ASIC² ARCHITECTURES SYSTEMS INTELLIGENT COMPUTING INTEGRATED CIRCUITS

הטכניון מכון טכנולוגי לישראל TECHNION Israel Institute of Technology

Accelerating DNN Applications with Emerging Memory Technologies

Tzofnat Greenberg-Toledo

Advisor: Prof. Shahar Kvatinsky

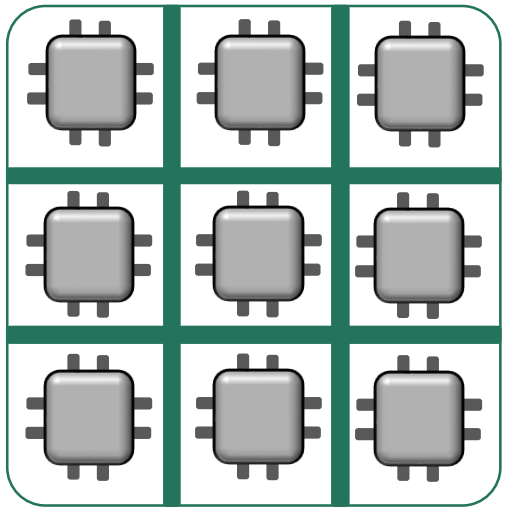
Viterby Faculty of Electrical Engineering

Technion – Israel Institute of Technology

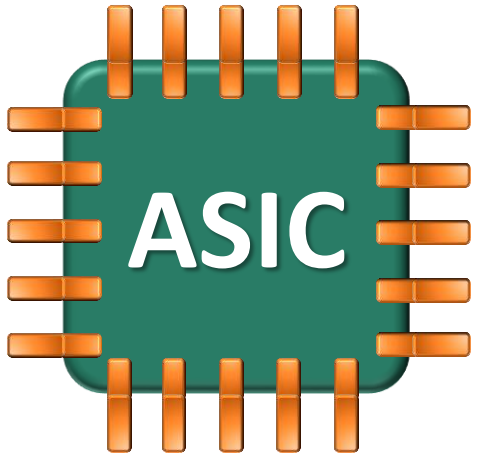
DNN Hardware

GPU

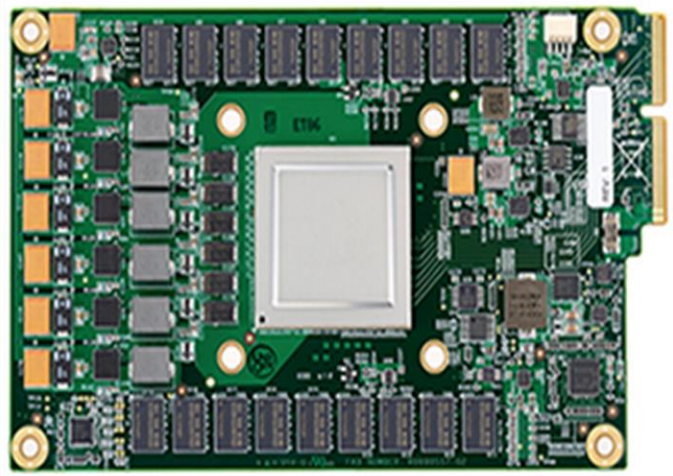
FPGA



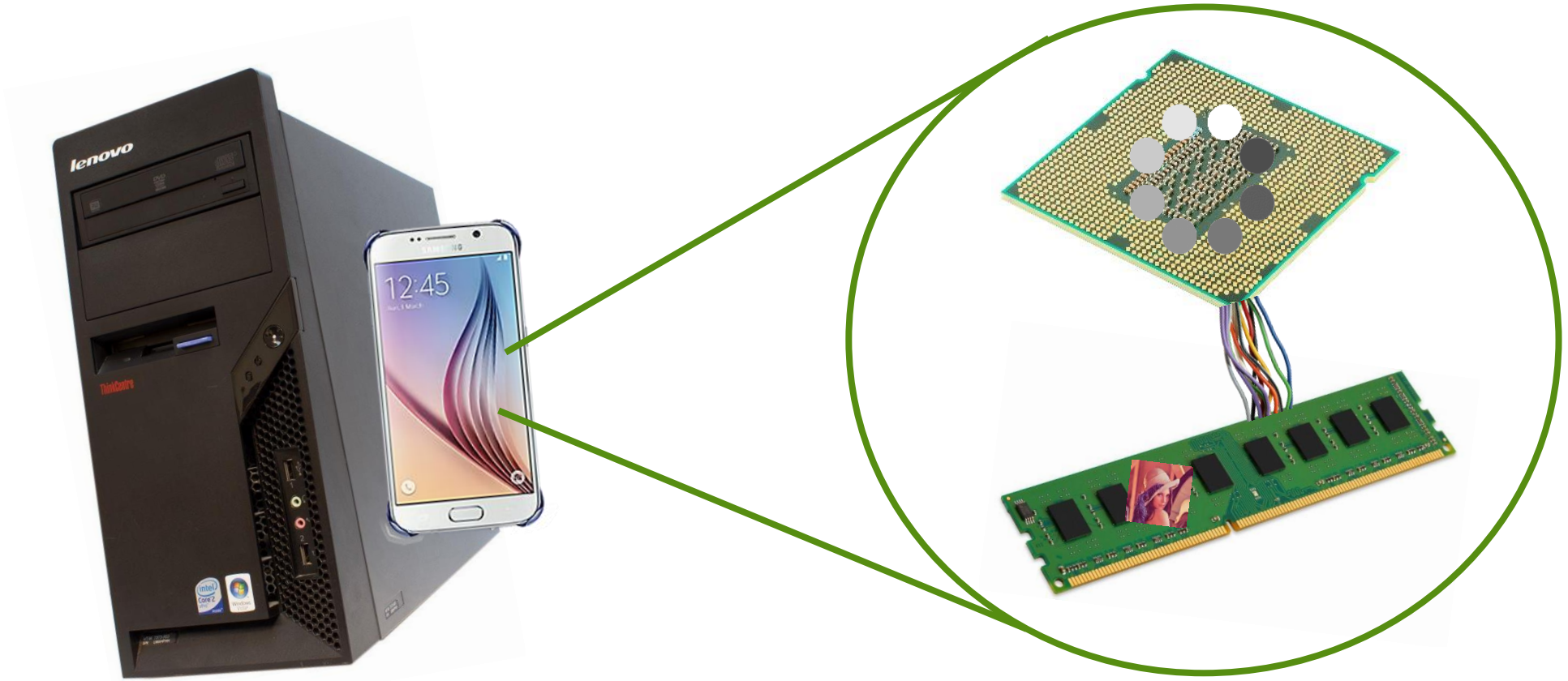
ASIC



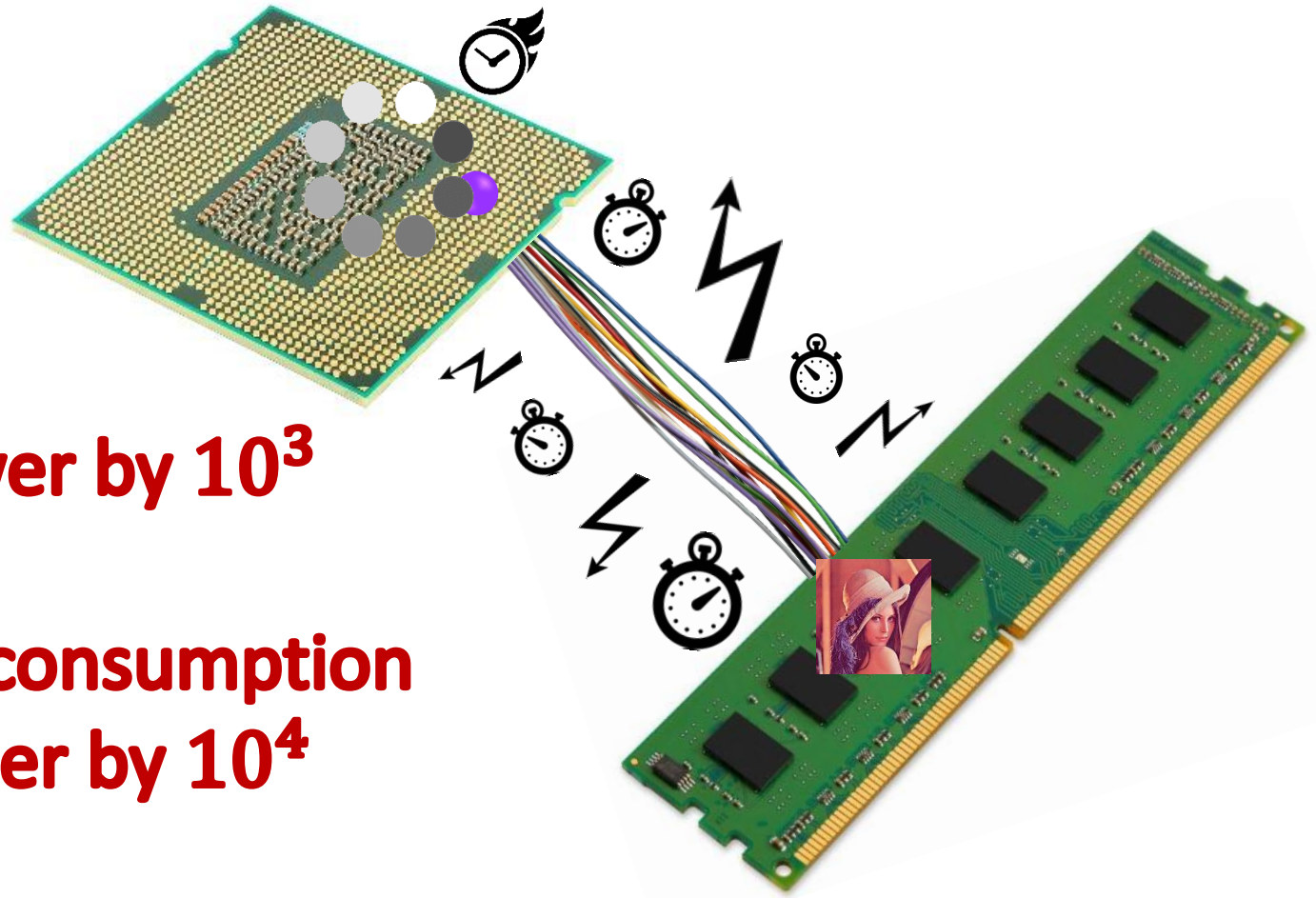
Google TPU



von Neumann Architecture



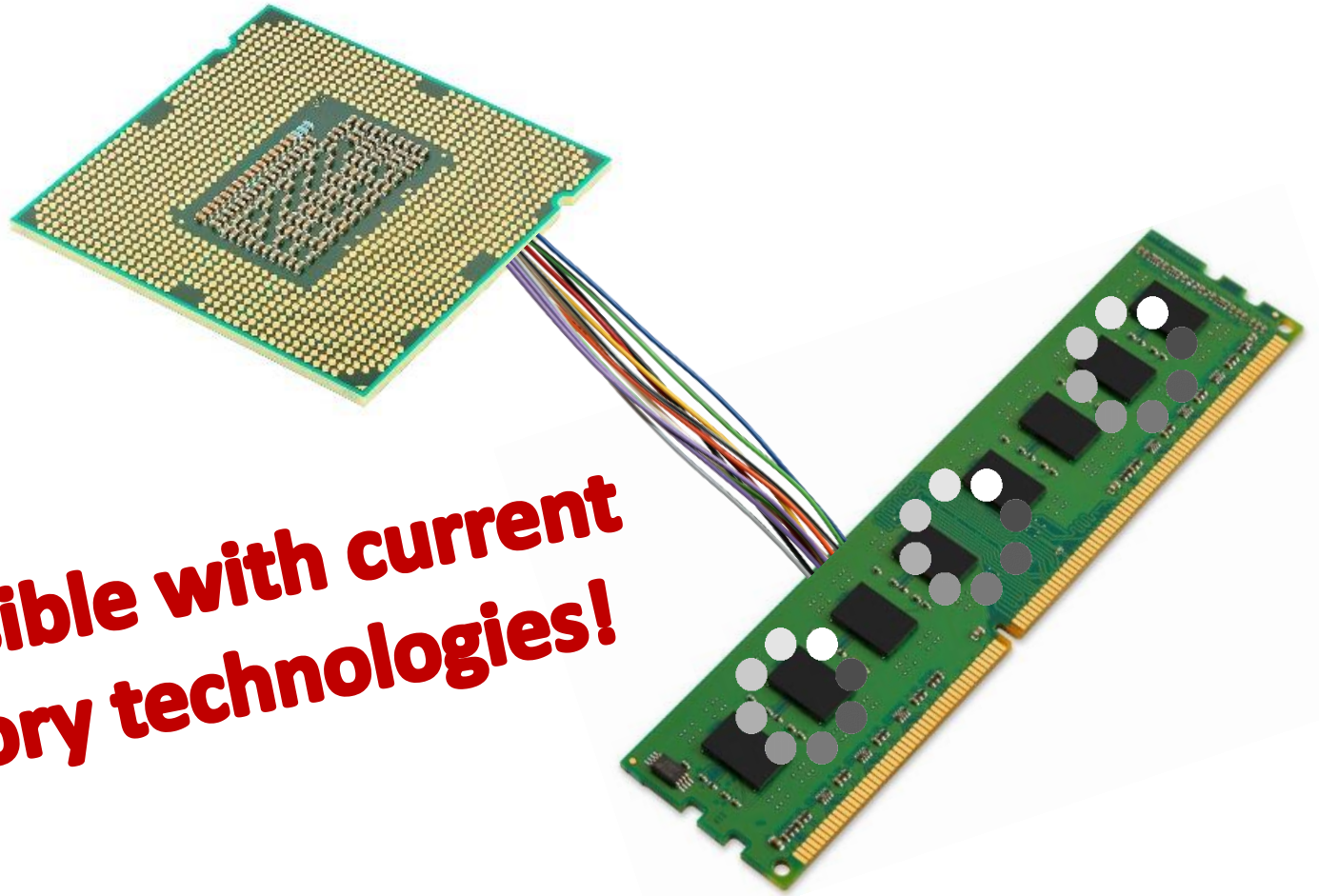
Processing Data in von Neumann Architecture



Slower by 10^3

Power consumption higher by 10^4

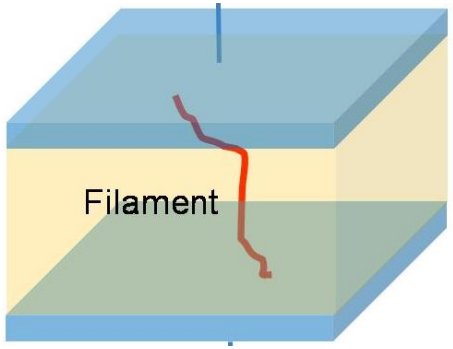
Solving the Bottleneck: Processing Data Within Memory



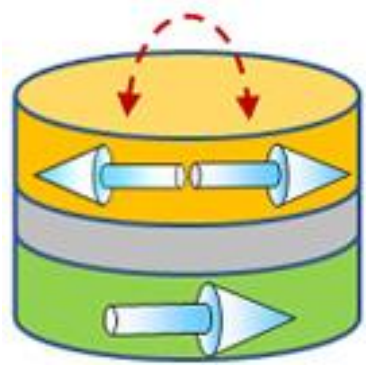
Impossible with current memory technologies!

Solution: Novel Memory Technologies

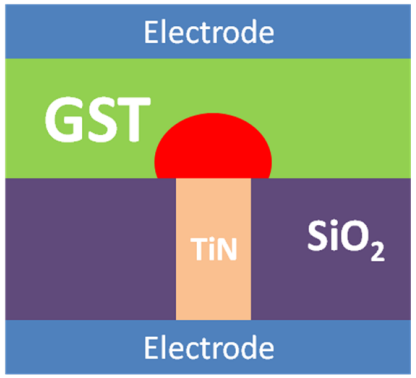
Resistive RAM (RRAM)



STT MRAM



Phase Change Memory (PCM)



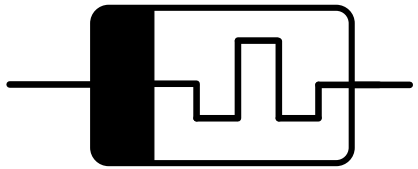
Industry investment (partial list)

TOSHIBA winbond intel Micron IBM

hp SAMSUNG QUALCOMM EVERSPIN TECHNOLOGIES WD

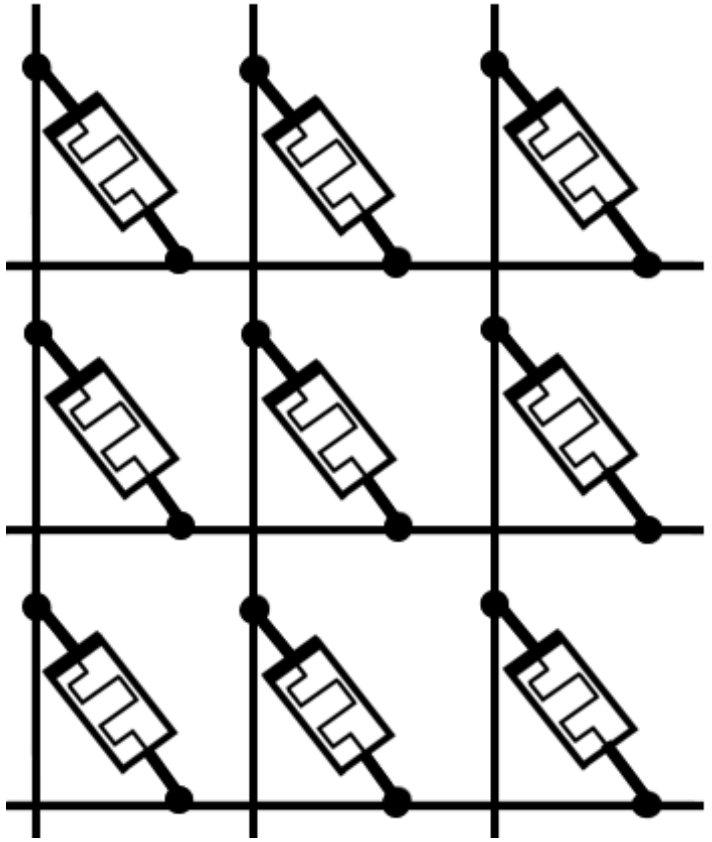
Processing within Memristive Memories

Basic memory cell

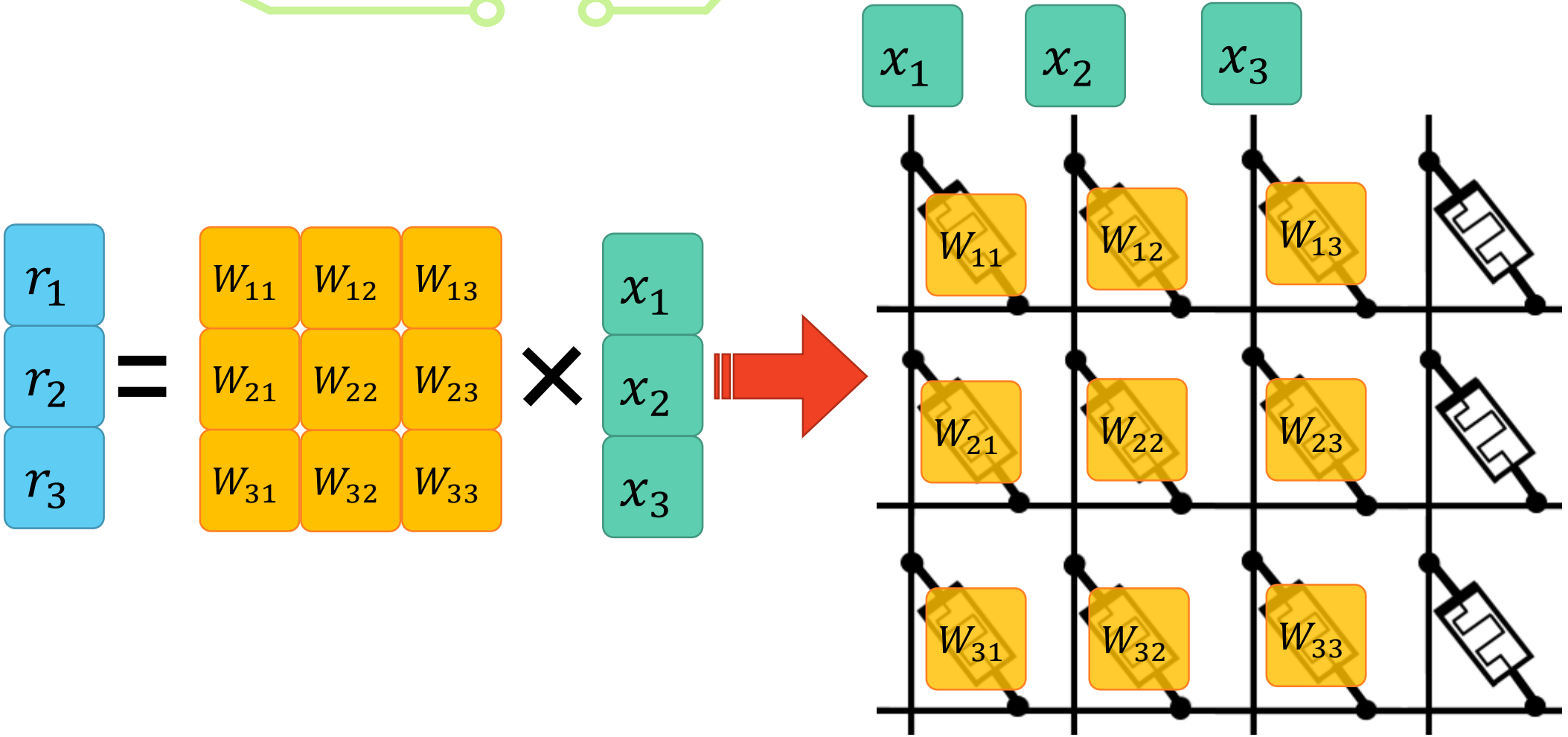


Memory resistor ⇒ **Memristor**

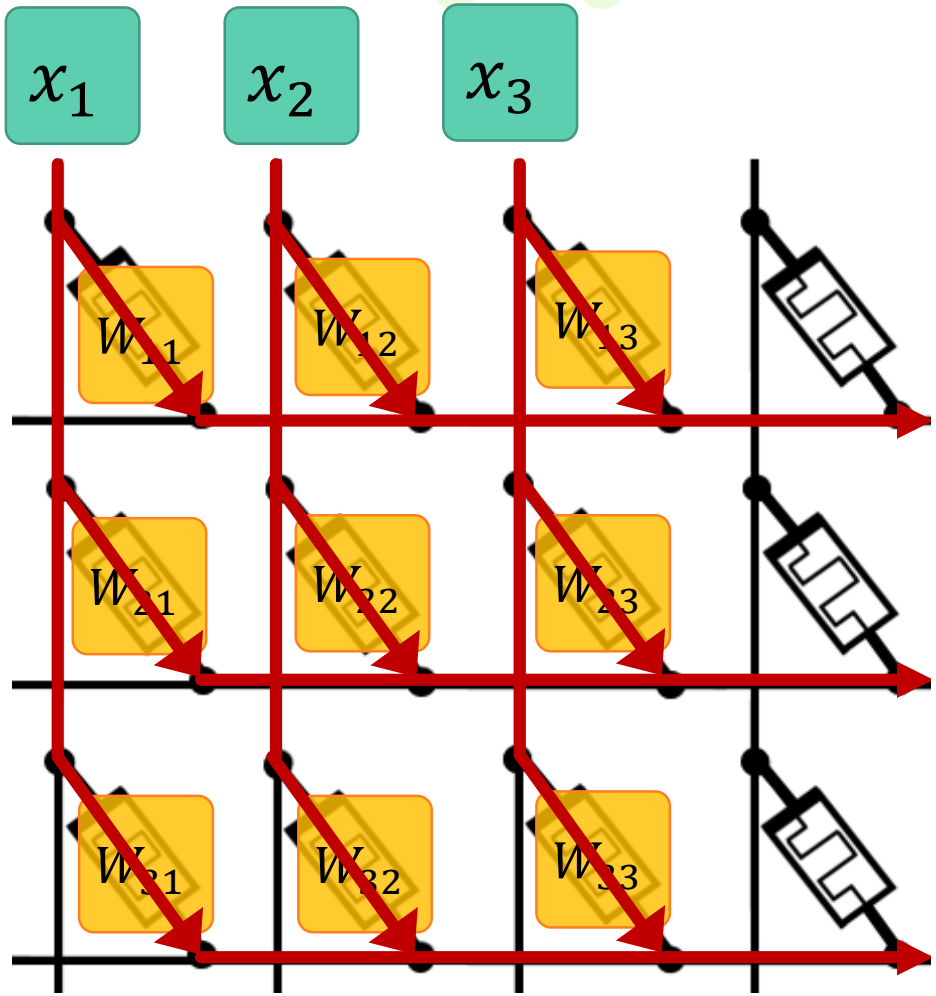
Memristive memory



Process in Memory for DNN



Process in Memory for DNN

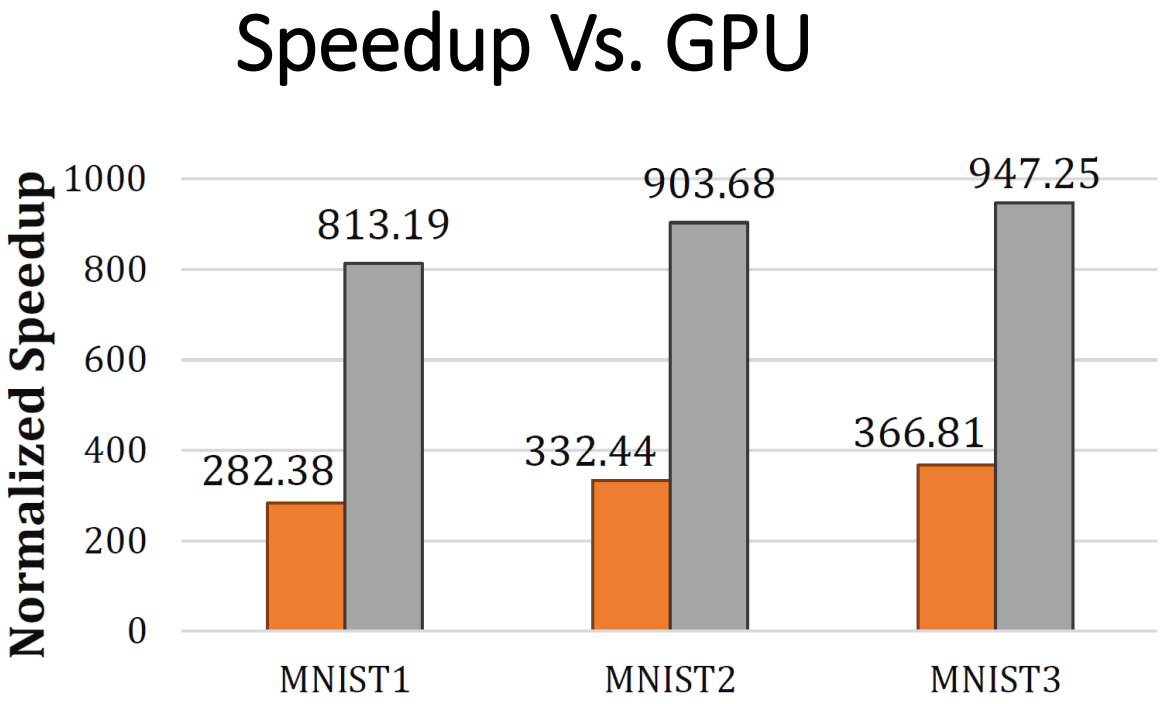


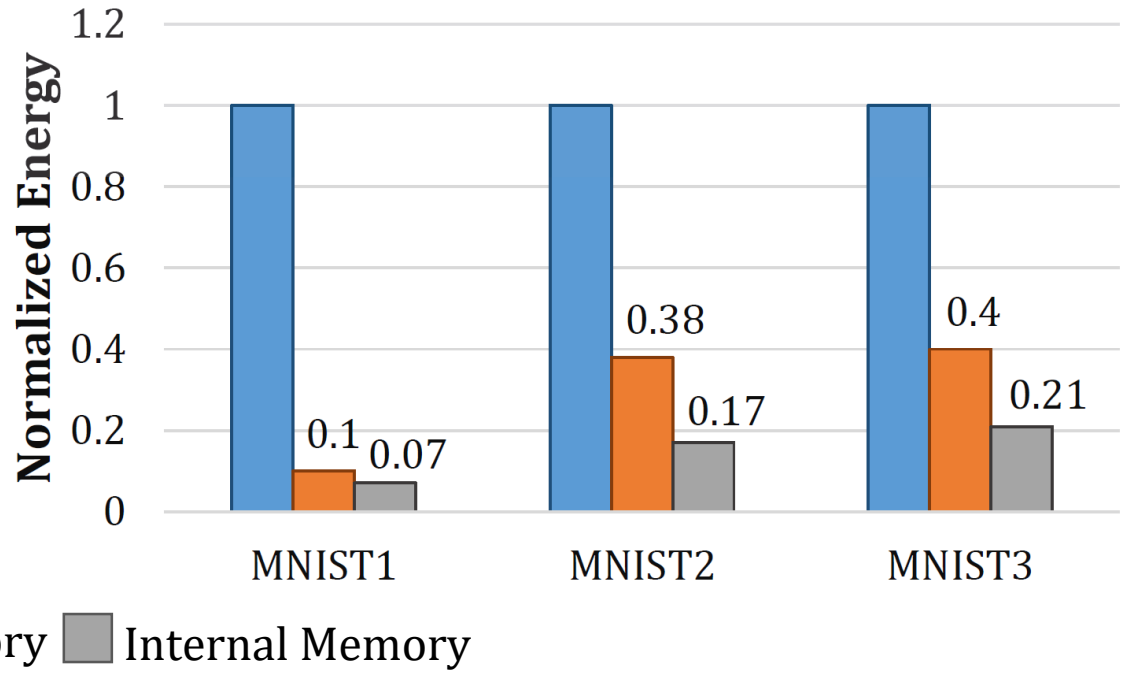
- **Reduce data movement**
- **Highly parallel operation**
- **Immediate results**
- **Low power consumption**

Example: Support SGD + Momentum

~1000 × Speedup

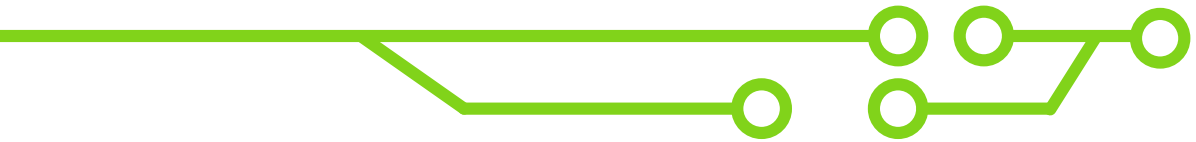
80% Energy reduction





T. Greenberg-Toledo, *et al,* "**Supporting the Momentum Training Algorithm Using a Memristor-Based Synapse**," in IEEE Transactions on Circuits and Systems I: Regular Papers, vol. 66, April 2019.

Thanks!




ASIC2
ARCHITECTURES
SYSTEMS
INTELLIGENT COMPUTING
INTEGRATED CIRCUITS