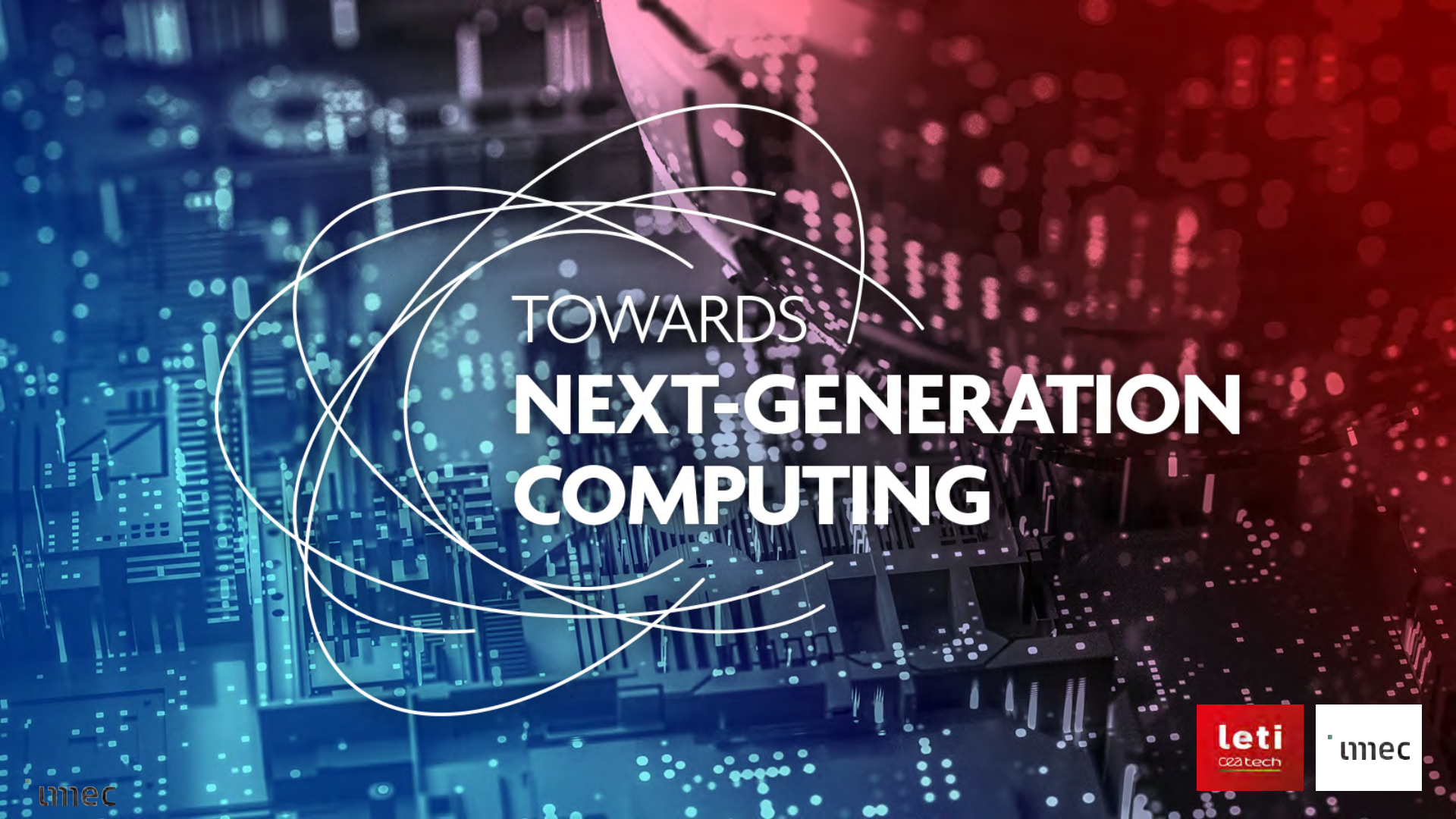

TOWARDS
NEXT-GENERATION
COMPUTING
leti
ceatech
imec
imec

IN-MEMORY COMPUTING TECHNOLOGIES FOR DEEP NEURAL NETWORK ACCELERATION

D. VERKEST

TOWARDS NEXT-GENERATION COMPUTING

SMART SERVICES

POWERED BY MACHINE LEARNING IN DATA CENTER

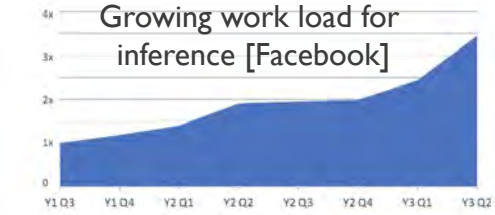


Figure 1: Server demand for DL inference across data centers

https://arxiv.org/pdf/1811.09886.pdf

Consumer & IOT

Infrastructure Provider

Cloud Provider

SMART SERVICES

POWERED BY MACHINE LEARNING IN DATA CENTER

LATENCY
ENERGY
PRIVACY

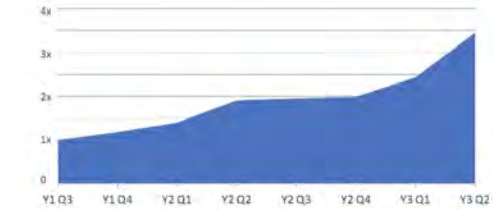


Figure 1: Server demand for DL inference across data centers

https://arxiv.org/pdf/1811.09886.pdf

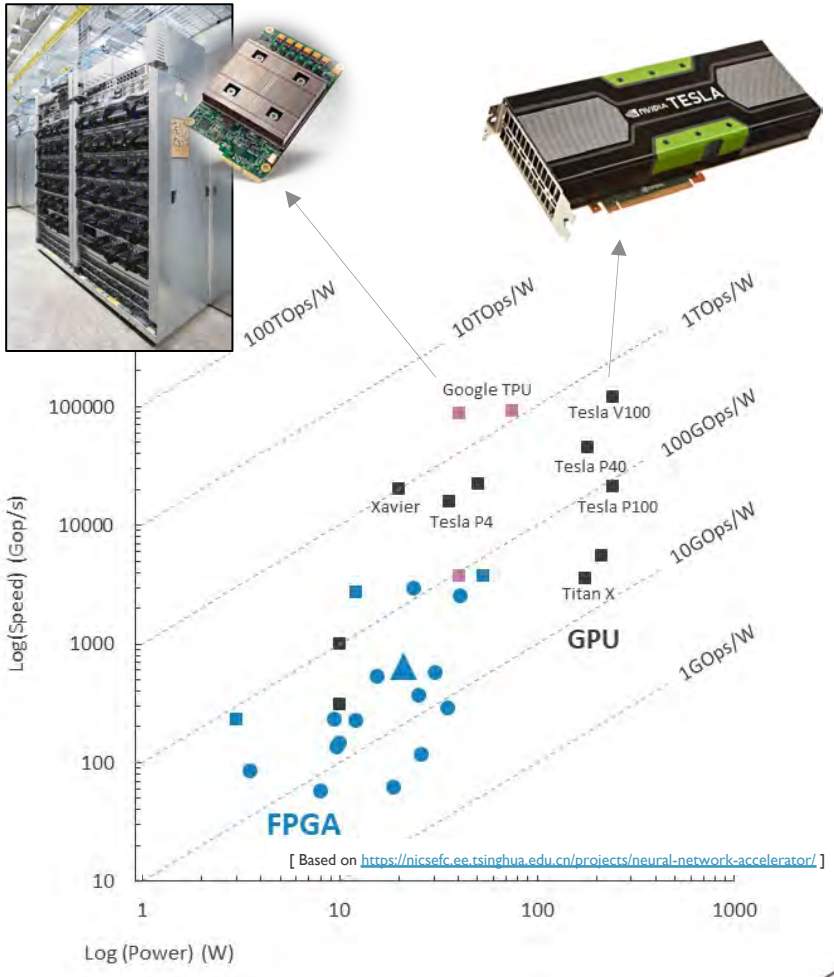
Consumer & IOT

Infrastructure Provider

Cloud Provider

TODAY'S DOMINANT SOLUTION
TARGET HIGH-PERFORMANCE MARKET

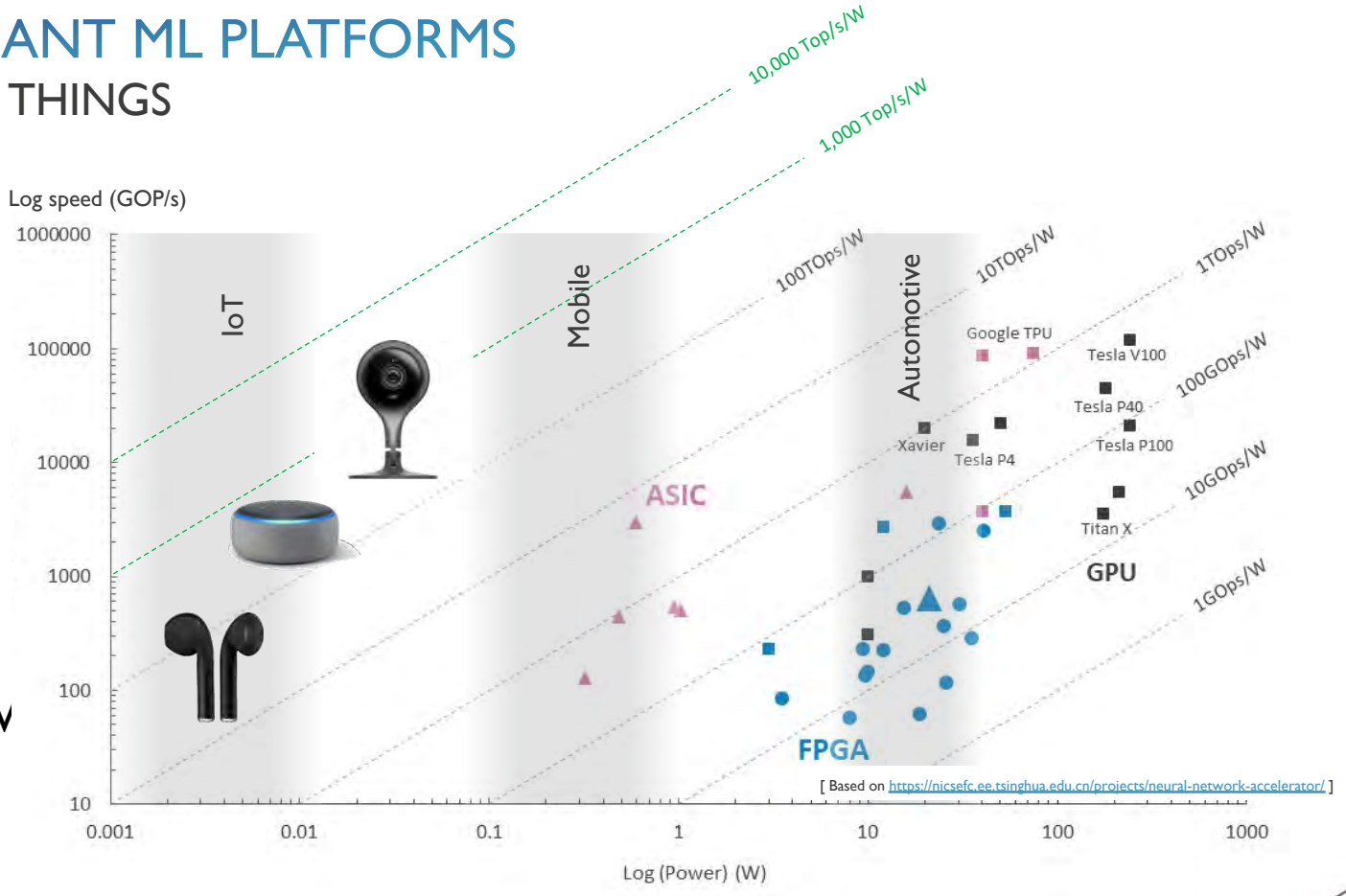
- Deep Neural Networks
  - CNN, LSTM
  - Vision, voice assistant
- General purpose platforms support
  - Training and inference
  - Flexibility in still evolving field
- Too power hungry for edge devices



TODAY'S DOMINANT ML PLATFORMS
NOT FIT FOR SMART THINGS

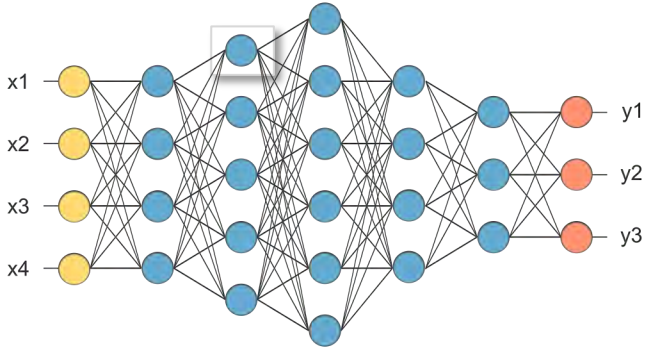
- Intelligence at extreme edge
  - Support high performance DNN inference
  - Always-on
  - On a tight power budget

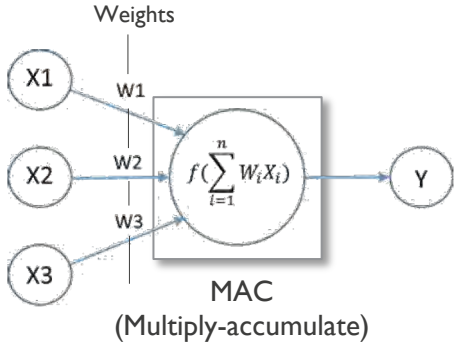
➔ 1 to 10 PetaOP/s/W



ANALOG COMPUTE-IN-MEMORY ACCELERATORS FOR ML

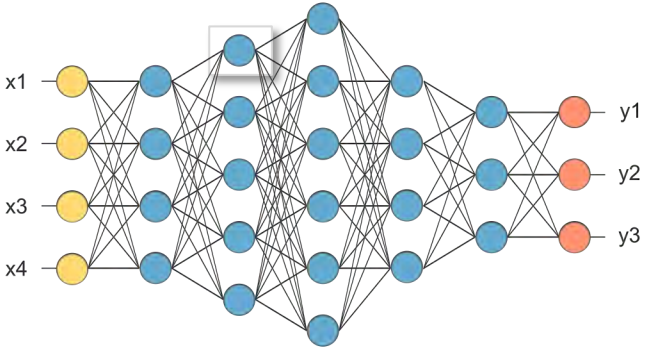
SUPPORTED BY NEW MEMORY TECHNOLOGY

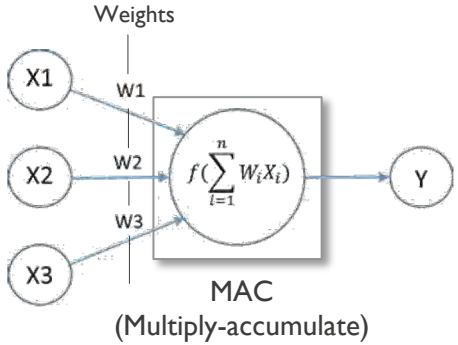


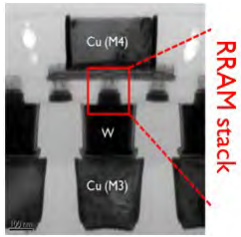


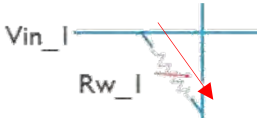
ANALOG COMPUTE-IN-MEMORY ACCELERATORS FOR ML

SUPPORTED BY NEW MEMORY TECHNOLOGY





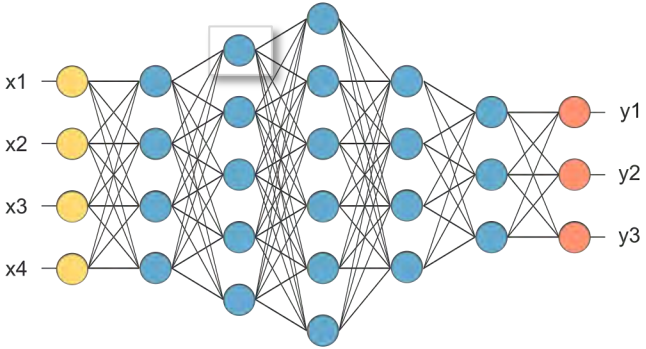


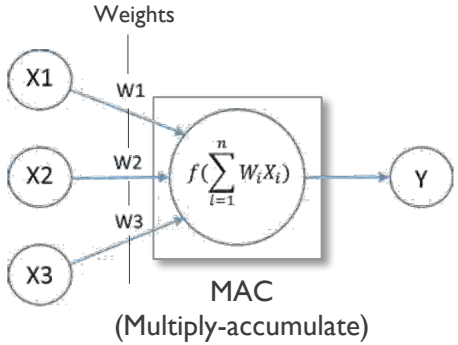


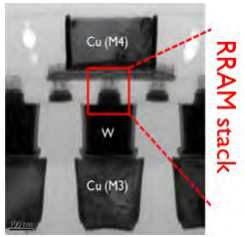
$I \sim x_1 w_1$

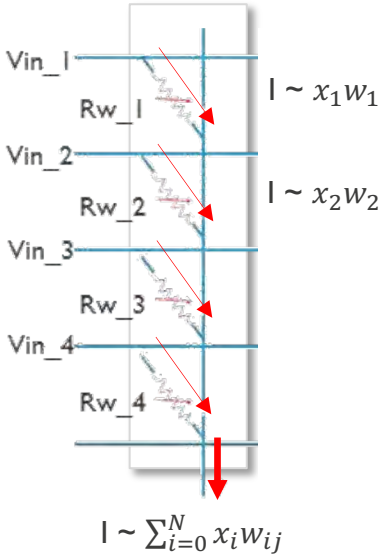
ANALOG COMPUTE-IN-MEMORY ACCELERATORS FOR ML

SUPPORTED BY NEW MEMORY TECHNOLOGY



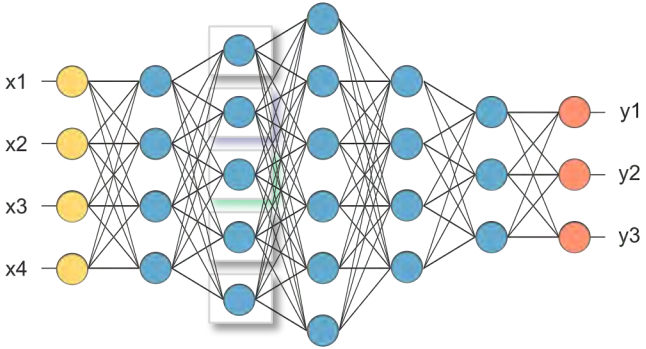


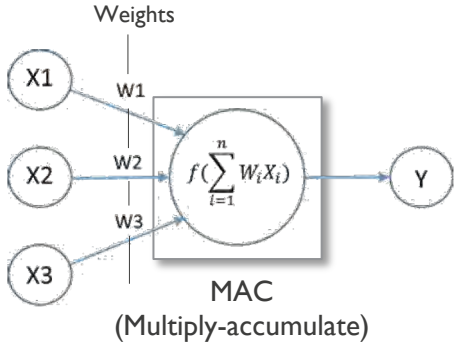


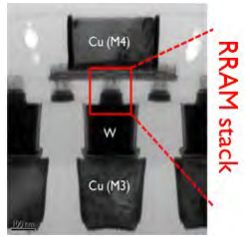


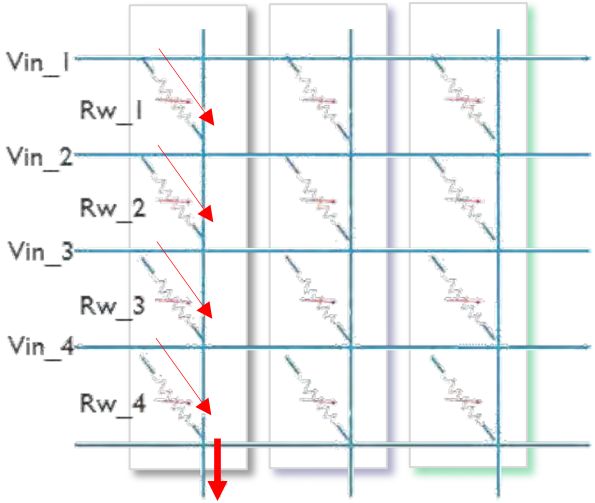
ANALOG COMPUTE-IN-MEMORY ACCELERATORS FOR ML

SUPPORTED BY NEW MEMORY TECHNOLOGY





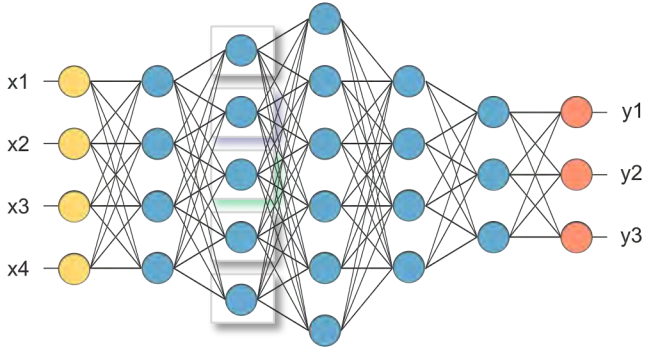


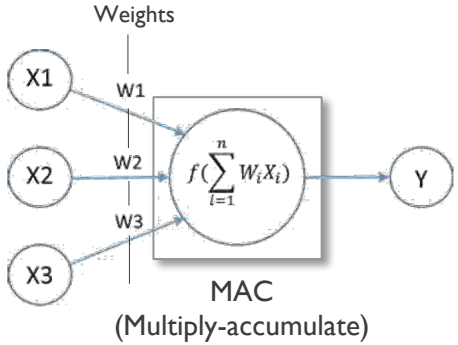


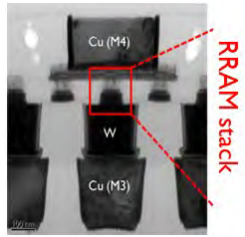
$I \sim \sum_{i=0}^{N} x_i w_{ij}$

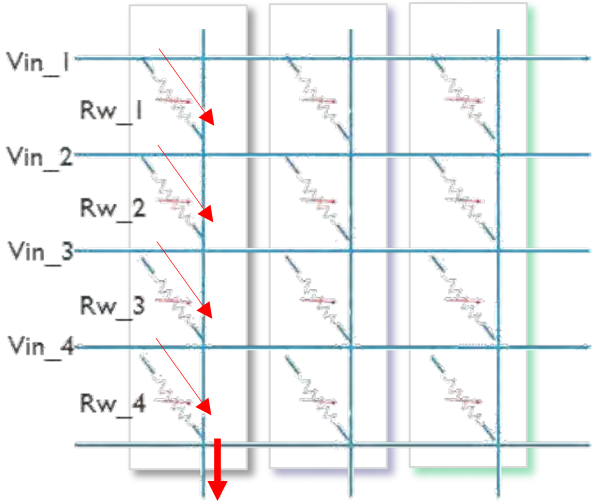
ANALOG COMPUTE-IN-MEMORY ACCELERATORS FOR ML

SUPPORTED BY NEW MEMORY TECHNOLOGY







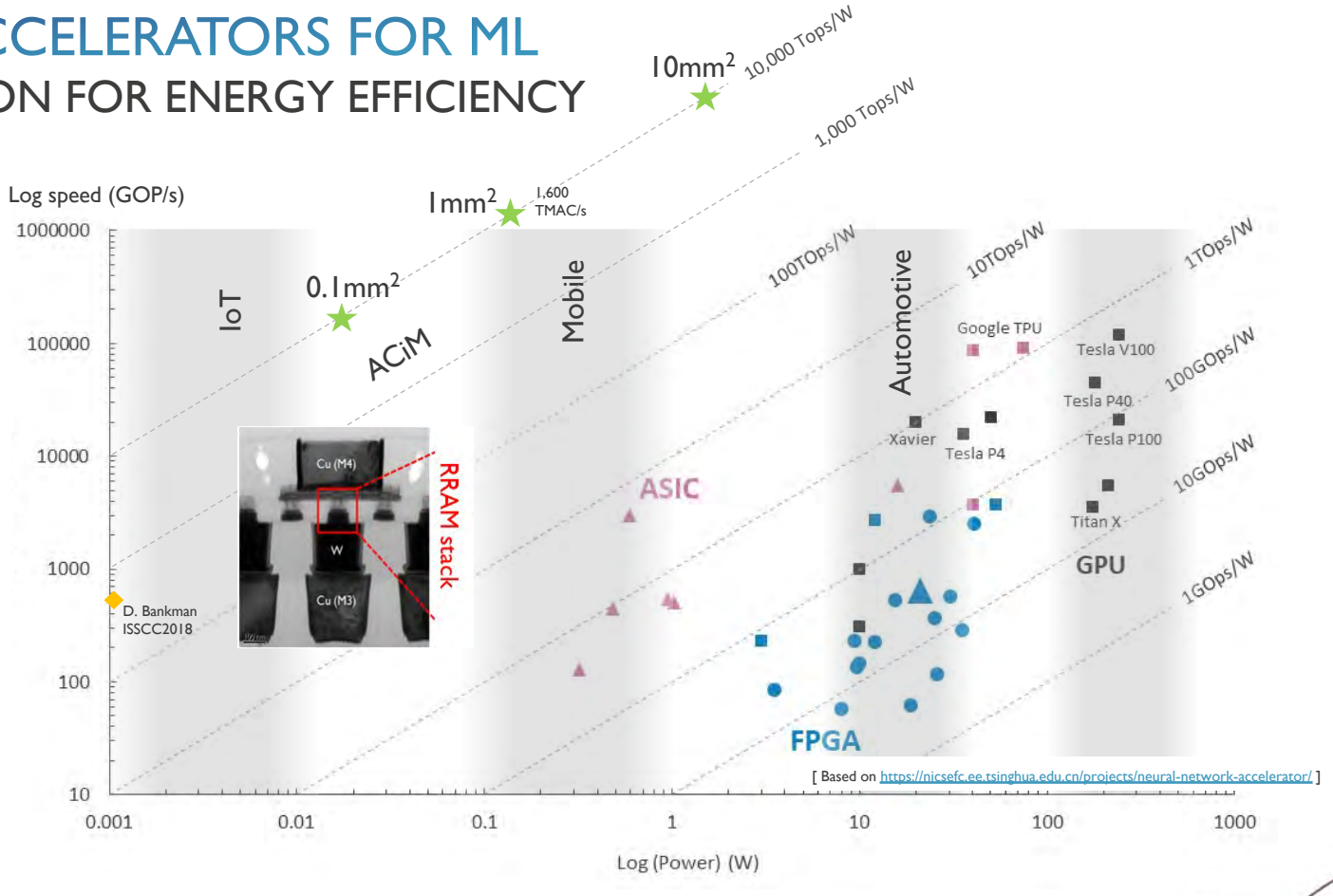


- Use memory array for massive parallel analog implementation of multiply-accumulate operations in DNN layer
- Memory array stores weights *and* implements the neural network layer matrix-vector multiplication in analog fashion
  - → analog compute-in-memory
  - → computational memory
  - → neuromorphic computing

ANALOG CIM ACCELERATORS FOR ML

PROMISING SOLUTION FOR ENERGY EFFICIENCY

- 10,000 TOP/s/W
  - Memory cell 0.25µm x 0.25µm
  - 16M cells/mm^2 at 100MHz
  - 1.6 PMAC/s/mm^2 for ~0.2fJ/MAC

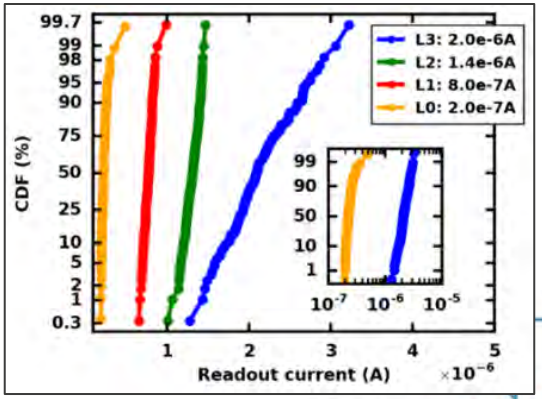


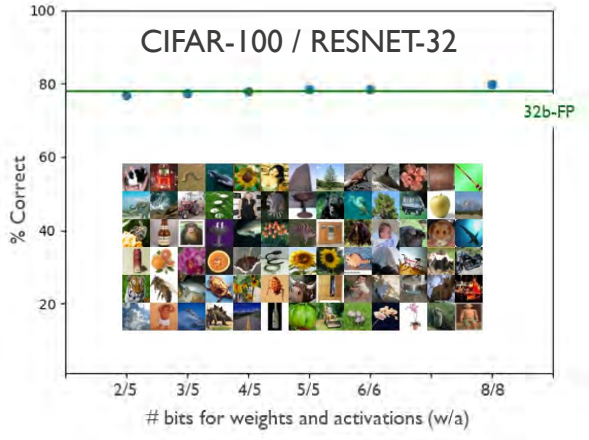
65th International Electron Devices Meeting

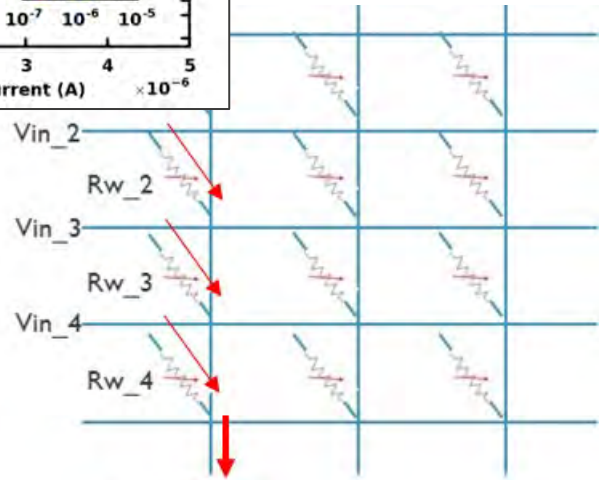
Tue Dec 10, 2019, Focus session: Emerging AI hardware, Paper 22.2 by S. Cosemans, et al.

HOW TO MAKE IT WORK?

SYSTEM-TECHNOLOGY CO-OPTIMIZATION







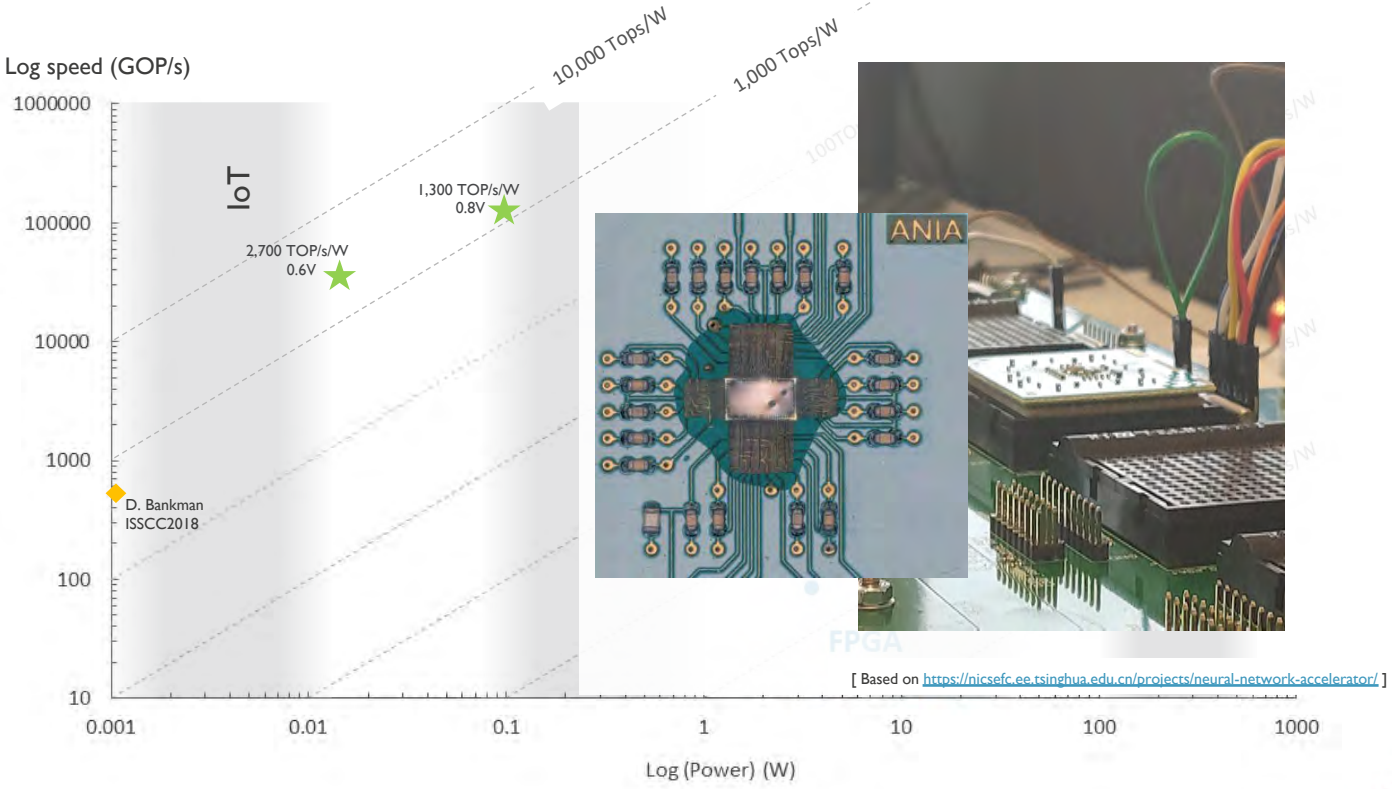
- Deep neural networks
  - Perform inference with lower precision
  - Can tolerate some weight variation
- Architecture
  - Use multiple memory cells per weight for more precision
- Circuit and device
  - Tune devices for neural network operation

imec

ANALOG COMPUTE-IN-MEMORY ACCELERATORS FOR ML

PROOF-OF-CONCEPT MATRIX-VECTOR MULTIPLICATION (MVM)

- SRAM-based compute cells for MVM of CNN and LSTM
- FDX22
  - 4mm2 incl.
  - 500k cells
  - Ternary weights
  - 7-bit activations
  - 6-bit ADC (MAC)
- Record efficiencies up to 2,700 TOP/s/W



[Based on https://nicsefc.ee.tsinghua.edu.cn/projects/neural-network-accelerator/]

imec

embracing a better life

THANK YOU FOR YOUR ATTENTION