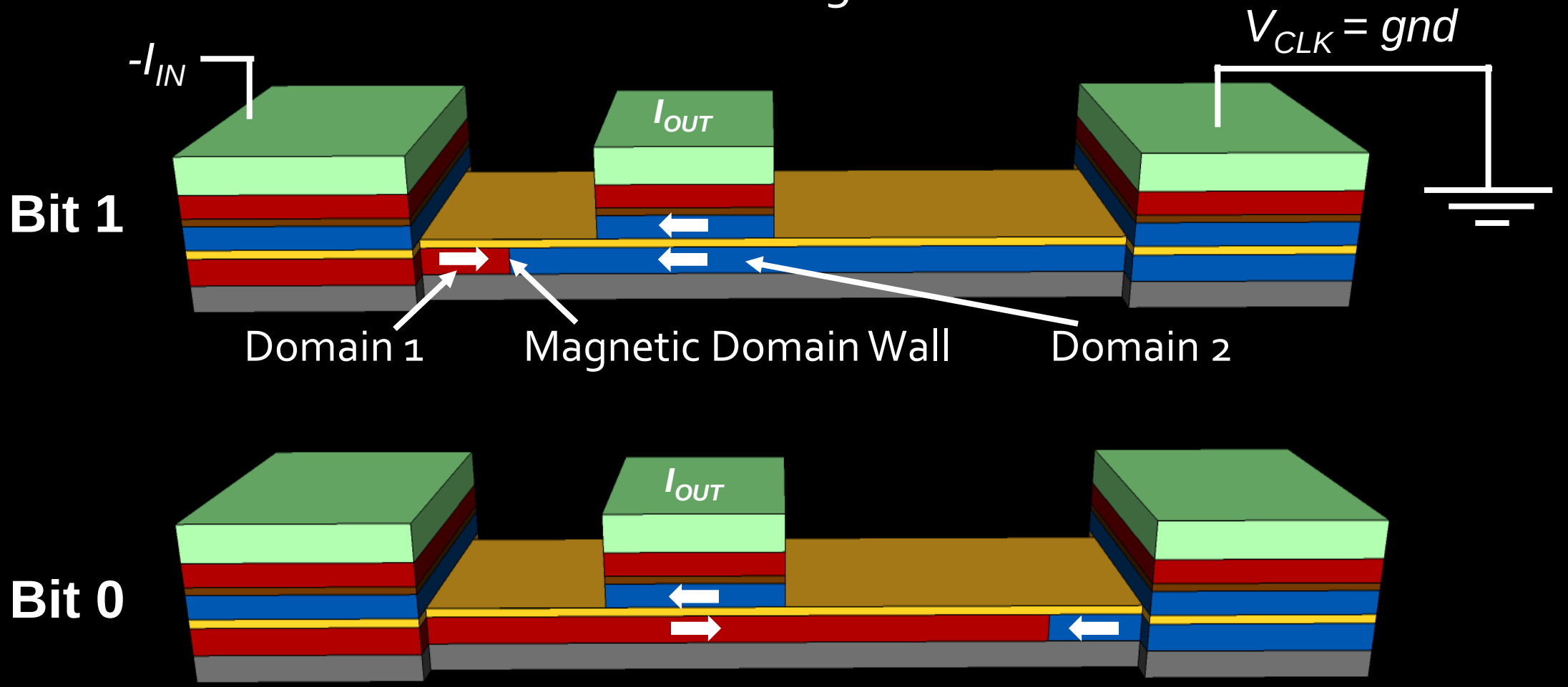
Wong et. al., **Nature Nano,** 191-194, 2015

Outline

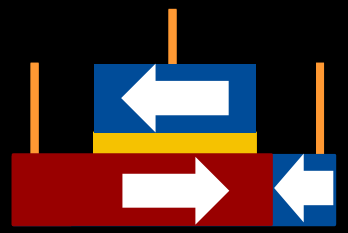
- Analog devices
  - Synaptic device
  - Thresholding device
- Digitized devices using magnetic tunnel junctions
  - Synaptic device
  - Thresholding device
- System design with analog devices

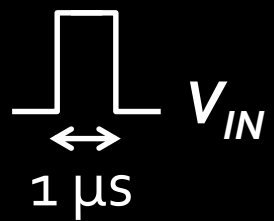
Novel logic device with domain wall-magnetic tunnel junction

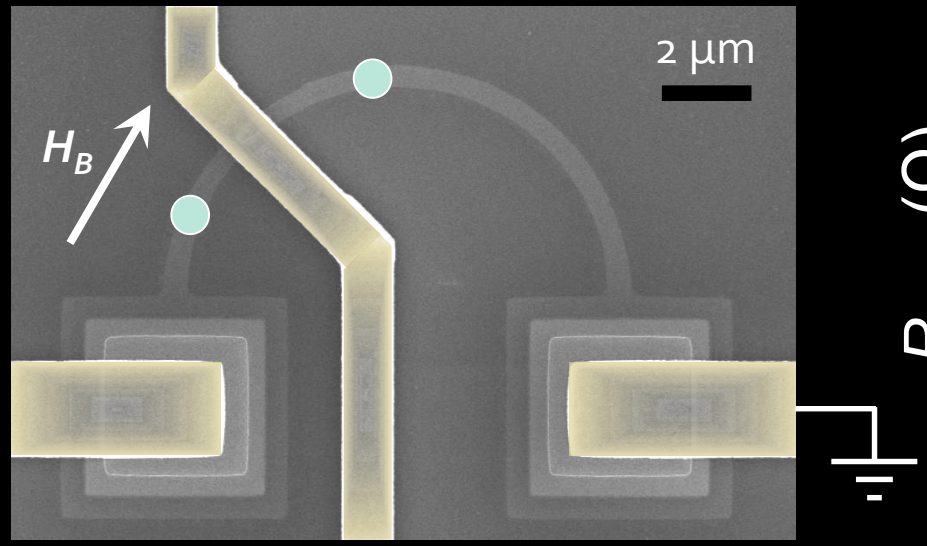
Position of domain wall defines the logic state of the device

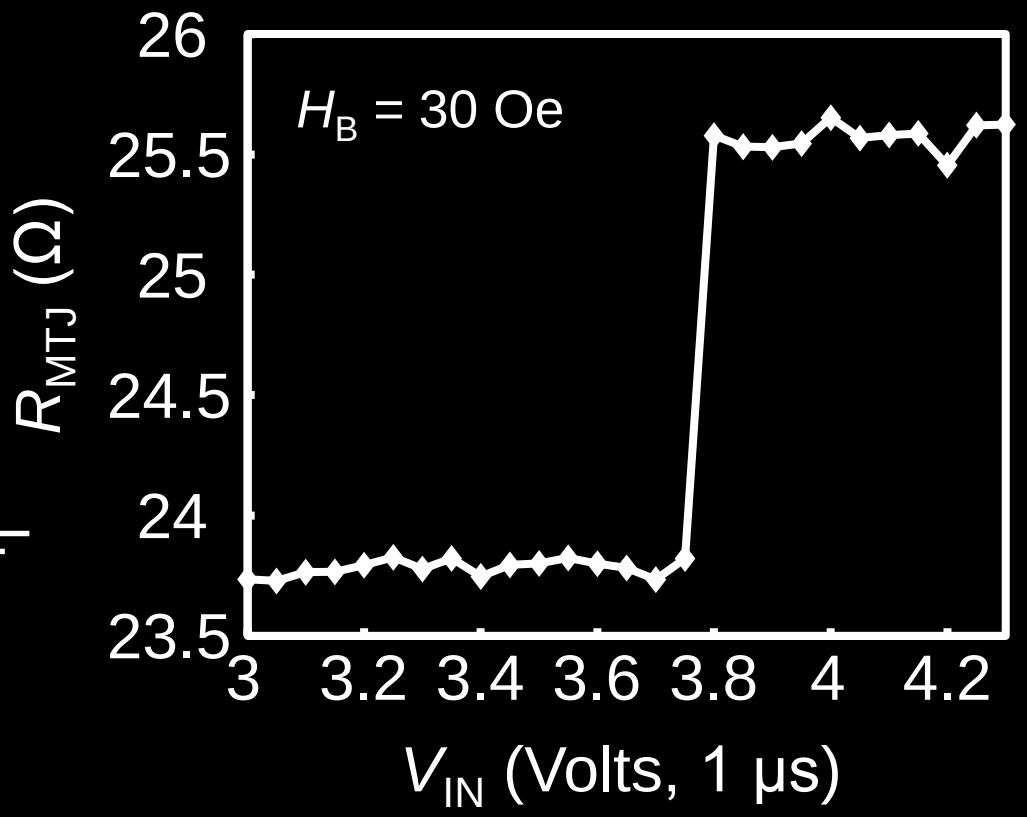


Inverter operation using Domain wall magnetic tunnel junction





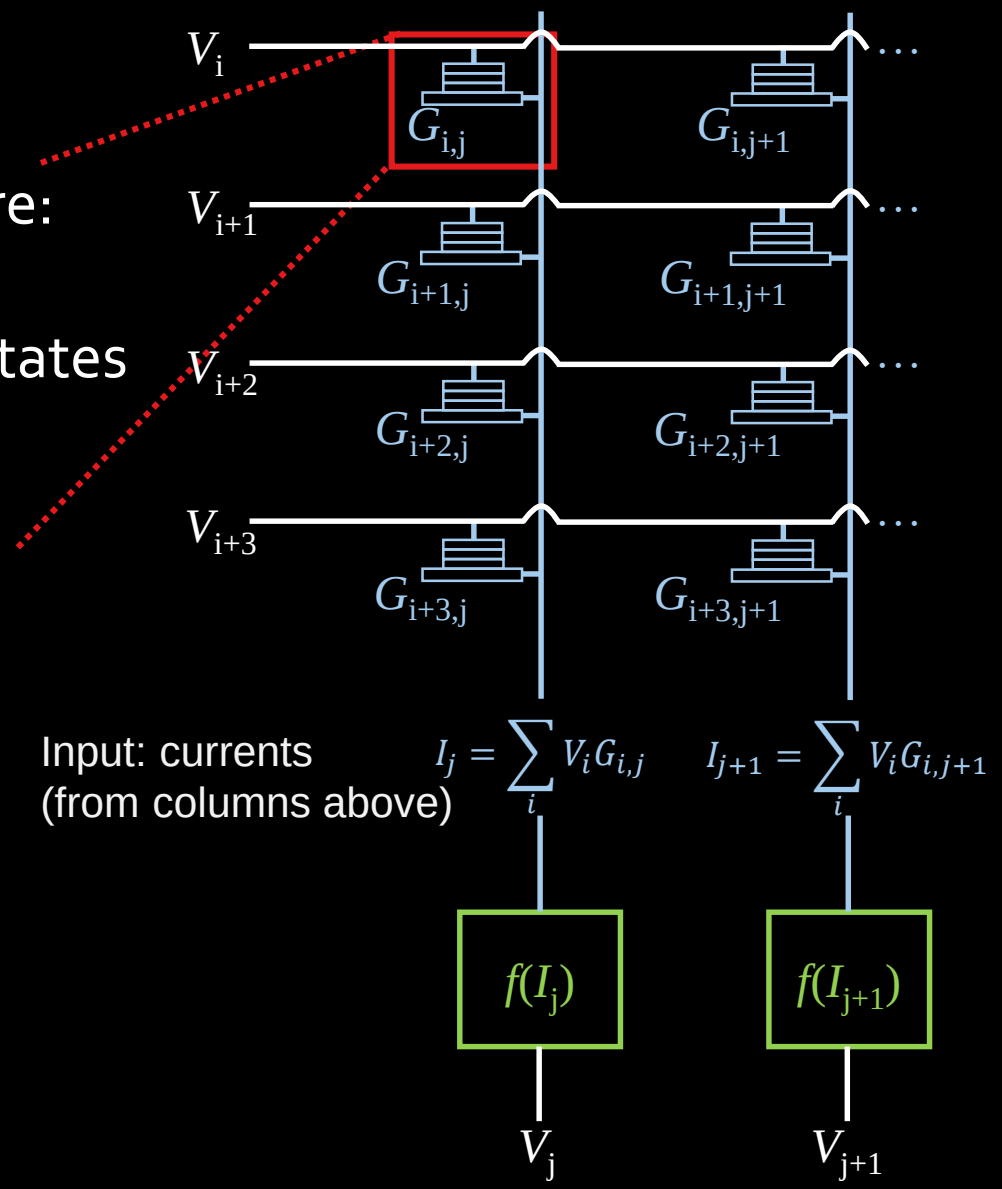




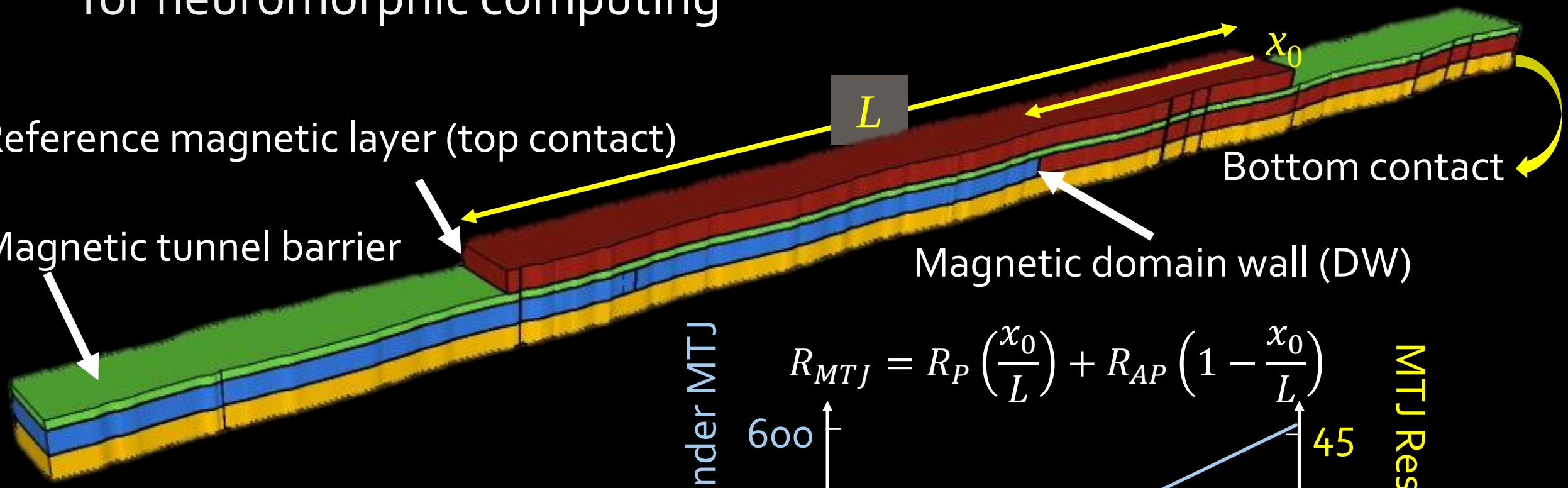
Architecture for in-memory computing: Neural network

Synaptic devices require:

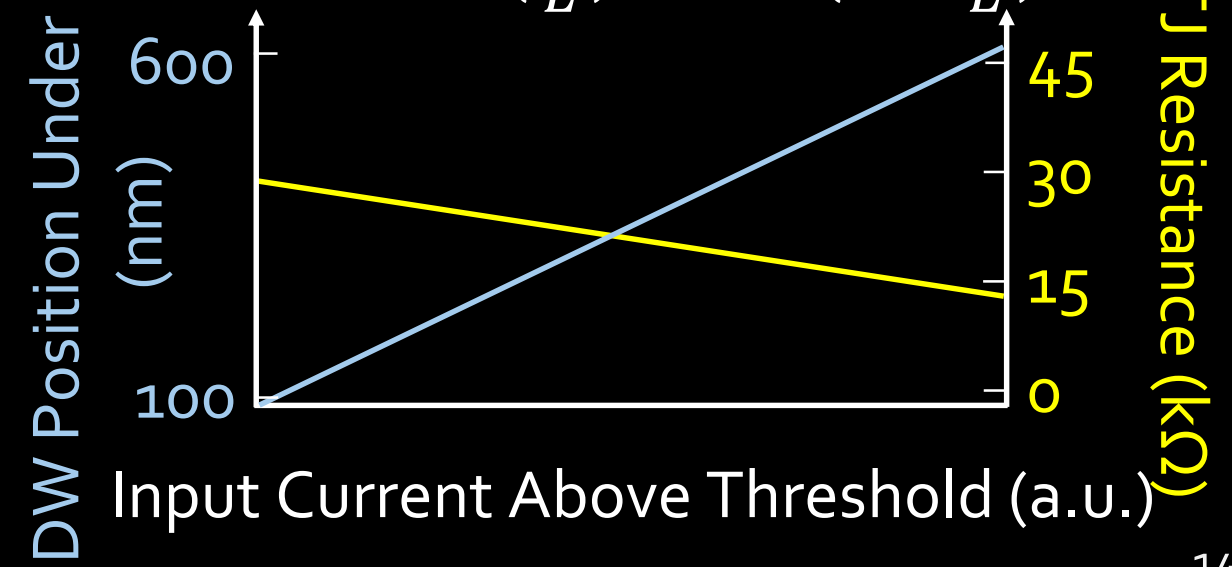
- Non-volatility
- Multiple resistance states
- Linear input-output characteristics



Analog circuitry (reconfigurable resistor as weighting device) for neuromorphic computing



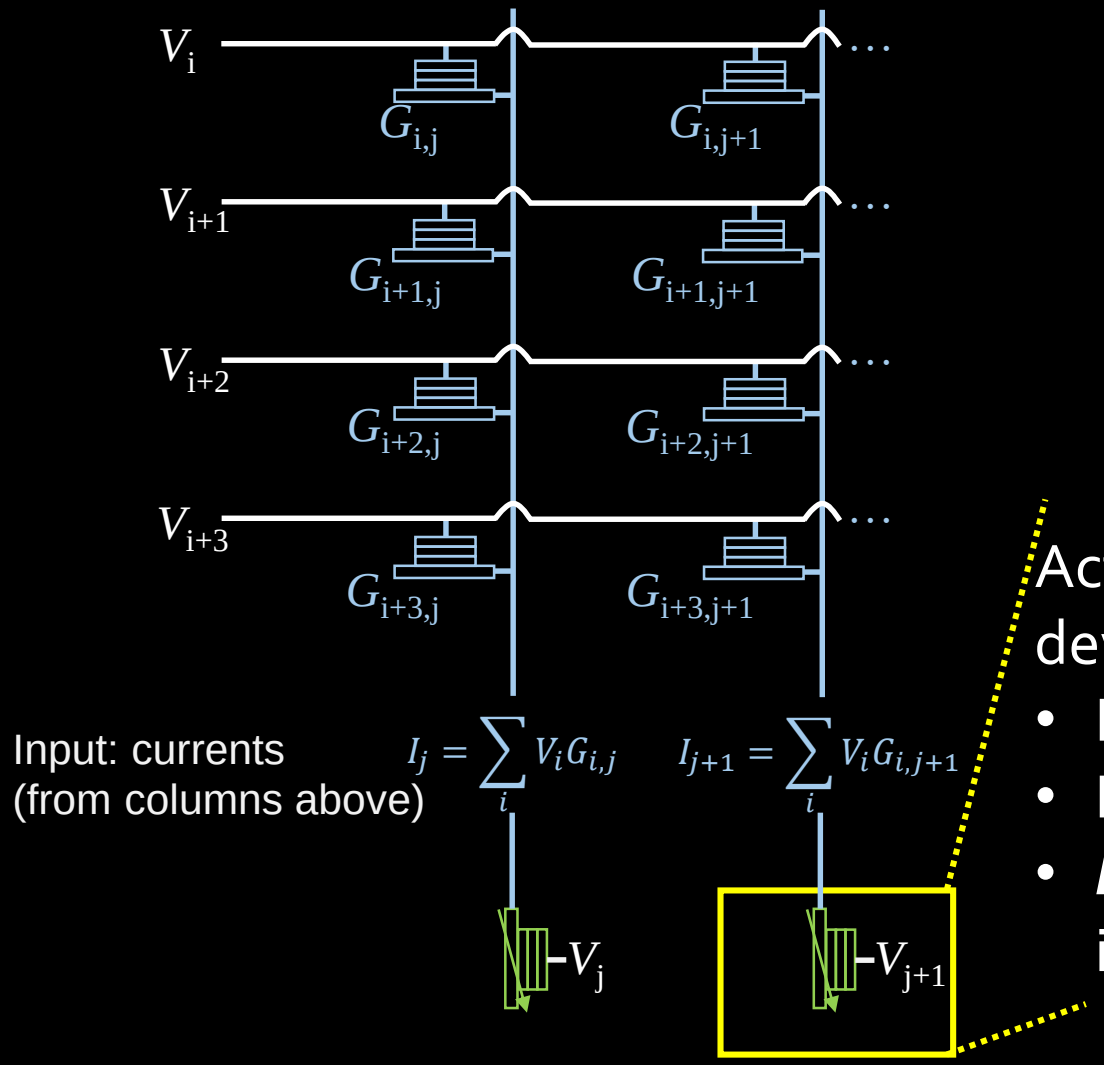
$$R_{MTJ} = R_P\left(\frac{x_0}{L}\right) + R_{AP}\left(1 - \frac{x_0}{L}\right)$$



Prof. Kaushik Roy, Jun 2016

Dutta, Siddiqui et al., IEEE Int Symp Nano, 83-88 (2017)

Architecture for in-memory computing: Neural network

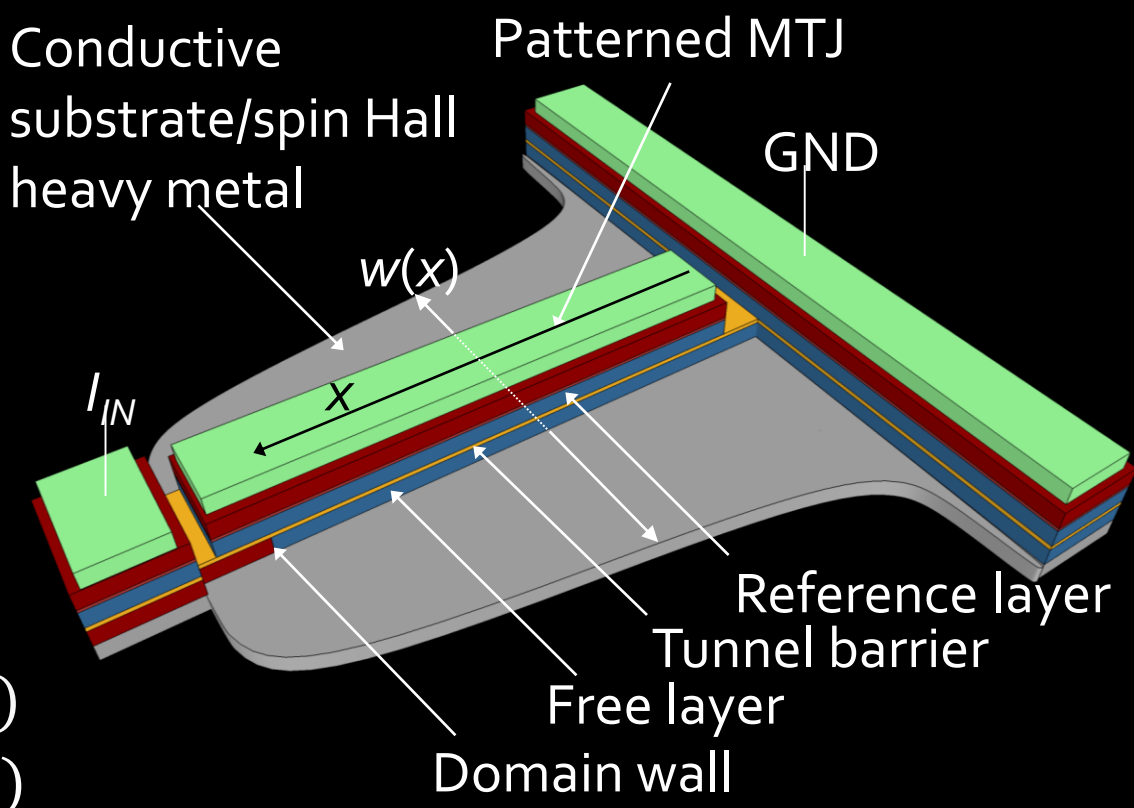


Activation/thresholding devices require:

- Non-volatility
- Multiple resistance states
- ***Programmable non-linear*** input-output characteristics

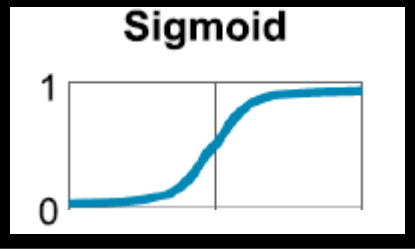
Analytical Model for Width $w(x)$ to generate activation function

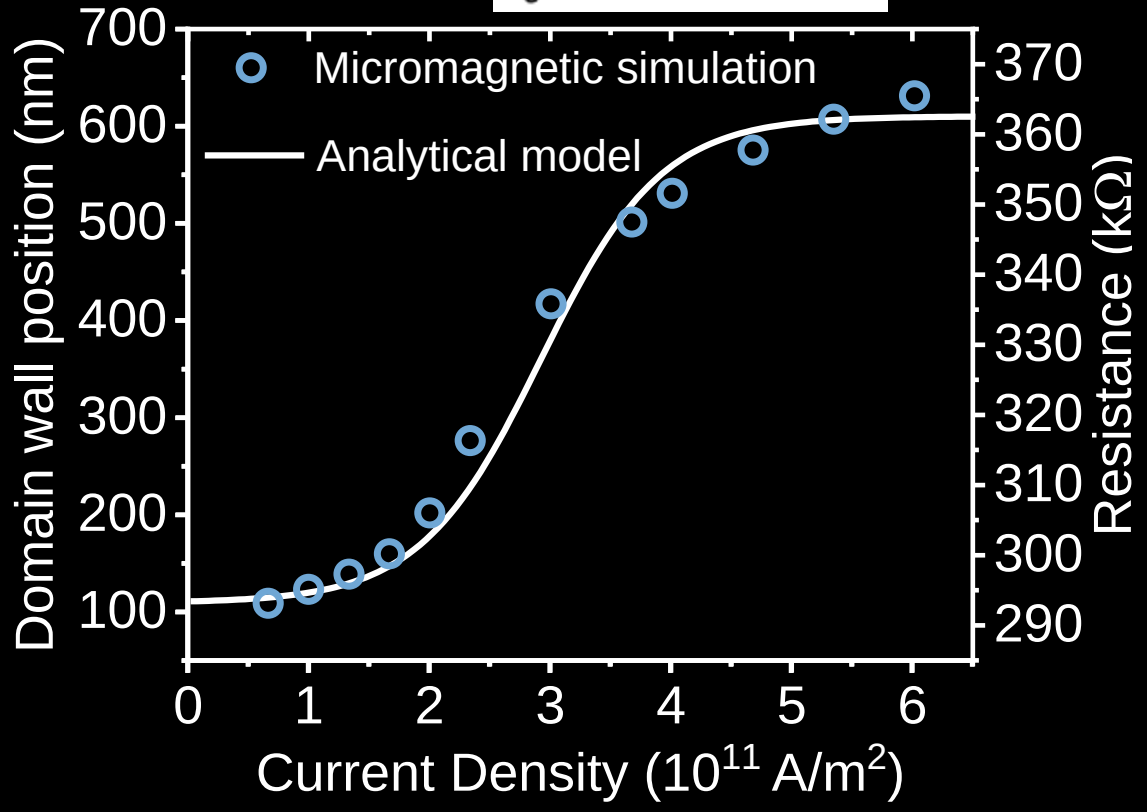
- Input: current I_{IN}
- Output: resistance R_{MTJ}
  - $R_{MTJ} = R_P\left(\frac{x_0}{L}\right) + R_{AP}\left(1 - \frac{x_0}{L}\right)$
  - x_0 = final domain wall position
  - L = length of the magnetic tunnel junction
- Given:
  - $\frac{dx}{dt} = v(x) = \begin{cases} 0, & J < J_C \text{ (i.e. } x = 0) \\ \eta(J(x) - J_C), & J \geq J_C \text{ (i.e. } x > 0) \end{cases}$
  - $\frac{dx}{dt} = \eta \frac{I(x_0) - I_C}{w(x)d}$
- Solution after integration: $w(x_0) = \frac{\eta t_0}{d} \frac{dI(x_0)}{dx_0}$



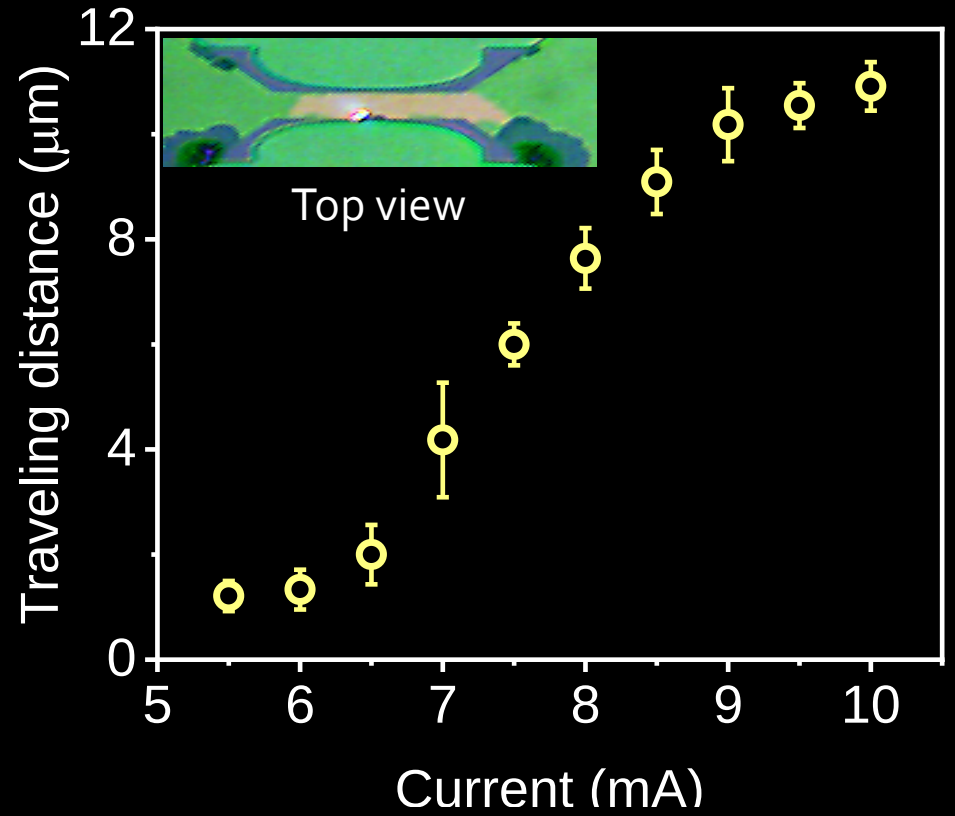
Dutta, Siddiqui et al., IEEE Int Symp Nano, 83-88 (2017).

Analog circuitry (threshold function generator) for neuromorphic computing

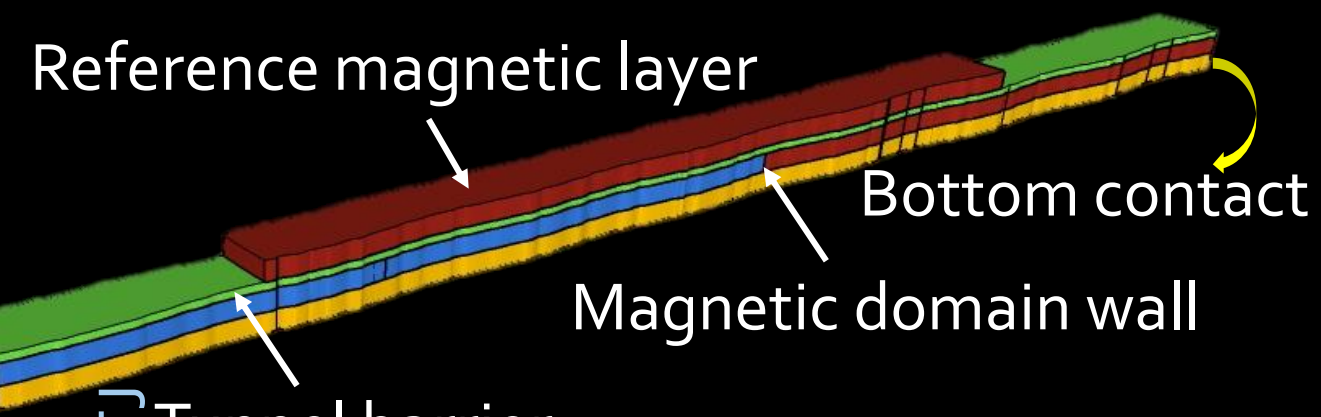


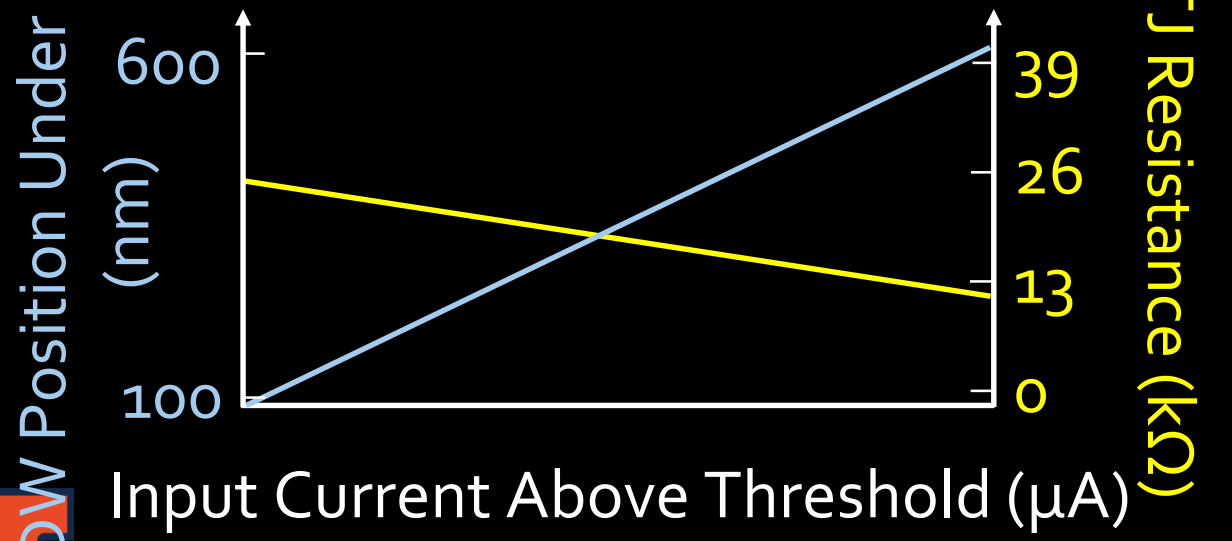


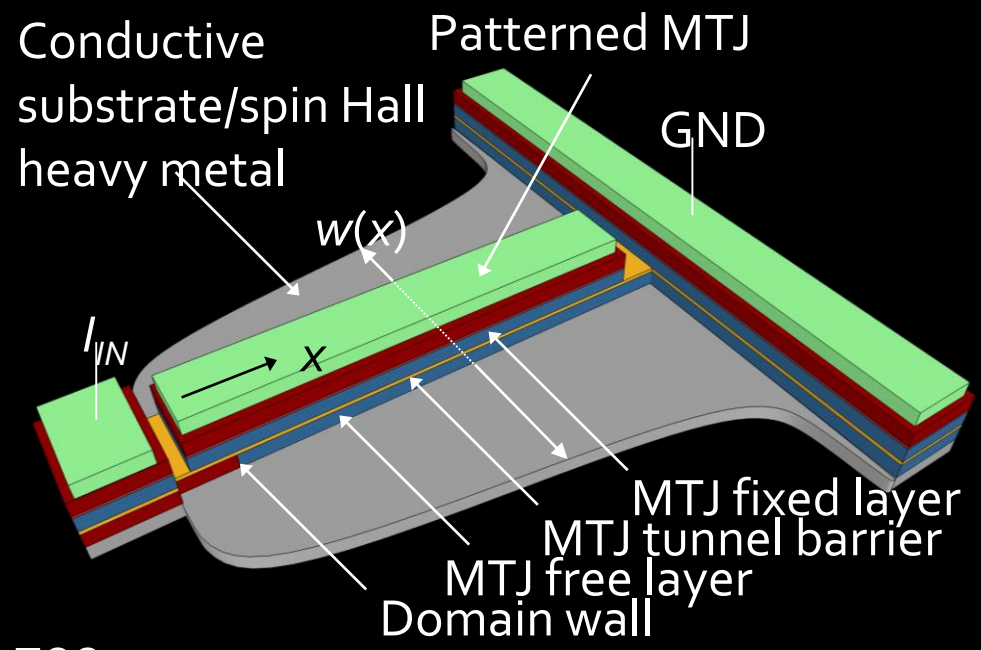
Electrical measurement with optical readout

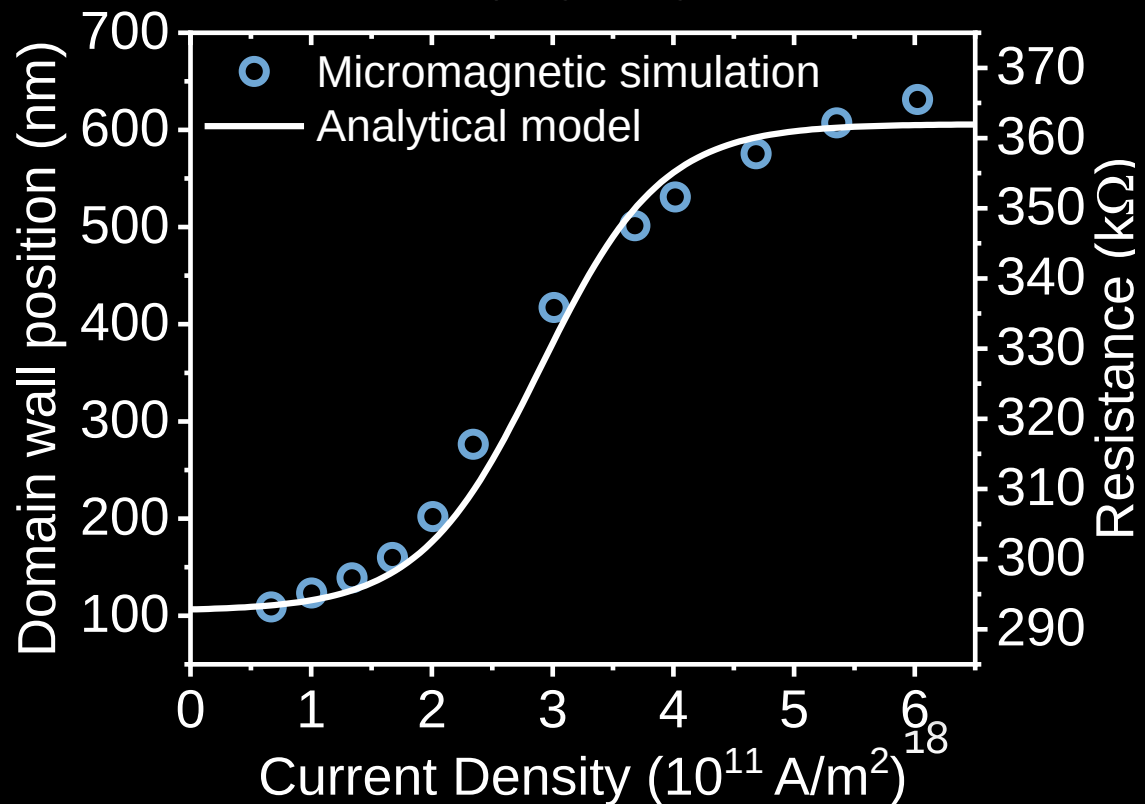


Analog synaptic and thresholding device





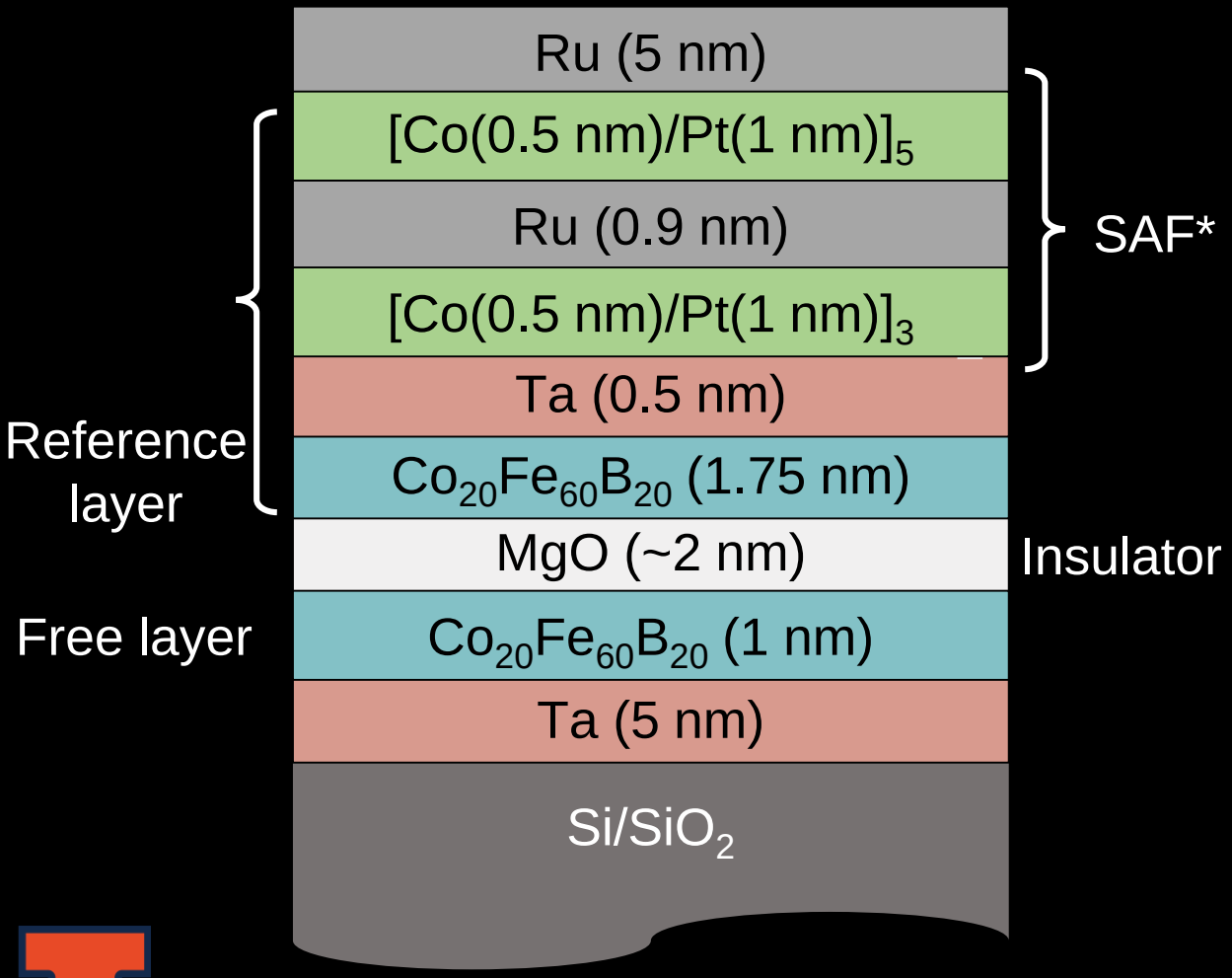


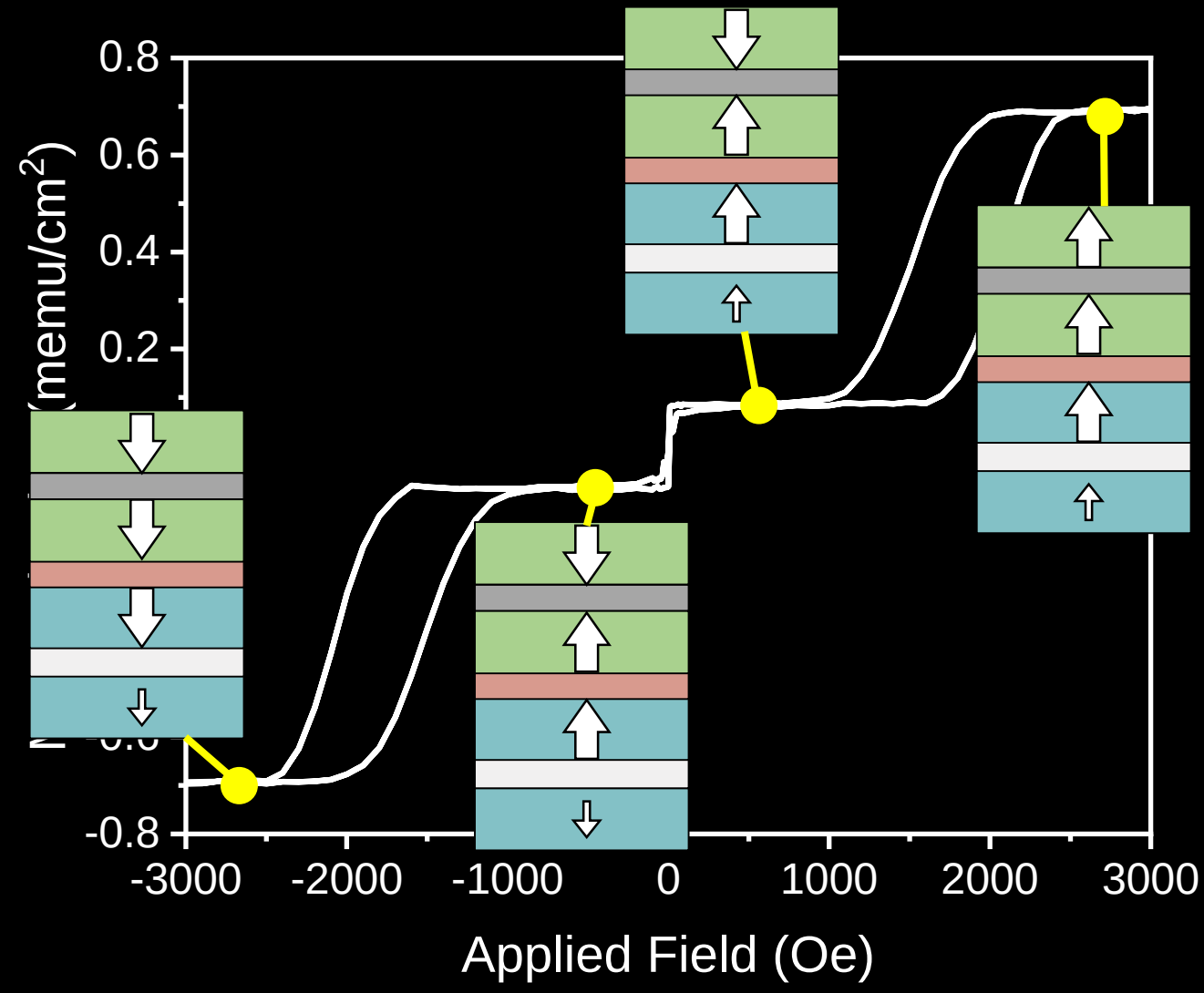


Outline

- Analog devices
  - Synaptic device
  - Thresholding device
- Digitized devices using magnetic tunnel junctions
  - Synaptic device
  - Thresholding device
- System design with analog devices

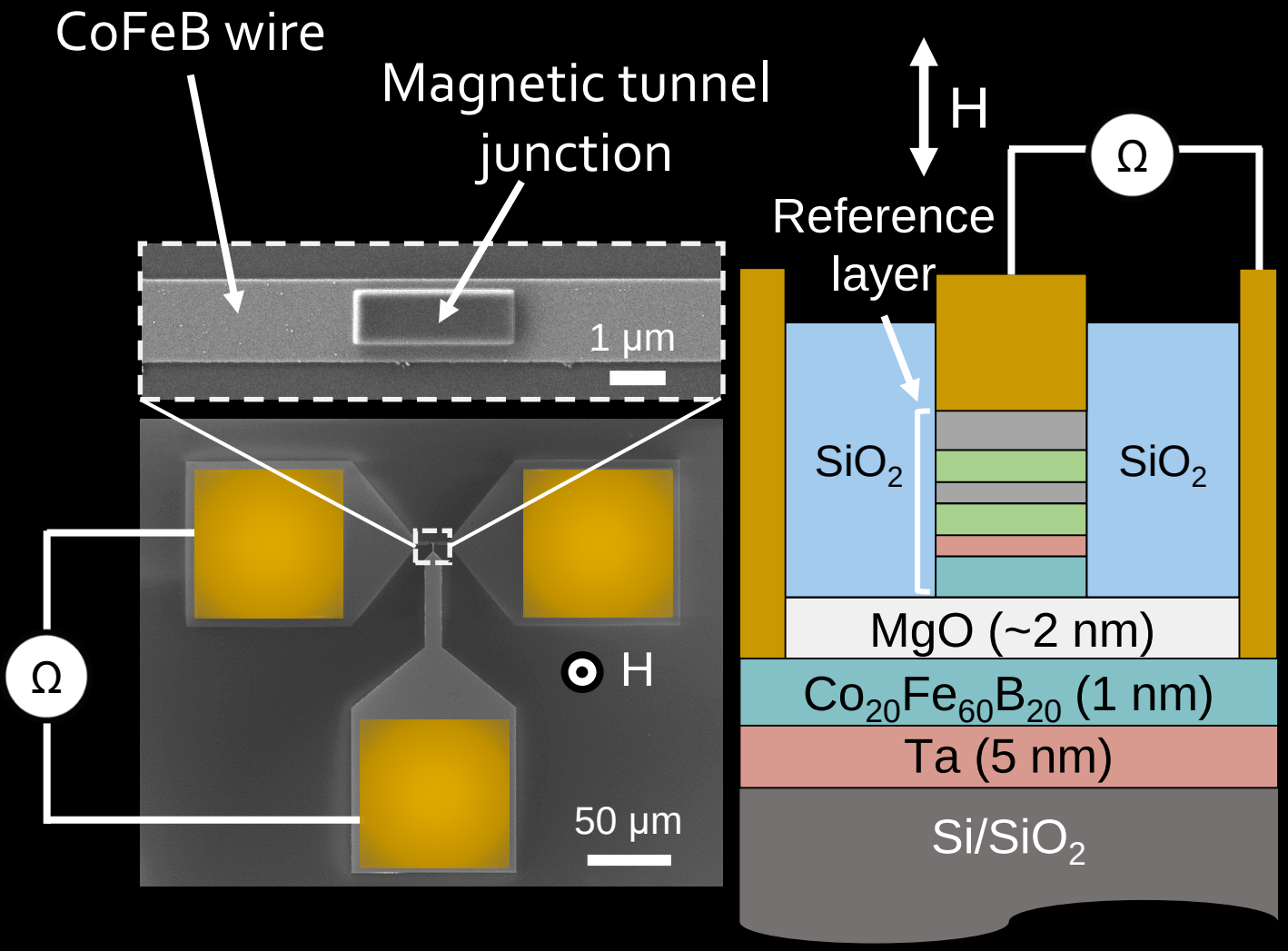
Development of magnetic tunnel junction



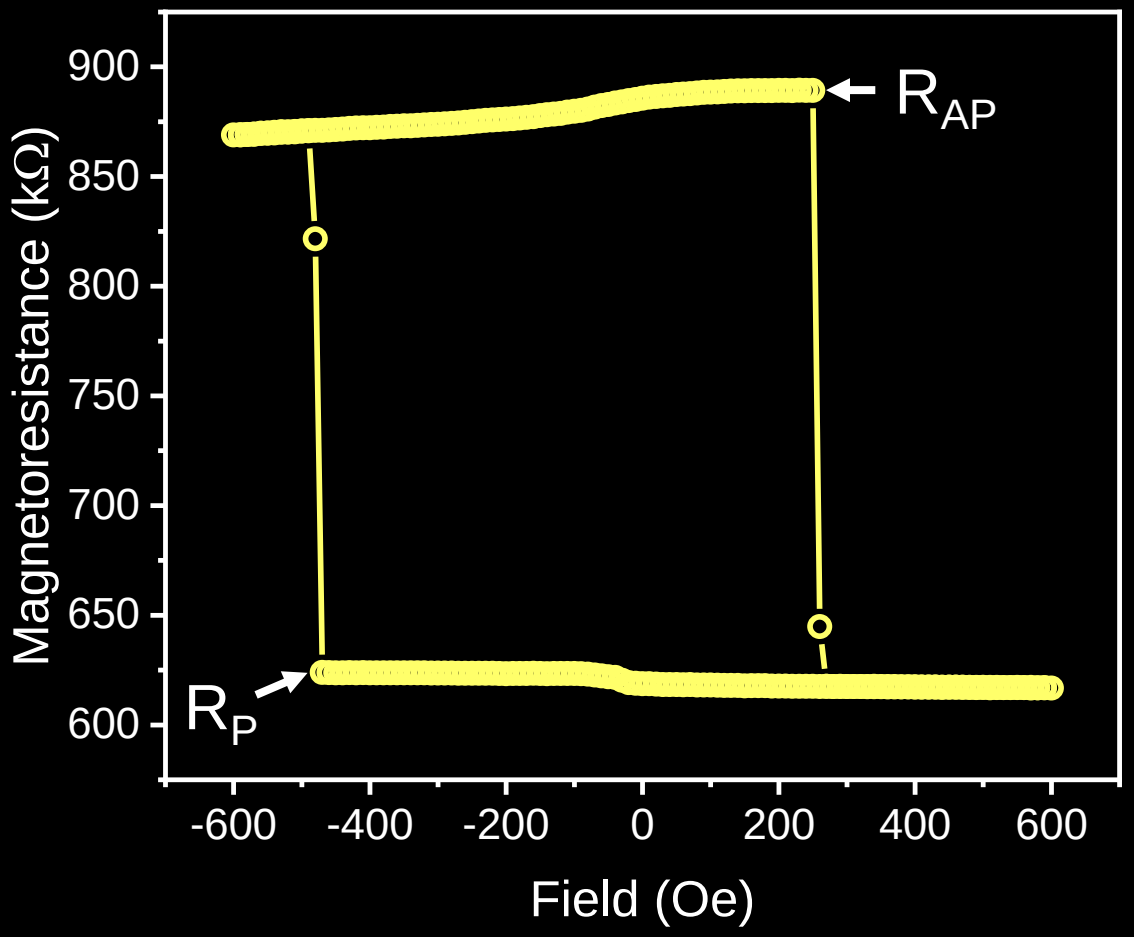


*SAF = Synthetic antiferromagnet

Development of magnetic tunnel junction

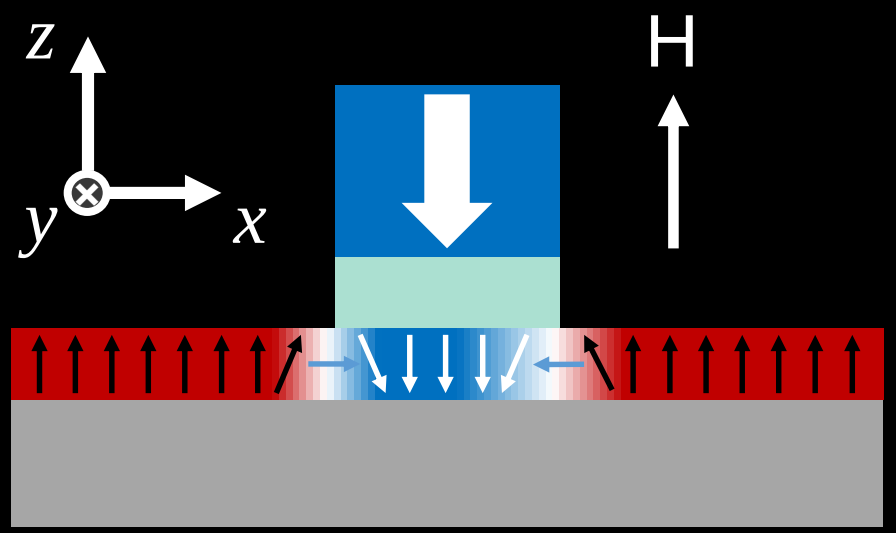


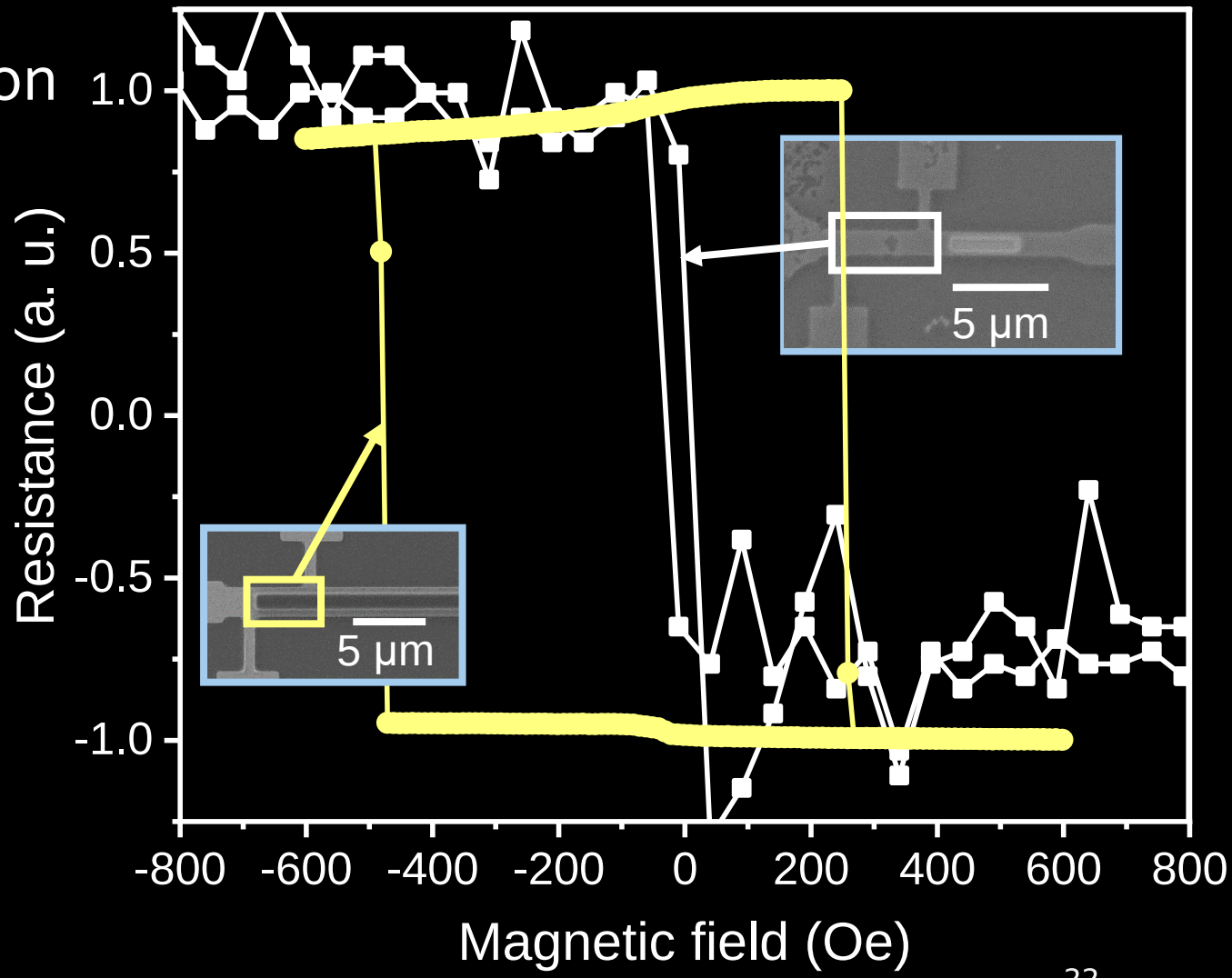
$$TMR = \frac{(R_{AP} - R_P)}{R_P} = 44\%$$



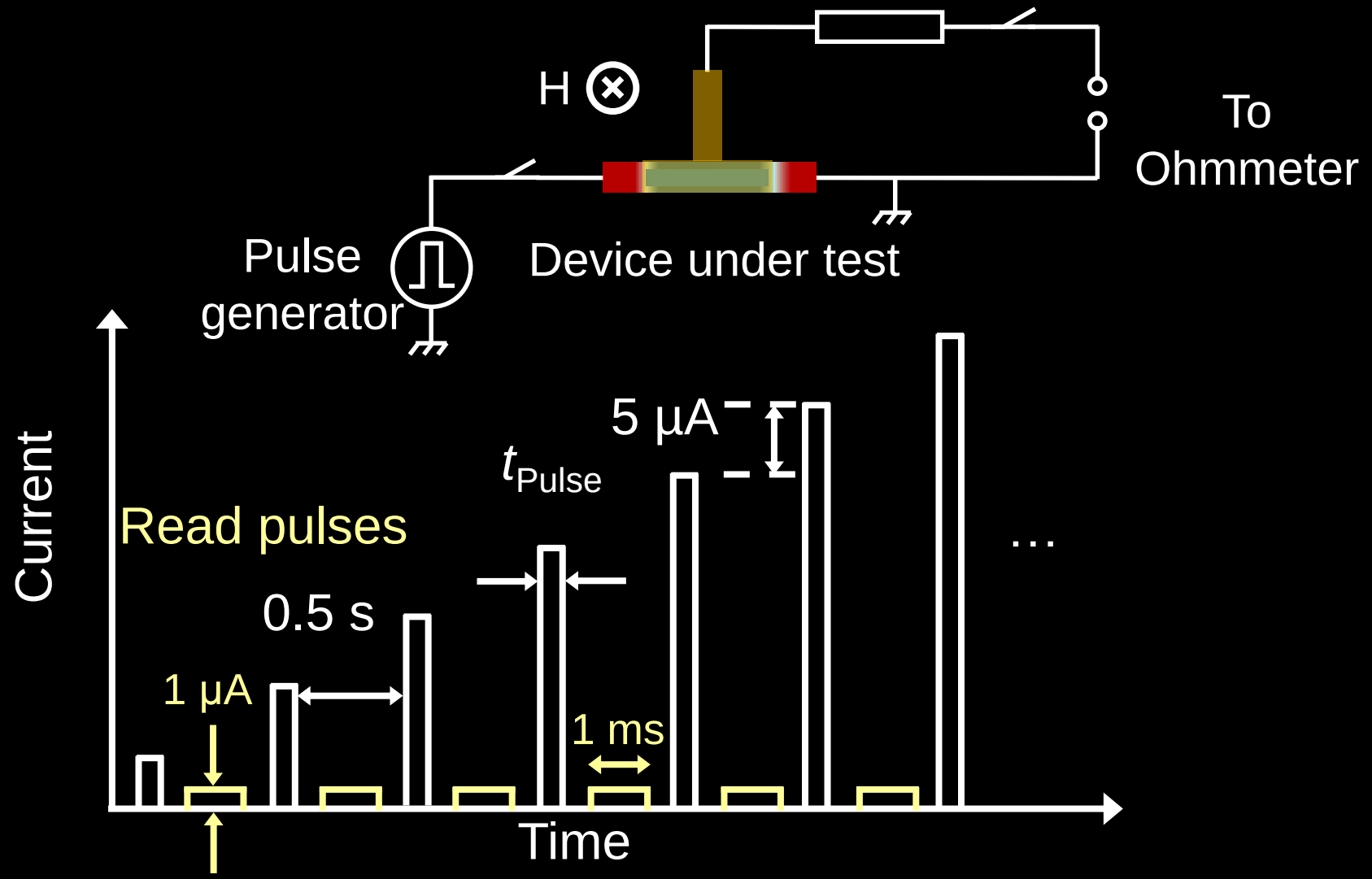
Different switching field of the CoFeB wire helps to nucleate domain walls on the edges of tunnel junctions

Side view of magnetic tunnel junction

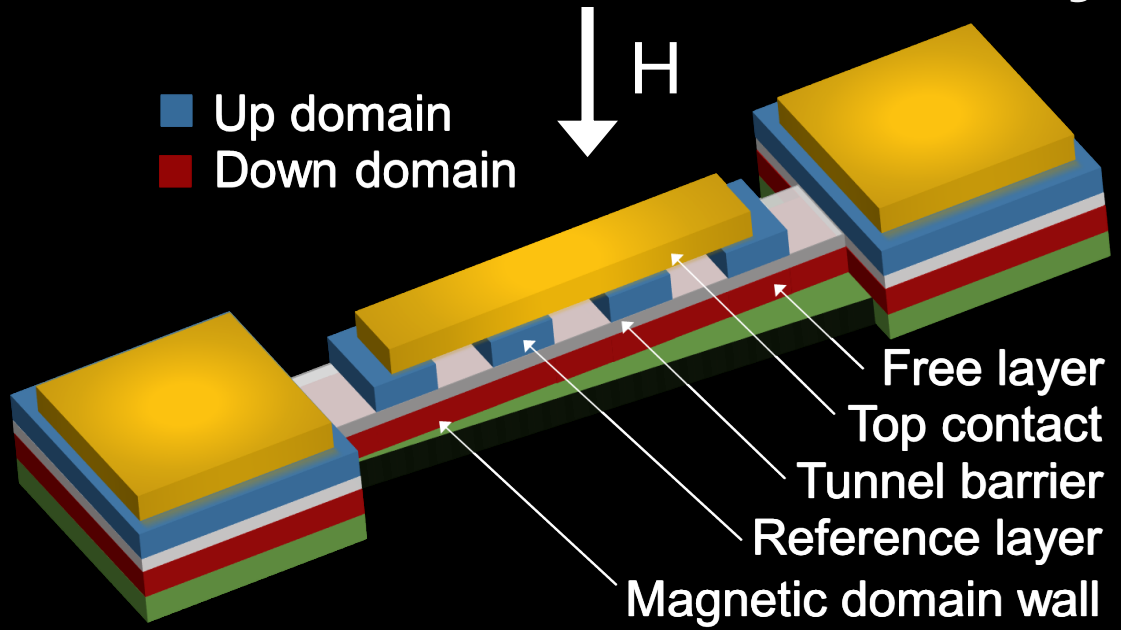




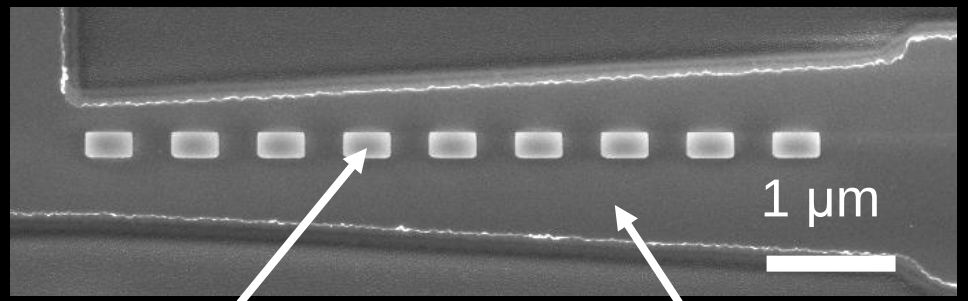
Measurement setup for current induced domain wall motion under magnetic tunnel junction

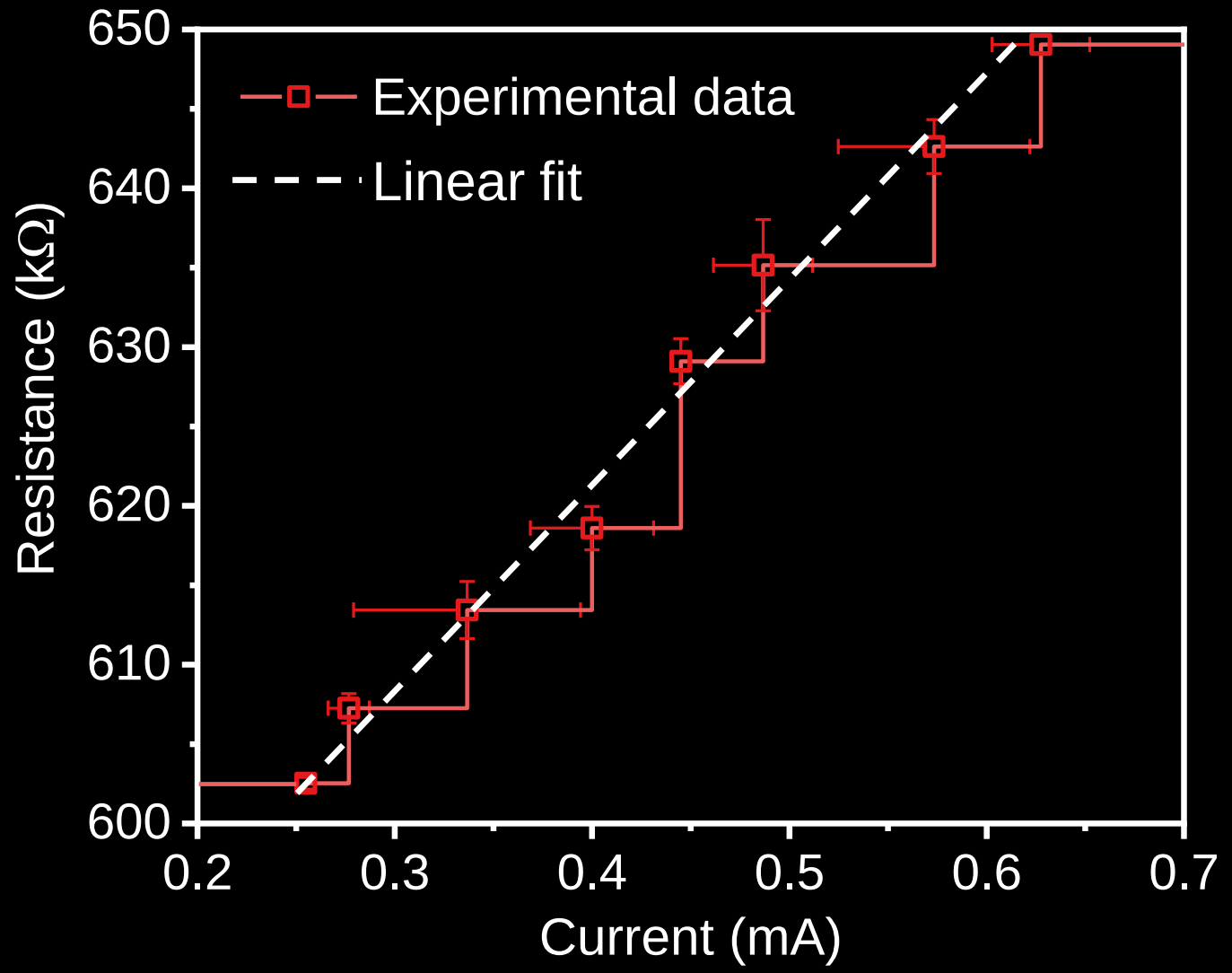


Synaptic weight generation using multiple tunnel junctions



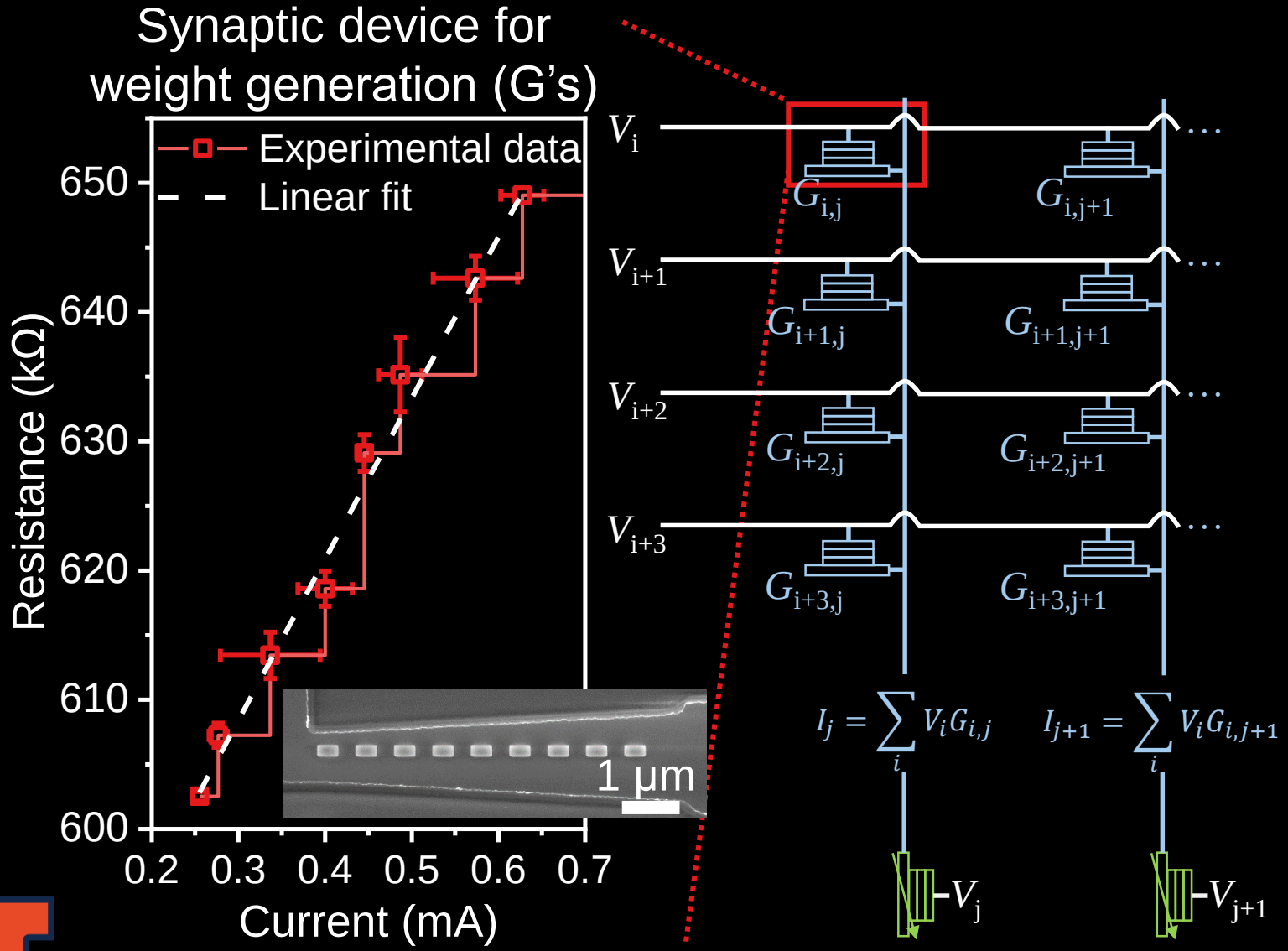
Synaptic device without contacts





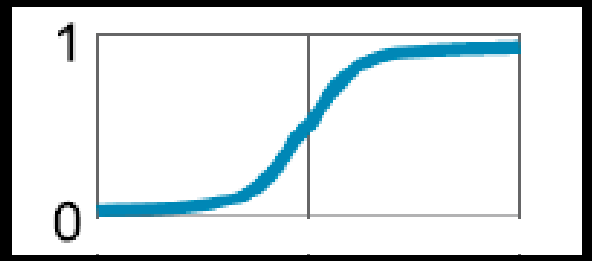
Siddiqui et al., Nano Lett. , 20, 2, 1033–1040 (2020)

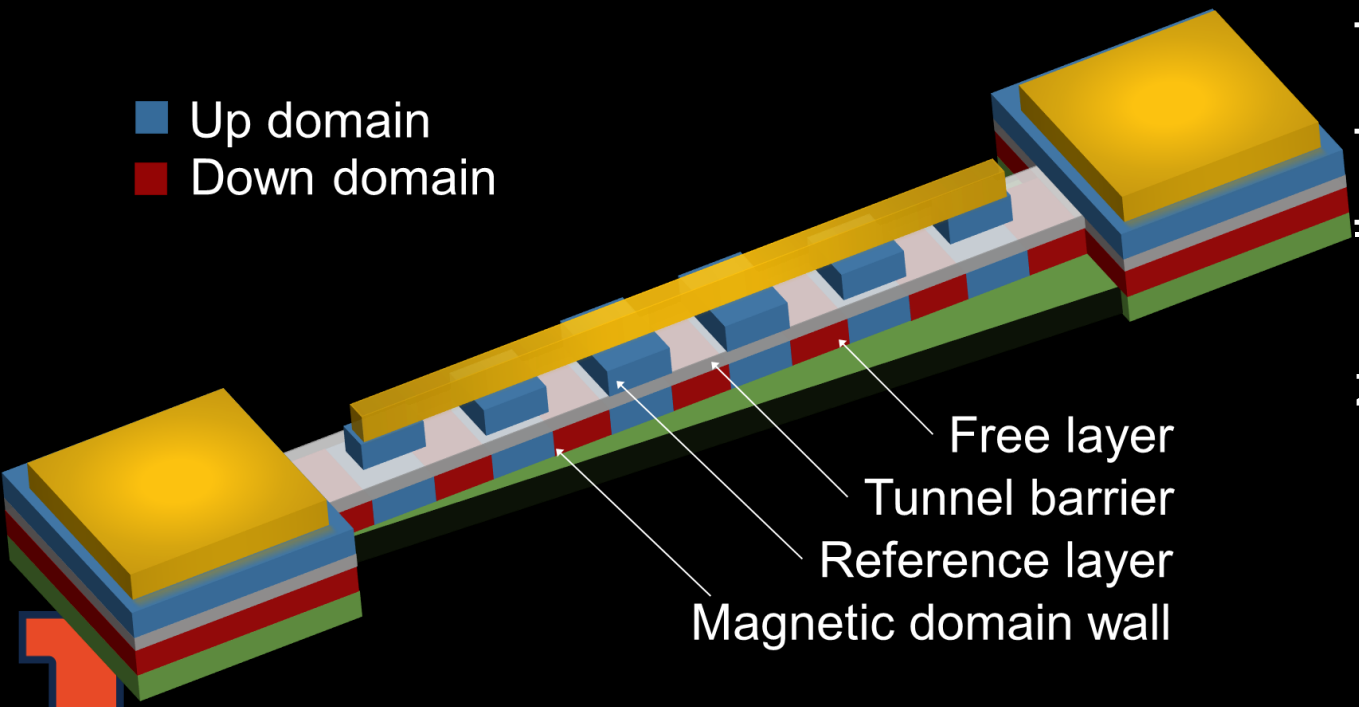
Summary of synaptic devices



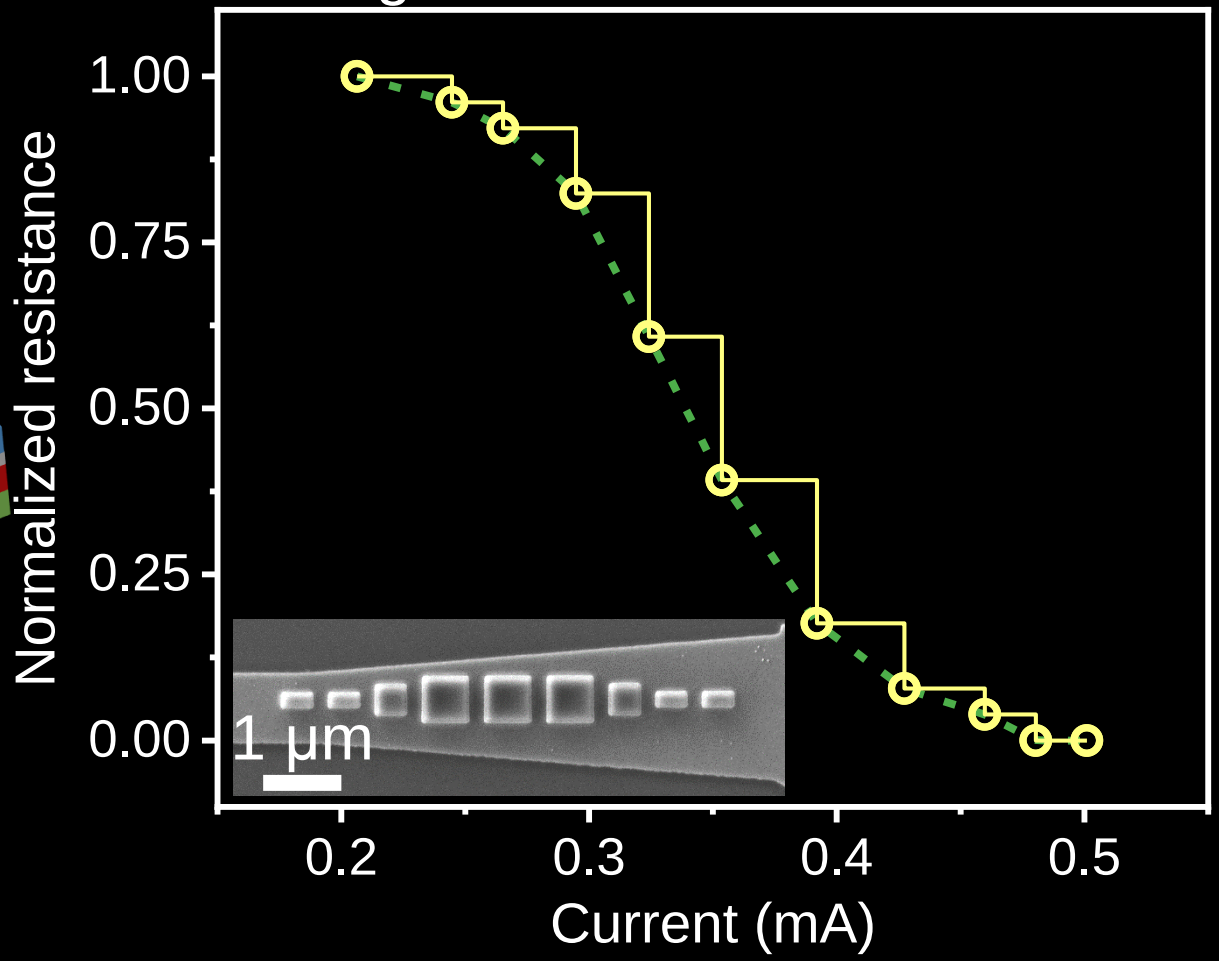
Activation function generation by engineering the area of the magnetic tunnel junctions

Sigmoid activation function



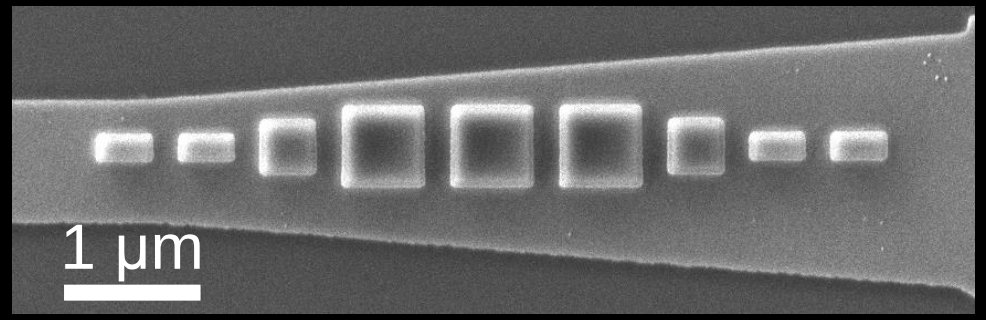


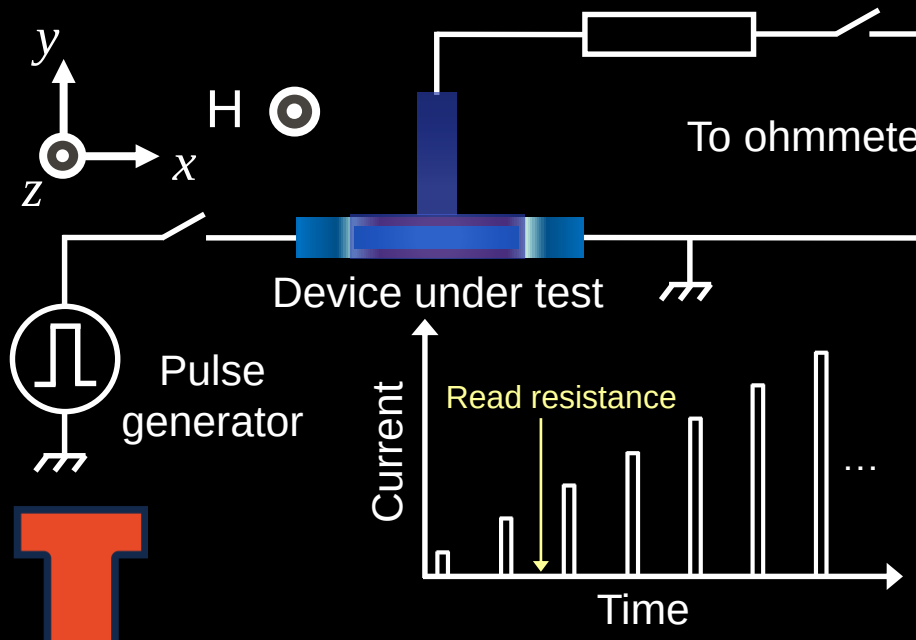
Design of mirrored and shifted sigmoid activation function

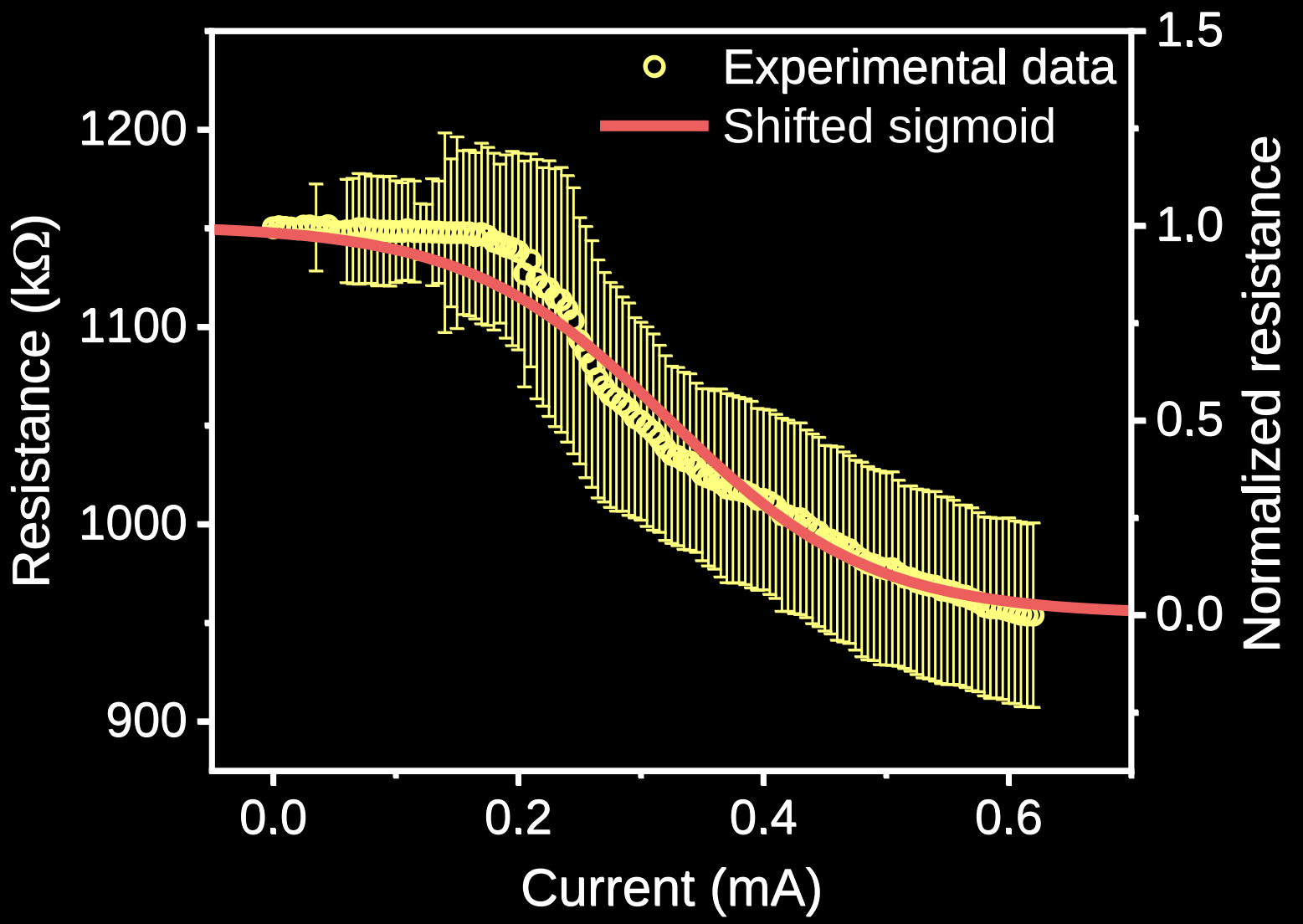


Siddiqui et al., Nano Lett. , 20, 2, 1033–1040 (2020)

Experimental demonstration of activation function generation by engineering the area of the magnetic tunnel junctions



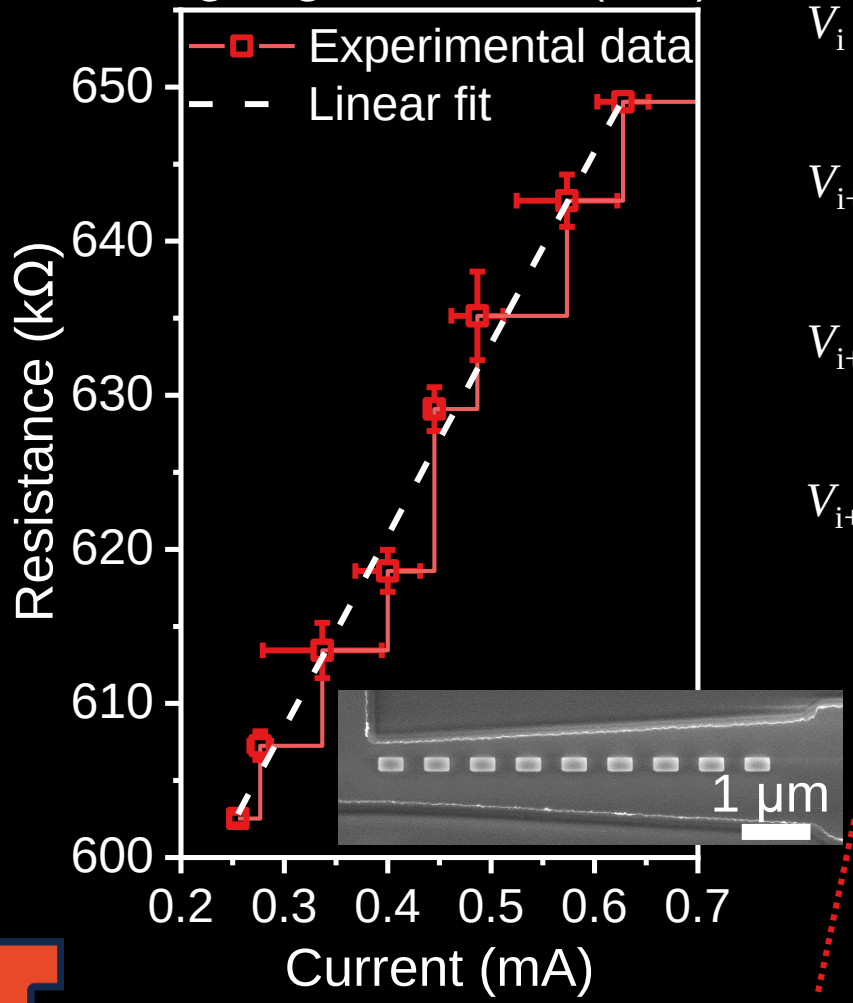


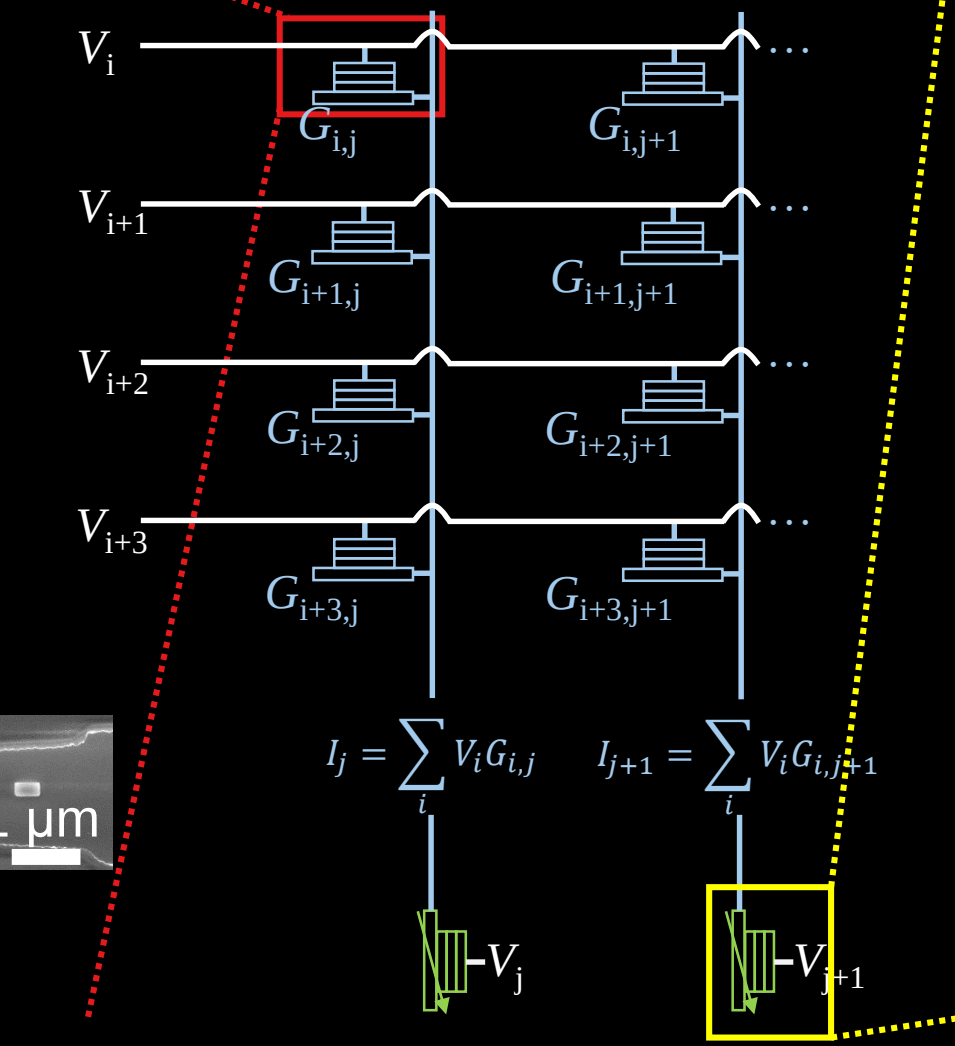


Siddiqui et al., Nano Lett. , 20, 2, 1033–1040 (2020)

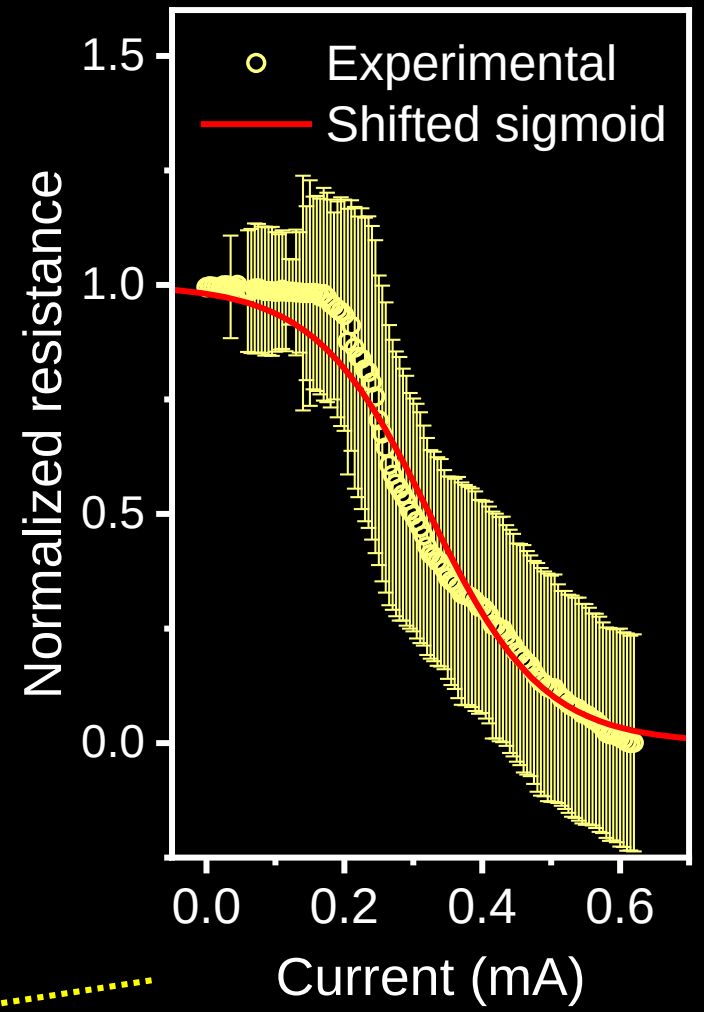
Summary of synaptic and thresholding devices

Synaptic device for weight generation (G's)



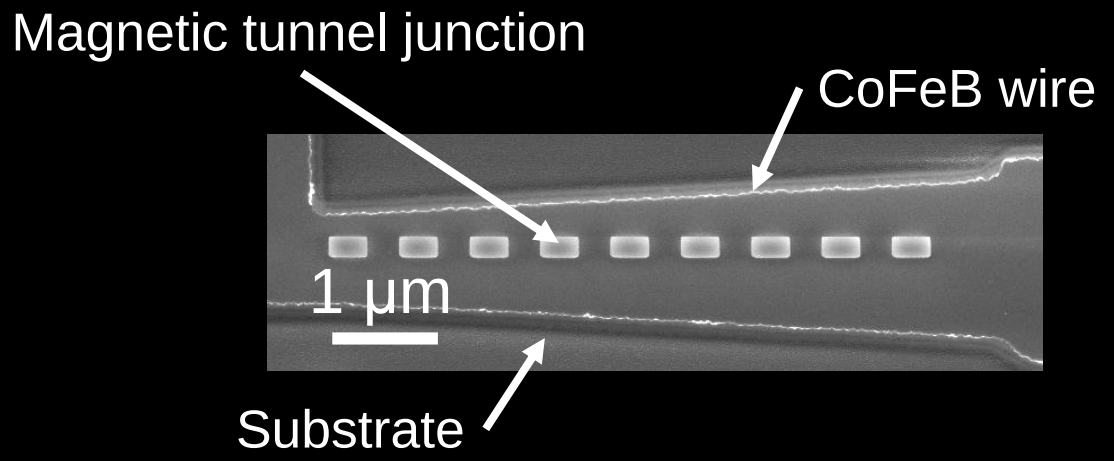


Thresholding device for activation



Required energy is very small compared to other non-volatile memories

- Prototype device
  - The power required to switch the prototype MTJs = 150 µW – 2000 µW.
  - Current pulse width = 8 nsec
  - Energy consumption per weight update = 1 pJ – 16 pJ ⟶ similar to resistive RAM
- Scaled device
  - MTJ width = 20 nm
  - Width of CoFeB wire = 25 nm
  - Current pulse width = 8 nsec
  - Energy consumption per weight update = 18 fJ – 36 fJ



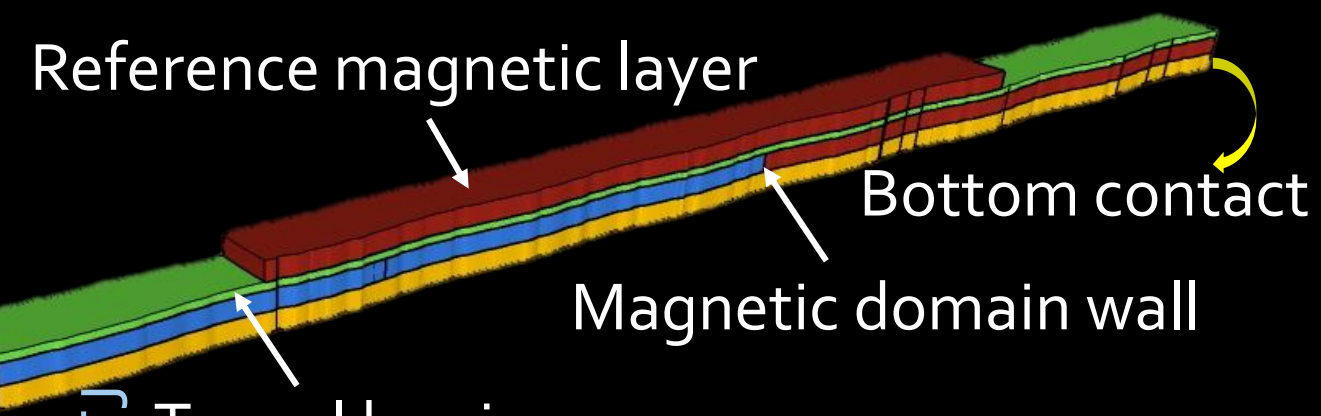
Biological synapses need ~fJ energy to change the weight

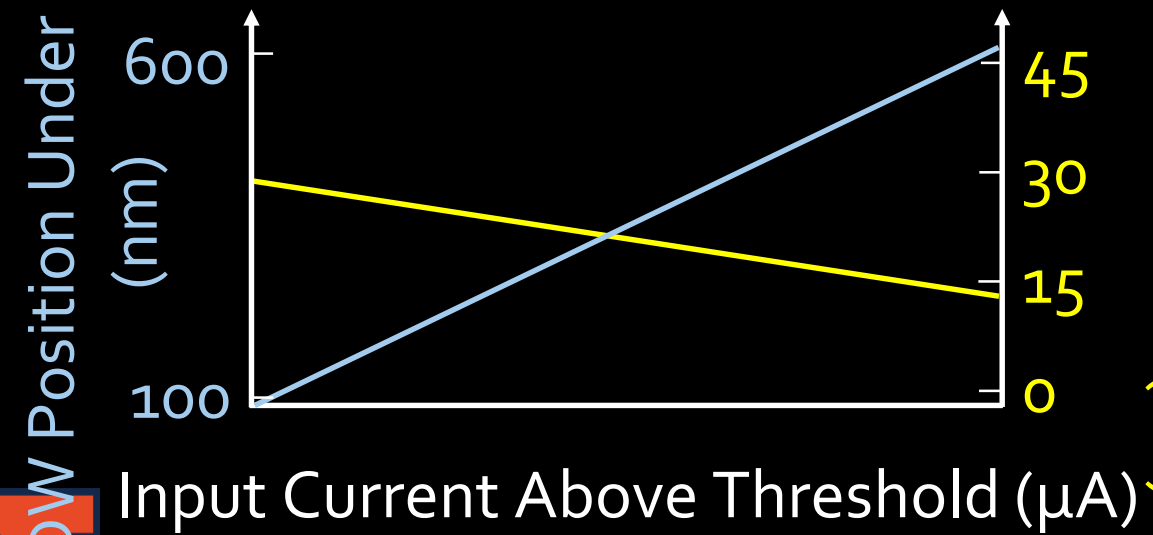
Siddiqui et al., Nano Lett. , 20, 2, 1033–1040 (2020)
Laughlin *et al*., Nat Neurosci, 1, 36-41, 1998

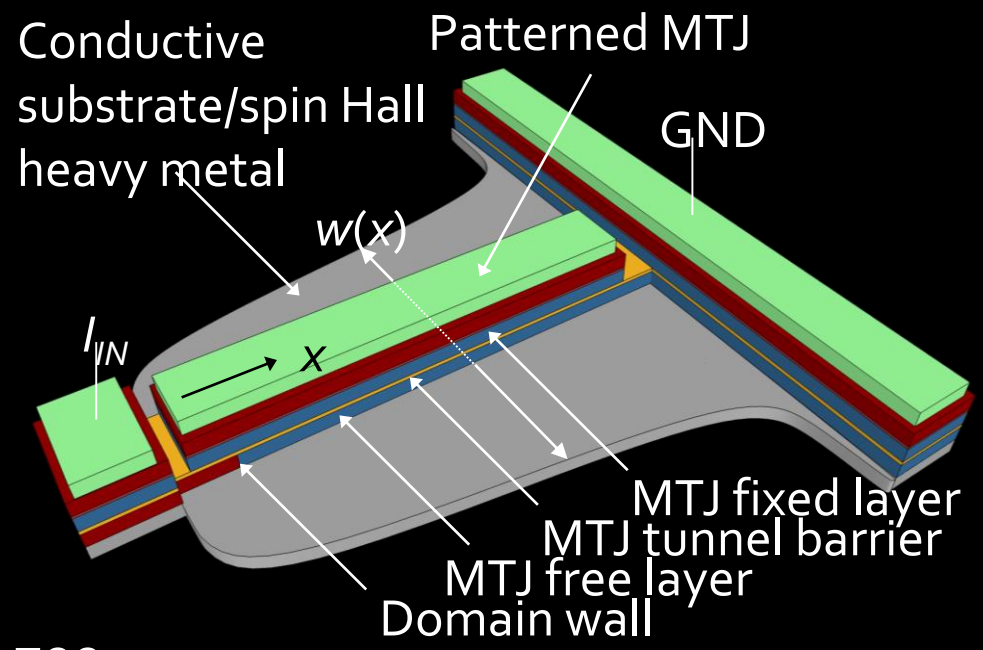
Outline

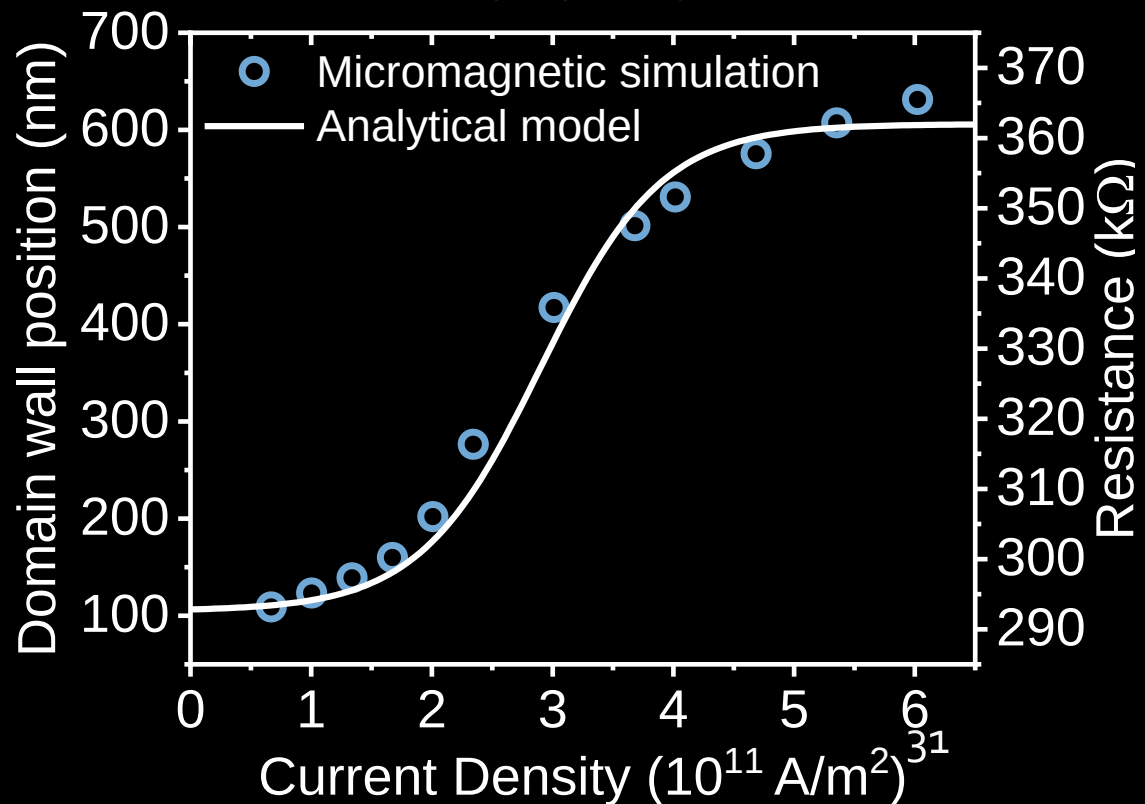
- Analog devices
  - Synaptic device
  - Thresholding device
- Digitized devices using magnetic tunnel junctions
  - Synaptic device
  - Thresholding device
- System design with analog devices

Analog synaptic and thresholding device







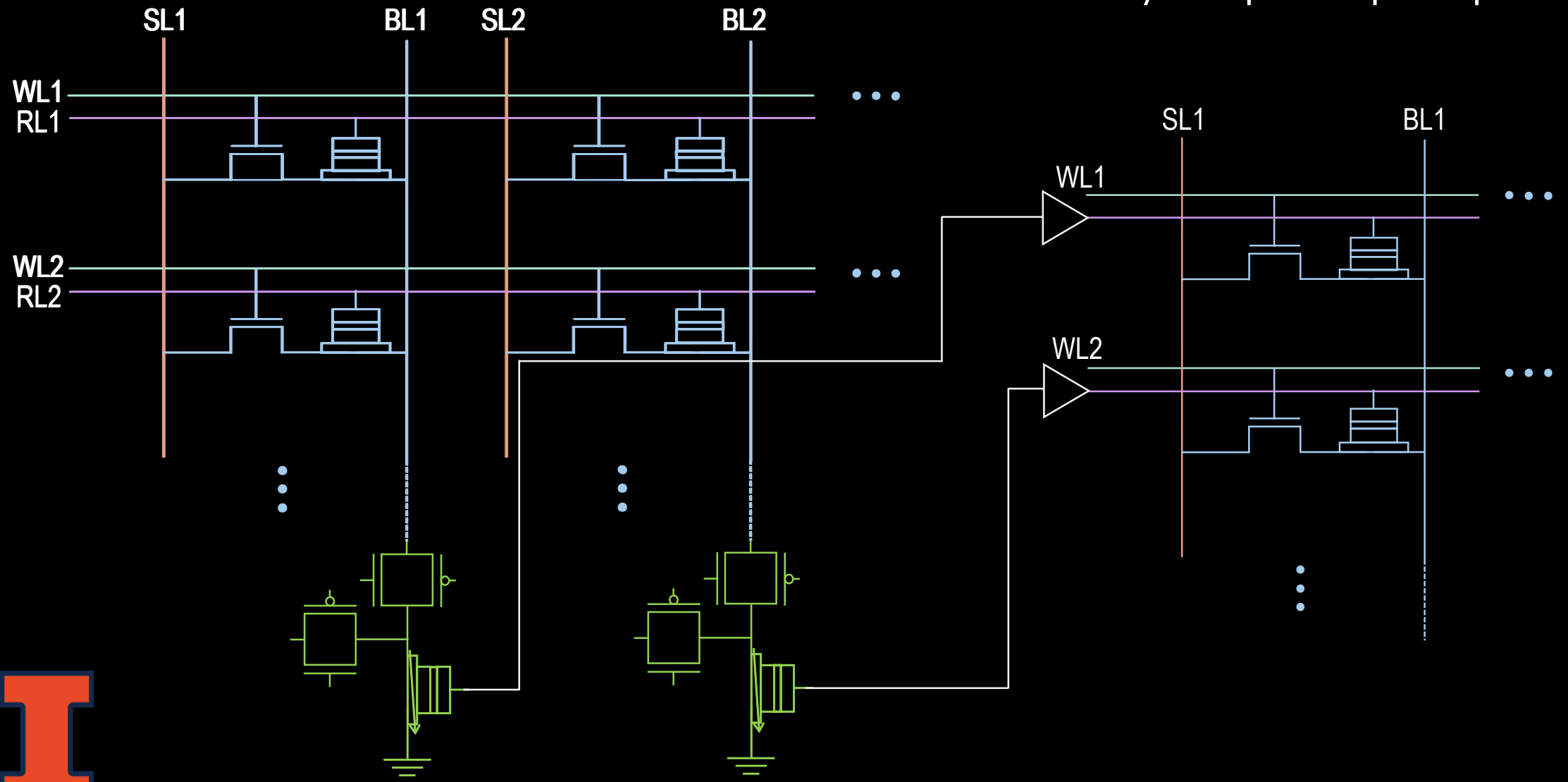


Logic-in-Memory System Design

Crosspoint array of a 3-terminal MTJs:

Operation: Weighted sum operations

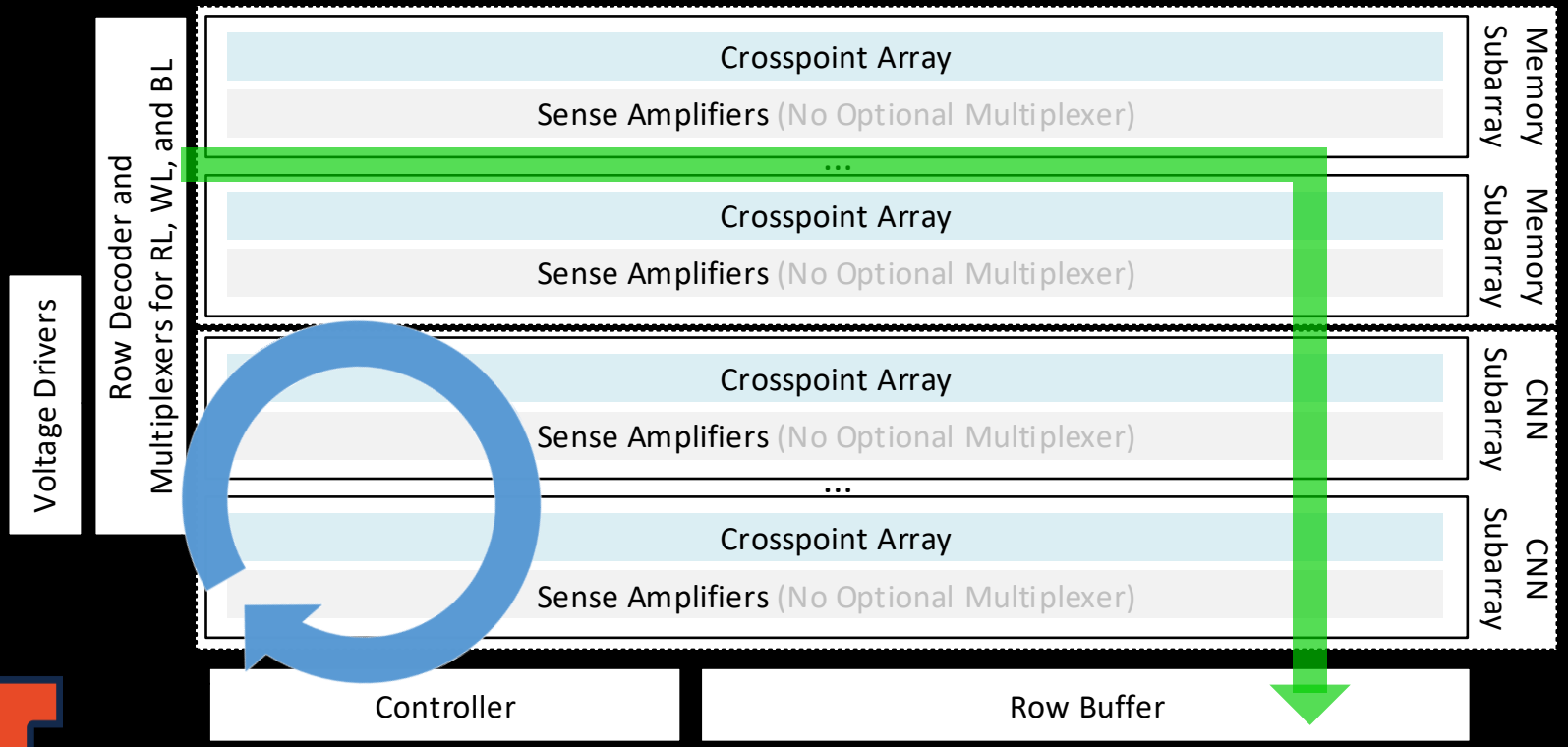
Direct connections between layers speed up deep CNNs

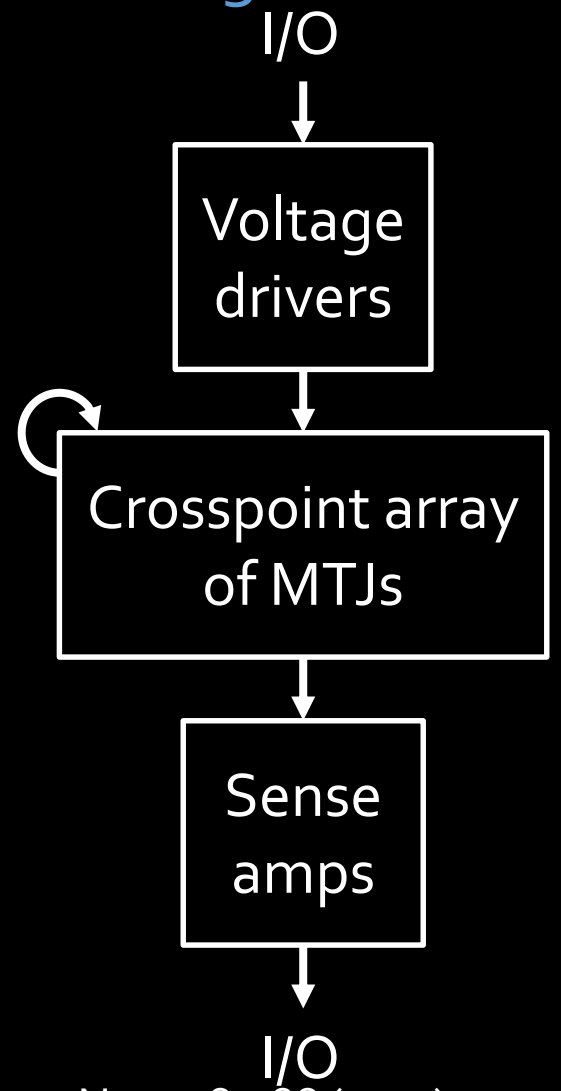


Architecture of the logic-in-memory system

Shared architecture for memory access and neural network logic

- Many neural network layers without conversion
- Similar to an RRAM architecture (Chi et al, ISCA 2016)





Dutta, Siddiqui et al., IEEE Int Symp Nano, 83-88 (2017).

Key Benefits of This Design

- Synaptic MTJ: accurate variable resistance
  - Overcome limitation of prior MTJ designs since binary weights lead to a 20% accuracy loss
  - Neural network can be trained to overcome process variations
- Thresholding MTJ: accurate function evaluation
  - No extra overhead to store results
  - Direct layer-to-layer flow means fast deep nets
- Power with a 2×2 array: 68.6 µW (static)
  129 µW for read, else 15 nW (dynamic)

Courbariaux, arXiv 2016

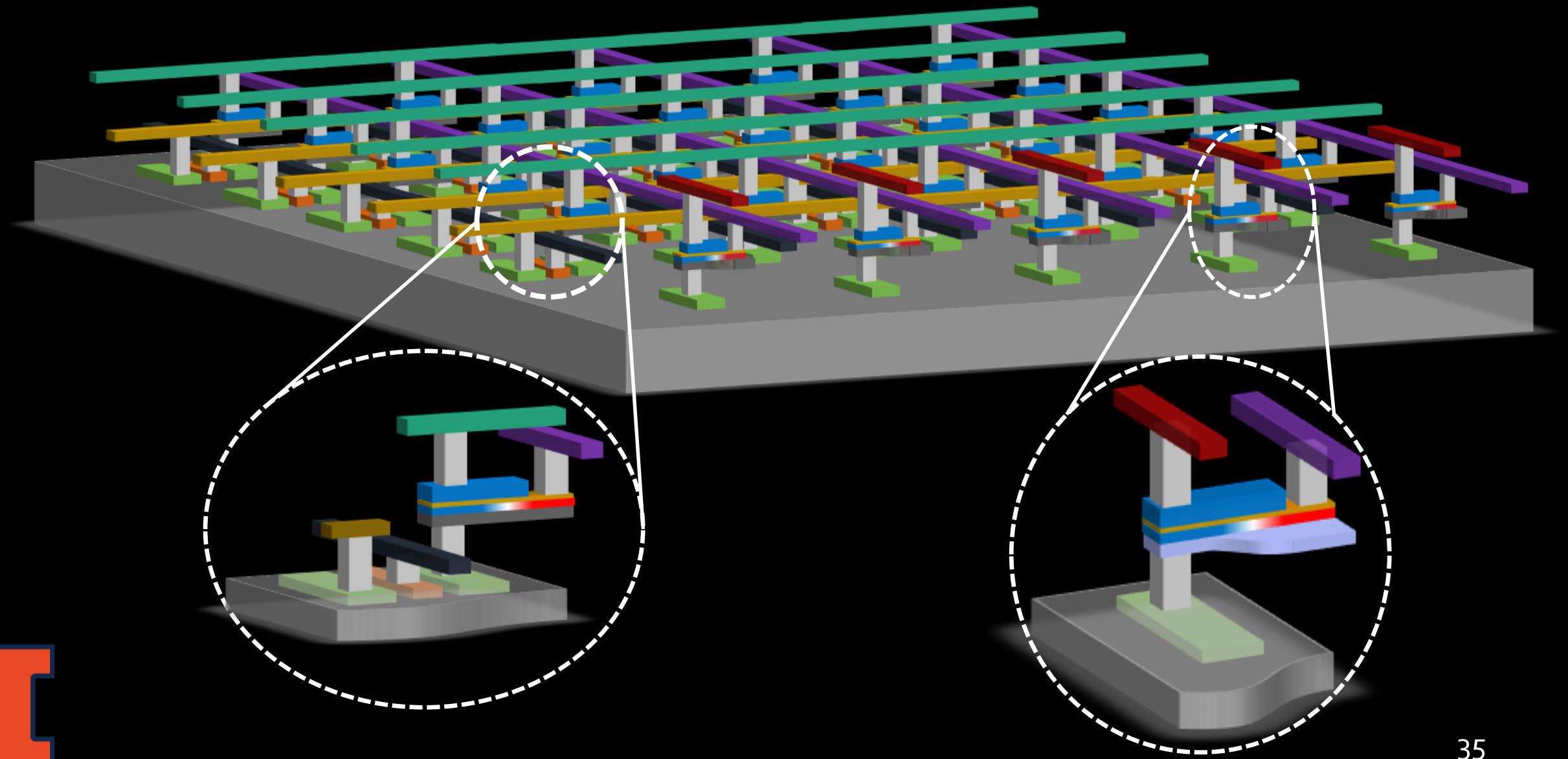
MTJ currents up to 10 µA (5× below other analog-only design)

Sengupta et al, IEEE Biomedical 2016

- Latency: 4 ns (per layer) Same as the clock period of a CMOS-only design (many cycles/layer)

Y. Chen et al, ISSCC 2016

Future direction

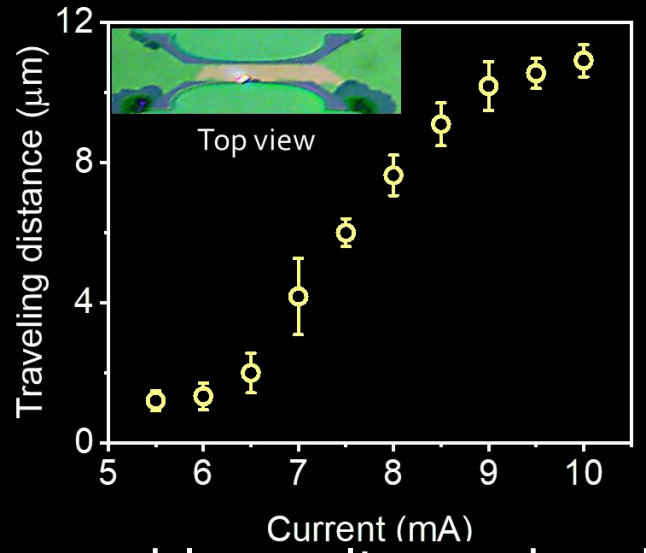


Summary

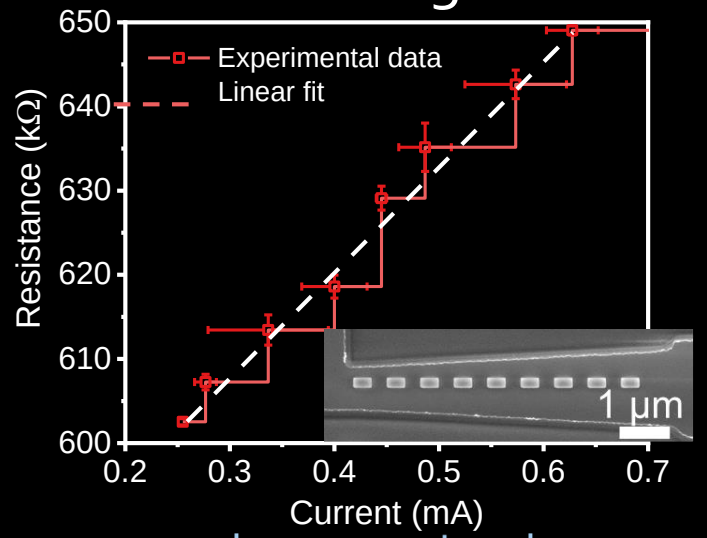
Siddiqui et al., Nano Lett. , 20, 2, 1033–1040 (2020)

Dutta, Siddiqui et al., IEEE Int Symp Nano, 83-88 (2017)

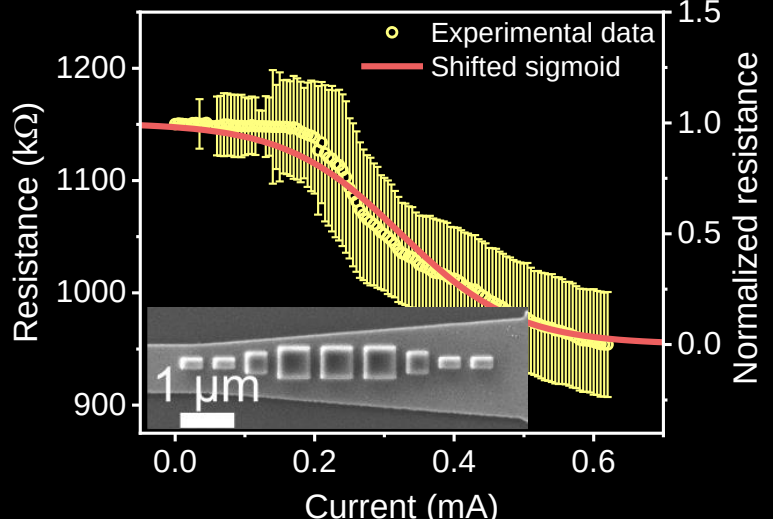
- Design of analog thresholding device



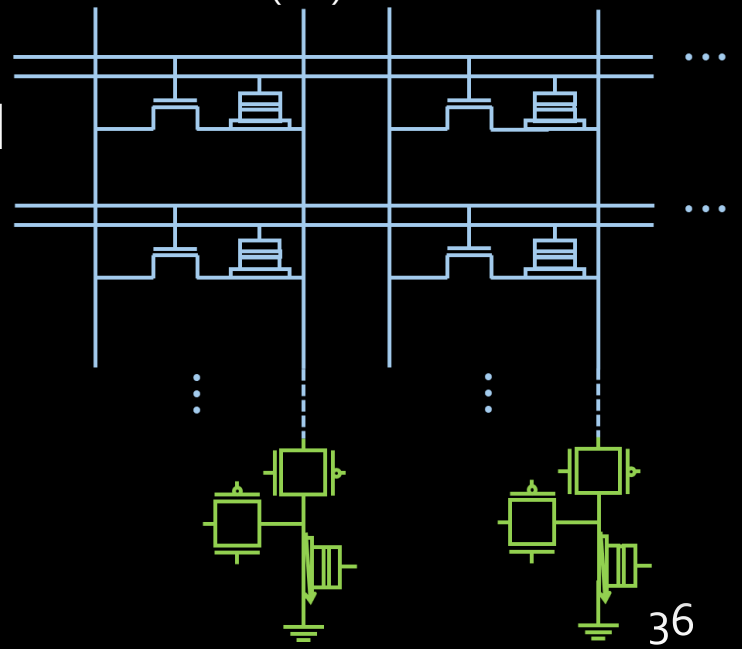
- Linear synaptic device with magnetic tunnel junctions



- Programmable nonlinear threshold device with electrical readout



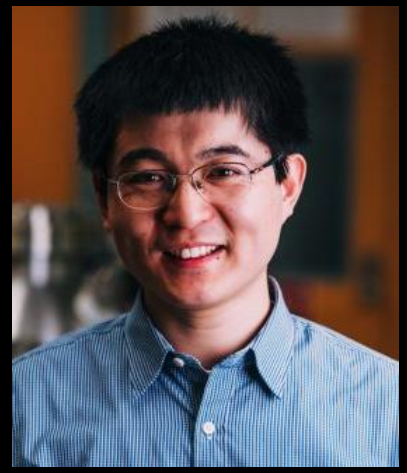
- Architecture of analog artificial neural network with magnetic domain wall synaptic and thresholding devices



Acknowledgements

Prof. Marc Baldo

Prof. Caroline Ross



Prof. Luqiao Liu

Prof. Axel Hoffmann

Dr. Sumit Dutta

Dr. Astera Tang