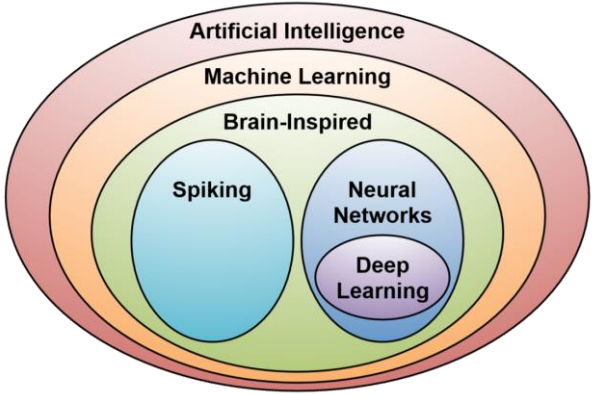
# SEMICONDUCTOR TECHNOLOGIES FOR ENERGY-EFFICIENT INTELLIGENCE

PETER.DEBACKER@IMEC.BE

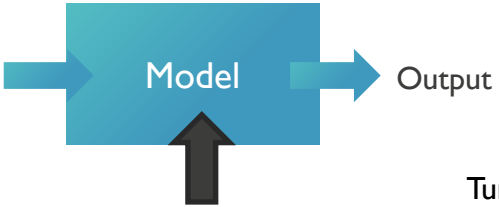
# INTELLIGENT HARDWARE?

## FROM DEEP NEURAL NETWORKS TO NEUROMORPHIC




V. Sze, et al. "Efficient processing of deep neural networks: A Tutorial and Survey", Proc. of the IEEE, Vol. 105, No. 12, Dec. 2017 [ https://arxiv.org/abs/1703.09039 ]

### MACHINE LEARNING





Tuning model parameters based on available data = "learning" without explicit programming

- Pattern Recognition
- Feature Extraction
- …

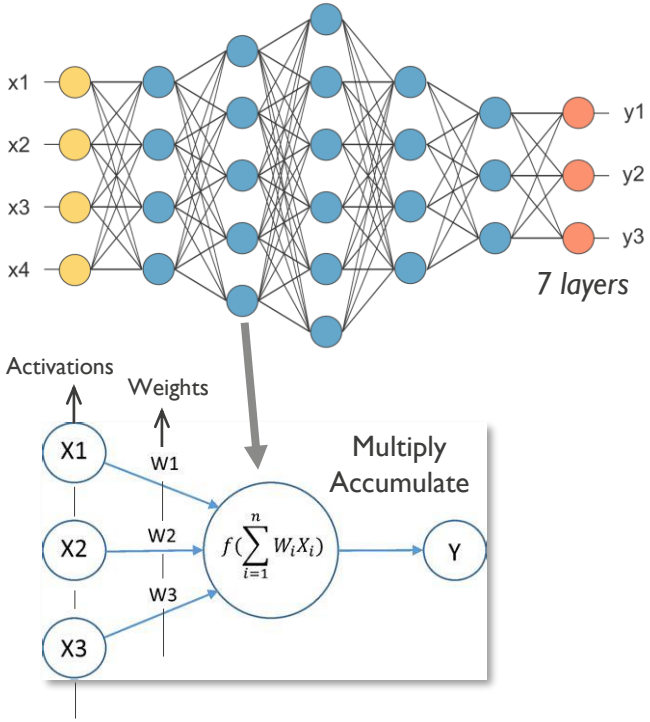
### BRAIN-INSPIRED

Using artificial neural networks as machine learning model

- Multilayer perceptron
- Convolutional Neural Networks (CNN)
- Long Short Term Memory (LSTM)
- Spiking Neural Networks (SNN)
- Hierarchical Temporal Memory (HTM)
- ...

imec

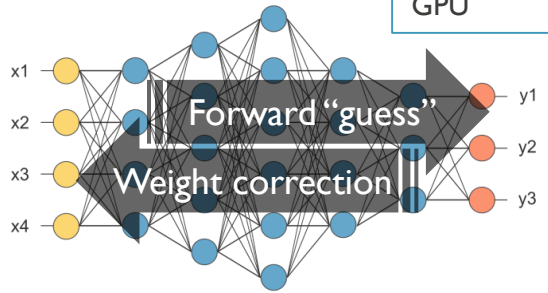
# MACHINE LEARNING
## DEEP NEURAL NETWORKS



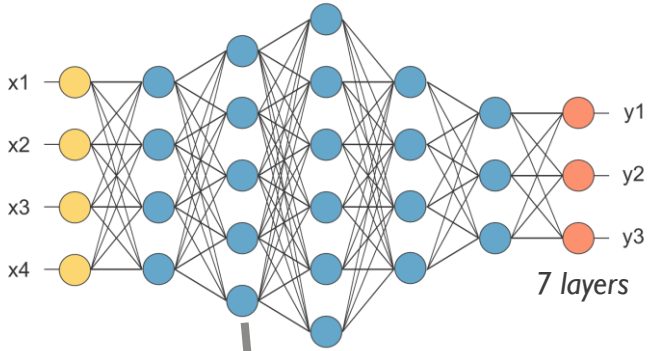

100,000's images

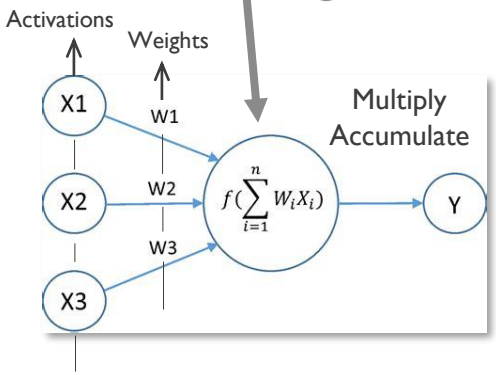
Training of CNN:
- Huge data set with known objects to determine value for millions of weights

Typically in data center on high-performance GPU



# MACHINE LEARNING
## DEEP NEURAL NETWORKS



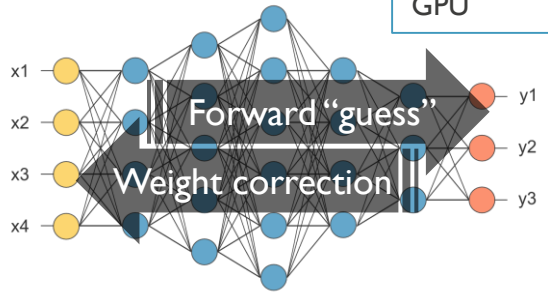


100,000's images

Training of CNN:

- Huge data set with known objects to determine value for millions of weights

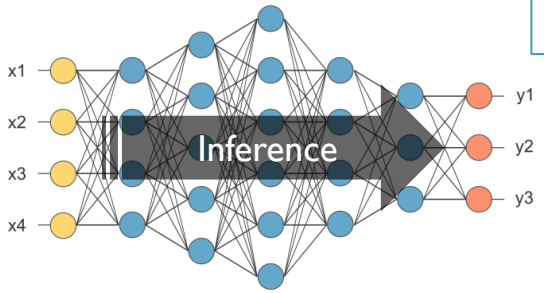
Typically in data center on high-performance GPU




Use of CNN for inference

- Use the trained network to classify objects

Face recognition
Smart driving
Surveillance
...
In embedded device or in the cloud



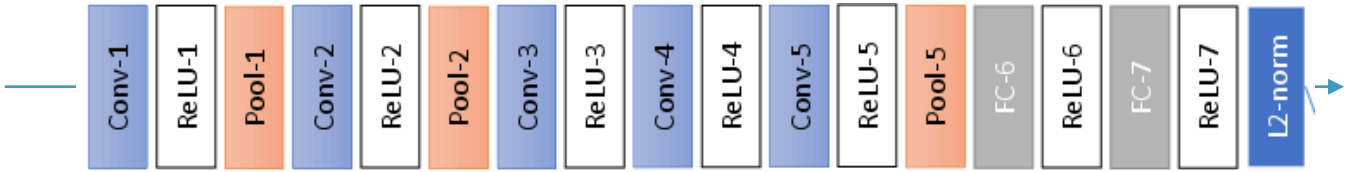

95%
container ship

4%
life boat

1%
toy

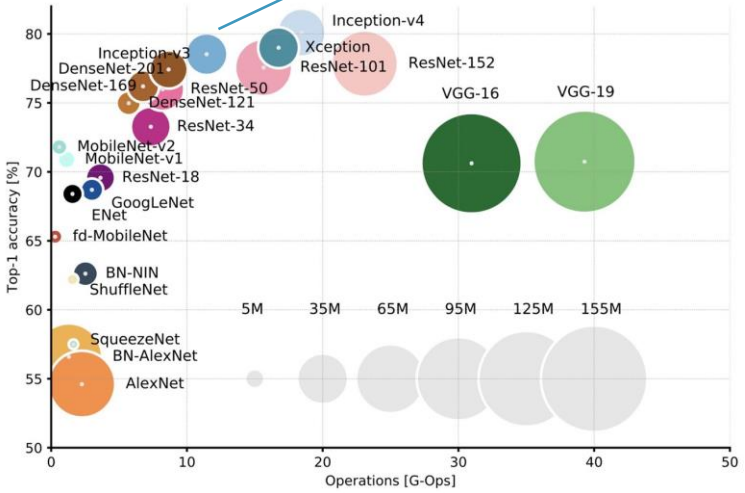
# MACHINE LEARNING USING DEEP NEURAL NETWORKS

## >95% OF OPERATONS ARE MULTIPLY-ACCUMULATE




- >95% of computations: Multiply-Accumulate (MAC)
  - Convolutional (Conv) layers: feature extractors
  - Fully-Connected (FC) layers: classifiers
- Remaining functions (pooling, threshold, normalization, ...) are critical for correct operation of DNN but do not dominate performance
- Number of layers and number of weights growing in pursuit of better classification accuracy.

E.g. ResNet-50:
- 50 layers
- 20+ Million weights
- 12 GOPs/inference




Source: https://medium.com/towards-data-science/neural-network-architectures-156e5bad51ba and arXiv:1605.07678v4

# DEEP NEURAL NETWORKS (CNN, LSTM)

## MAINSTREAM DEEP LEARNING

- Convolutional Neural Networks: visual processing and object classification
  - Character recognition
  - Medical image analysis
  - People / face identification
  - ...
- LSTM
  - Time-series prediction, e.g. predict the weather or stocks
  - Action recognition, e.g. event detection in security cameras
  - Robotic movements, e.g. drive various joints based on sensory inputs
  - Speech recognition, e.g. ask Siri, Alexa, Cortana or Google assistant questions
  - Language models, e.g. Google translate
  - Anomaly detection, e.g. firewall applications in network traffic

# SMART SERVICES REQUIRE ENERGY-EFFICIENT MACHINE LEARNING

## ENABLE SMART EDGE DEVICES ... GROWING DATACENTER INFERENCE WORKLOAD


**LATENCY**
**ENERGY**
**PRIVACY**

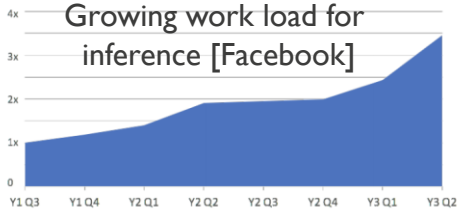



**Figure 1: Server demand for DL inference across data centers**

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

Consumer & IOT

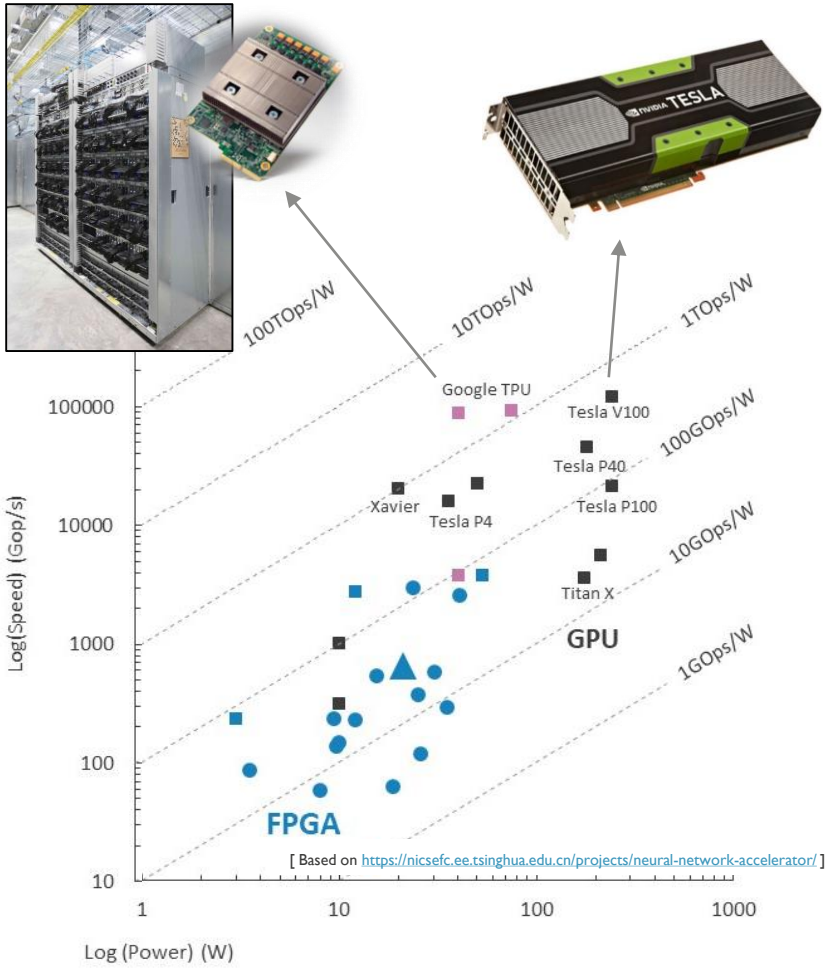
Infrastructure Provider

Cloud Provider

# HARDWARE PLATFORMS FOR MACHINE LEARNING

# TODAY'S DOMINANT PLATFORMS
## TARGET HIGH-PERFORMANCE MARKET

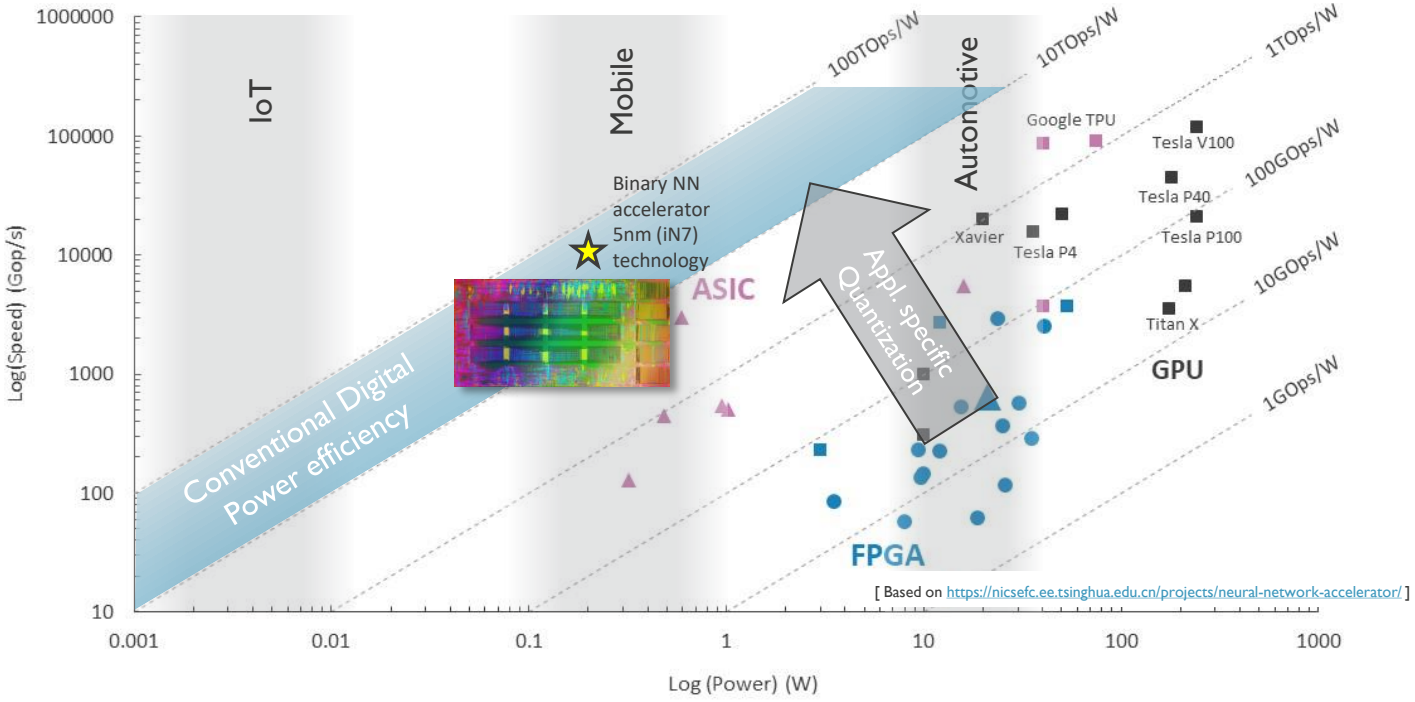
- Focus on <u>training</u> of deep learning algorithms
- General purpose platforms (GPU, FPGA) support flexibility in still evolving field
- First application-specific IC (Google TPU) emerge



# NEURAL NET ACCELERATORS (ASIC) EMERGE FOR INFERENCE

## CONVENTIONAL DIGITAL CMOS

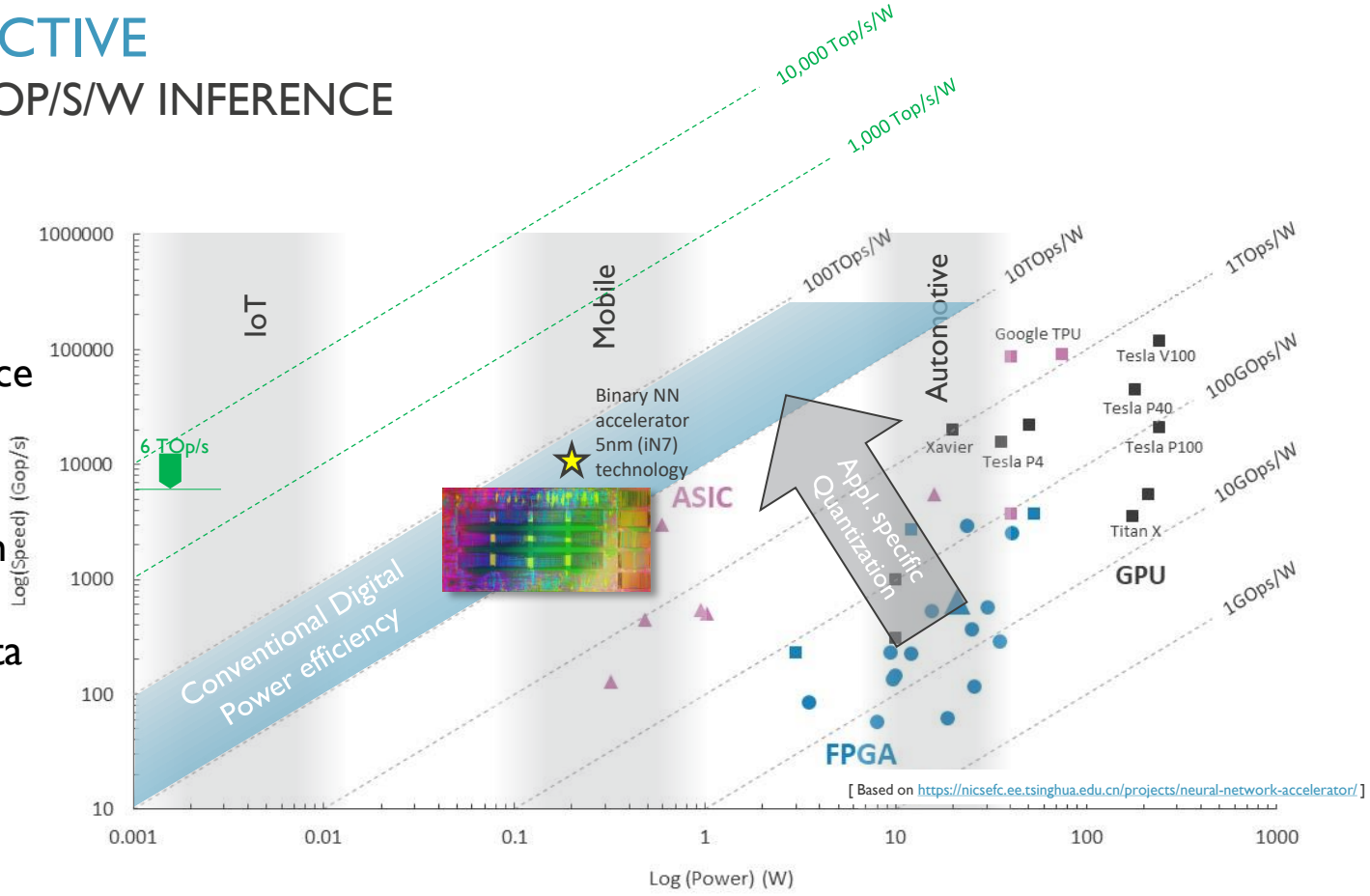
- Power efficiency < 100 TOP/s/W
  - application specific HW
  - large embedded memories
  - aggressively quantized NN



# PROGRAM OBJECTIVE

## TOWARDS 10,000 TOP/S/W INFERENCE

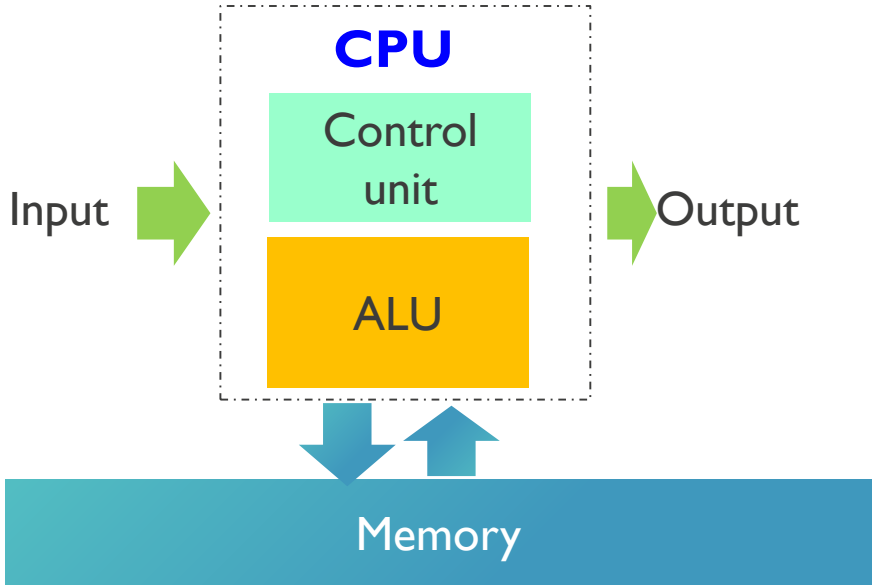
- Sustainable smart services
  - Smart IOT devices: inference in edge devices
  - Support growth of inference workload in data centers




imec

# OPPORTUNITIES FOR ANALOG IN-MEMORY COMPUTING

## CHALLENGES OF VON NEUMANN COMPUTING



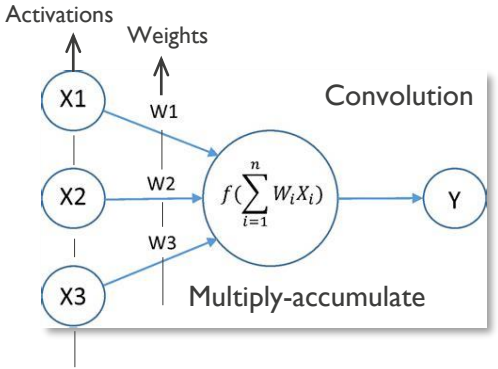

**Von-Neumann architectures**

- **Challenges:** Fundamental difficulties in scalability of memory bandwidth and capacity
- **Emerging Opportunities**
  - Performing calculation in memory
  - Non-volatile memory by limiting the data transfer between memory and CPU

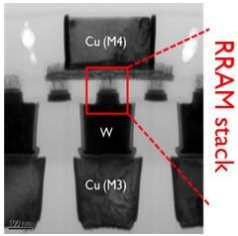
# ANALOG COMPUTE IN-MEMORY

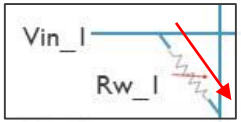
# ANALOG COMPUTE-IN-MEMORY ACCELERATORS FOR ML

## SUPPORTED BY NEW MEMORY TECHNOLOGY




1. Memristor (programmable resistor) stores weight value $W_I$ as an analog quantity: conductance $Rw_I$
2. Activation $X_I$ applied as analog voltage $Vin_I$
3. Ohm's law: memristor cell current ~ $X_I.W_I$
4. Kirchoff's law: bit-line current ~ $\sum X_i.W_i$



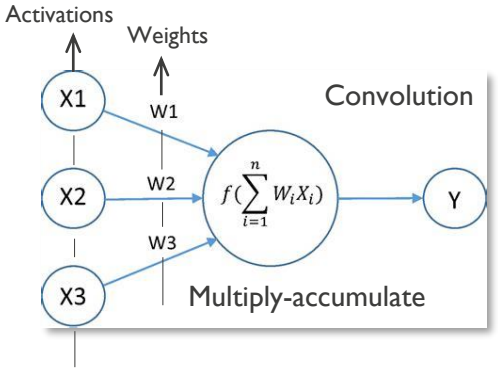




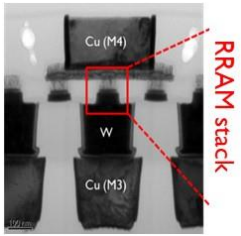
imec

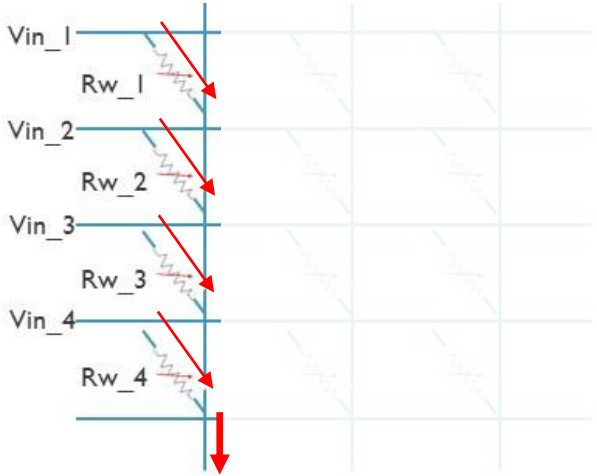
# ANALOG COMPUTE-IN-MEMORY ACCELERATORS FOR ML

## SUPPORTED BY NEW MEMORY TECHNOLOGY




1. Memristor (programmable resistor) stores weight value $W_I$ as an analog quantity: conductance $Rw_I$
2. Activation $X_I$ applied as analog voltage $Vin_I$
3. Ohm's law: memristor cell current ~ $X_I.W_I$
4. Kirchoff's law: bit-line current ~ $\sum X_i.W_i$



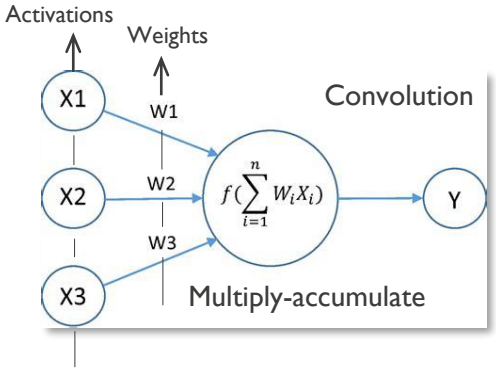



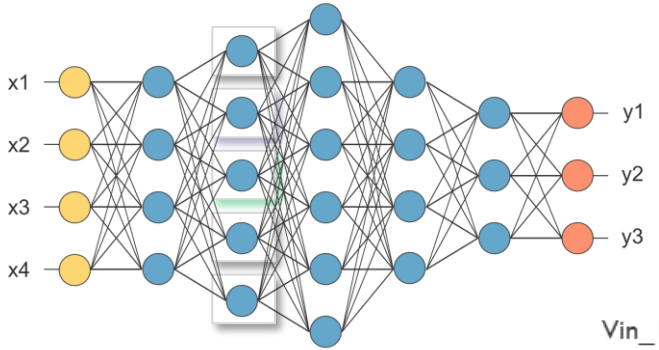

$I \sim \sum_{i=0}^{N} x_i w_{ij}$

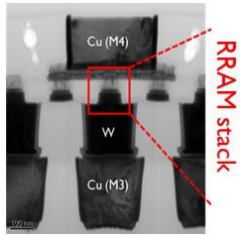
imec

# ANALOG COMPUTE-IN-MEMORY ACCELERATORS FOR ML

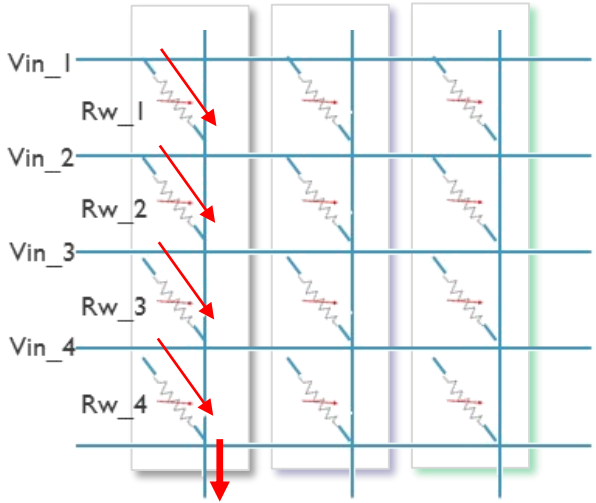
## SUPPORTED BY NEW MEMORY TECHNOLOGY



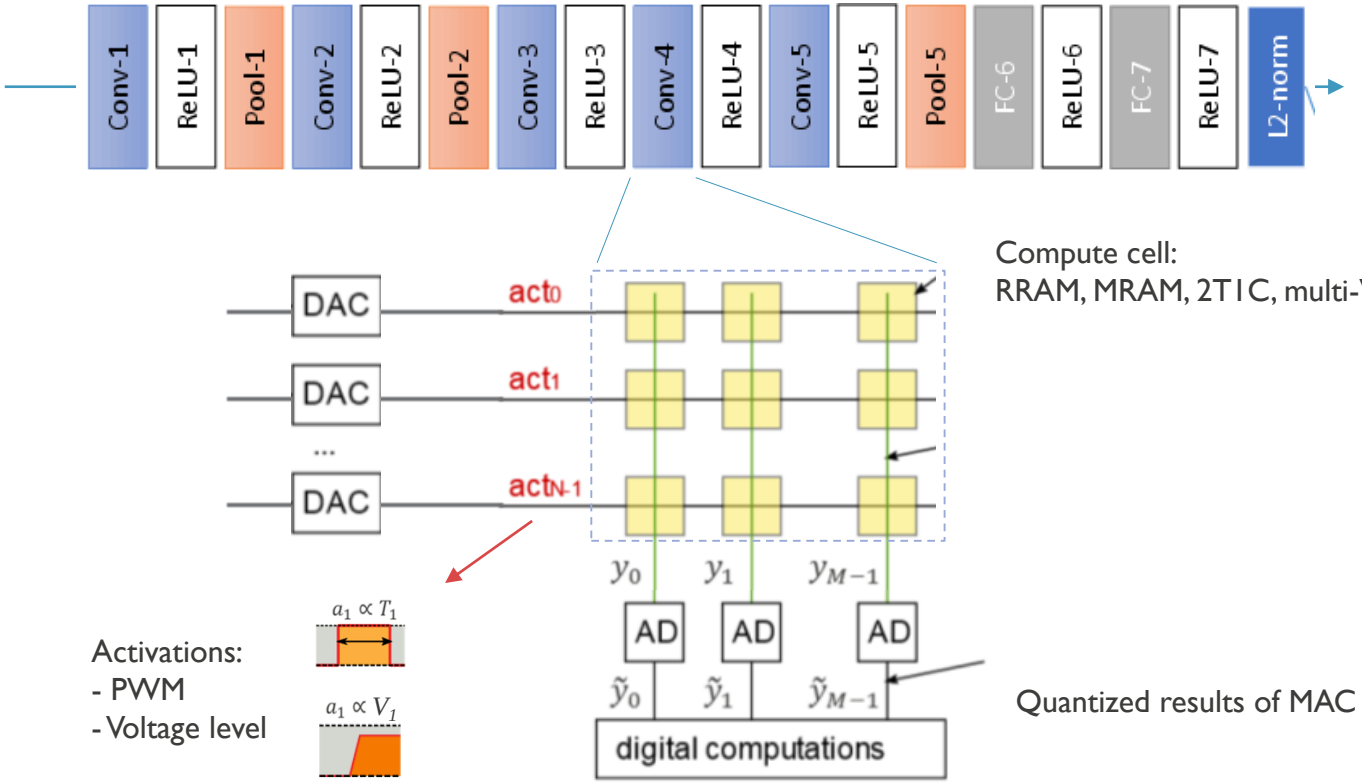

- Use memory array for massive parallel analog implementation of multiply-accumulate operations in DNN layer
- Memory array stores weights *and* implements a logic function (MAC) in analog fashion
  - → compute-in-memory
  - → computational memory
  - → neuromorphic computing




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

# POOLING, THRESHOLDING, ... IN DIGITAL DOMAIN

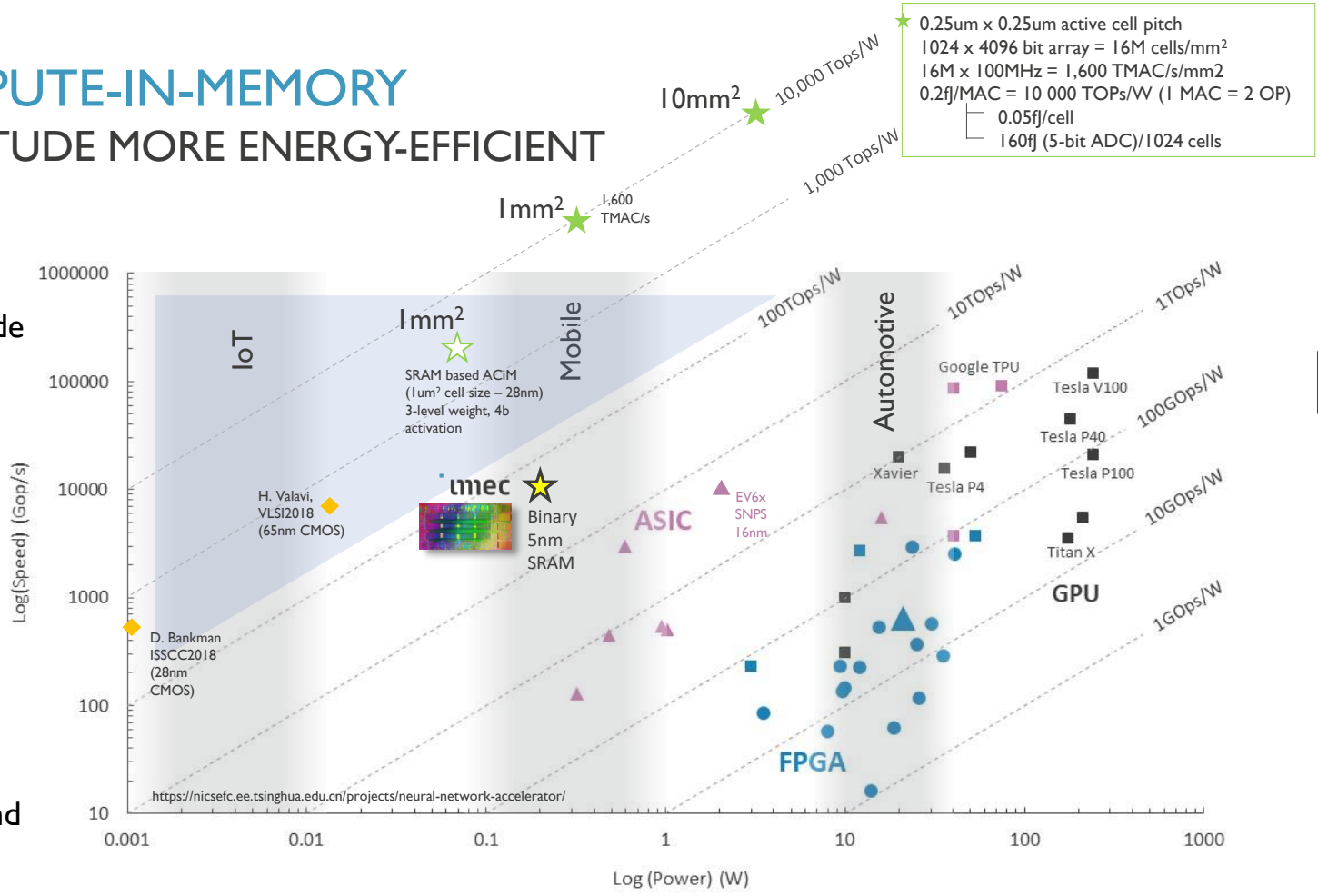
## DAC AND ADC REQUIRED



# ANALOG COMPUTE-IN-MEMORY

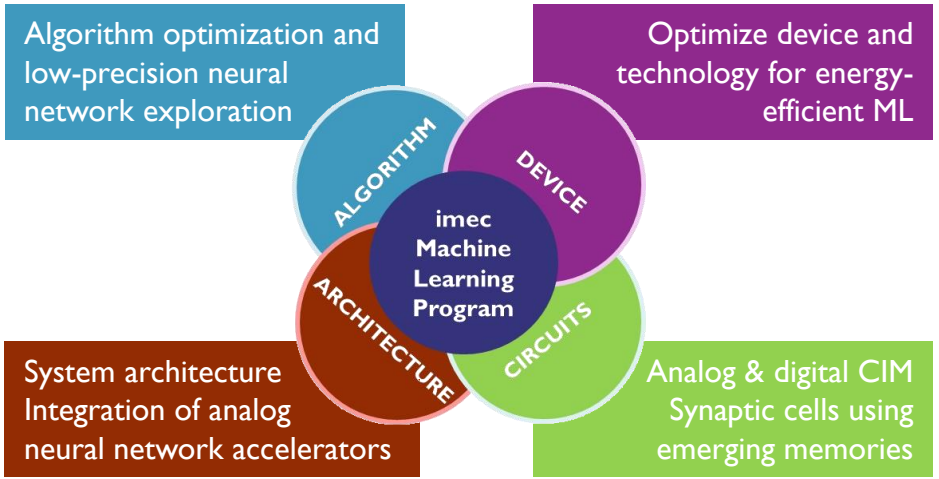
## ORDER OF MAGNITUDE MORE ENERGY-EFFICIENT

- Tune devices for analog MAC
  - Order of magnitude energy efficiency gain.
- Challenges:
  - Limited precision for weights
  - Stochastic behavior, variability, noise
  - Scale-up
  - Integration with digital functions and logic technologies



# PROGRAM OBJECTIVE

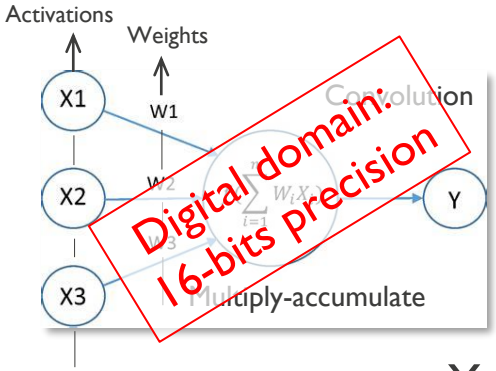
Combine algorithm and architecture optimization with semiconductor technology elements to enable energy-efficient implementation of DNNs and beyond

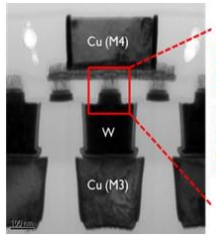


# SYSTEM DESIGN-TECHNOLOGY CO-OPTIMIZATION FOR ANALOG COMPUTE-IN-MEMORY

# ANALOG COMPUTE-IN-MEMORY ACCELERATORS FOR ML

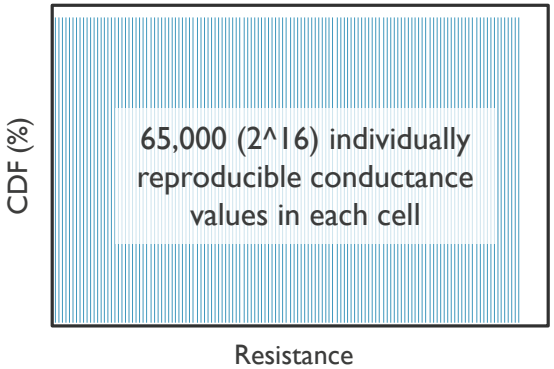
## CHALLENGES …

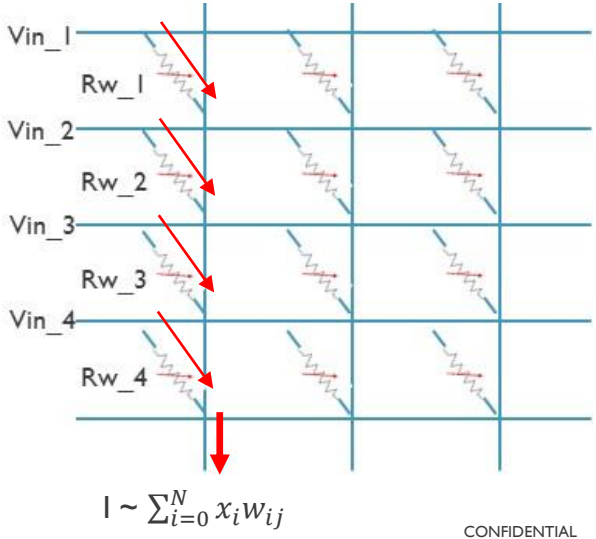



1. Memristor (programmable resistor) stores weight value $W_I$ as an analog quantity: conductance $Rw_I$
2. Activation $X_I$ applied as analog voltage $Vin_I$
3. Ohm's law: memristor cell current ~ $X_I.W_I$
4. Kirchoff's law: bit-line current ~ $\sum X_i.W_i$

You want



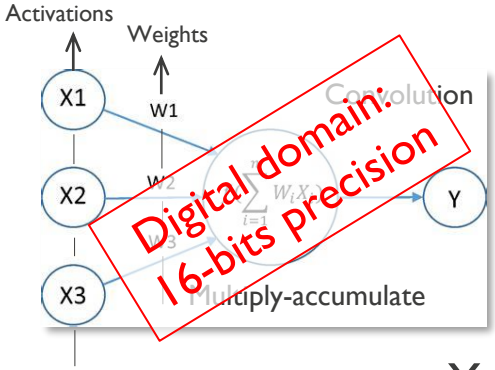




$I \sim \sum_{i=0}^{N} x_i w_{ij}$

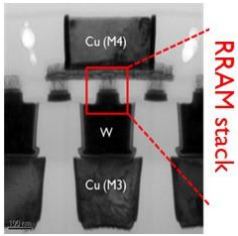
imec

# ANALOG COMPUTE-IN-MEMORY ACCELERATORS FOR ML

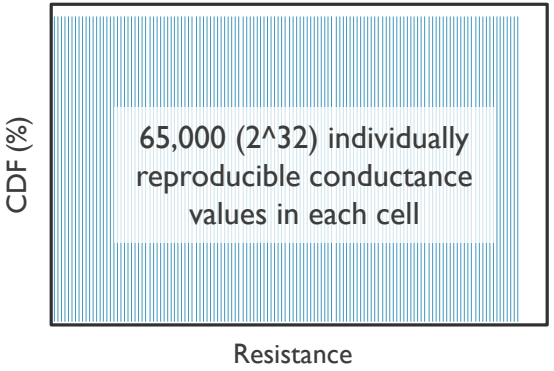
## CHALLENGES …



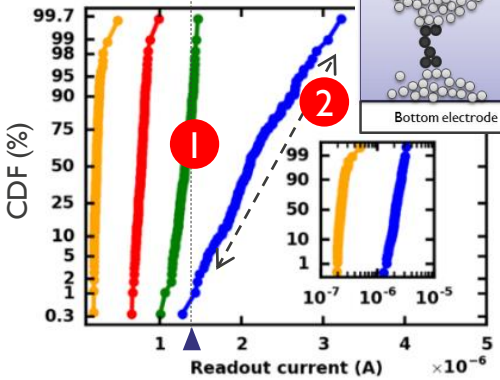

1. Memristor (programmable resistor) stores weight value $W_I$ as an analog quantity: conductance $Rw_I$
2. Activation $X_I$ applied as analog voltage $Vin_I$
3. Ohm's law: memristor cell current ~ $X_I.W_I$
4. Kirchoff's law: bit-line current ~ $\sum X_i.W_i$

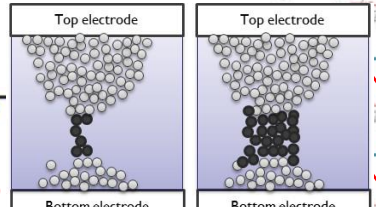



You want




You get

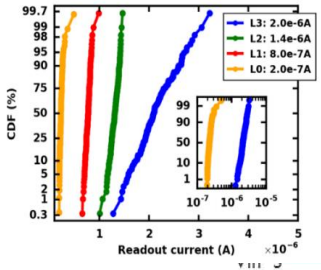


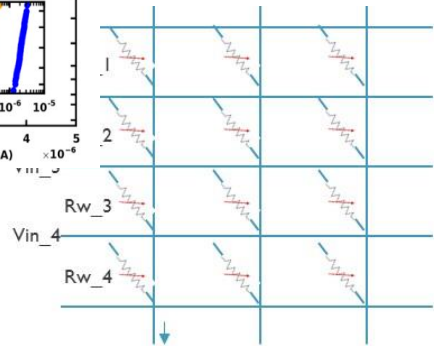



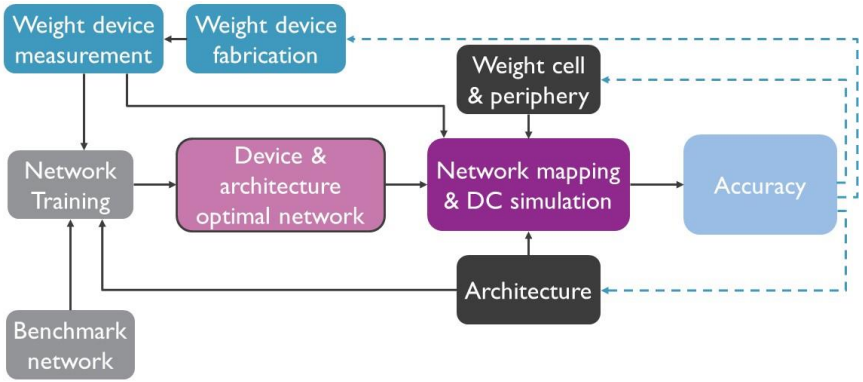
# HOW TO MAKE IT WORK?

## SYSTEM-TECHNOLOGY CO-OPTIMIZATION

- Deep neural networks
  - Perform inference with lower precision
  - Can tolerate some weight variation
- Neural network optimization
  - Train neural network with information of implementation limitations
  - Increase size of neural network to limit accuracy loss
- Architecture
  - Use multiple memory cells per weight for more precision
- Circuit and device
  - Tune devices for neural network operation

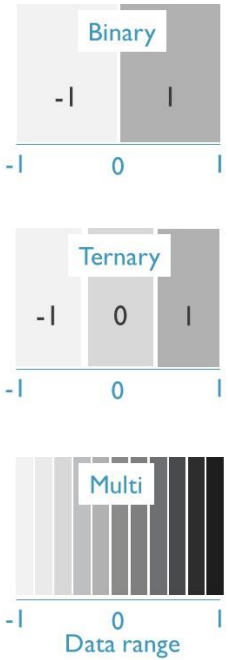


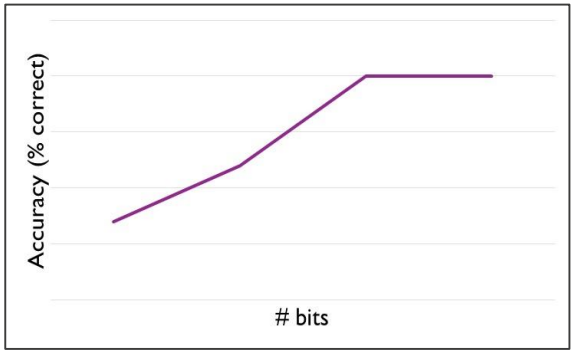
# NEURAL NETWORK OPTIMIZATION FOR ACiM

## QUANTIZATION BRINGS PRECISION IN RANGE OF ACiM




- Accuracy loss due to quantization can be mitigated
  - Modified training procedure
  - Modified network structure:
    - Higher #operations
    - Simpler operations
  - No accuracy loss for key benchmark networks using
    - 3-levels for weight
    - 5-levels for activations

→ In range of analog storage elements and analog MAC



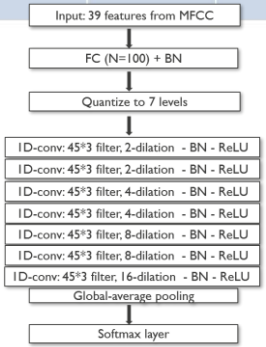
# NEURAL NETWORK OPTIMIZATION FOR ACiM

## TRAIN WITH ACiM IMPERFECTIONS IN MIND

- Up to 10% noise on weights and activations, and 50% LSB noise on MAC can be mitigated by training.

| Keyword spotting (KWS) Model | # param | Size | # Macs | #ACIM rows | #ACIM columns | Accuracy |
|---|---|---|---|---|---|---|
| LSTM (FP) | 800K | 3.2MB | 79.2M | 100-400 | 1600 | 93% |
| QLSTM | 800K | 200kB | 79.2M | 100-400 | 1600 | 93.2% |
| CNN (FP) | 50K | 200kB | 3.5M | 135 | 315 | 94.3% |
| QCNN | 50K | 12.5kB | 3.5M | 135 | 315 | 94.2% |
| ResNet-15 (Tang et al.2018) | 238K | 952kB | 894M | 405 | 585 | 95.8% |

- All quantization with 3-level weight, 7-level activation, 15-level accumulation and additive noise
- CNN: 1D dilated, 45 filters
- ResNet-15 is state-of-the-art for KWS; uses different preprocessing of data: filtering, more cepstral components, different window length and no input quantization layer.



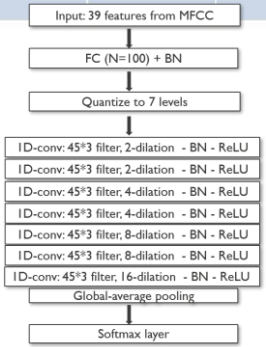
# NEURAL NETWORK OPTIMIZATION FOR ACiM

## TRAIN WITH ACiM IMPERFECTIONS IN MIND

- Up to 10% noise on weights and activations, and 50% LSB noise on MAC can be mitigated by training.

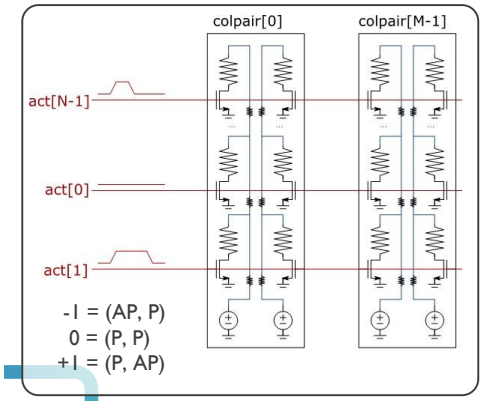
| Keyword spotting (KWS) Model | # param | Size | # Macs | #ACIM rows | #ACIM columns | Accuracy |
|---|---|---|---|---|---|---|
| LSTM (FP) | 800K | 3.2MB | 79.2M | 100-400 | 1600 | 93% |
| QLSTM | 800K | 200kB | 79.2M | 100-400 | 1600 | 93.2% |
| CNN (FP) | 50K | 200kB | 3.5M | 135 | 315 | 94.3% |
| QCNN | 50K | 12.5kB | 3.5M | 135 | 315 | 94.2% |
| ResNet-15 (Tang et al.2018) | 238K | 952kB | 894M | 405 | 585 | 95.8% |

- All quantization with 3-level weight, 7-level activation, 15-level accumulation and additive noise
- CNN: 1D dilated, 45 filters
- ResNet-15 is state-of-the-art for KWS; uses different preprocessing of data: filtering, more cepstral components, different window length and no input quantization layer.

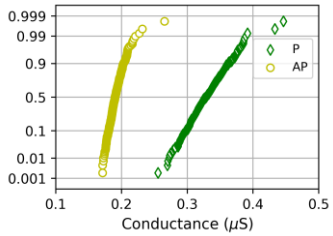



- Process variability can be accounted for during training

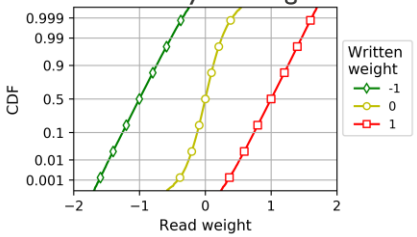
MRAM based synaptic cell with 3-level weight encoding



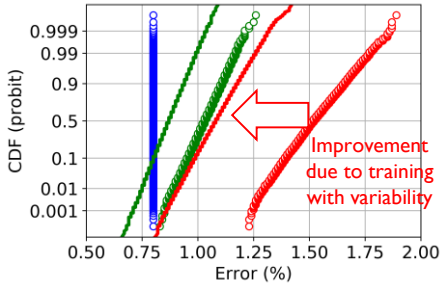

MRAM measured data




Variability on weight




Application accuracy

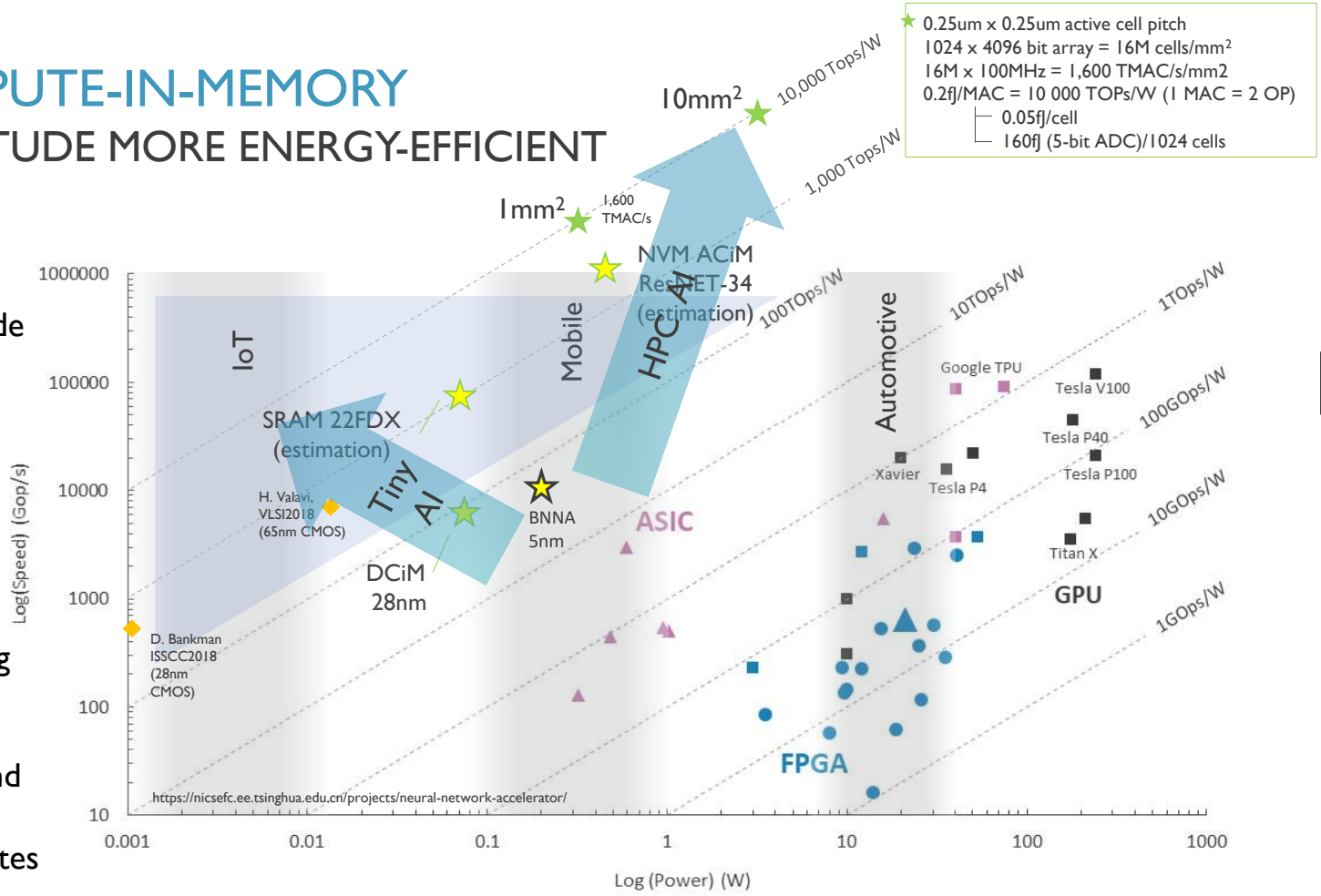



- Baseline
- With σ of 10%
- With σ of 20%

# ANALOG COMPUTE-IN-MEMORY

## ORDER OF MAGNITUDE MORE ENERGY-EFFICIENT

- Tune devices for analog MAC
  - Order of magnitude energy efficiency gain.
- Challenges
  - Limited precision
  - Variability, noise, stochastic behavior
  - Algorithm mapping flexibility
  - Integration with digital functions and logic technologies
  - Upscale to Gigabytes



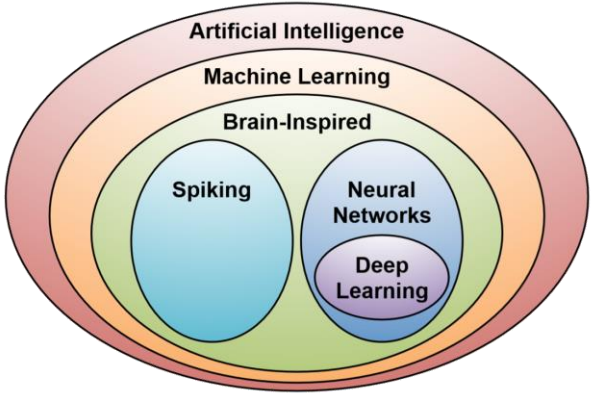
# CONCLUSIONS

# CONCLUSIONS

- CNN and LSTM are the work horses for deep learning
  - Huge computational complexity (Mult-Acc) and storage requirements (Weights)
  - High-performance compute architectures dominant platform for development and training phase
    - GPU: 1 TOP/s/W, >100W
- Quantization of neural networks trade-off complexity vs. accuracy
  - Key for low-energy digital neural network inference accelerators
    - Binary neural network accelerators can approximate 100 TOP/s/W
  - Necessity for mixed-signal and analog compute-in-memory implementations
    - Inherent variability, stochastic behavior, noise limits precision of operations
    - Promise large gain in energy efficiency to >1,000 TOP/s/W
    - 3-level weights and 5b-activations provide required accuracy and are in range of Analog Compute-in-Memory
    - Several cell architectures (MRAM, PCM, DRAM, SRAM-based) can enable ACiM accelerators
    - Training procedures and network topologies are adapted to limit/eliminate accuracy loss

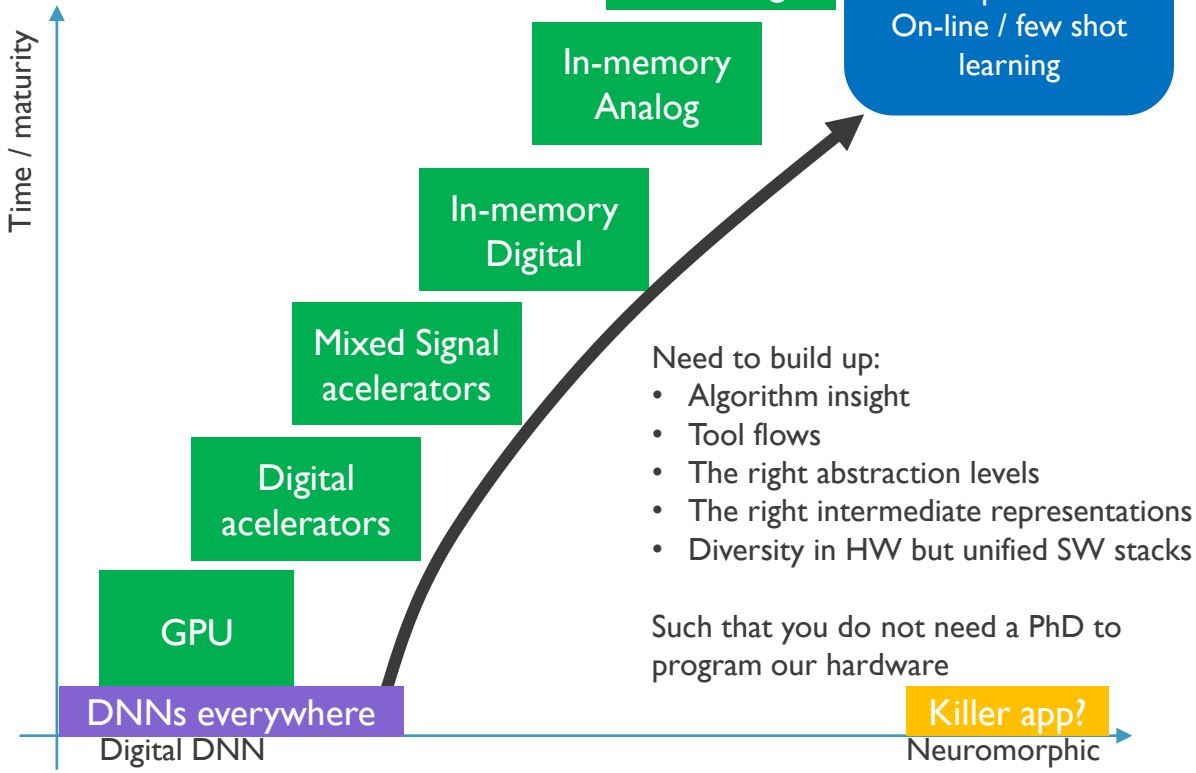
# INTELLIGENT HARDWARE?

## FROM DEEP NEURAL NETWORKS TO NEUROMORPHIC




V. Sze, et al. "Efficient processing of deep neural networks: A Tutorial and Survey", Proc. of the IEEE, Vol. 105, No. 12, Dec. 2017 [ https://arxiv.org/abs/1703.09039 ]

The lines are blurring!




Need to build up:

- Algorithm insight
- Tool flows
- The right abstraction levels
- The right intermediate representations
- Diversity in HW but unified SW stacks

Such that you do not need a PhD to program our hardware

imec


embracing a better life