ReRAM for Implementing Neural Network Synapses

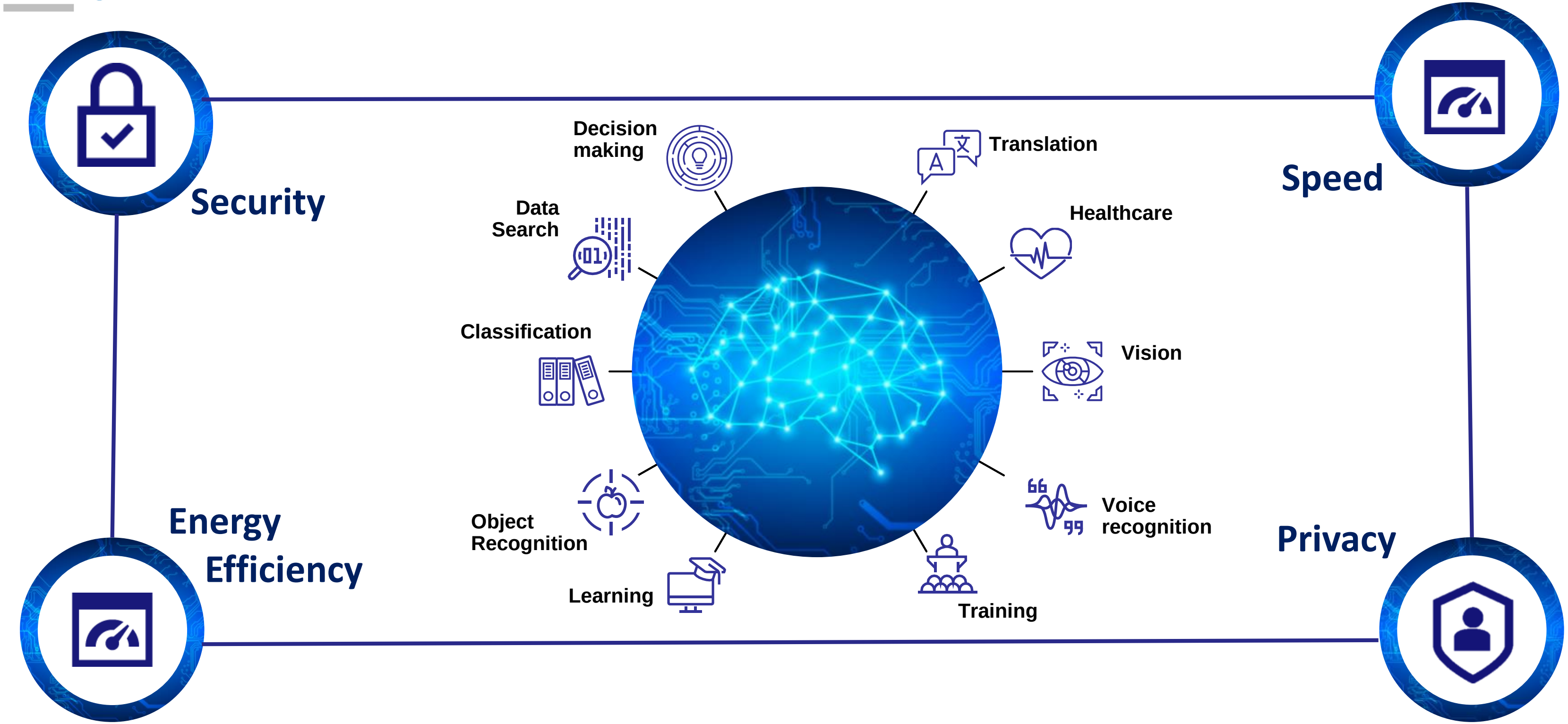
Amir Regev

CTO

Outline

- **Introduction**
- Towards AI Edge computing
- Accelerating AI
- Neuromorphic computing using ReRAM
- Weebit-Leti SPIRIT SNN demonstration
- Conclusions

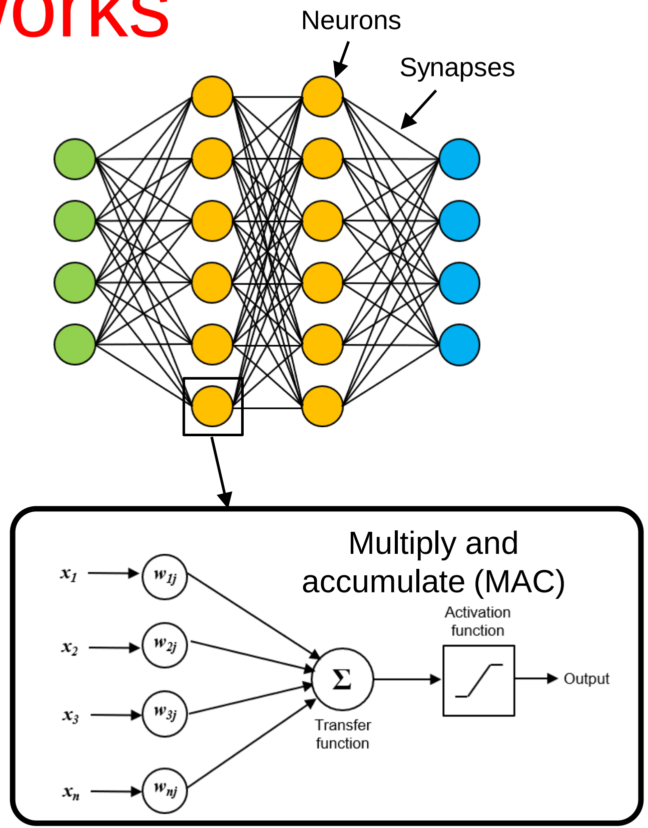
Towards AI Edge computing



Artificial Intelligence generations

1st	2nd	3rd
Machine Learning	**Deep Learning**	**Neuromorphic computing**
Algorithms identify patterns in data, and use them to make predictions	**Uses Artificial Neural Networks for learning**	**Fully biologically-inspired computing**
Learning through mathematical models Linear regression Decision Trees	**Networks with topology inspired by the human brain but not related implementation**	**Implements spiking behavior similar to the human brain**
Not brain-inspired	**Partially brain-inspired**	**Best exploited with neuromorphic hardware**

Artificial Neural Networks



Partially brain-inspired networks, using neurons and synapses for computation

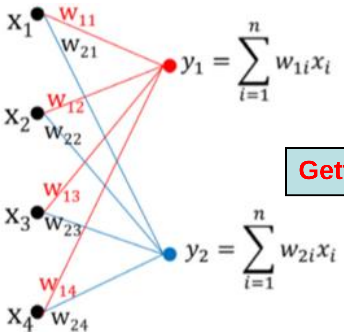
- Inputs are weighted through synapses and then summed (MAC)
- Weights are uploaded externally in DRAM chips
- Synchronous operation

Moving data between GPU and external memory cost 200x than staying inside the chip

Tremendous power consumption mainly due to data movement between computing cores and memory

Accelerating AI with ReRAM

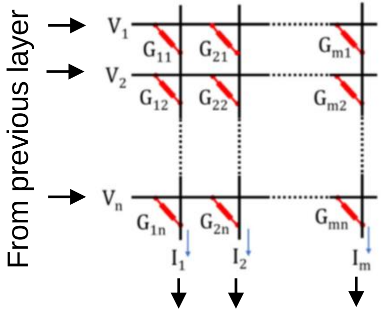
ReRAM for MAC operation naturally achievable using Ohm's and Kirchhoff laws



R.Islam, IOP J. Appl. Phys 2019

$x = V$
$w = G$
$y = I$

Getting rid of Multiplication – just accumulation



To next layer

- Non-Volatile
- CMOS compatible
- Scalable
- Fast

Co-location of memory / computing → boosts performance and reduces consumption

Still not optimal: this is <u>not</u> how the really brain works

Why is our brain so special?

- Massively parallel
- Three-dimensionally organized and extremely compact
- Extremely Power efficient
- Combines storage and computation
- Fault and variation tolerant
- Self-learning and adaptive to changing environments

Biological brain – towards efficient computing architecture

Why neuromorphic computing

Conventional computing:

- Already facing scaling challenge (Moore's law)
- Excessive power consumption – 4-6 orders of magnitude than the brain
- Physical separation between CPU and memory – Von Neuman bottleneck

VS

Neuromorphic computing:

- Mimic neuro-bio architecture of nervous system
- Highly energy efficient - Asynchronous event-driven algorithms
- Localization of the memory and processing units synapse and neurons

To fully exploit brain like capabilities new architectures are needed

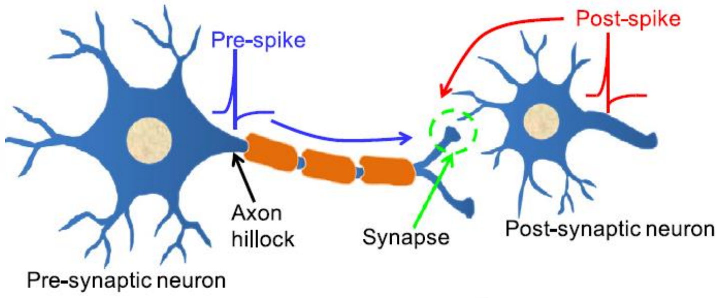
How does the brain work?

Neurons communicate through spikes – discrete events, robust to noise

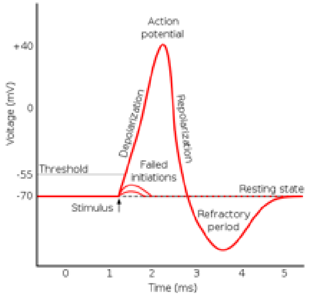
10^{11} Neurons
10^{15} Synapses

Massively parallel, highly energy efficient

neurons integrate incoming Pre-synaptic spikes



neuron fires only If the total integration spikes are above threshold



Action potential = spike

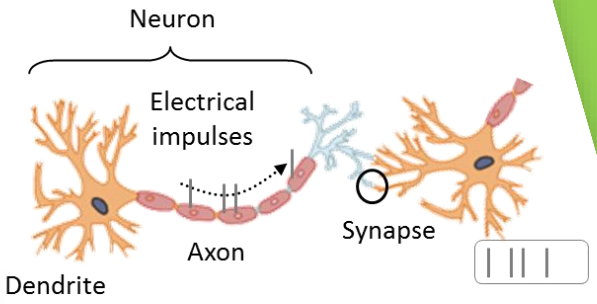
Biological brain – the most efficient computing architecture

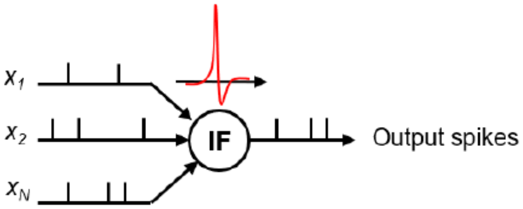
Getting closer to the brain - SNN

Spiking Neural Networks (SNNs)

Fully brain-inspired, use integrate & fire (IF) spiking neurons connected by analog synapses

- Each neuron integrates the incoming spikes, weighted through the synapses
- The neurons spike when the membrane potential threshold is exceeded



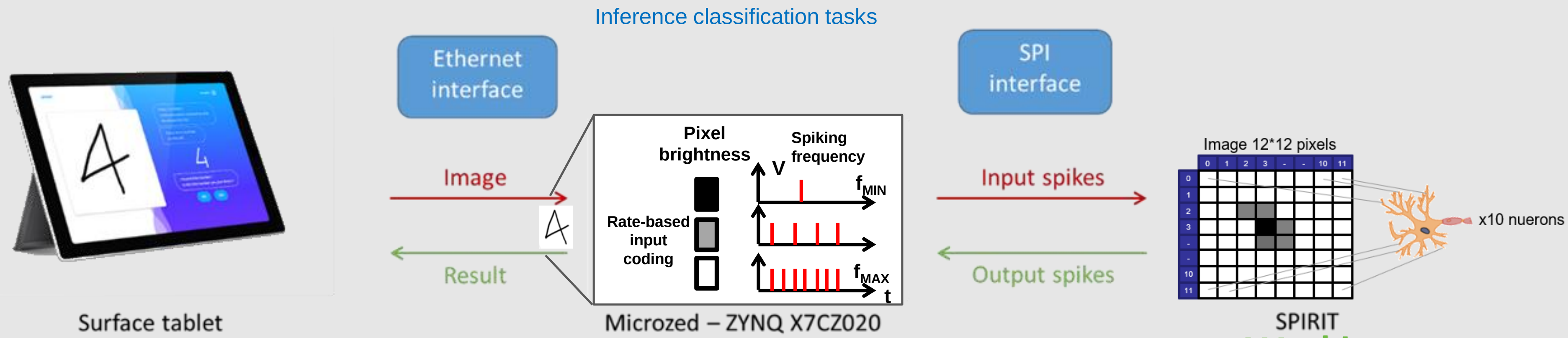


Spiking implementation allows for significant power reduction

ReRAM will allow to integrate dense non-volatile synapses for huge connectivity

Weebit-Leti SPIRIT demonstration

1st co-integration of analog spiking neurons and ReRAM based synapses for inference task

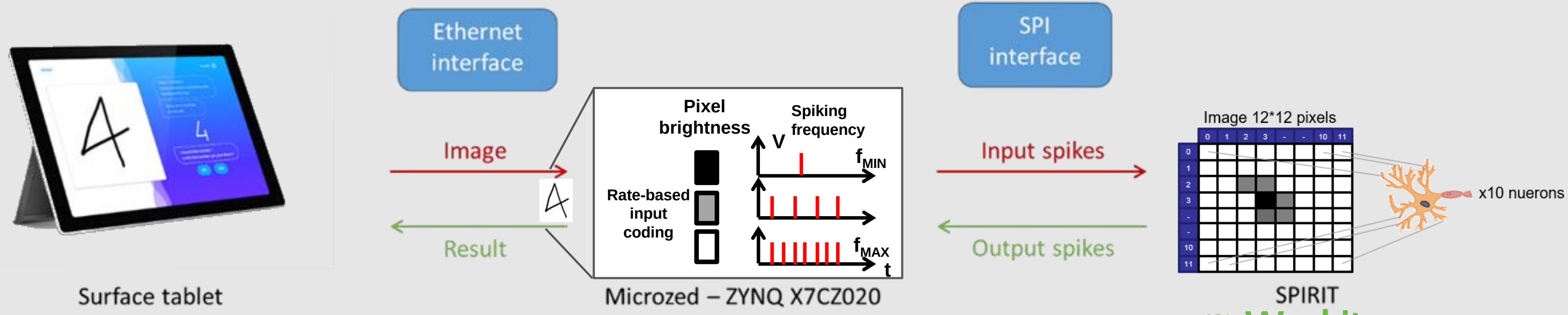


Weebit-Leti SPIRIT demonstration

SNN combining analog neurons and Weebit SiOx ReRAM synapses for MNIST digits recognition

- Fully connected - each neuron is connected to the entire image through ReRAM synapses
- Greyscale converted to input spikes frequencies
- Integrate & Fire (IF) analog neurons integrate the incoming spikes and fire
- Neuron with highest firing rate becomes the winner

Inference classification tasks



Conclusions

- Weebit – Leti demonstrates 1st ever analog spiking neurons and ReRAM based synapses
- Neuromorphic computing will enable efficient AI-dedicated hardware
- ReRAMs can be used to implement:
  - Analog accelerators for common deep learning neural networks
  - Brain-inspired spiking neural networks with resistive elements and analog neurons

THANK YOU

Weebitnano
The Future Memory