CMOS-Compatible Protonic Synapse Technology

for Analog AI Training Accelerators

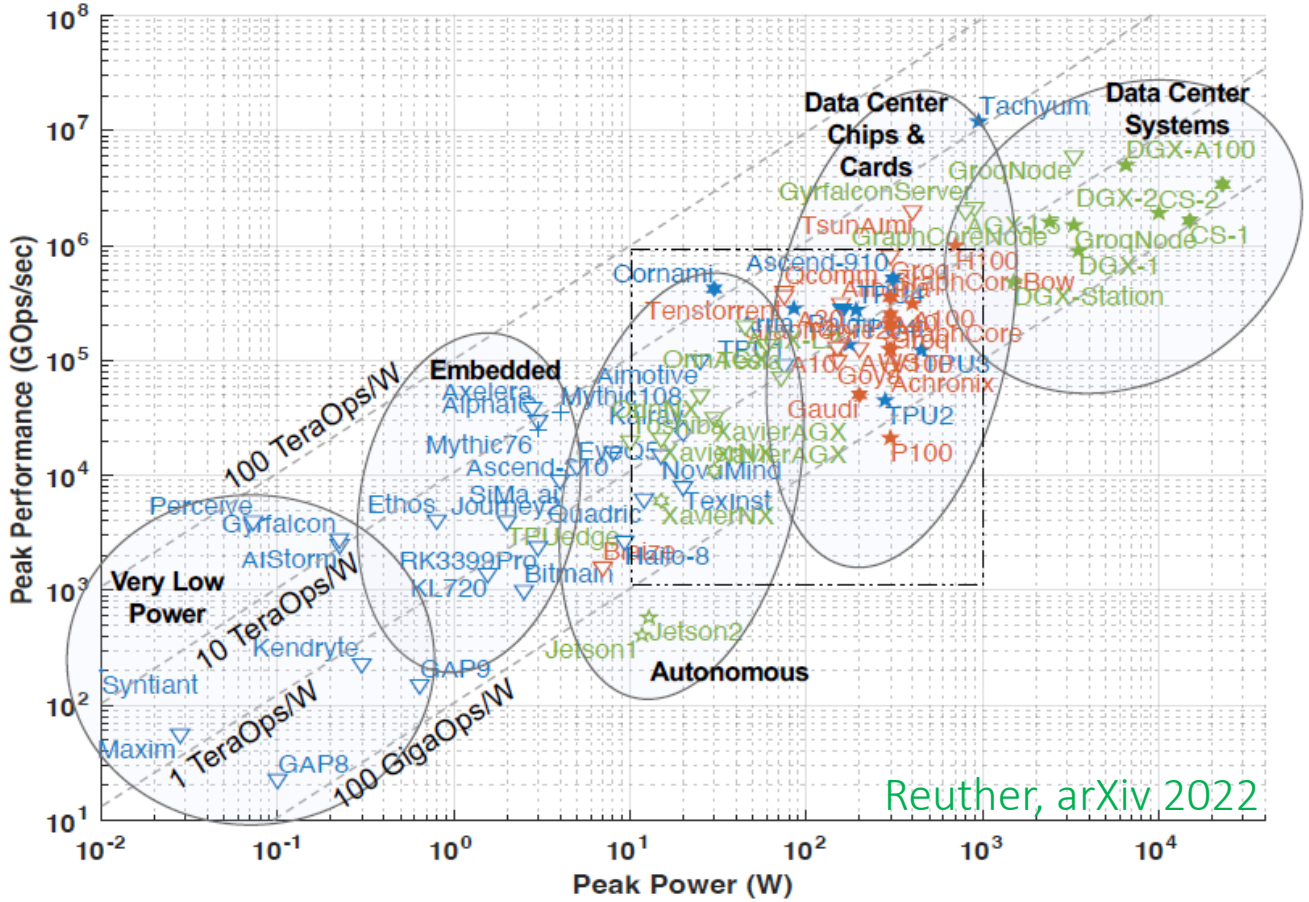
Dingyu Shen

Supervisor: Jesús A. del Alamo

Massachusetts Institute of Technology

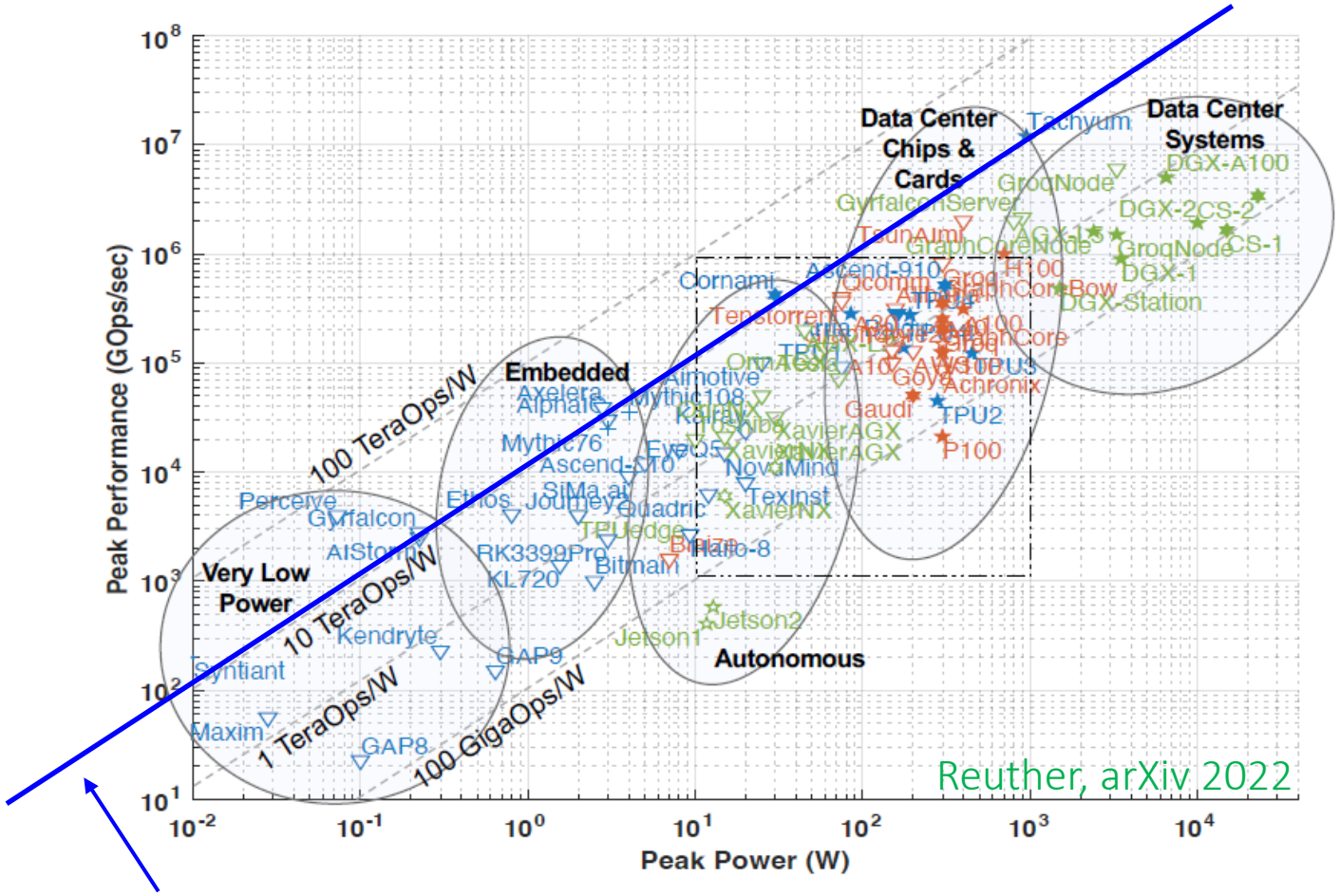
Acknowledgment: **MIT-IBM Watson AI Lab**, Ericsson, MIT Quest for Intelligence

AI accelerators and processors: peak performance vs. power



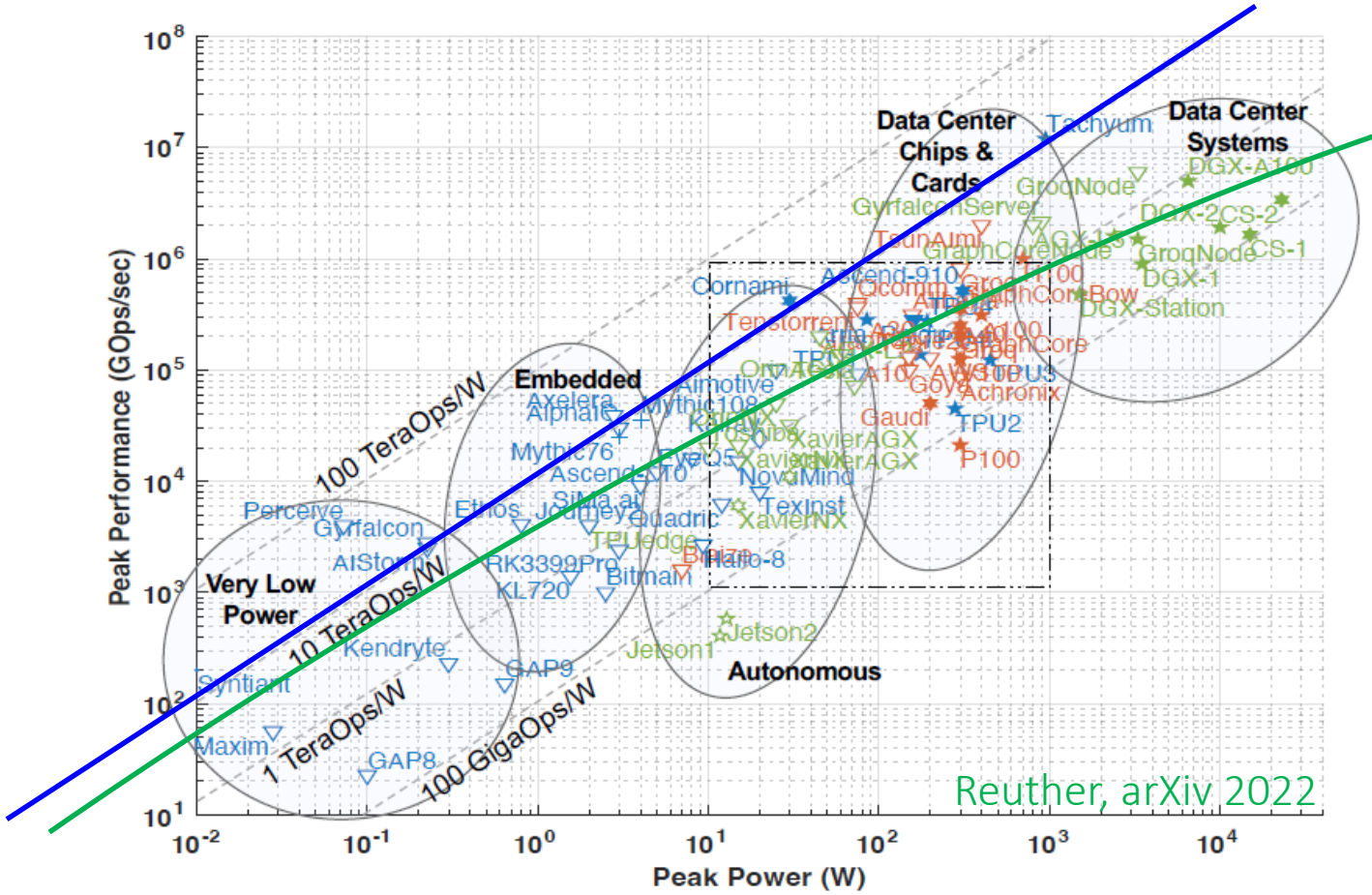
Trade-off between performance and power

Existing AI hardware based on Si CMOS technology

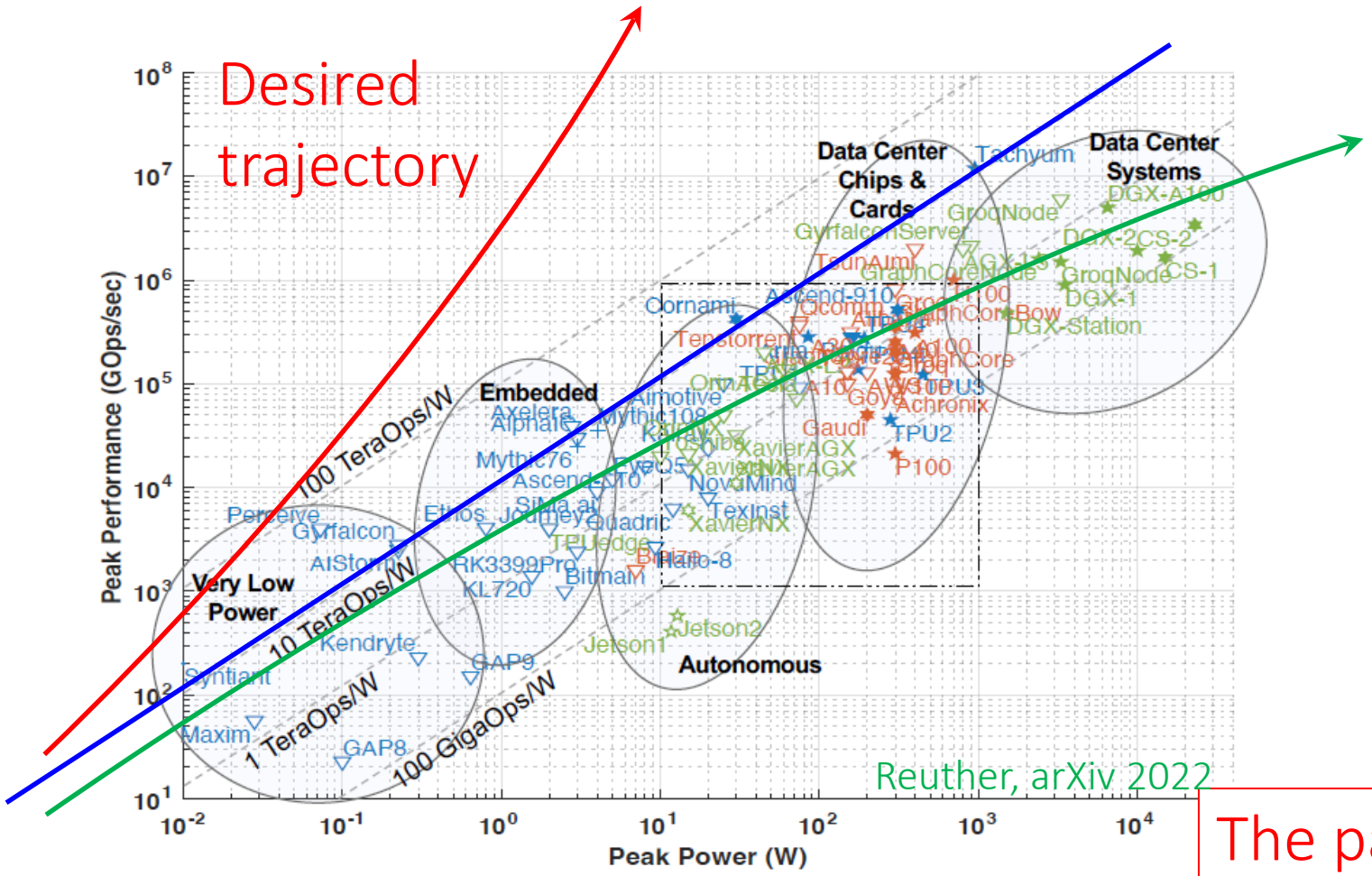


Conventional CMOS limit: ~10 TeraOps/W

Unsustainable trajectory of existing AI hardware



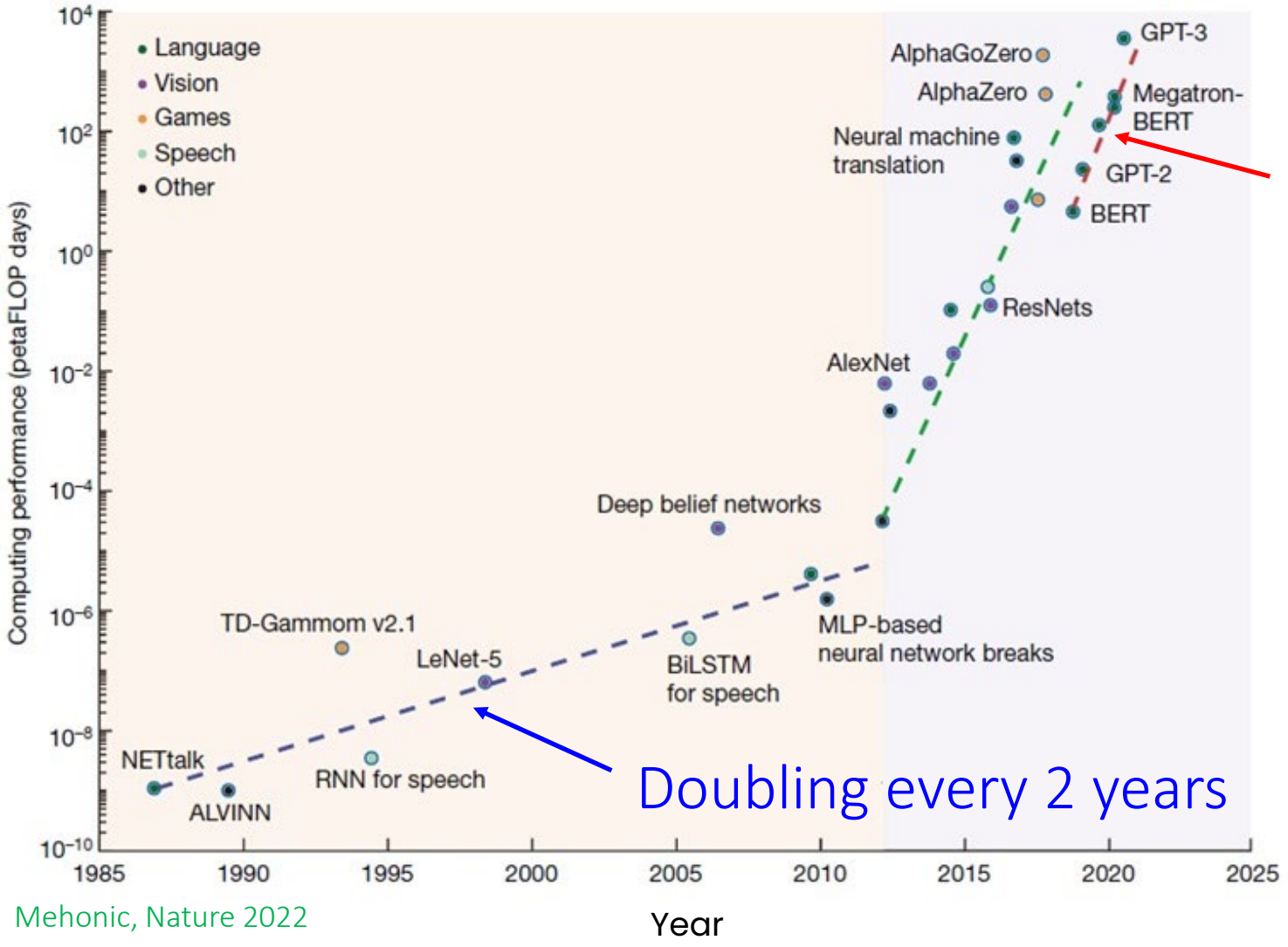
AI hardware: new concepts needed



The path forward:

- New AI hardware

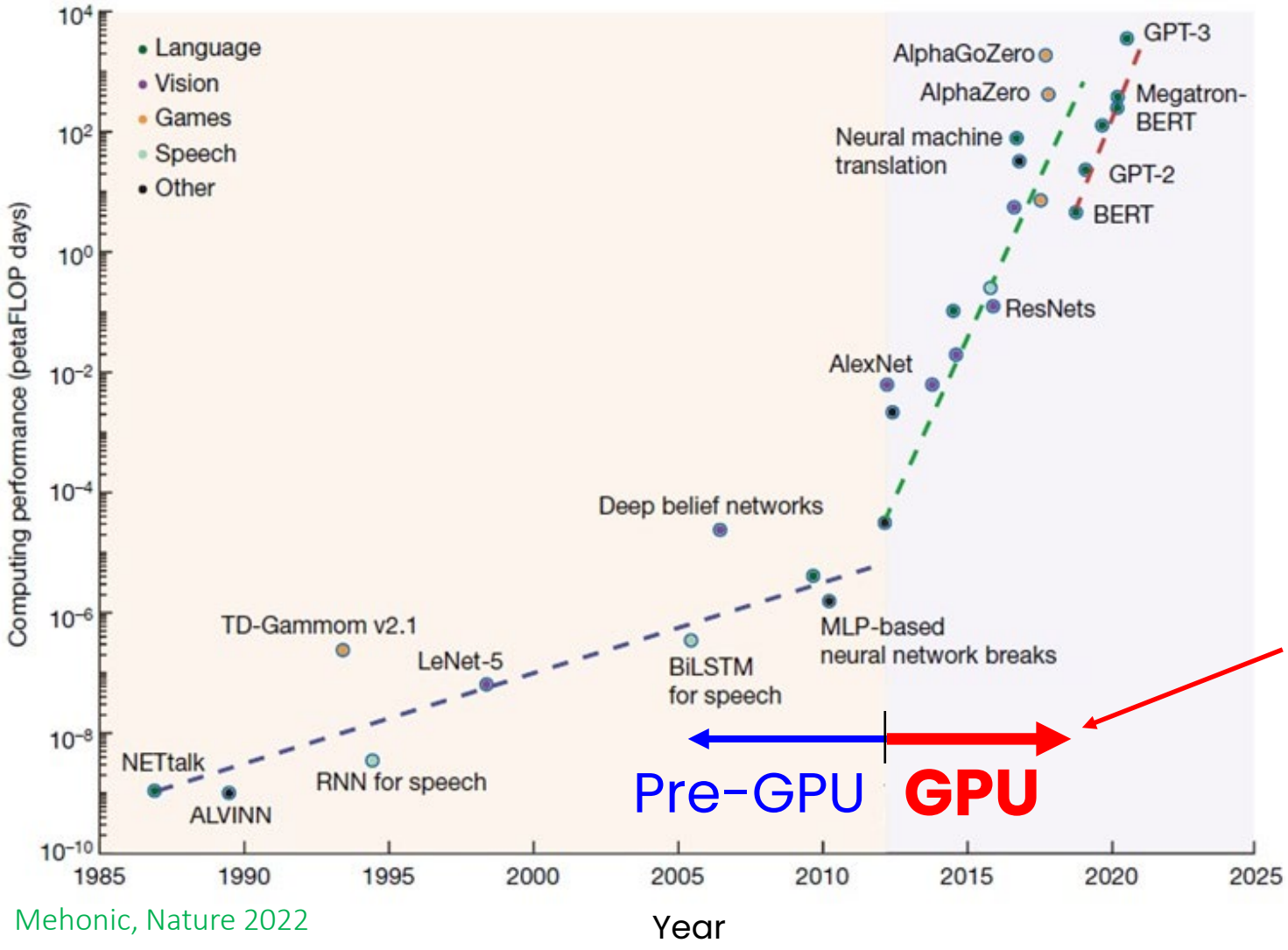
AI compute demands growing exponentially



Doubling every 2 months!

Mehonic, Nature 2022

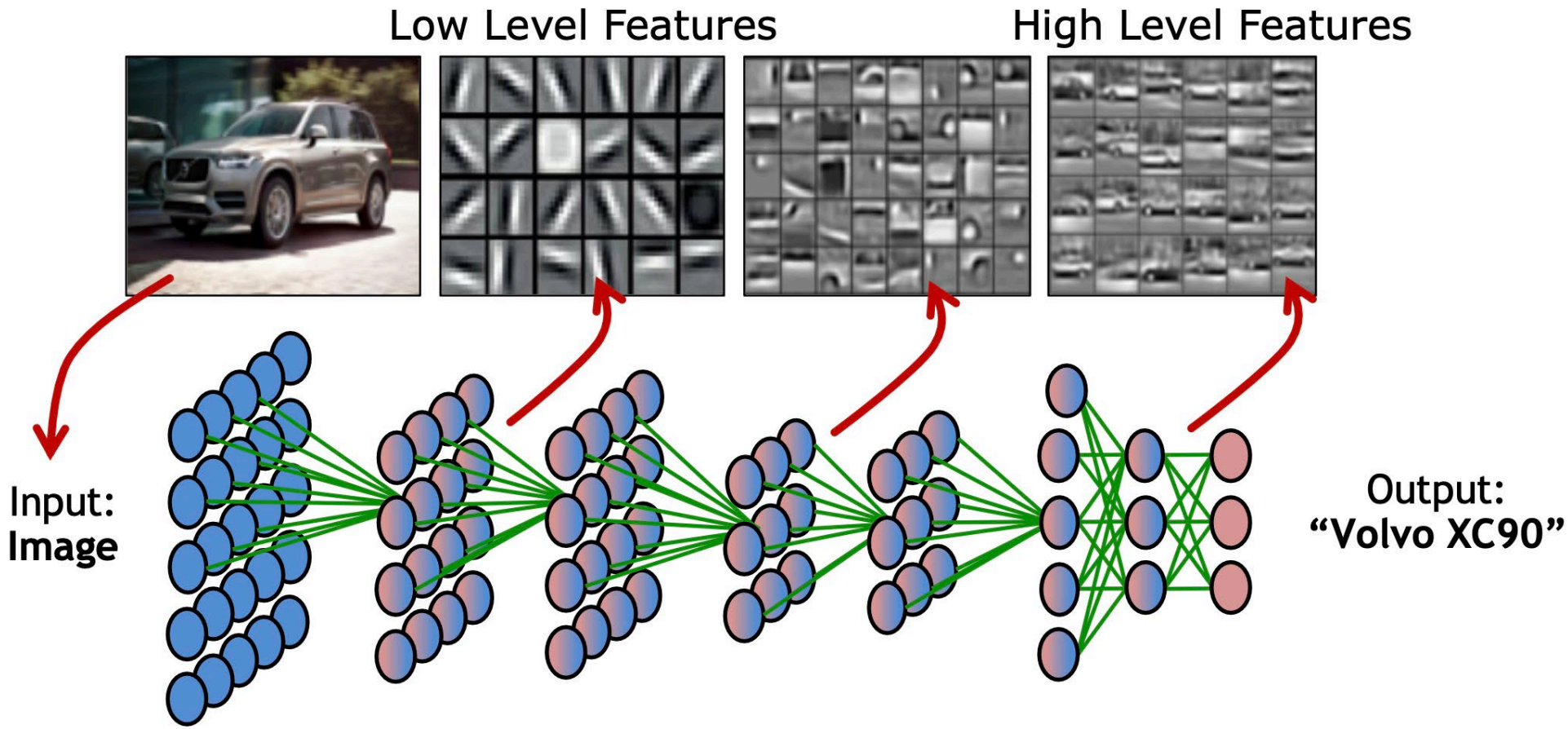
AI compute demands growing exponentially



Mehonic, Nature 2022

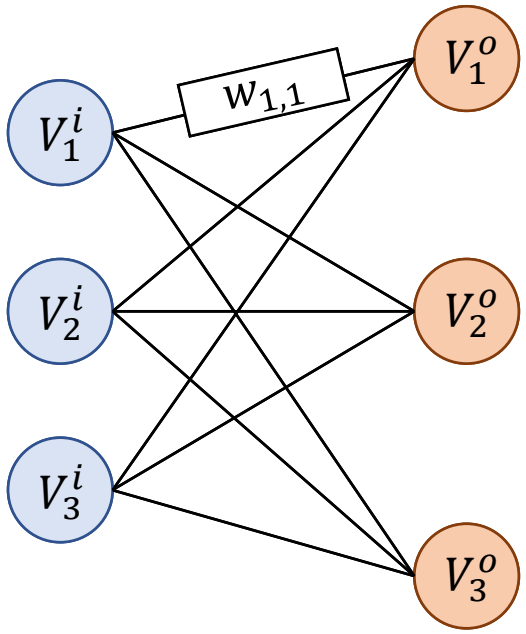
Specialized hardware makes a huge difference!

Deep Neural Networks: Data represented at multiple levels of abstraction



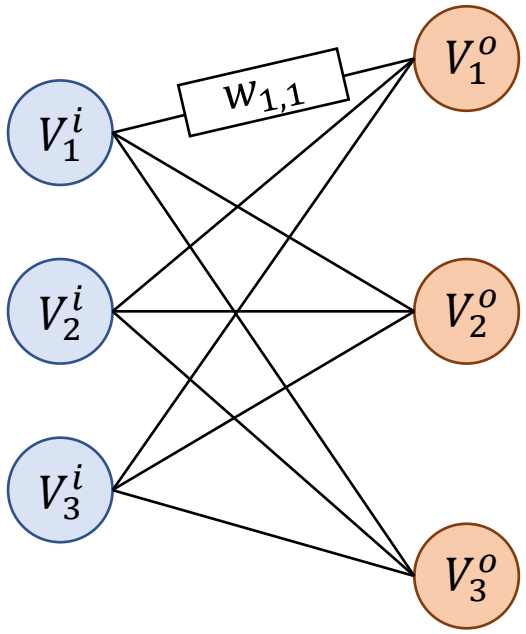
Lee, CACM 2011

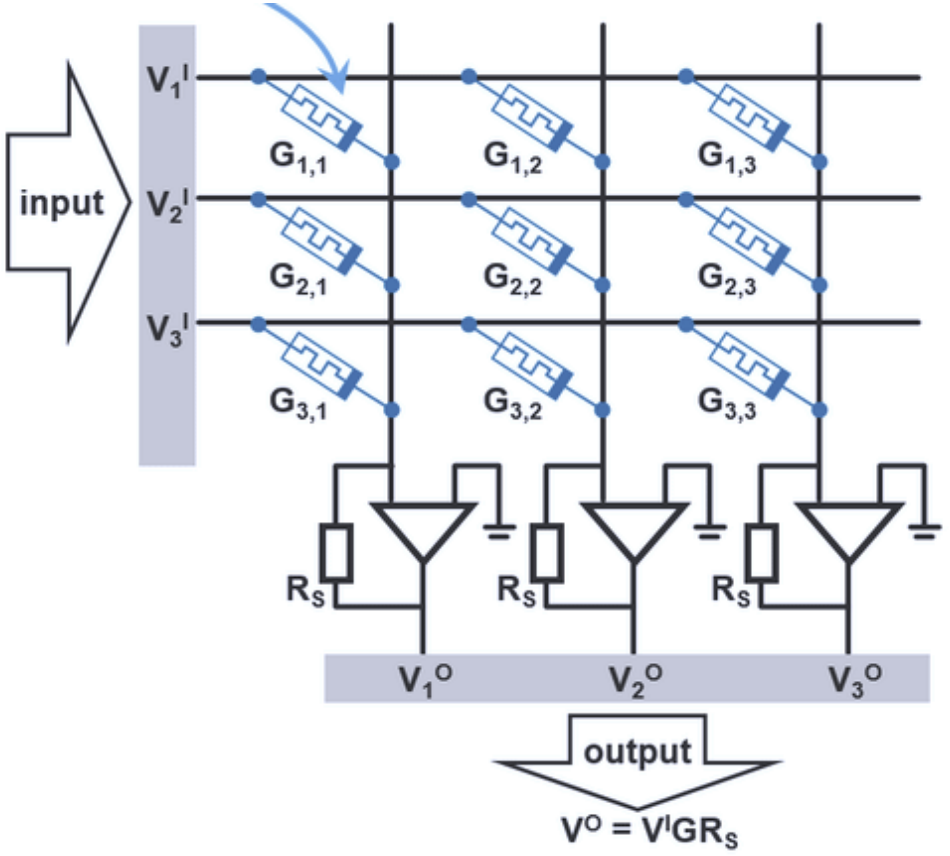
DNN heart: multiplication and accumulation (MAC)



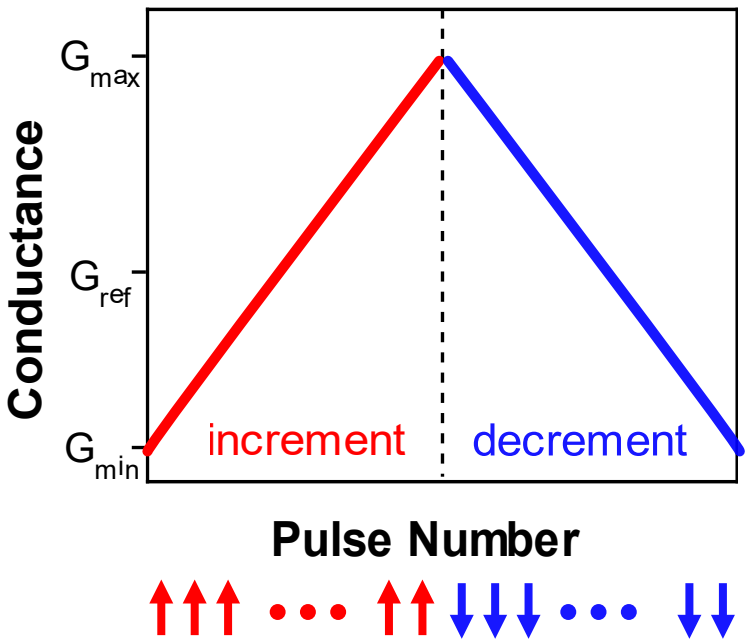
Analog Neural Networks for deep learning

Programmable non-volatile resistor (“synapse”)





Desired characteristics for programmable resistors



Conductance modulated through application of pulses

- Desired characteristics are relatively well understood:

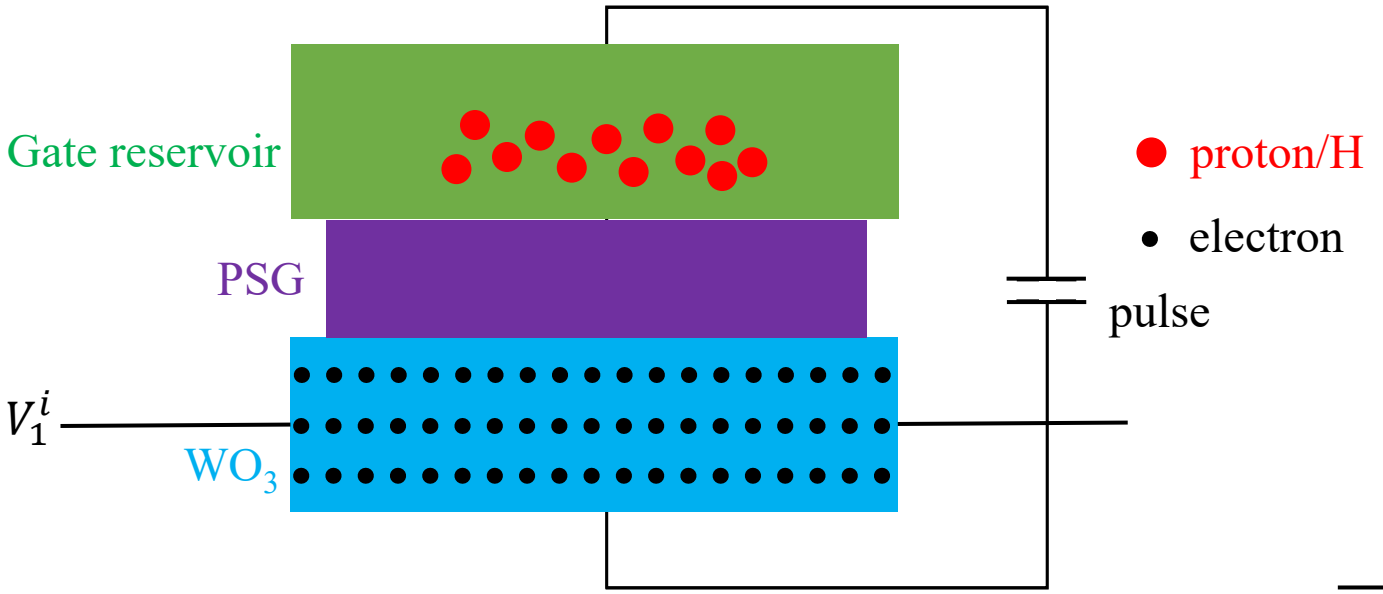
• G~0.1 μS	• Symmetric modulation
• Large dynamic range (~10x)	• Low variability
• Fine-tunable (~1000 levels)	• High endurance
• High speed (ns)	• Good retention
• High efficiency (sub-pJ)	• CMOS compatible

Gokmen, Front Neurosci 2016

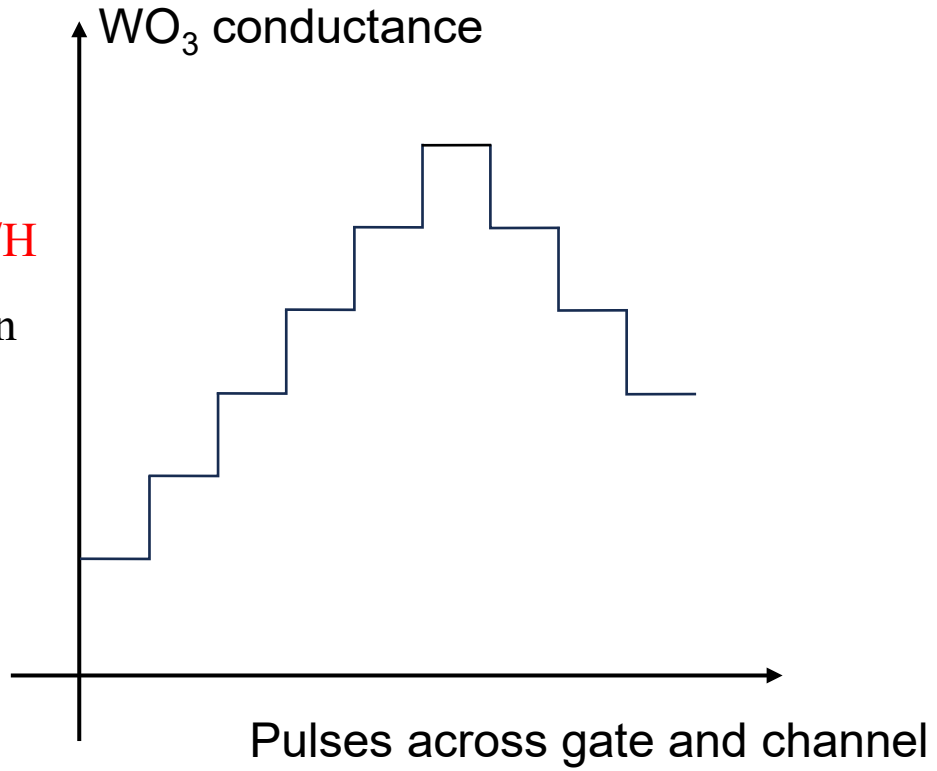
- No known technology today meets all requirements

Ion-Intercalation non-volatile programmable resistor

del Alamo group at MIT: protonic synapse based on WO_3

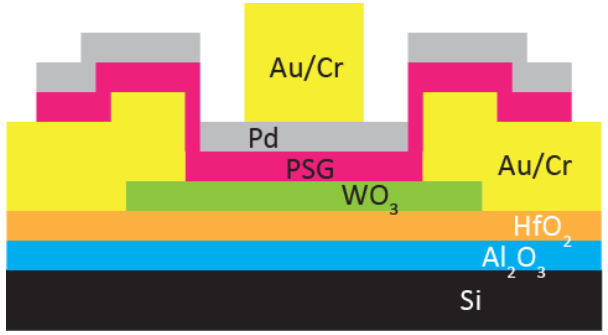


PSG=Phosphosilicate glass

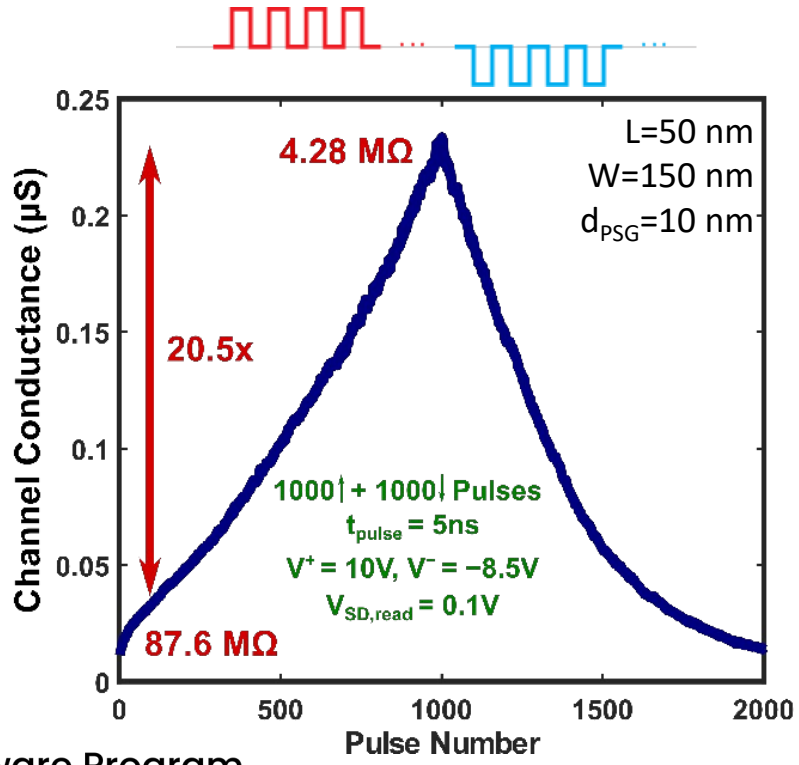


Pd/PSG/WO3 self-aligned nanoscale device

CMOS-compatible materials and thermal budget

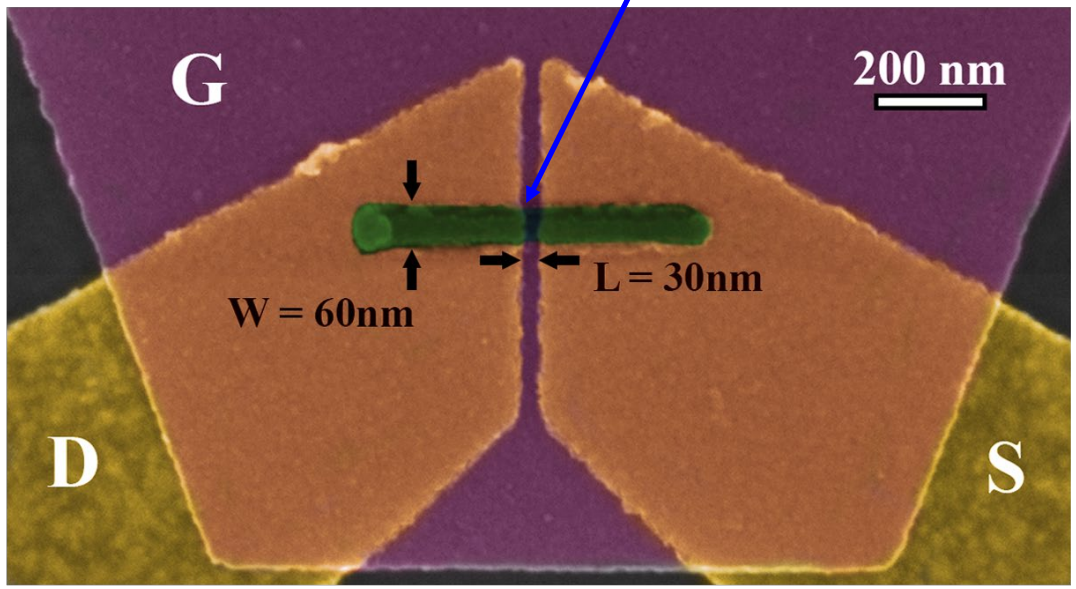


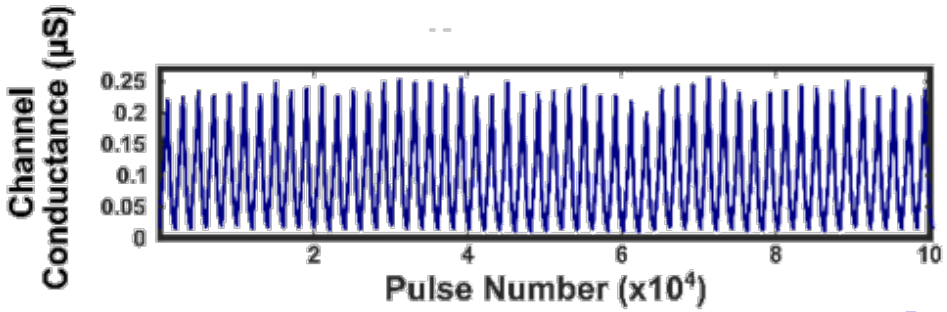
Fast, nearly linear, symmetric conductance modulation



Onen, Science 2022

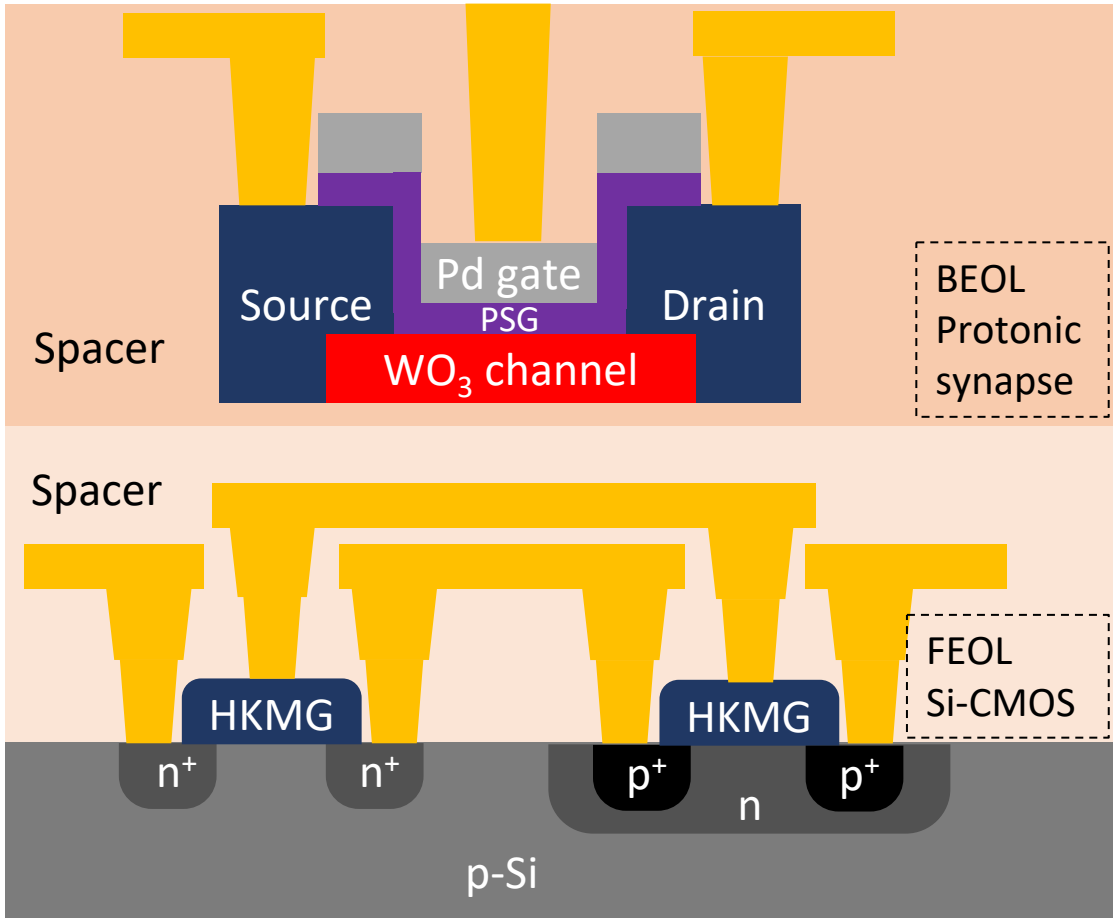
Nanoscale active area





Endurance and repeatability through 10^5 pulses conducted over 30 hours

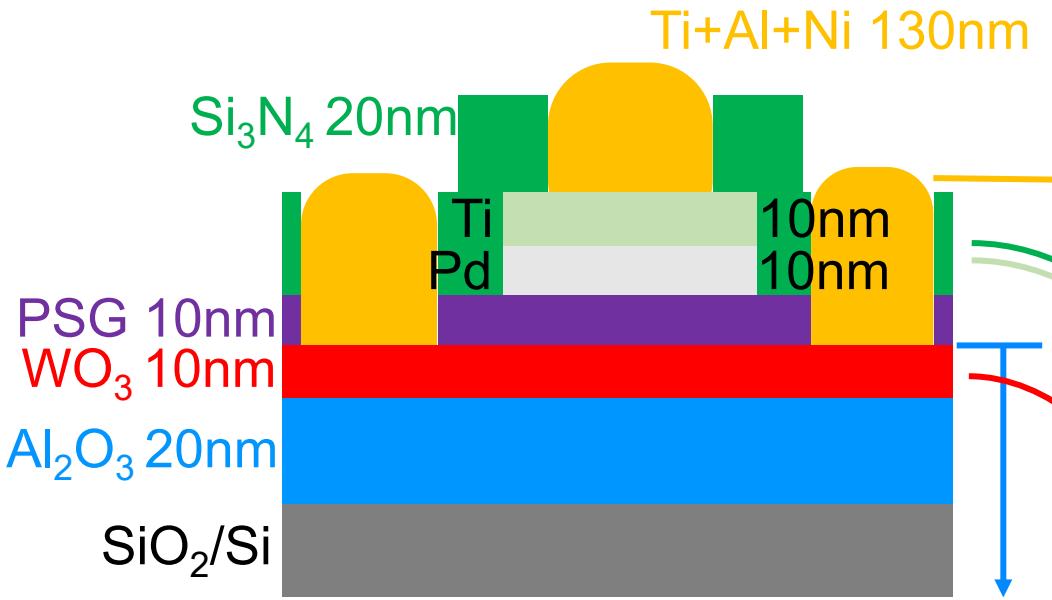
Goals for ongoing research

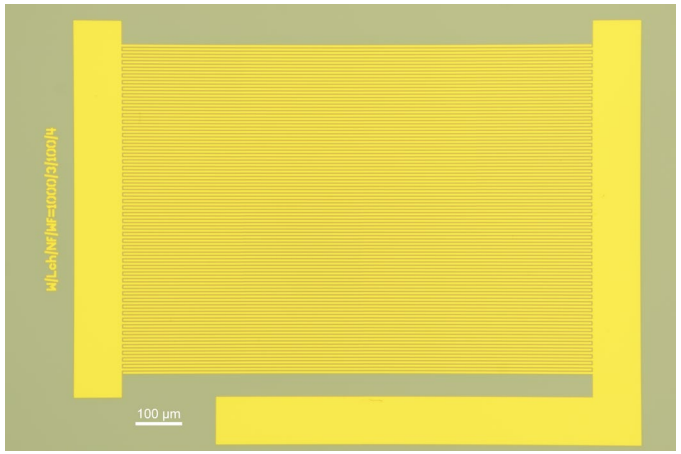


Vertical integration of protonic synapses on top of Si CMOS

- CMOS-compatible process
- Device fully encapsulated
- protonation as part of process
- Joint 200 mm process with IBM

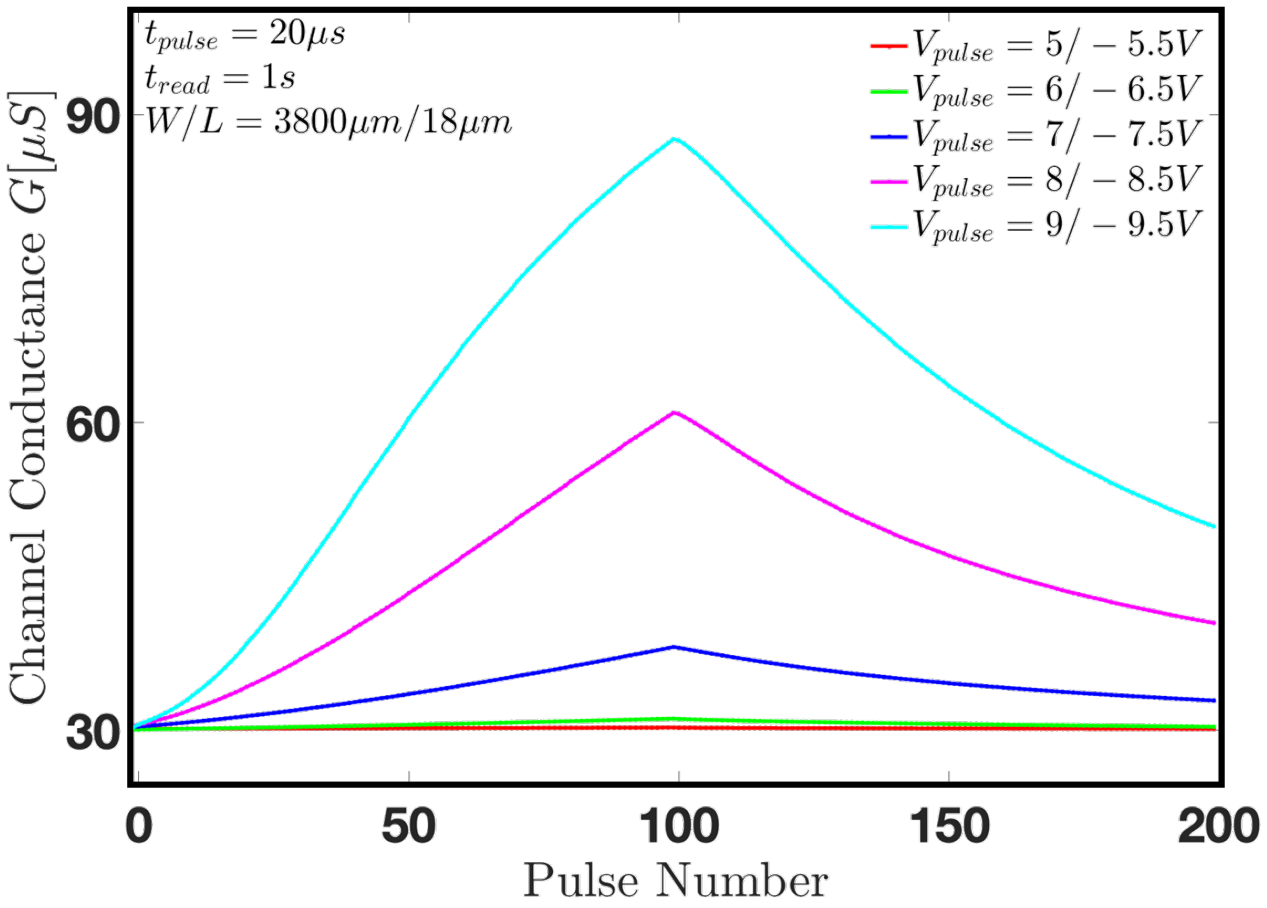
Devices scaled onto 200 mm wafers





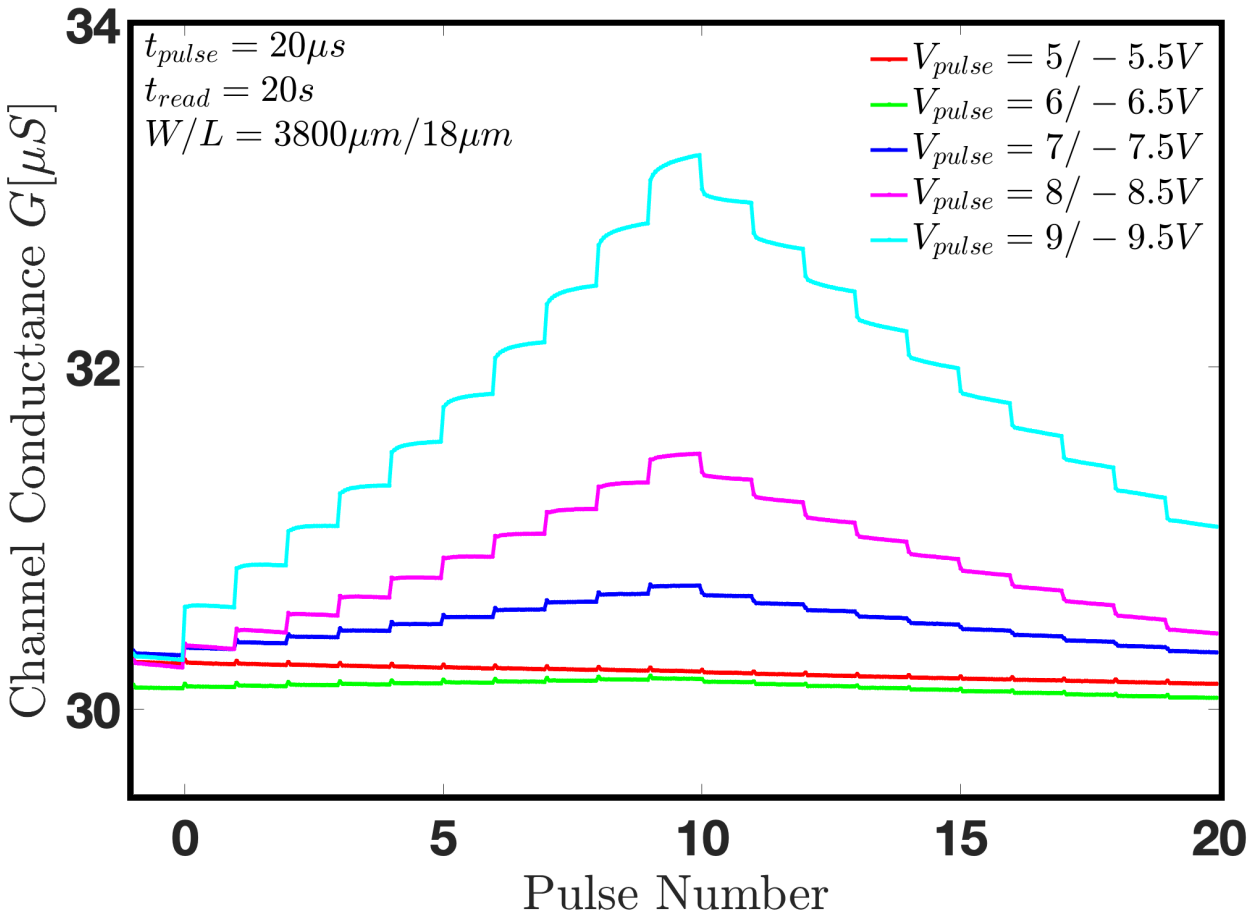
- CMOS-compatible process
  - Ti + Al + Ni contacts
- Device fully encapsulated
  - Si_3N_4 deposited in PECVD
  - Ti on top of Pd
- protonation as part of process
  - H_2 plasma treatment in PECVD
- Joint 200 mm process with IBM
  - 200 mm wafer from IBM

Test results for devices scaled onto 200 mm wafers



- Strong conductance modulation: ratio of 3:1
- Short pulses: 20us
- Nearly symmetric modulation

Test results for devices scaled onto 200 mm wafers



Good retention over 20s

Conclusions and future work

Conclusions:

- CMOS-compatible process designed
- In-situ protonation and encapsulation method developed
- Joint process of 200mm wafers between MIT.Nano and IBM demonstrated

Future work:

- High performance nanoscale device
- Demonstration of arrays and integration with CMOS circuits with IBM

Continuous progress in AI hardware will demand:

- new electronics
- based on new devices
- built with new materials
- operating under new physical principles

Wonderful new opportunities for hardware innovations!

MIT AI Hardware Program