Energy-Efficient Hardware for Computing at the Edge

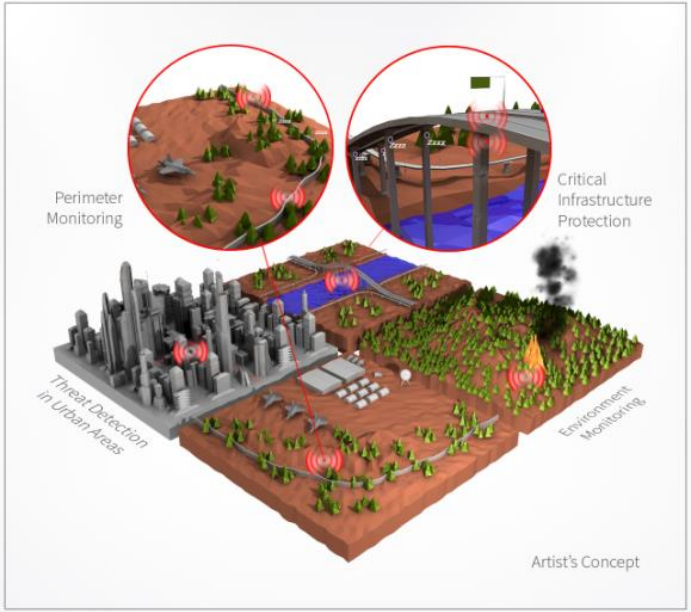
Dr. Sahil Shah , Assistant Professor, Electrical Engineer.

UNIVERSITY OF MARYLAND

Edge Intelligence Application

Autonomous drones

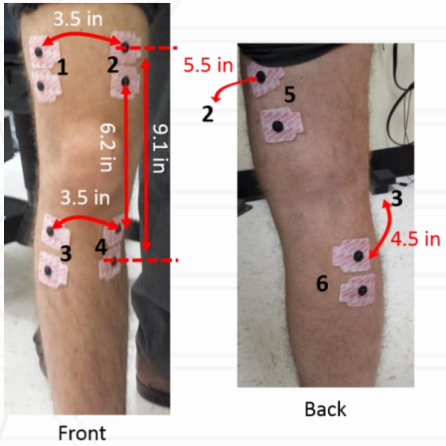
Augmented Reality



Remote Sensing

Edge Intelligence Application

Implantable Devices

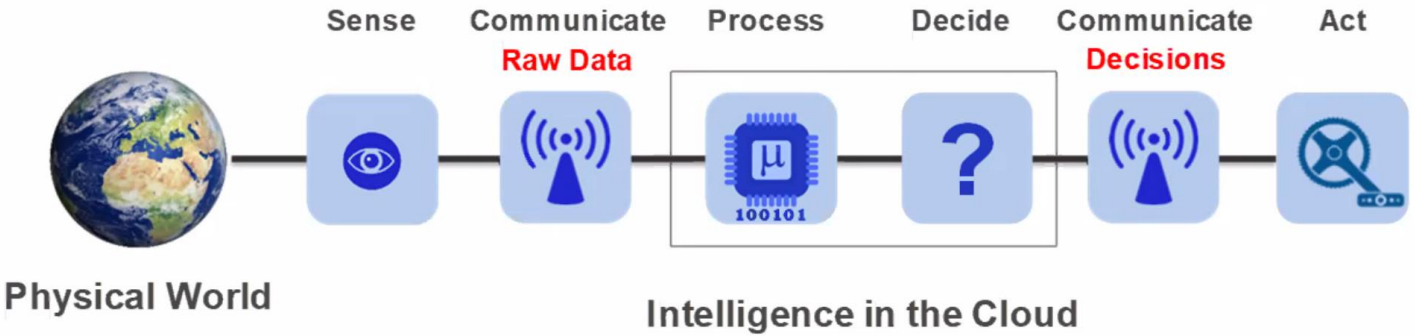


Continuous monitoring
Wearable devices

Edge Device Specifications

- Low-power consumption
- Low Latency in decision
- Smaller Area
- On-Device learning (Intelligence)

Current Edge Computing Framework

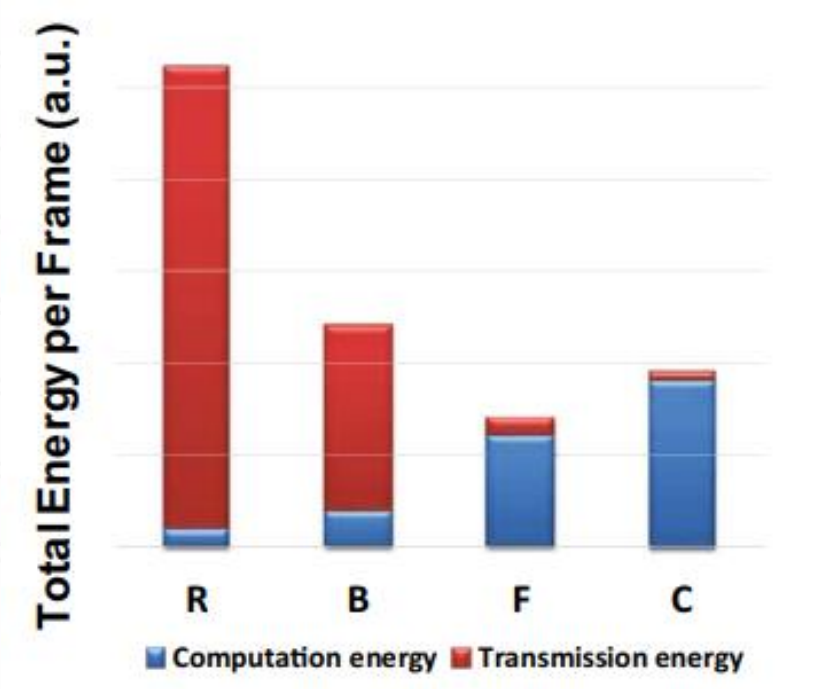


Drawbacks

- Latency
- Privacy Concerns
- Power Consumption

A. Keshavarzi and W. van den Hoek, "Edge Intelligence—On the Challenging Road to a Trillion Smart Connected IoT Devices," in IEEE Design & Test, vol. 36, no. 2, pp. 41-64, April 2019, doi: 10.1109/MDAT.2019.2899075.

Local Computing vs Communication



R: Communicating Raw data

B: Background computation

F: Feature extraction on-chip

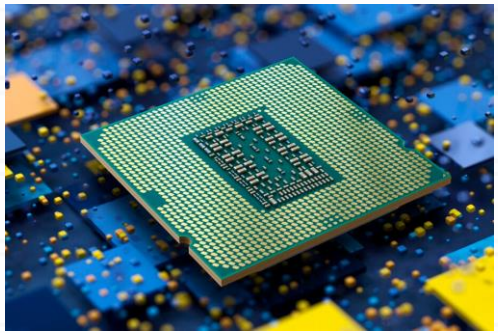
C: Classification on-chip

On-chip compute enables:

- Reducing latency
- Learning on-chip
- Increase security

A. Keshavarzi and W. van den Hoek, "Edge Intelligence—On the Challenging Road to a Trillion Smart Connected IoT Devices," in IEEE Design & Test, vol. 36, no. 2, pp. 41-64, April 2019, doi: 10.1109/MDAT.2019.2899075.

Current Computing Paradigm



CPUs

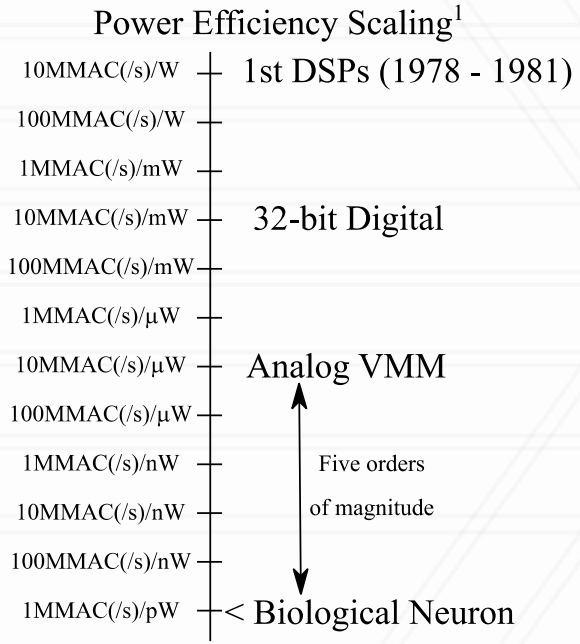
GPUs

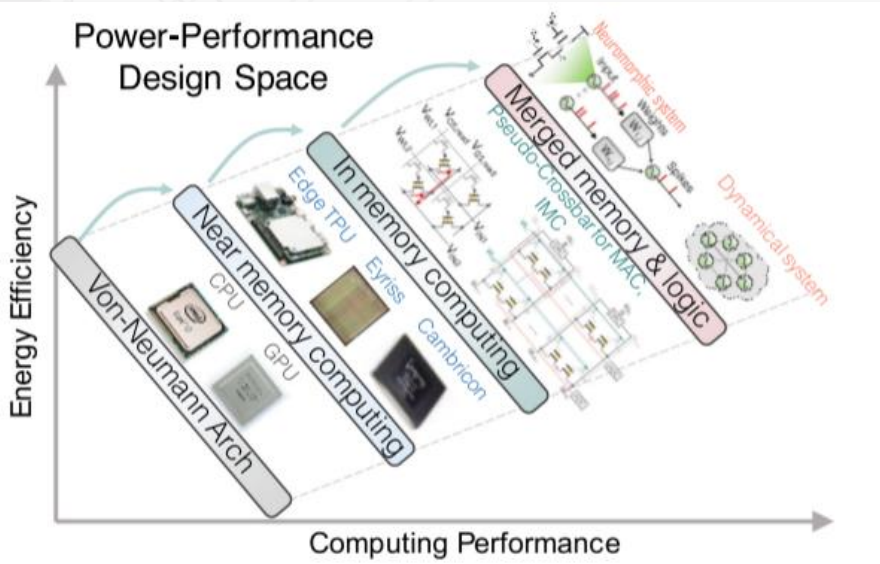
Power for computing Neural Network

Device	Power(W)
GPU GTX 1080 Ti	95.9
i7-8700K	35.6

1. Chang Gao et. al. “DeltaRNN: A Power-efficient Recurrent Neural Network AcceleratorFPGA’18”, February 25–27, Monterey, CA, USA

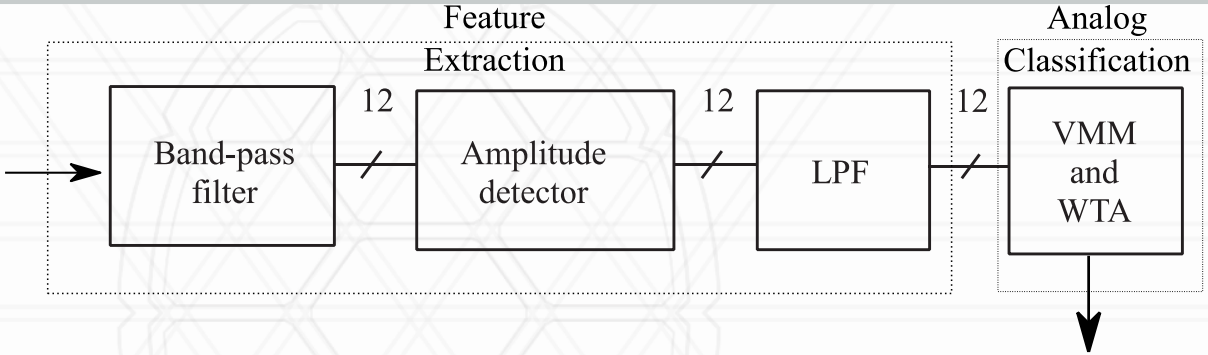
Investigating Energy-Efficient Computation



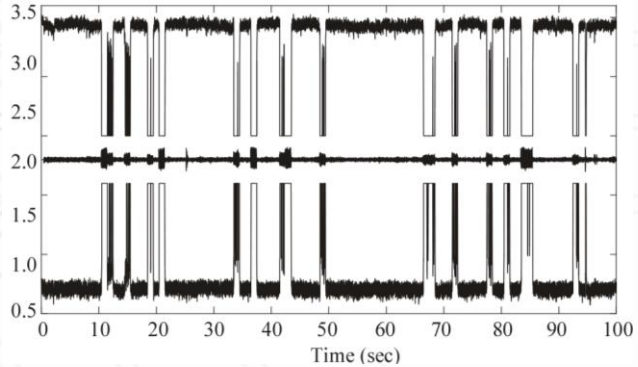


1. Jennifer Hasler and Bo Marr "Finding a roadmap to achieve large neuromorphic hardware systems" Frontiers in Neuroscience
2. A. Keshavarzi and W. van den Hoek, "Edge Intelligence—On the Challenging Road to a Trillion Smart Connected IoT Devices," in IEEE Design & Test, vol. 36, no. 2, pp. 41-64, April 2019, doi: 10.1109/MDAT.2019.2899075.

Application: Edge Computing for Remote Sensing

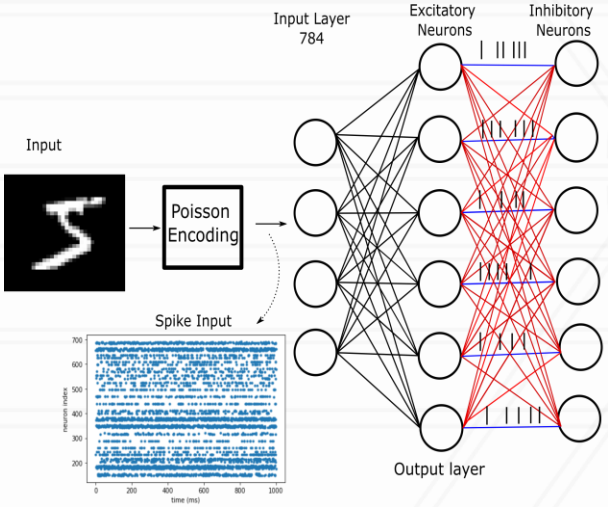


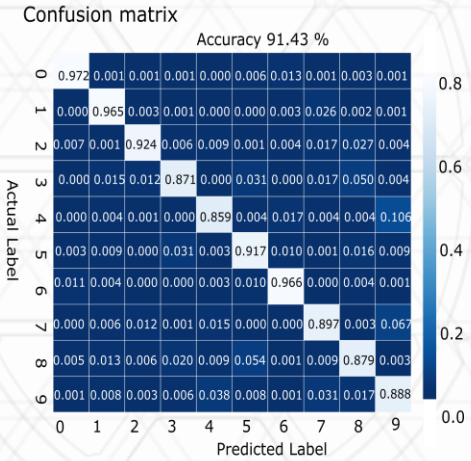
- ADC less classification
- Power 23μW



Application: Edge Computing for Image Classification

MNIST classification



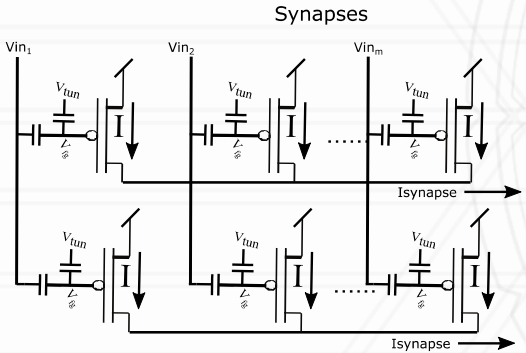


Advantages

- Low spike rate leads to sparsity
- Local learning rules are energy-efficient

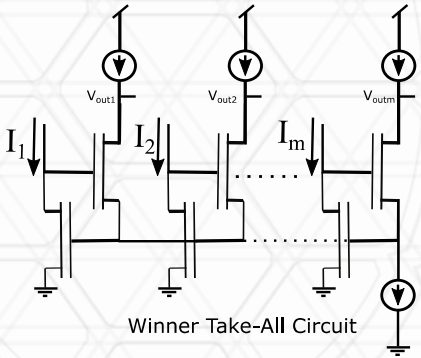
Building Blocks

Matrix Multiplication

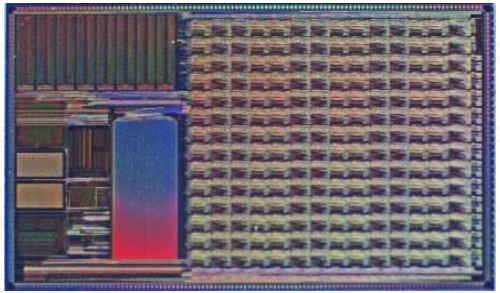


Non-volatile FG transistors

Winner-Take All Circuit (Soft-Max)



Die Photo



Future Directions

- Scaling it for Complex Tasks (eg. Object Detection)
- Reconfigurable and Programmable Hardware
- On-Chip Learning (Adapt based on data)

Questions?

UNIVERSITY OF
MARYLAND