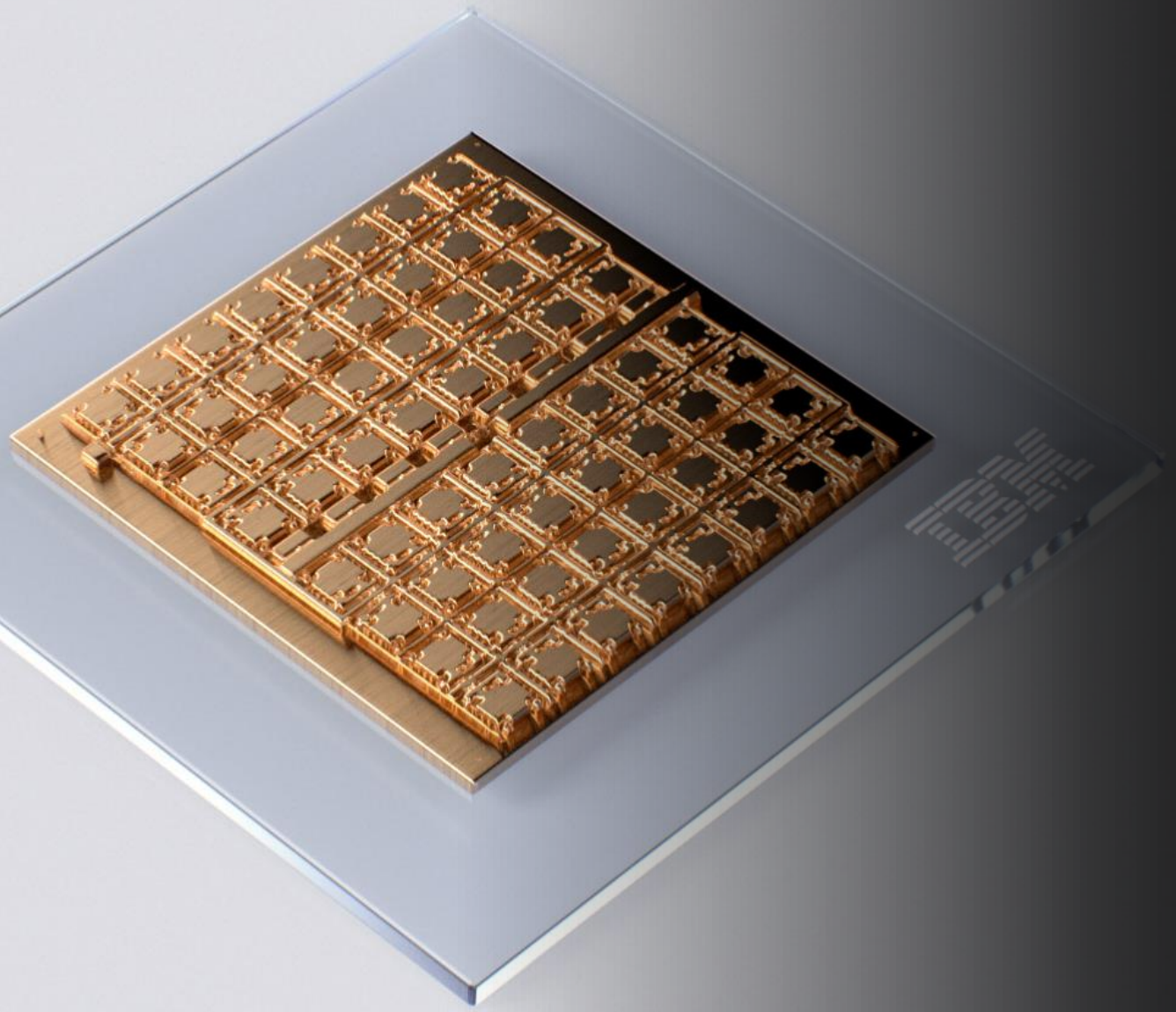


Deep learning acceleration: A killer application for in-memory computing?

Abu Sebastian
Distinguished Research Staff Member
IBM Research - Zurich

MiM webinar, TU Vienna, June 21st 2021

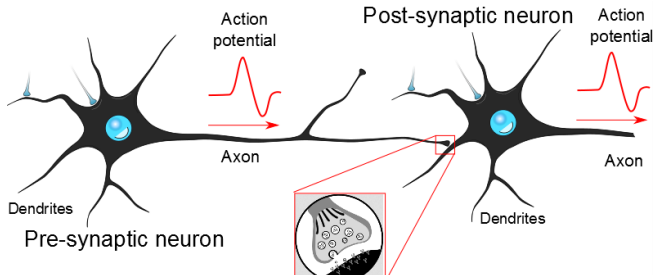
Outline

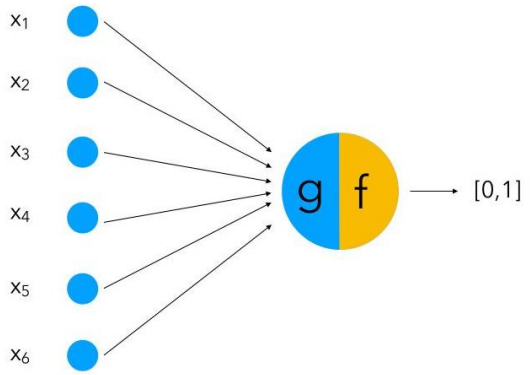
- **Introduction**
  - Deep learning
  - In-memory computing
- **Deep learning based on computational phase-change memory**
  - Phase-change memory and synaptic emulation
  - DL inference and training
  - In-memory compute core
  - Device-level innovations
- **Applications beyond conventional DL**
  - DNN + “something”
  - Spiking deep neural networks

Outline

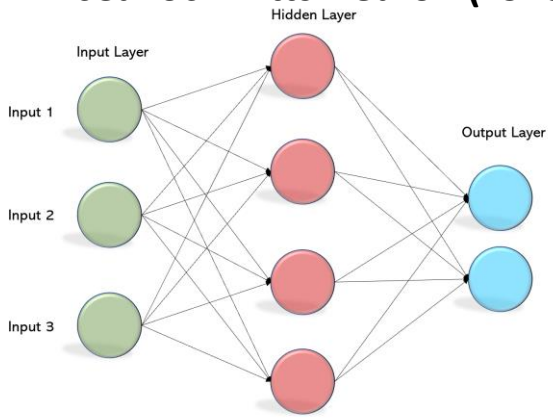
- **Introduction**
  - Deep learning
  - In-memory computing
- **Deep learning based on computational phase-change memory**
  - Phase-change memory and synaptic emulation
  - DL inference and training
  - In-memory compute core
  - Device-level innovations
- **Applications beyond conventional DL**
  - DNN + “something”
  - Spiking deep neural networks

Deep (artificial) neural networks



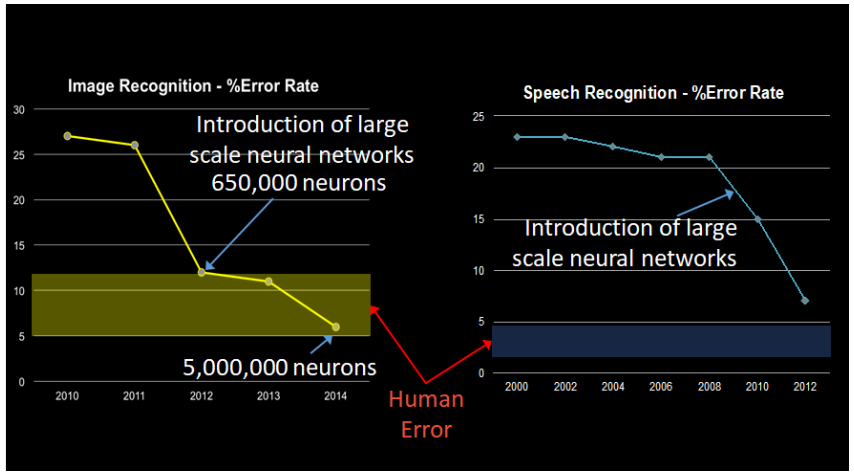


McCulloch-Pitts Neuron (1943)



Multi-layer perceptrons

DNNs meet the IT revolution



Mainstay of the AI portfolio of almost all IT companies

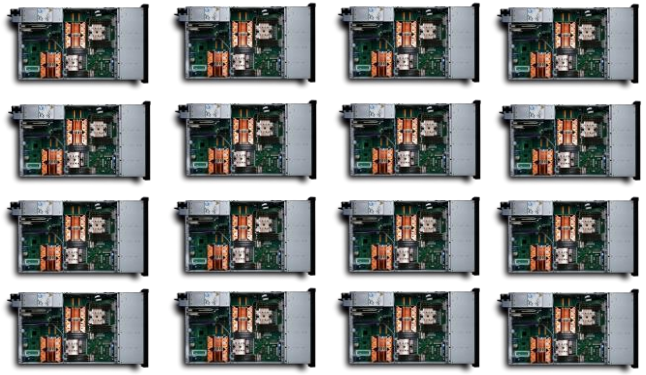
- Translation
- Search ranking
- News feed
- Face recognition
- Content understanding

DNN's Computational Efficiency Problem

Training Image recognition model
Dataset: ImageNet-22K
Network: ResNet-101

4 GPUs
16 days
~385 kWh

256 GPUs
7 hours
~450kWh



For reference: 1 model training run is
Approx. 2 weeks of home energy consumption
https://arxiv.org/abs/1708.02188

- Deep learning is computationally intensive
- Time consuming even with high-performance computing resources
- Power consumption prohibitive for applicability in domains such as IoT

Key driver for innovations in computing systems

General purpose computing systems

STORAGE
(e.g. Flash, HDD)
(nonvolatile, slow)

STORAGE-CLASS MEMORY

MEMORY (e.g. DRAM)
(volatile, fast)

NEAR MEMORY COMPUTING

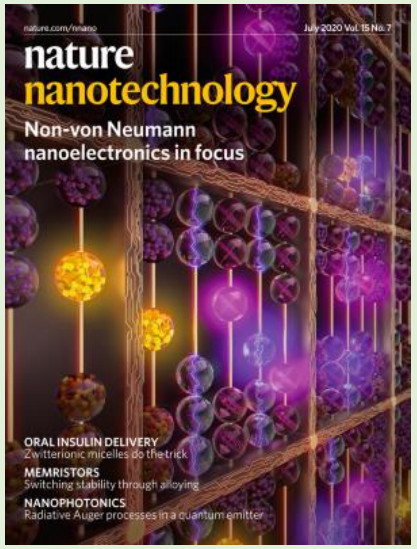
CPU

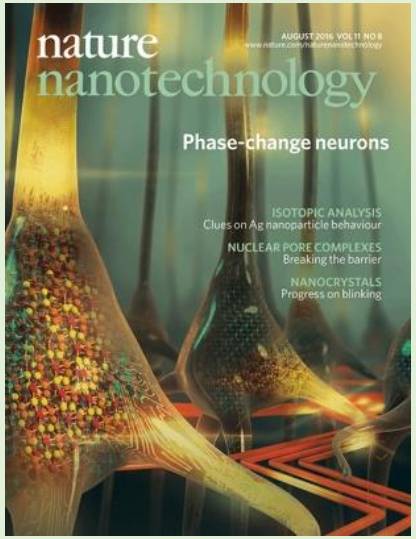
Conventional digital accelerators

GPUs, TPUs,
Graphcore, Inferentia,
Kunlun, Hanguang etc.

High-bandwidth memory

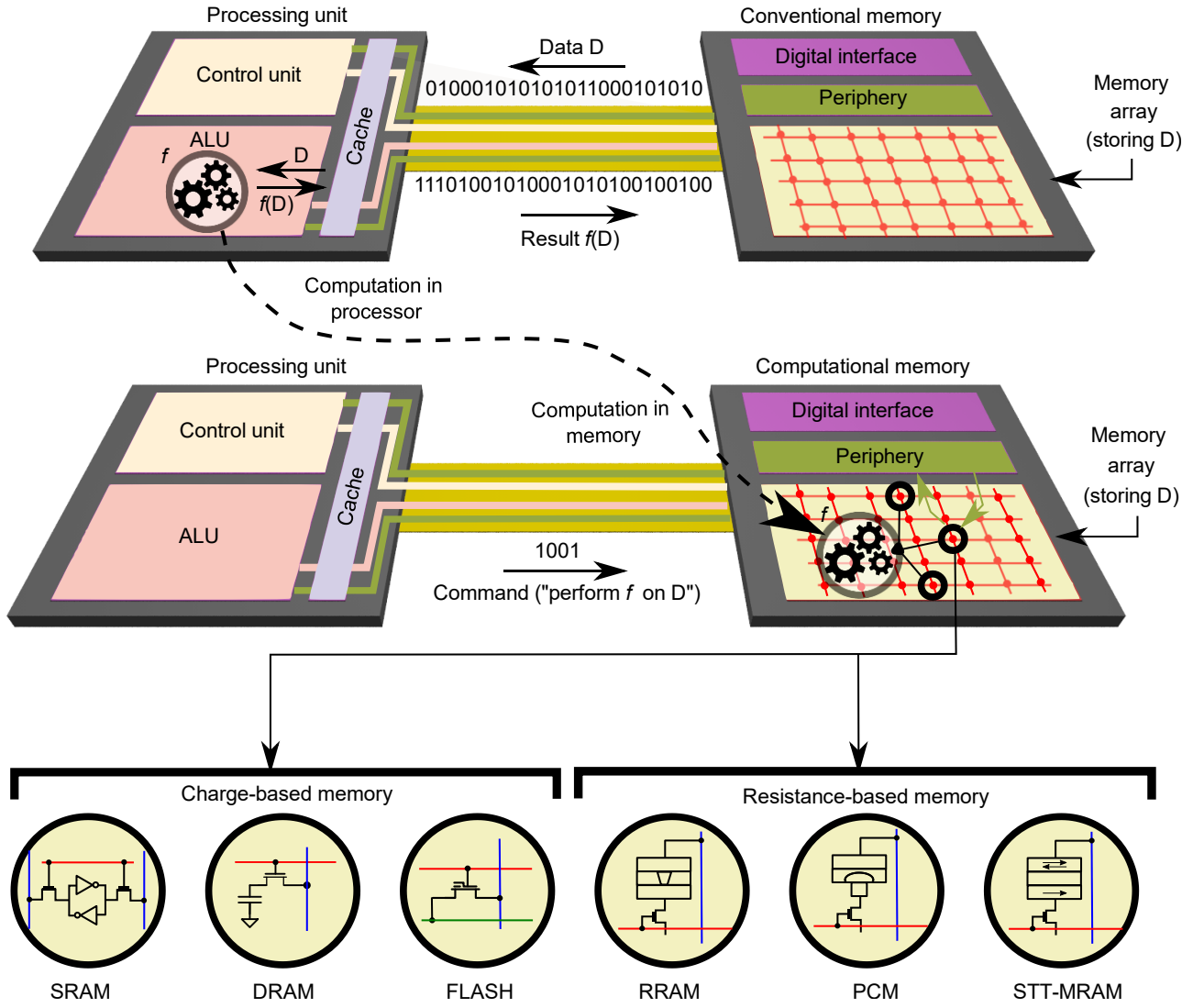
Non-von Neumann accelerators with post-CMOS devices





In-memory computing

- Perform “certain” computational tasks **in place in memory**
- Achieved by exploiting **the physical attributes of memory devices**
- Can we viewed as a sub-category of processing in memory (PIM) or compute in memory (CIM)
- At **no point during computation**, the memory content is read back and **processed at the granularity of a single memory element**



nature nanotechnology
FOCUS | REVIEW ARTICLE
https://doi.org/10.1038/s41565-020-0655-z
Check for updates

Memory devices and applications for in-memory computing

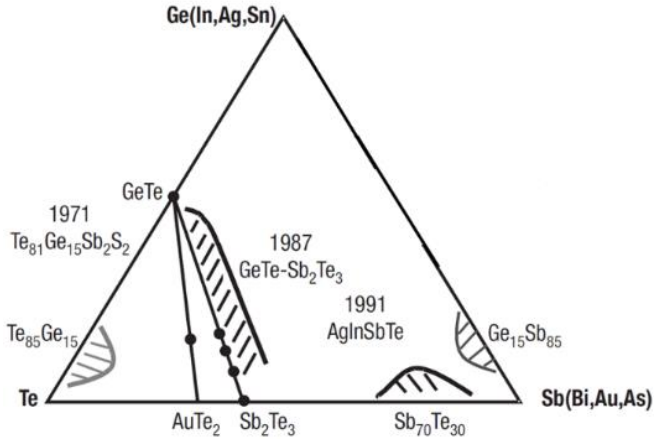
Abu Sebastian, Manuel Le Gallo, Riduan Khaddam-Aljameh and Evangelos Eleftheriou

Outline

- Introduction
  - Deep learning
  - In-memory computing
- **Deep learning based on computational phase-change memory**
  - Phase-change memory and synaptic emulation
  - DL inference and training
  - In-memory compute core
  - Device-level innovations
- Applications beyond conventional DL
  - DNN + “something”
  - Spiking deep neural networks

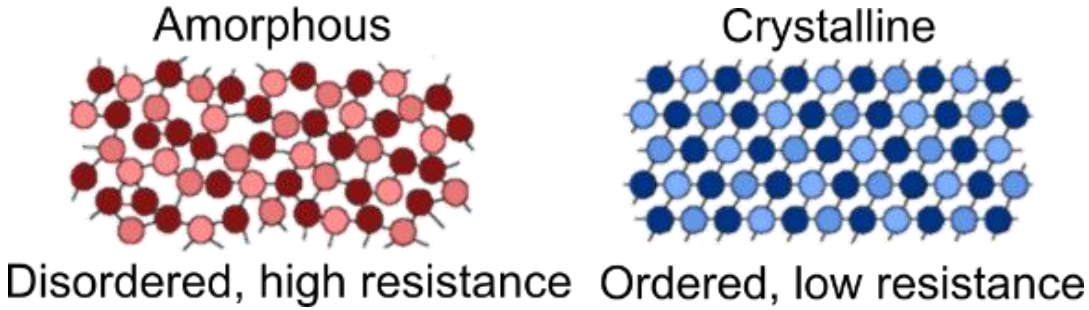
Phase-change memory

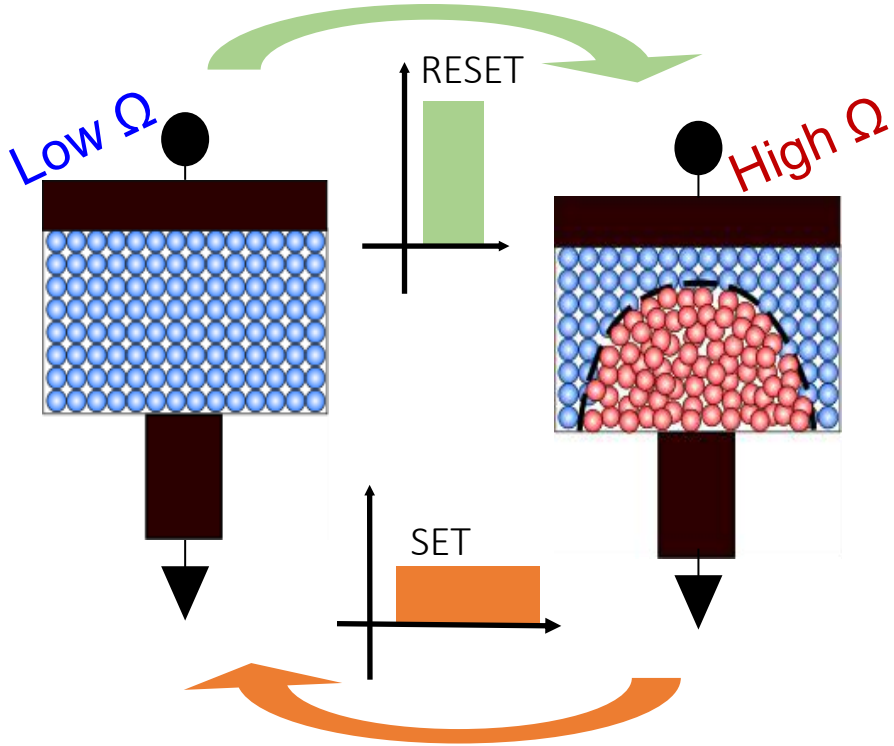
Commonly used phase change materials



Wuttig & Yamada, Nature Materials (2007)
Burr et al., JETCAS (2016)

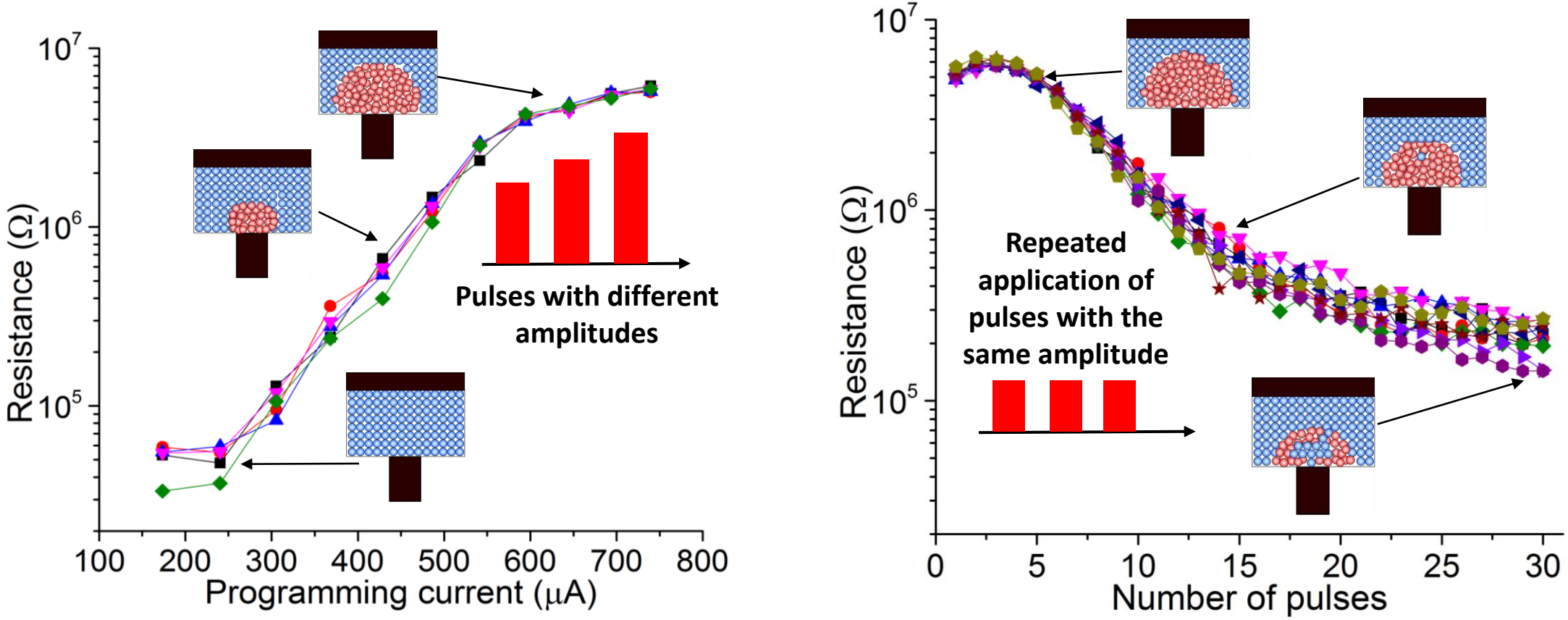
- A nanometric volume of phase change material between two electrodes
- "WRITE" Process
  - By applying a voltage pulse the material can be changed from crystalline phase (SET) to amorphous phase (RESET)
- "READ" process
  - Low-field electrical resistance





Le Gallo & Sebastian, An overview of phase-change memory device physics, J. Phys. D: Appl. Phys.

Analog storage and accumulation behavior



Sebastian et al., J. Appl. Phys. (2018)

- Can achieve a continuum of conductance states
- A non-volatile integrator of pulses

Phase-change synapses

Synaptic efficacy (Inference)

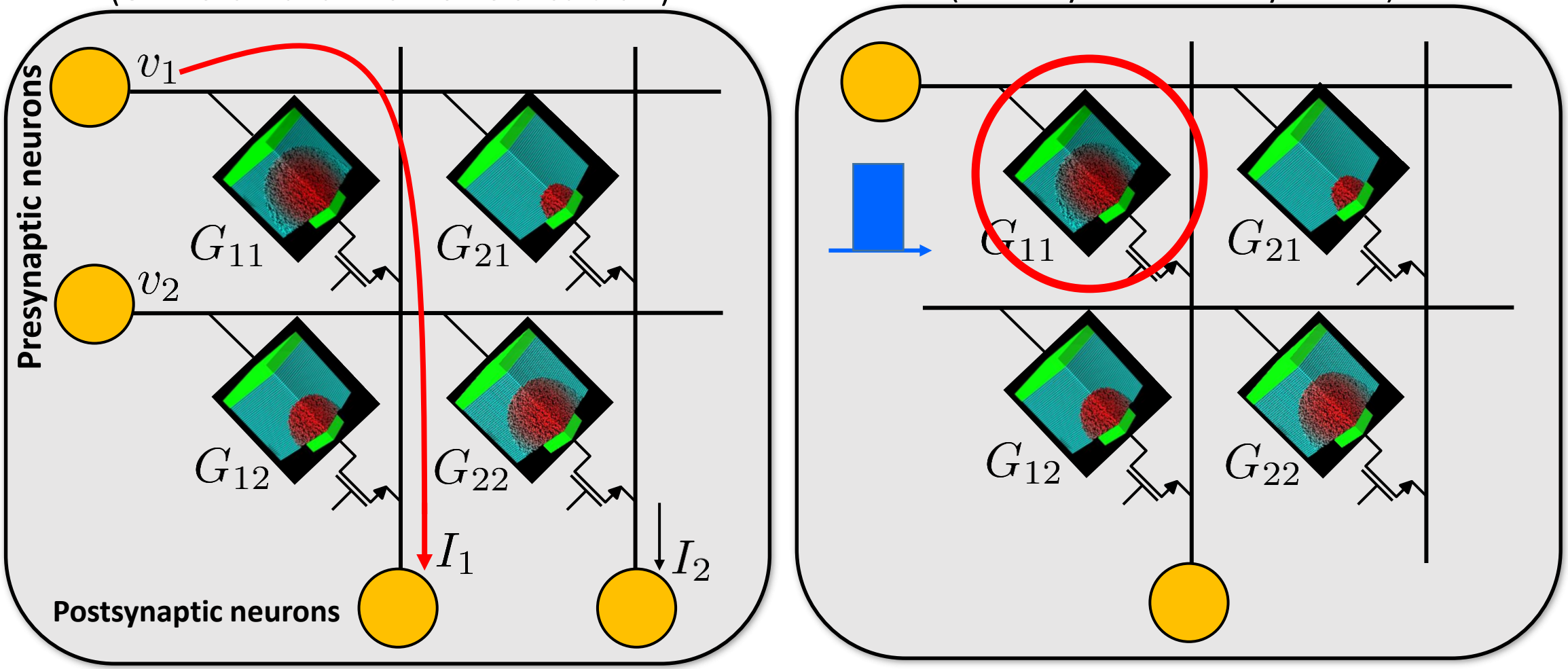
Analog Storage

(Ohm's law and Kirchhoff's circuit law)

Synaptic plasticity (Training)

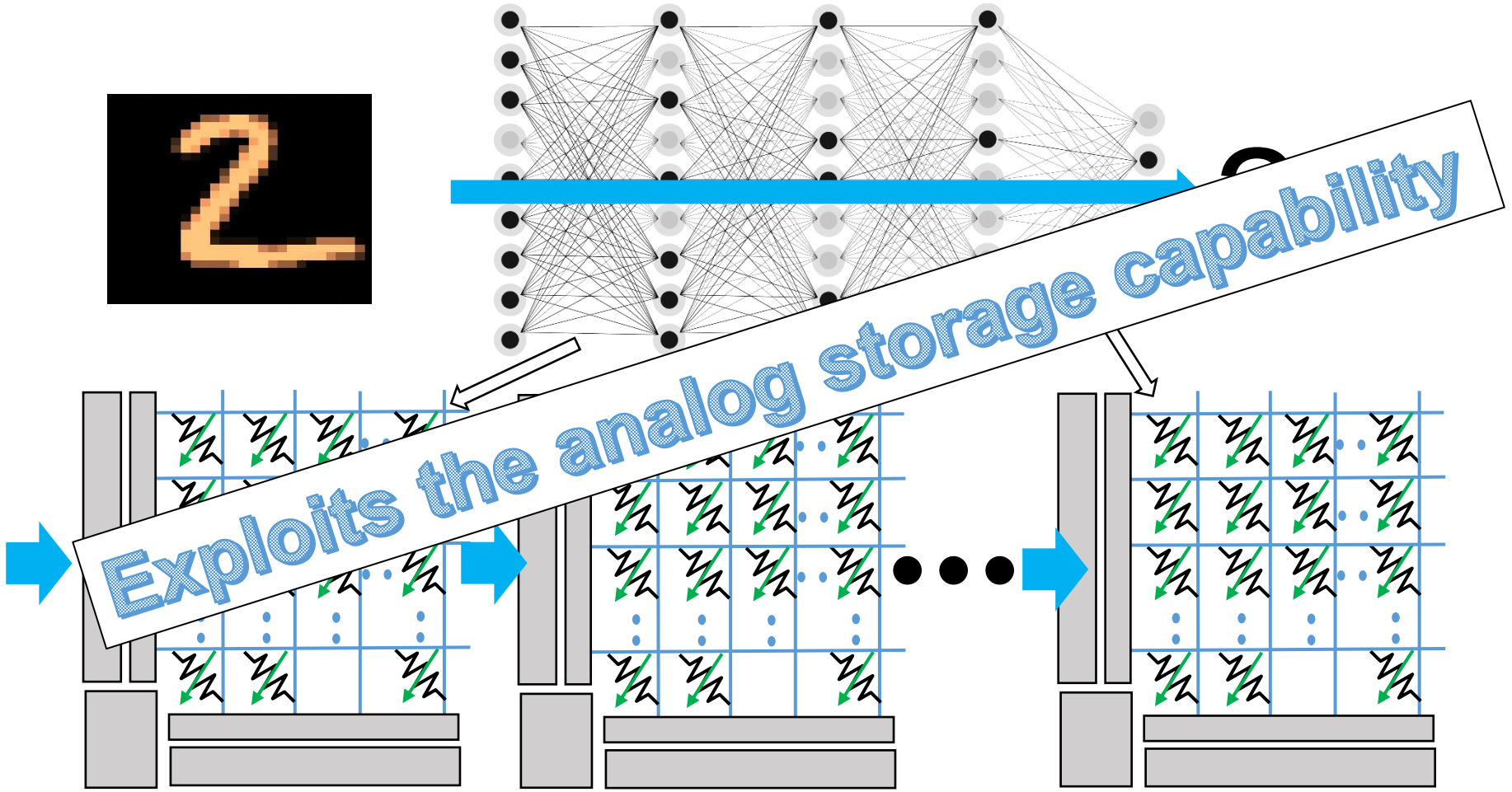
Accumulation behavior

(PCM crystallization dynamics)



Sebastian et al., "Brain-inspired computing using phase-change memory devices", J. Appl. Phys. (2018)

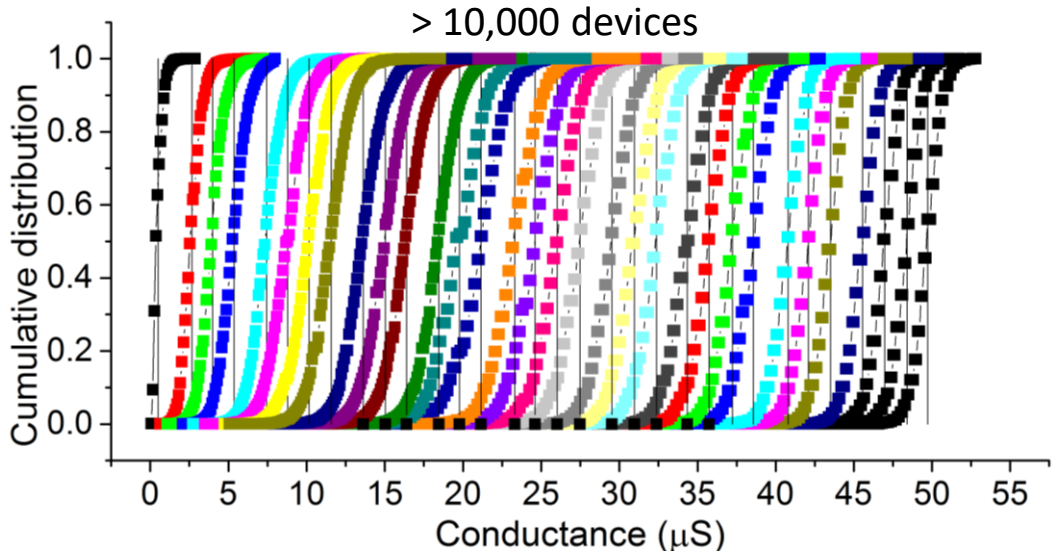
DNN inference with in-memory computing



Sebastian et al., VLSI (2019), Tsai et al., J. Phys. D: Appl. Phys. (2018)

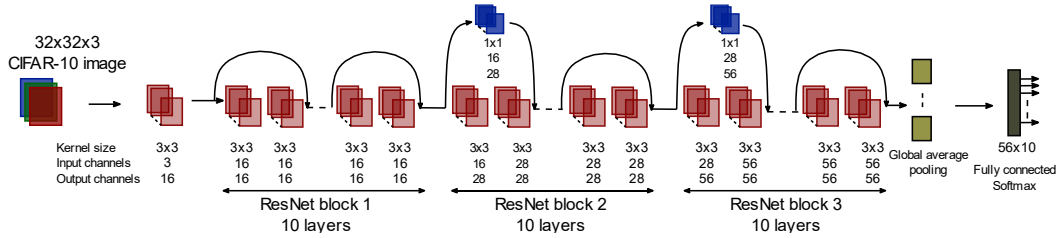
The trained synaptic weights are mapped to an array of computational memory cores performing matrix vector multiply operations corresponding to each layer

DNN inference with in-memory computing

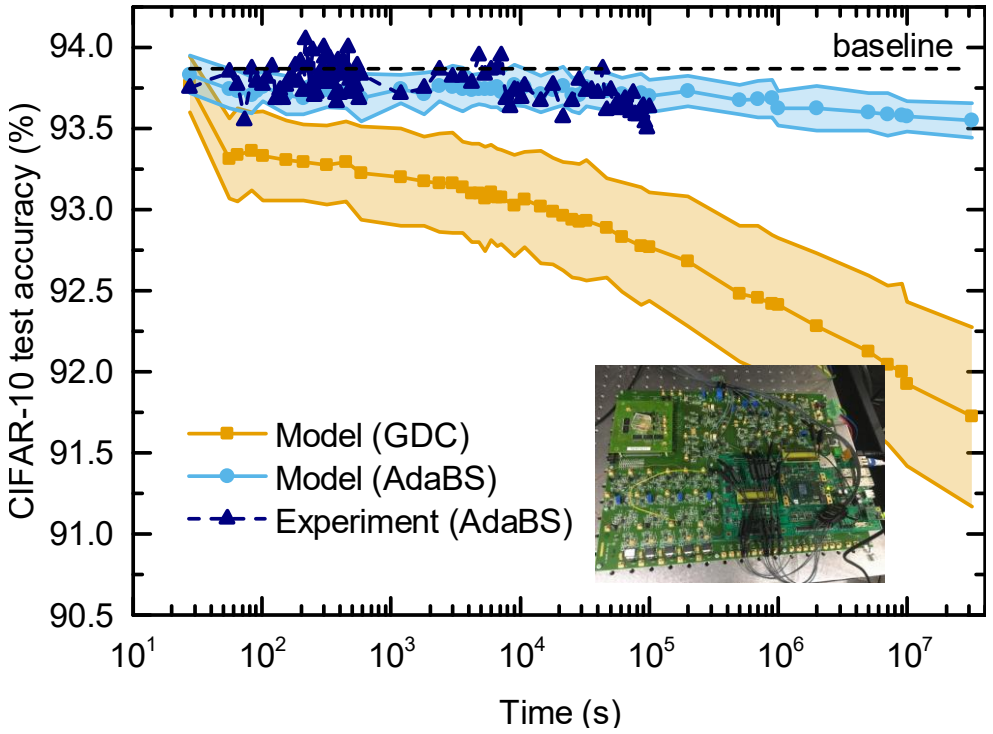


- Key challenge is the imprecision arising from conductance fluctuation, drift etc.
- A custom "**additive noise training**" procedure is essential to overcome this
- Experimental demonstration using PCM devices fabricated in 90nm CMOS technology

ResNet-32 on CIFAR-10

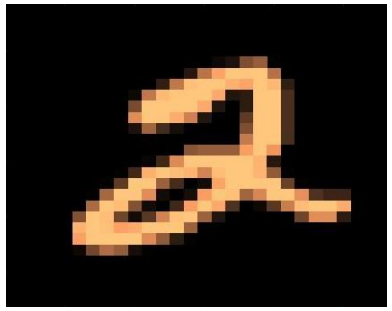


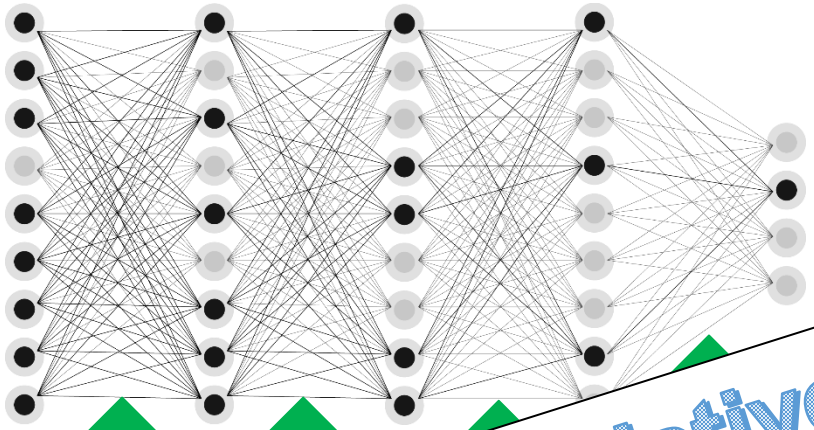
723,444 PCM devices (1T1R)



Joshi et al., "Accurate deep neural network inference using computational phase-change memory", Nature Comm. (2020)

DNN training with in-memory computing

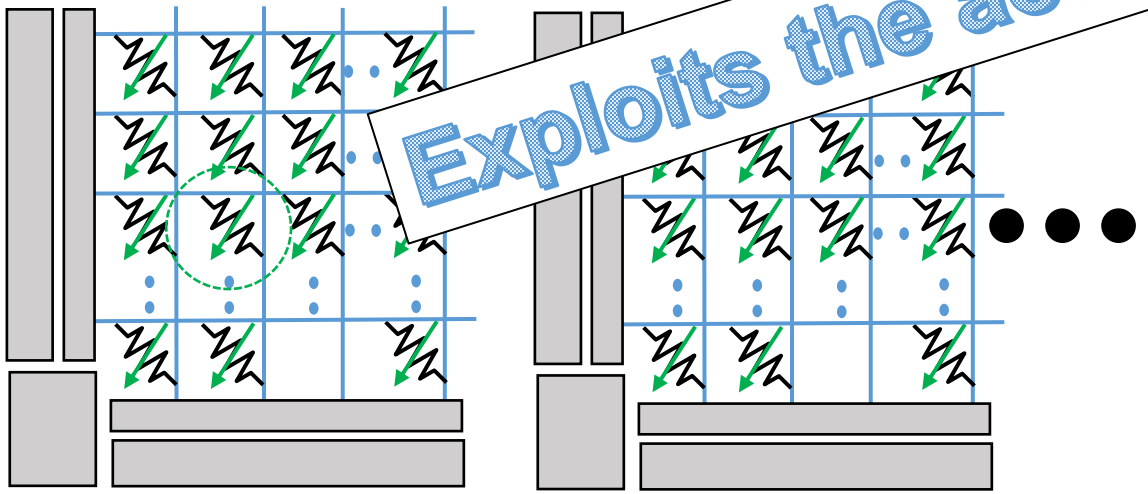


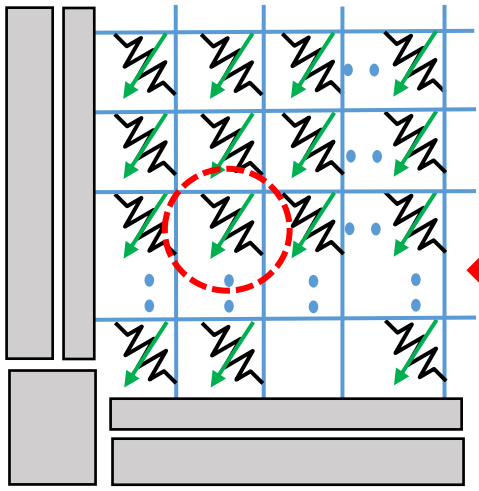


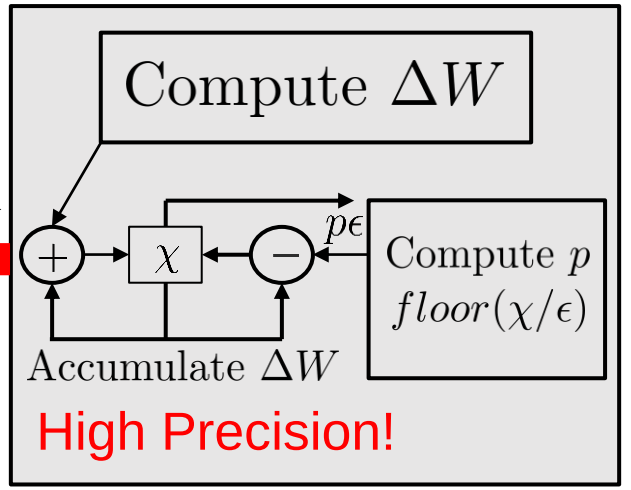
Exploits the accumulative behavior

weights

Apply acc. pulses sporadically & blindly!



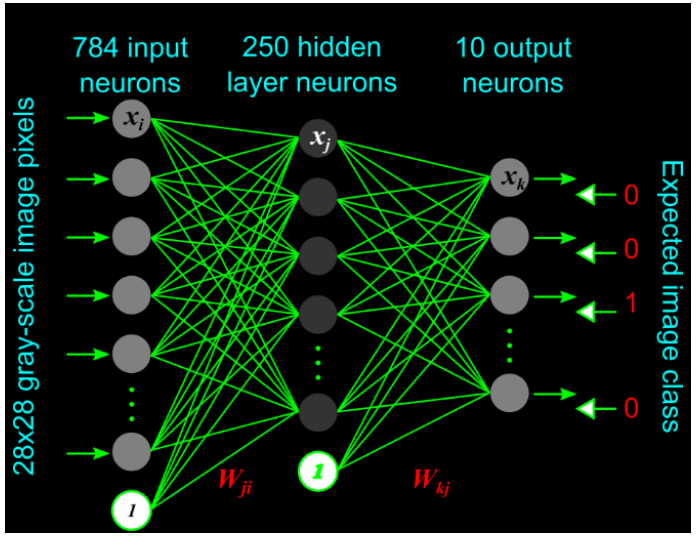


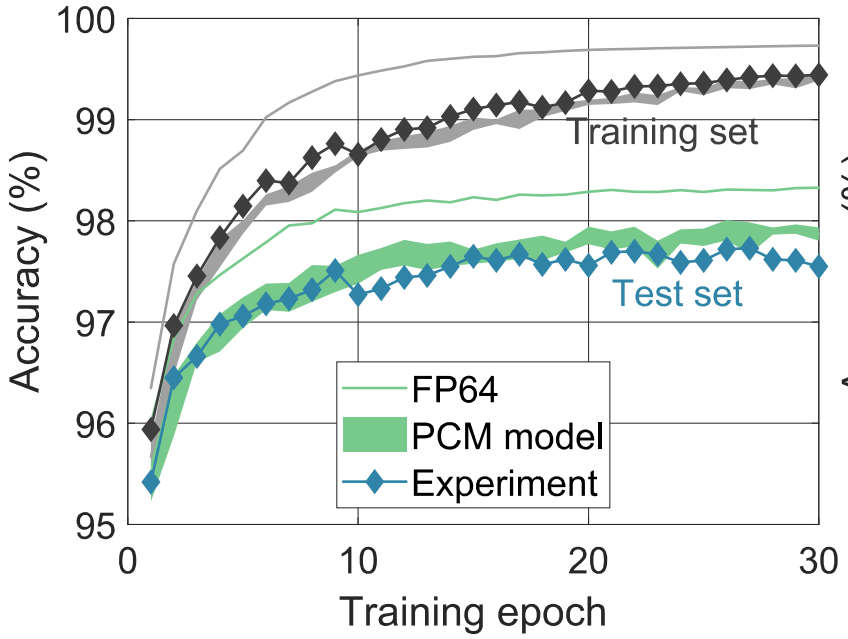


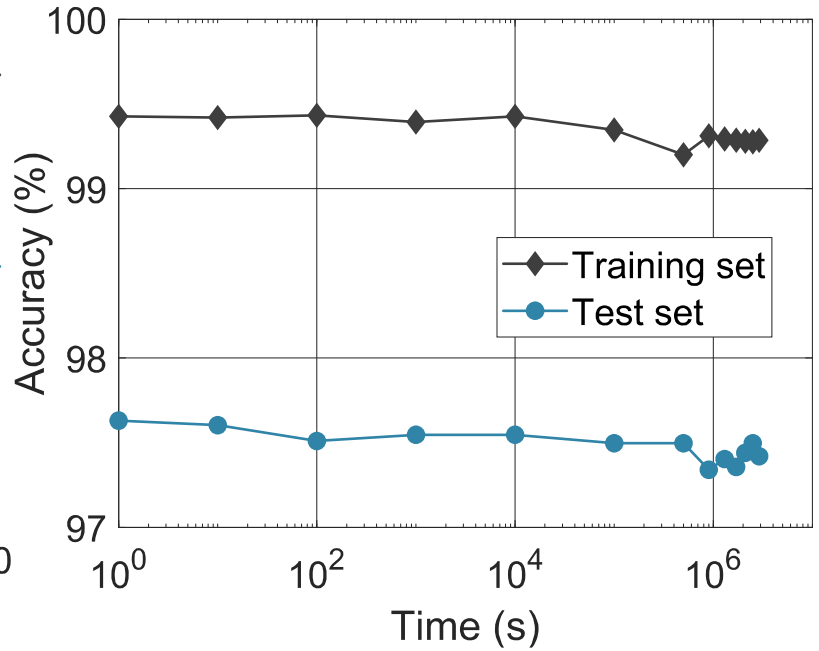
Nandakumar et al., ArXiv, 2017

Sebastian et al., VLSI, 2019

DNN training with in-memory computing

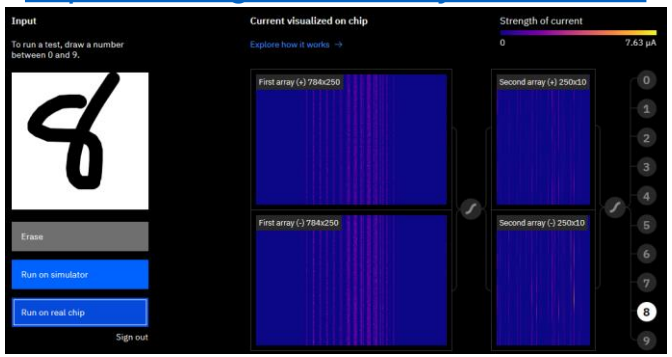






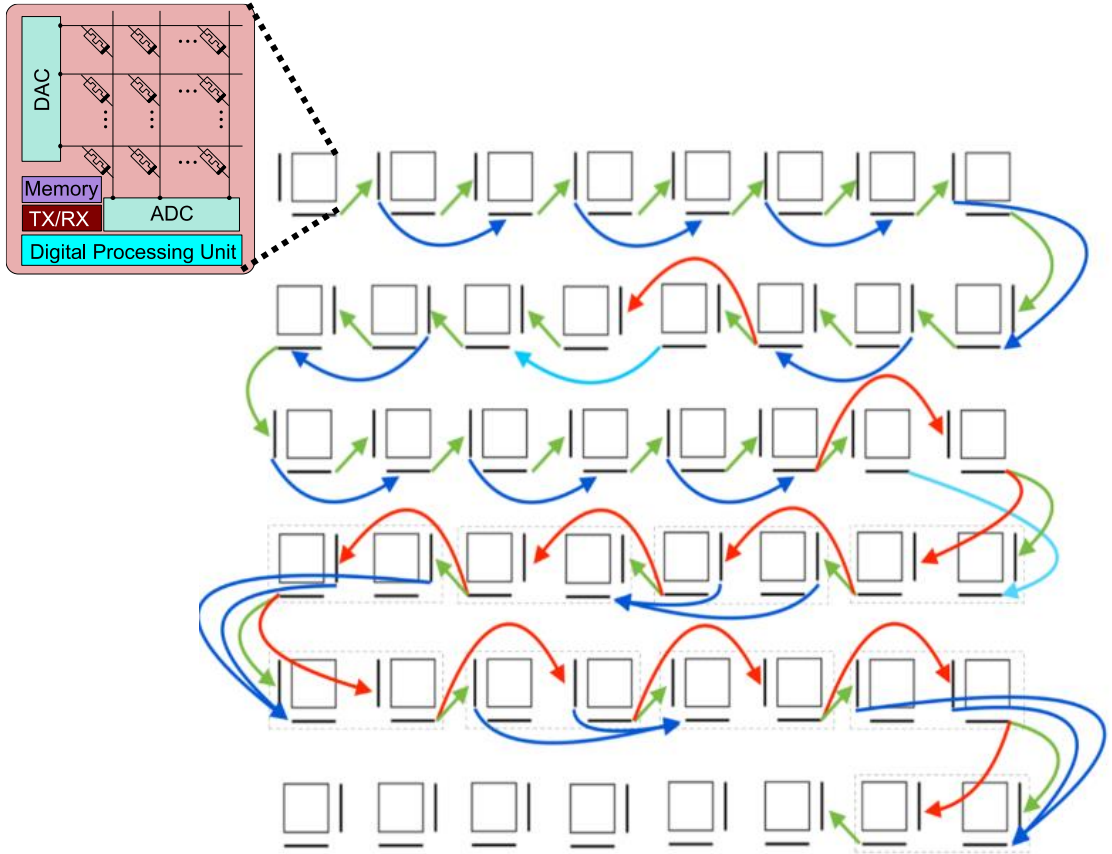
- Each synaptic weight mapped to two PCM devices (~400,000 PCM devices)
- Comparable test accuracy as FP32 training
- Negligible accuracy drop during inference after training

https://analog-ai-demo.mybluemix.net

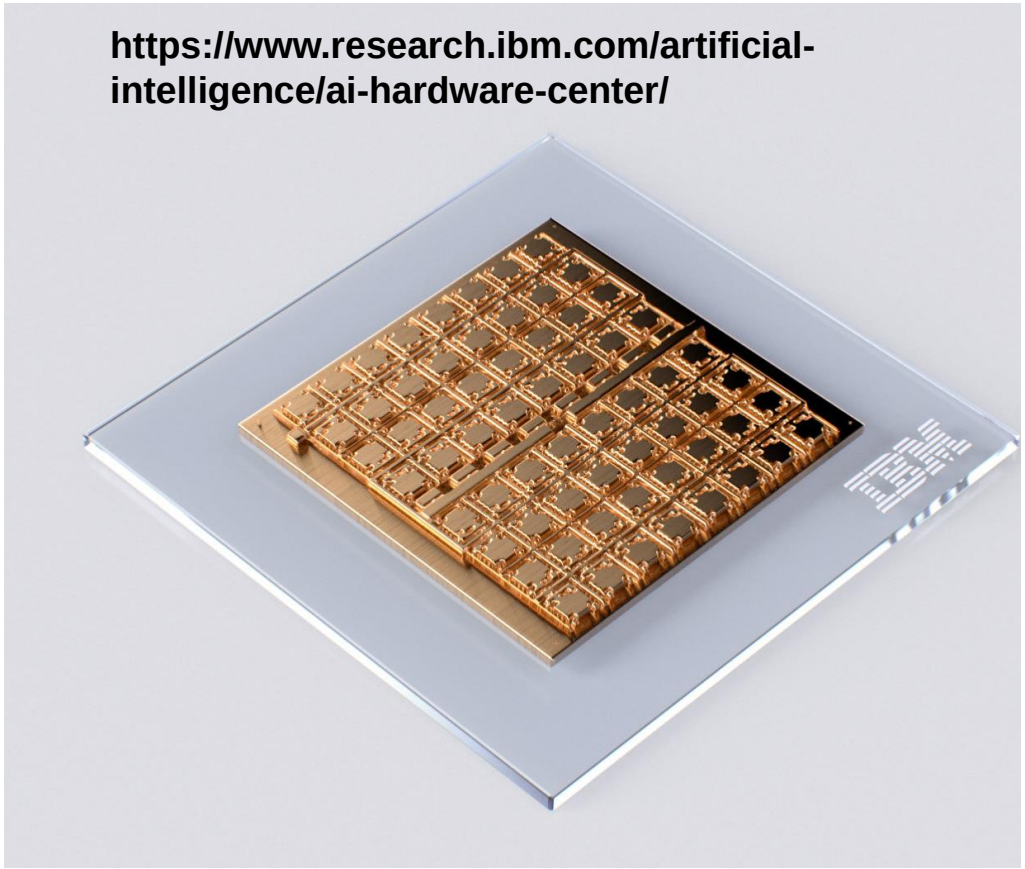


Nandakumar et al., "Mixed-precision deep learning based on computational memory", Front. Neuroscience (2020)

Architecture of an IMC-based accelerator



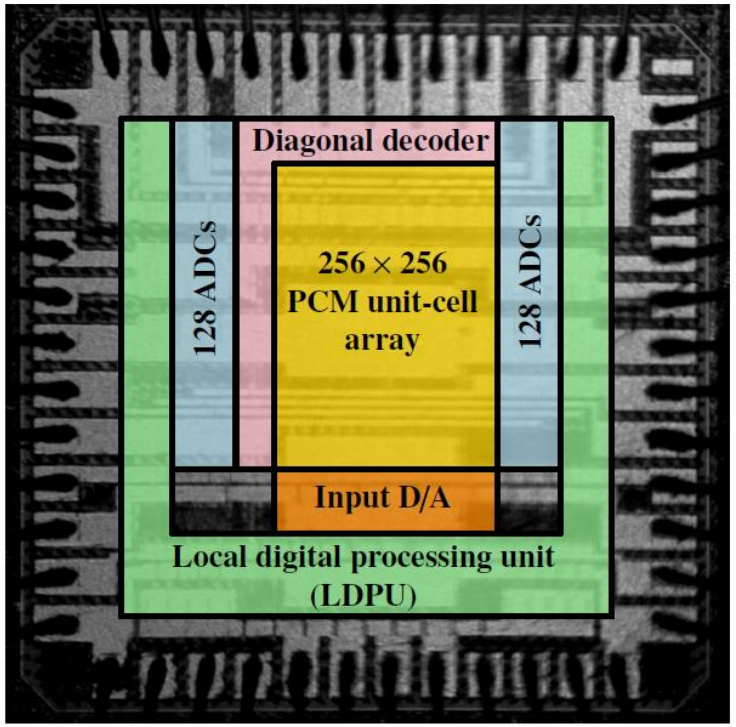
https://www.research.ibm.com/artificial-intelligence/ai-hardware-center/

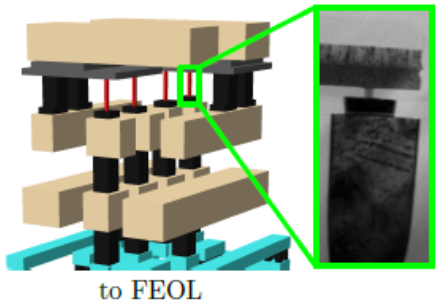


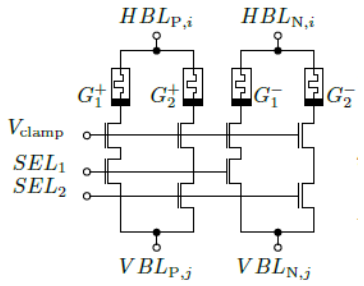
- Mixed-signal in-memory compute cores and system integration

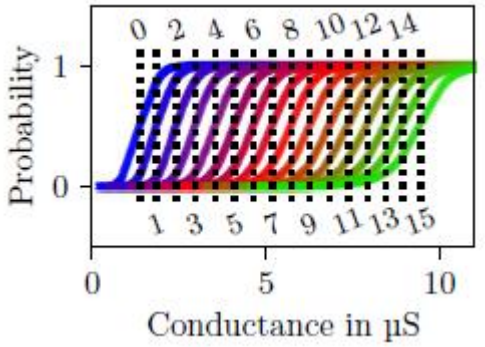
Eleftheriou et al., IBM JRD (2019),
Dazzi et al., MLSys Workshop @NeurIPS (2019)
Boybat et al., Nature Comm. (2018)
Khaddam-Aljameh et al., Proc. VLSI (2021)

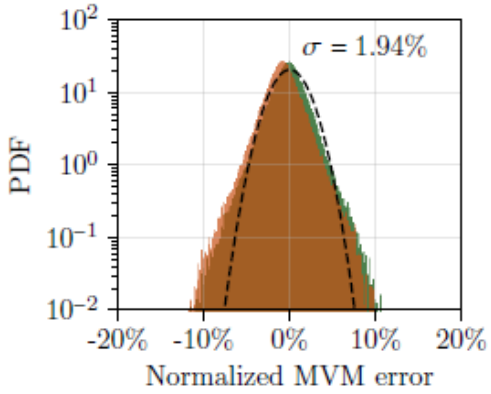
Hermes: IMC compute core in 14nm CMOS technology







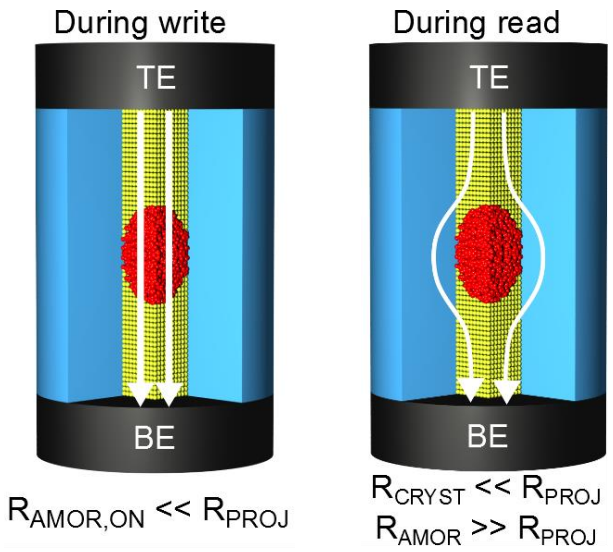


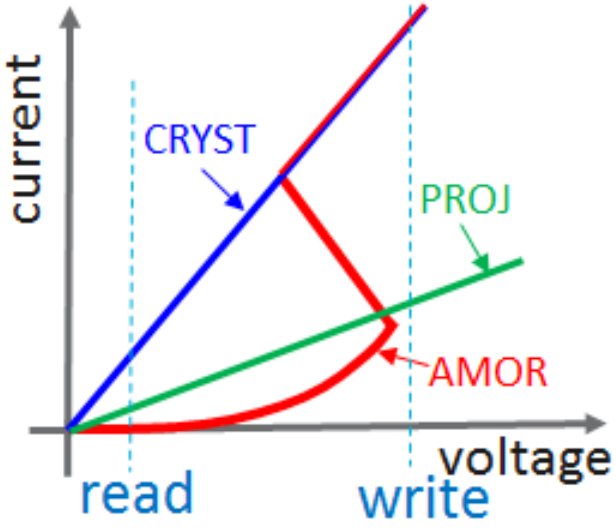


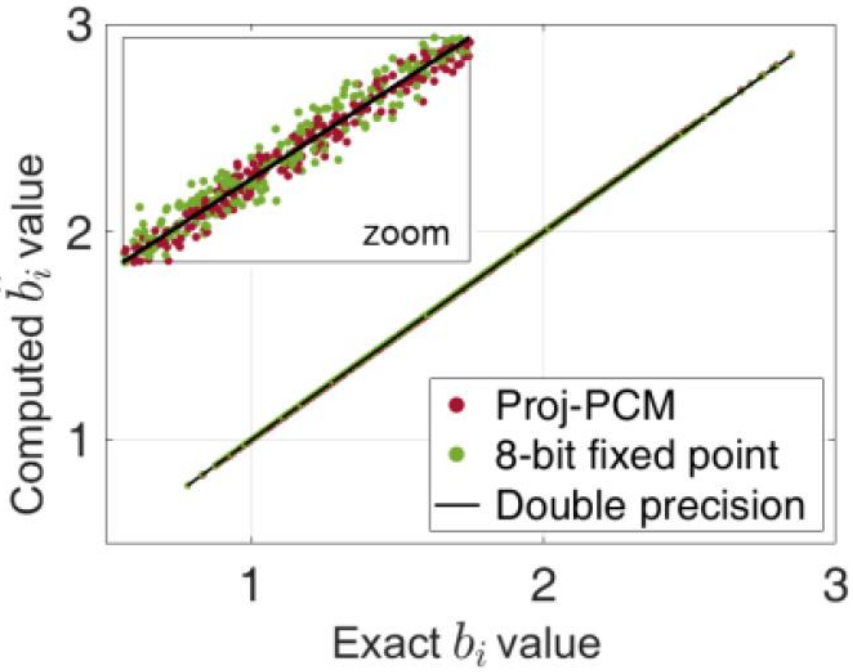
- 256x256 array of 8T4R unit cells
- PCM devices integrated in the back-end of 14nm CMOS chip
- Compact current controlled oscillator-based ADCs
- Local digital processing
- MVM performance: 10.5 TOPS/W and 16.5 TOPS/W/sq.mm.

Khaddam-Aljameh et al., Proc. VLSI (2021) (highlight paper)

Towards higher precision: Projected memory



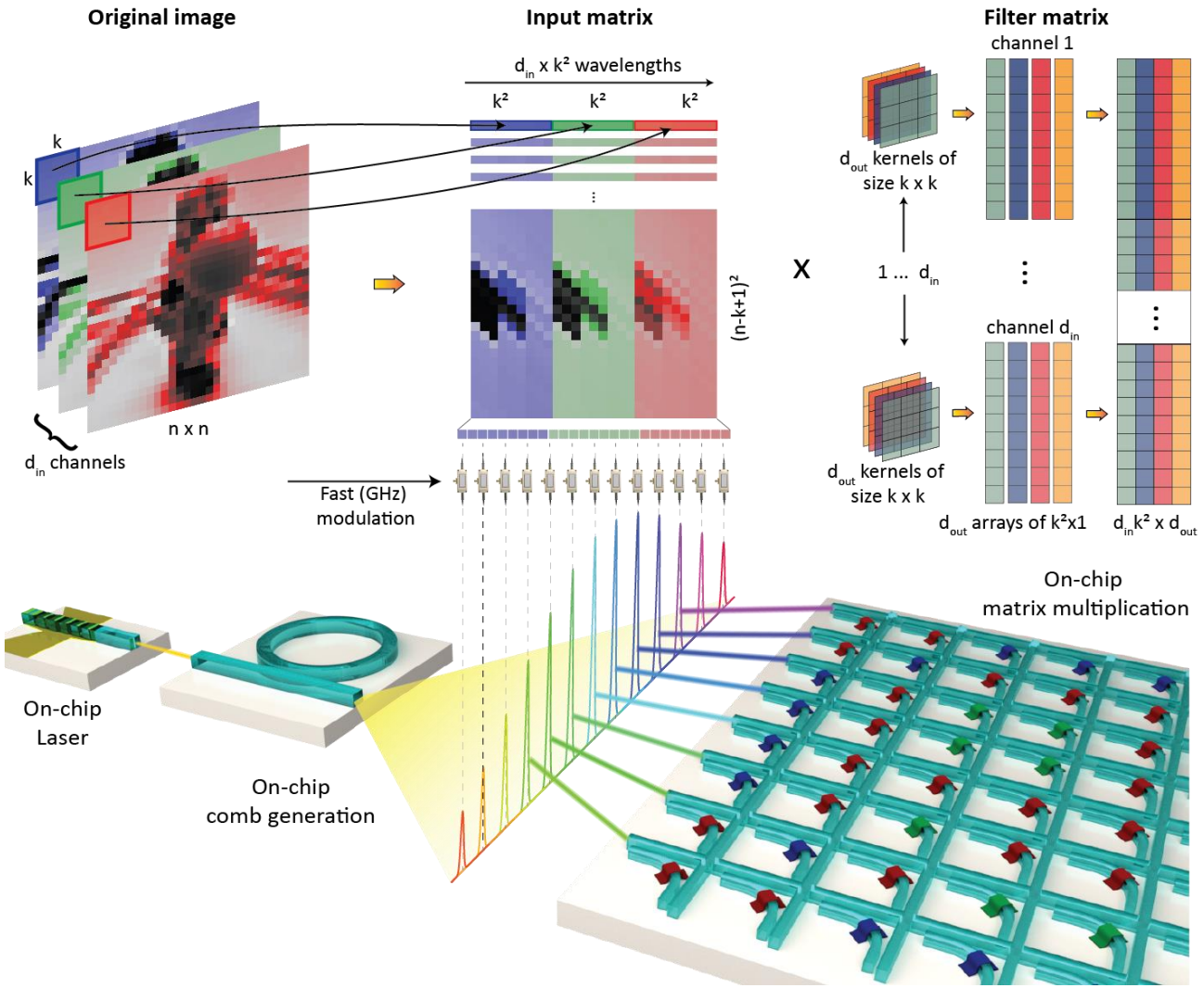


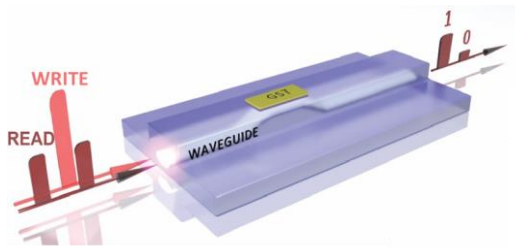


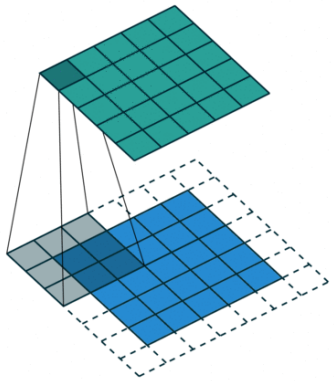
- Modified PCM device concept
- Exploits the I-V characteristic of phase change materials
- Substantially lower drift and conductance fluctuations arising from 1/f noise
- **Precision equal to 8-bit fixed-point arithmetic**

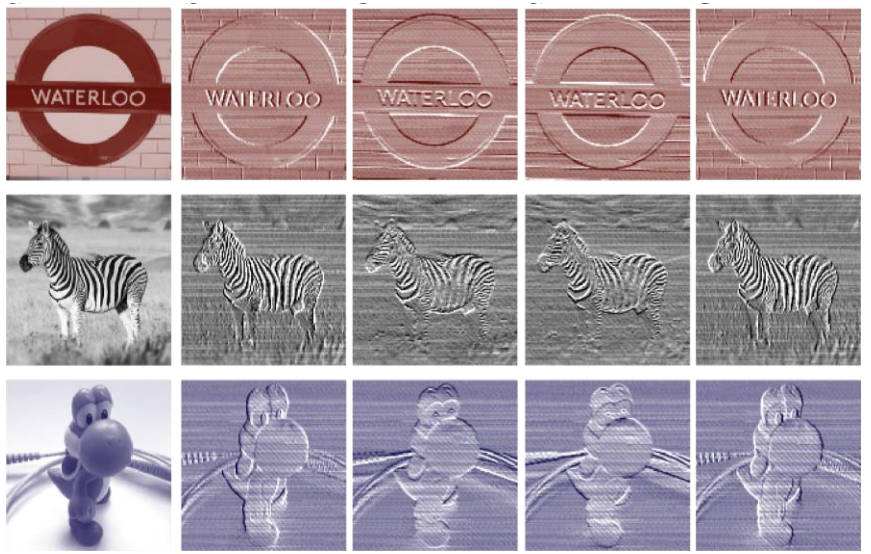
Kim et al., Proc. IEDM (2013), Koelmans et al., Nature Comm. (2015), Giannopoulos et al., IEDM (2018)

Towards higher speed: Photonic in-memory computing









Feldmann et al., Parallel convolution processing using an integrated photonic tensor core, Nature (2021)

Outline

- **Introduction**
  - Deep learning
  - In-memory computing
- **Deep learning based on computational phase-change memory**
  - Phase-change memory and synaptic emulation
  - DL inference and training
  - In-memory compute core
  - Device-level innovations
- **Applications beyond conventional DL**
  - DNN + “something”
  - Spiking deep neural networks

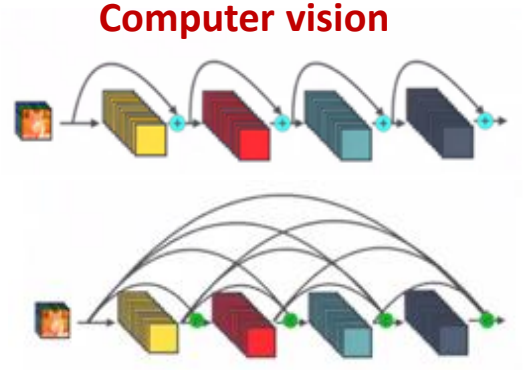
Hardware lottery

- Most of the algorithmic components for DNNs were in place decades earlier
  - Backprop (1963, 1976, 1988 (Rummelhart et al.)
  - CNNs (Fukushima & Miyake, 1982, LeCun et al., 1989)
- Several decades lost due to the lack of adequate hardware
- We need to ensure that good ideas are not lost or delayed this way
- What is next in deep learning?
- What role can IMC play going forward?

Hooker et al., The hardware lottery, ArXiv, 2020

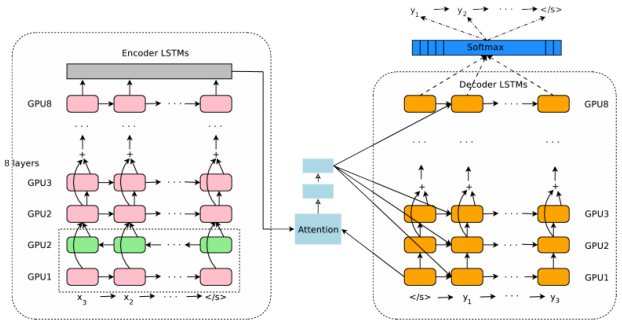
The deep learning landscape

Computer vision



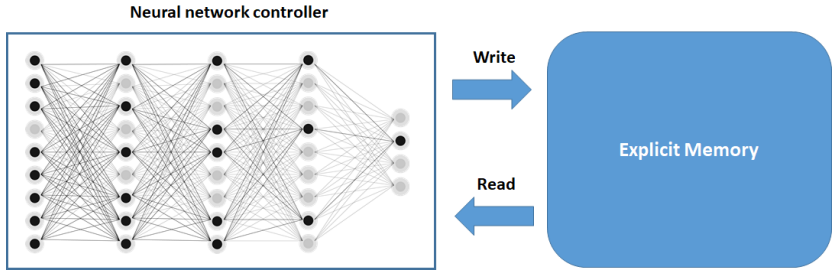
He et al, Deep residual learning for image recognition (2015)

Machine translation



Vaswani et al, Attention is all you need (2017)

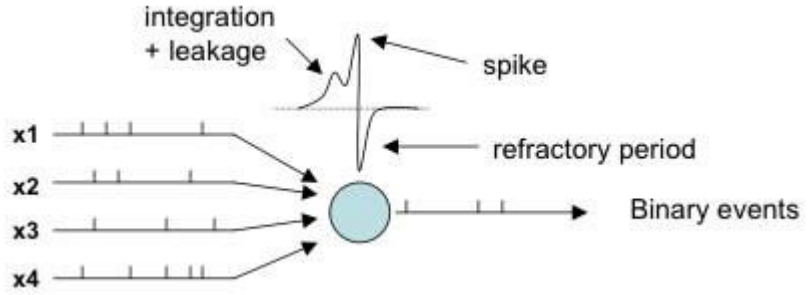
DNNs + "something"



Graves et al., Neural Turing machines (2014)

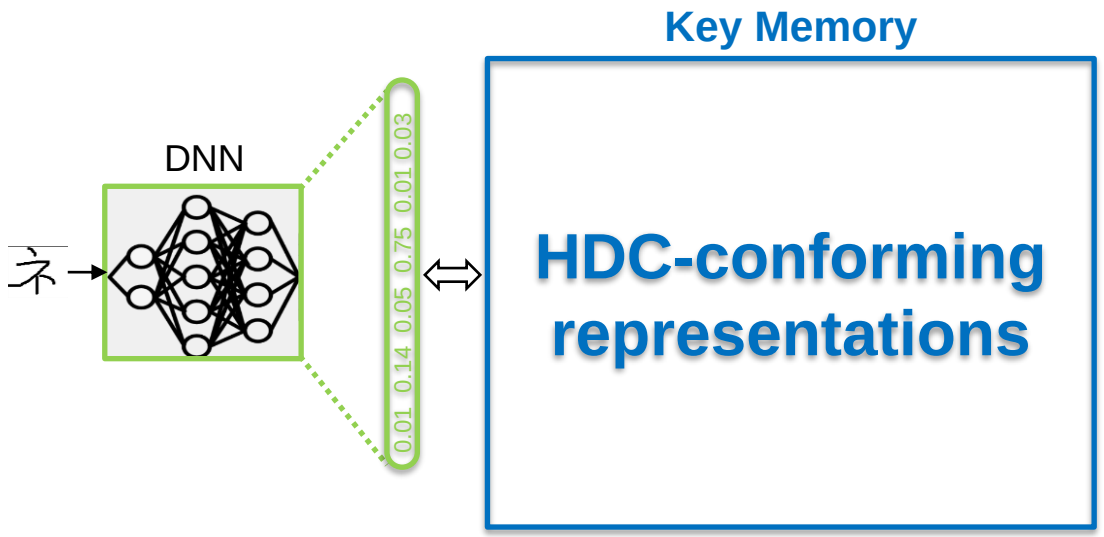
Hassabis et al., Neuroscience-inspired AI (2017)

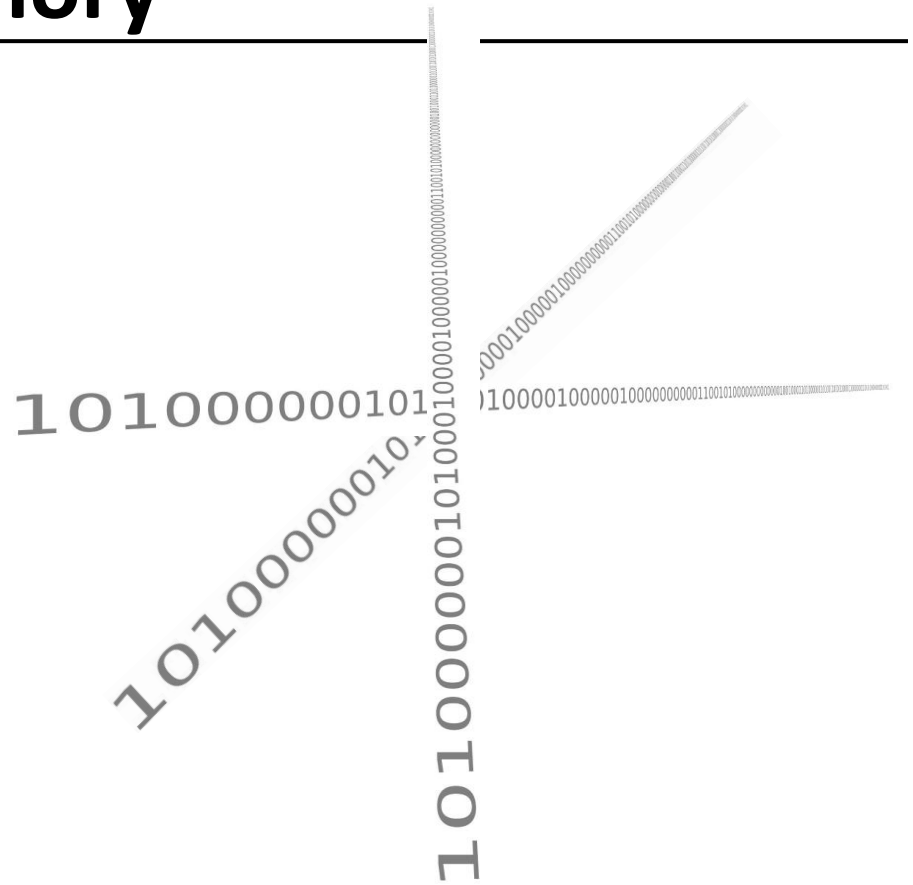
More biologically plausible DNNs



Pfeiffer and Pfeil, Deep learning with spiking neurons: Opportunities and challenges (2018)

DNN + Explicit HD associative memory





- Recent work on realizing explicit memory in terms of high dimensional vectors
- Powerful tool for few-shot learning
- High-dimensional algebraic operations for variable binding?

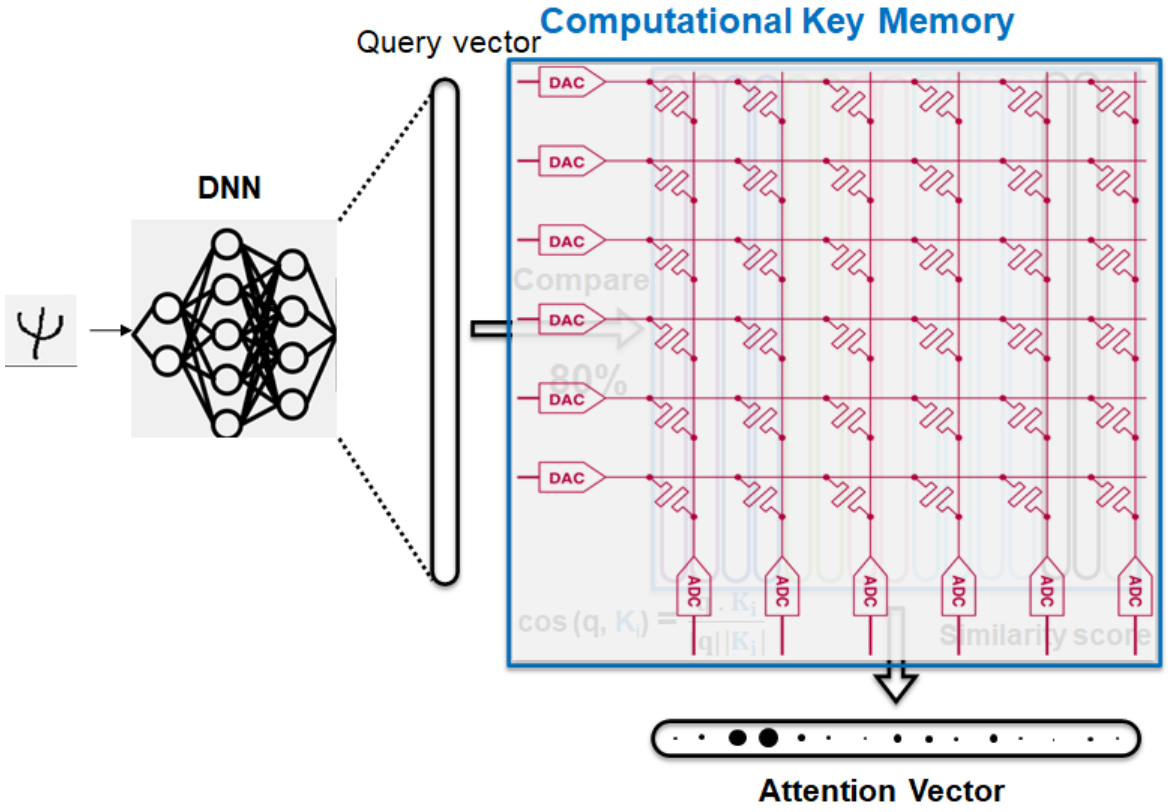
Mimicking the brain: Deep learning meets vector-symbolic AI

To better simulate how the human brain makes decisions, we've combined the strengths of symbolic AI and neural networks.

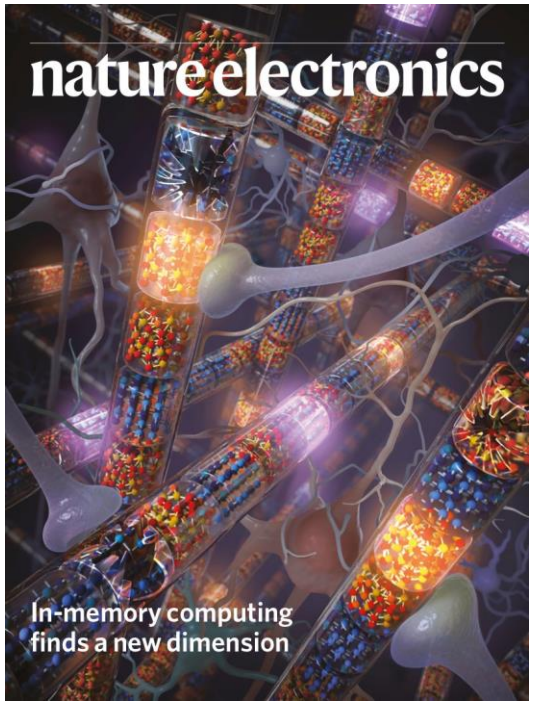
Abbas Rahimi & Abu Sebastian, IBM Research Blog

Karunaratne et al., "Robust high-dimensional memory-augmented neural networks", Nature Comm. (2021)

Efficient realization using IMC



Karunaratne et al., "Robust high-dimensional memory-augmented neural networks", Nature Comm. (2021)



Karunaratne et al., "In-memory hyperdimensional computing", Nature Electronics (2020)

- The high-dimensional explicit memory content (support vectors) stored in a PCM crossbar array
- The similarity between the input query vectors and support vectors computed through in-memory dot product operations

Spiking Neural Networks (SNNs)

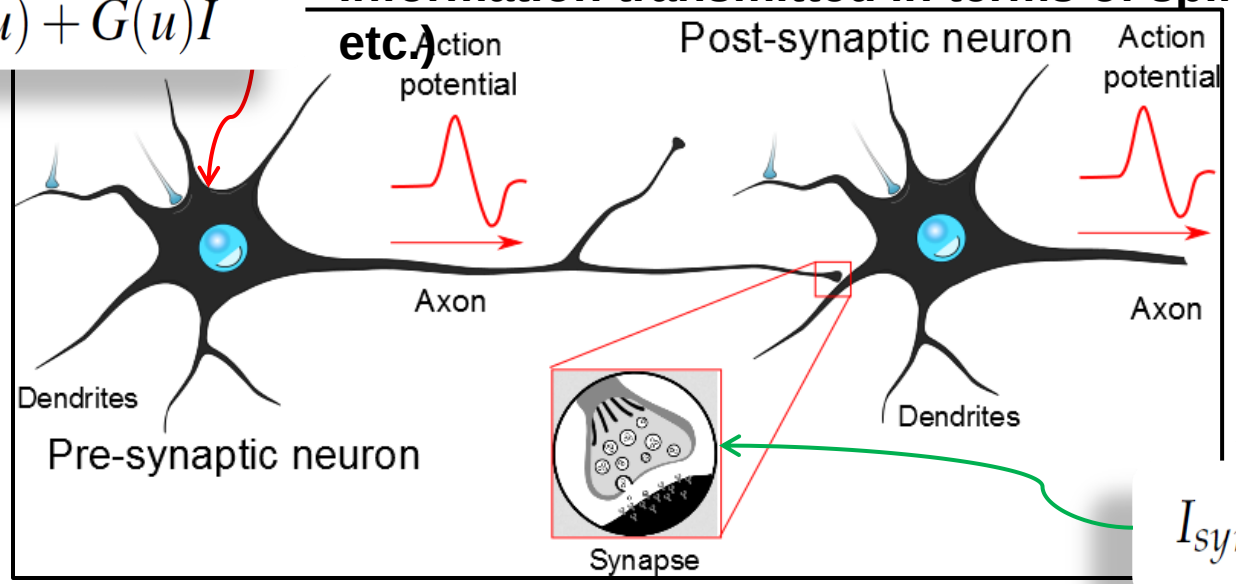
Information transmitted as floating point numbers!

Conventional Neural Networks

Neuronal dynamics

$$du/dt = F(u) + G(u)I$$

Information transmitted in terms of spikes (rate, timing etc.)



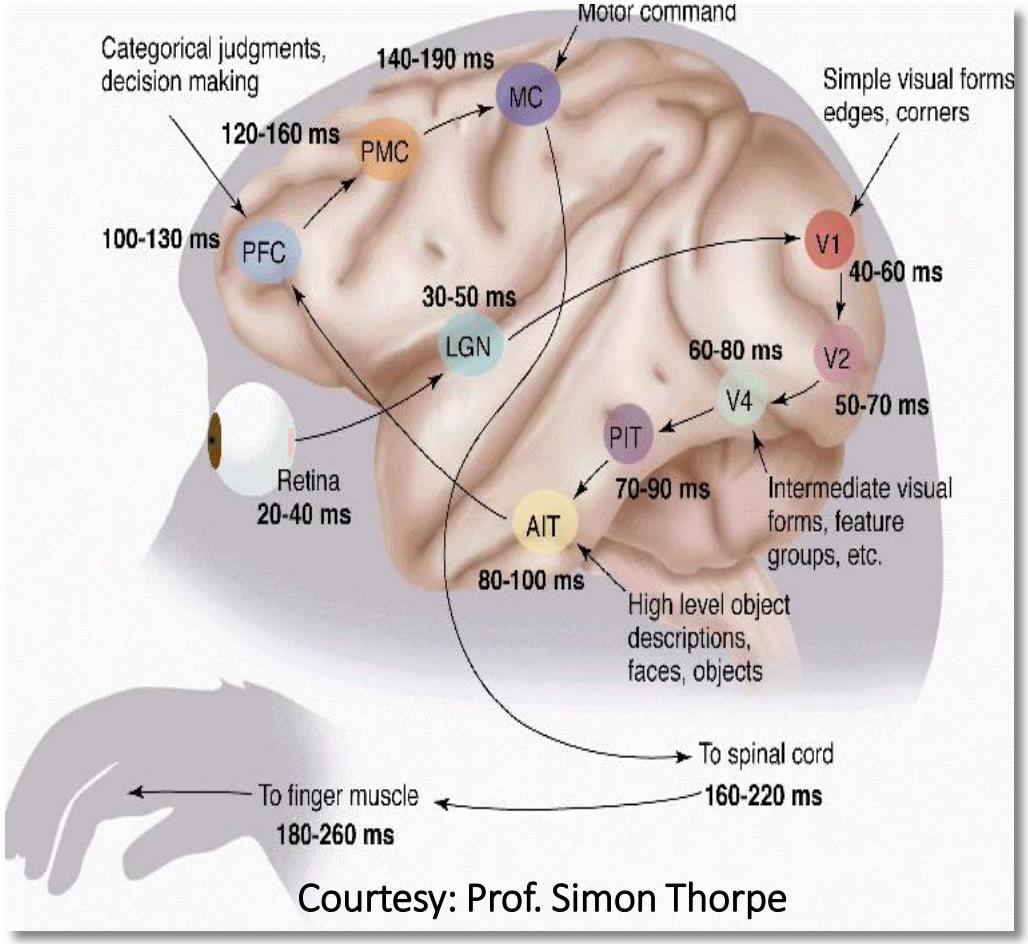
- Asynchronous
- Local, event-based learning
- Employed by the brain

Synaptic dynamics

$$I_{syn} = g_{syn} S(V - E_{syn})$$

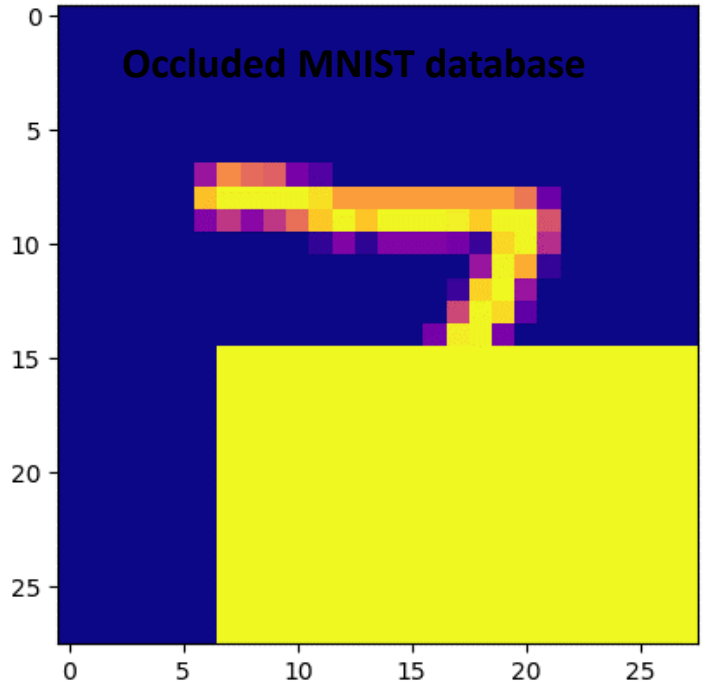
Why SNNs?

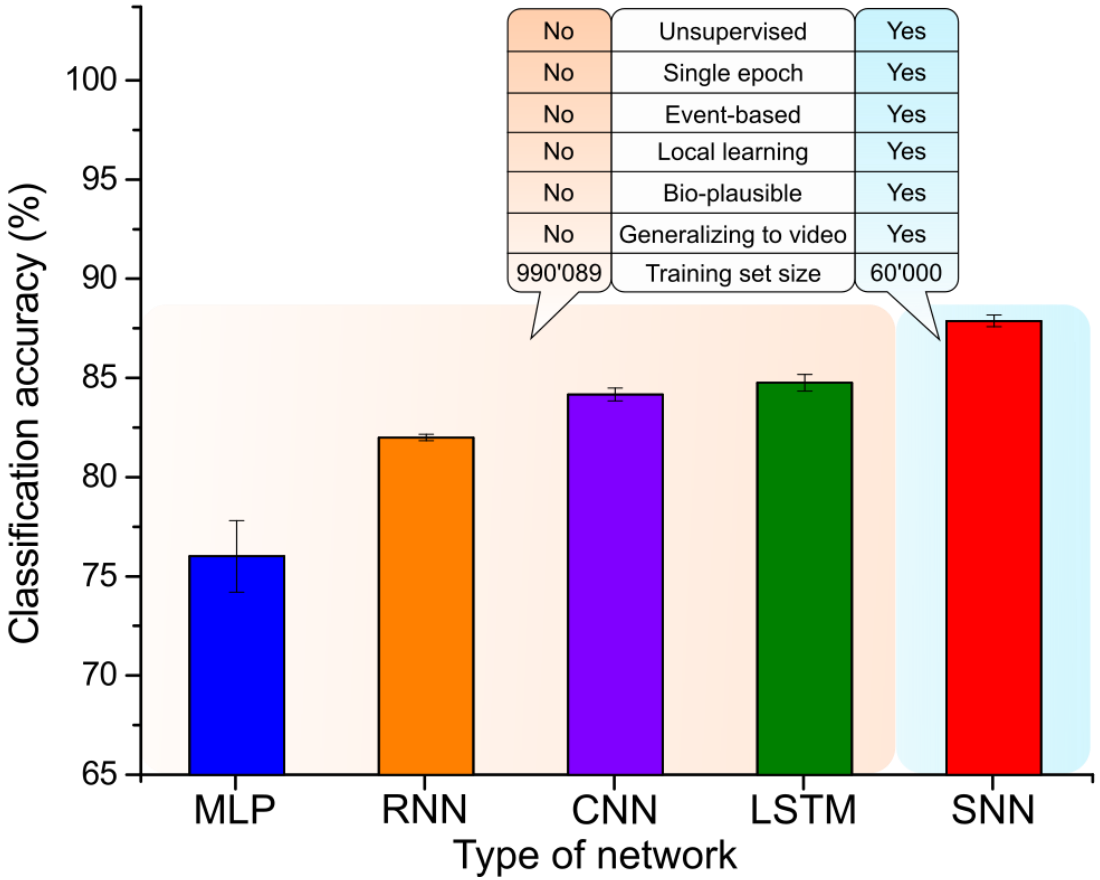
- Asynchronous processing (Energy efficiency)
- Temporal codes (Ultra-low latency)
- Local event-based learning (Energy efficiency)
- **Synaptic dynamics (Computational superiority in specific AI tasks?)**



Pfeiffer and Pfeil, Front. Neuroscience (2018)
Rajendran, Sebastian et al., IEEE SP Magazine (2019)

Inference in Dynamically Changing Environments

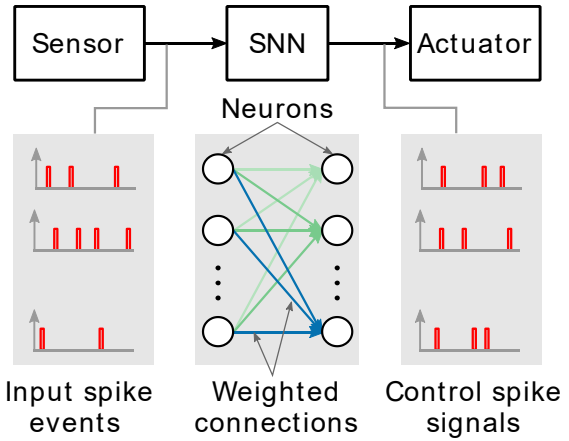


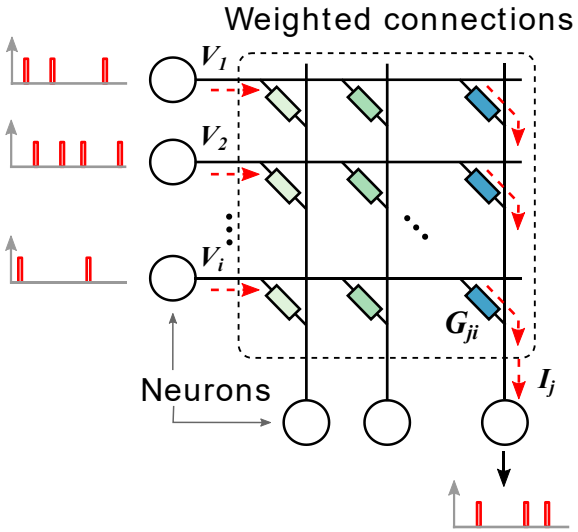


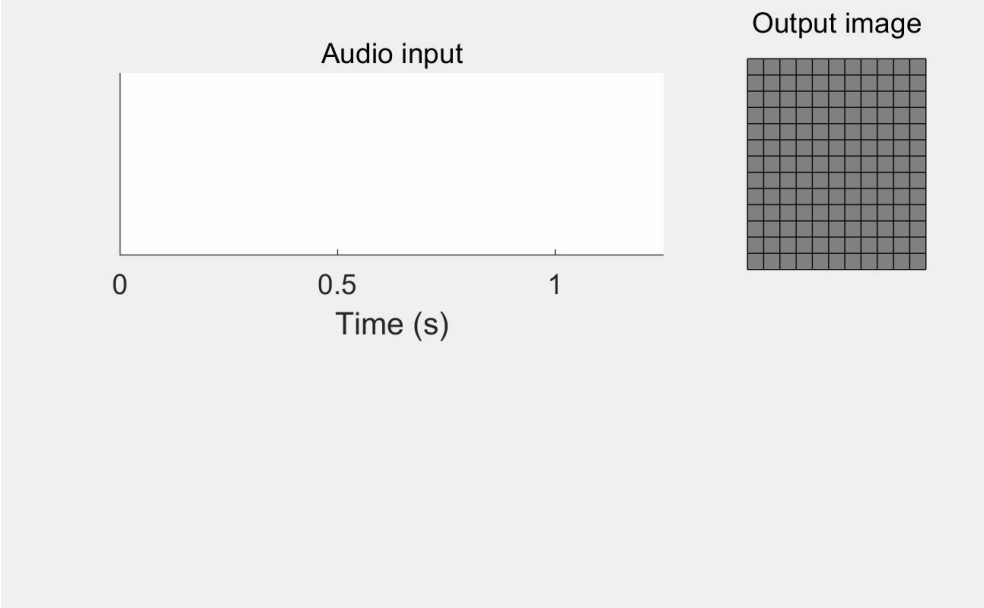
- Demonstration of an SNN surpassing ANNs in a specific task
- Purely neuromorphic and biologically modeled
- Cortical-like circuits can perform Bayesian inference on dynamic environments

Moraitis, Sebastian, Eleftheriou, Short-term synaptic plasticity optimally models continuous environments, ArXiv (2020)

IMC for SNNs







- IMC-based DNN acceleration can be easily extended to Spiking Neural Networks
- Synaptic efficacy and plasticity efficiently realized with physically instantiated synaptic arrays
- Potential to implement even more intricate synaptic dynamics and update rules

Nandakumar et al., "Experimental demonstration of supervised learning in spiking neural networks with phase-change memory synapses", Sci. Report, 2020

Wozniak et al., "Deep learning incorporating biologically inspired neural dynamics and in-memory computing", Nature Machine Intelligence, 2020

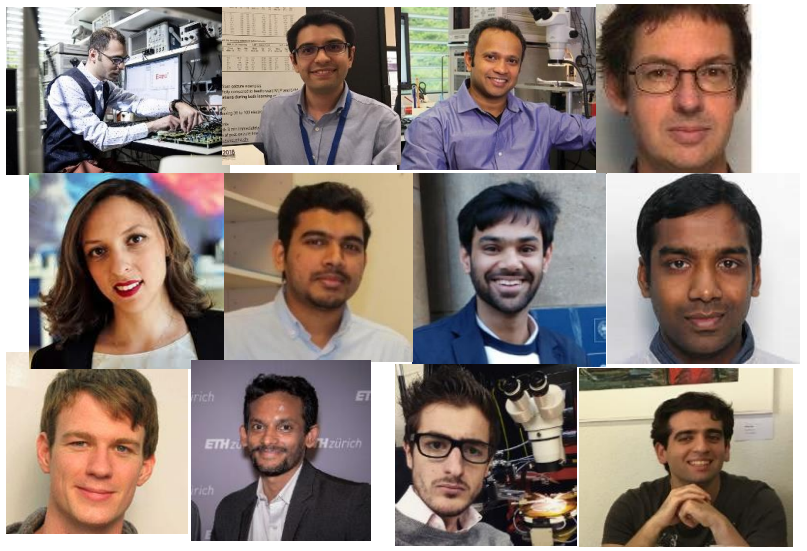
Moraitis et al., "Short-term synaptic plasticity optimally models continuous environments", ArXiv, 2020

Summary

- Deep learning is a key driver for innovations in computing systems
- New forms of computing such as in-memory computing (IMC) are being explored
- Attributes such as synaptic efficacy and plasticity can be implemented in-memory by exploiting the physical attributes of memory devices such as phase-change memory
- Iso-accuracy DNN inference and training is possible with IMC
- Recently fabricated mixed-signal IMC cores demonstrate the promise of this technology
- Concepts such as projected memory and photonic in-memory computing could significantly improve the computational precision and performance
- The IMC approach could also impact applications that transcend conventional DL such as memory-augmented neural networks and spiking neural networks

Acknowledgements

- In-memory computing group, IBM Research – Europe
- Several other groups @ IBM Research – Europe
- IBM AI Hardware Center (https://www.research.ibm.com/artificial-intelligence/ai-hardware-center/)
- New Jersey Institute of Technology, University of Patras, ETH Zürich, École polytechnique fédérale de Lausanne, RWTH Aachen, Oxford University, University of Münster



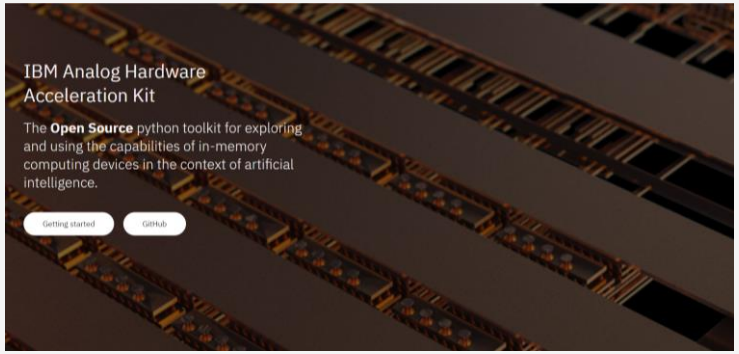
European Research Council
Established by the European Commission

THE FRAMEWORK PROGRAMME FOR RESEARCH AND INNOVATION
HORIZON 2020

FNSNF
FONDS NATIONAL SUISSE
SCHWEIZERISCHER NATIONALFONDS
FONDO NAZIONALE SVIZZERO
SWISS NATIONAL SCIENCE FOUNDATION

Analog AI Hardware Acceleration Toolkit

https://analog-ai.mybluemix.net/



Current Capabilities Include:

- Simulate analog MVM operation including analog backward/update pass
- Simulate a wide range of analog AI devices and crossbar configurations by using abstract functional models of material characteristics with adjustable parameters
- Abstract device (update) models
- Analog friendly learning rule
- Hardware-aware training for inference capability
- Inference capability with drift and statistical (programming) noise models

Roadmap:

- Integration of more simulator features in the PyTorch interface
- Tools to improve inference accuracy by converting pre-trained models with hardware-aware training
- Algorithmic tools to improve training accuracy
- Additional analog neural network layers
- Additional analog optimizers
- Custom network architectures and dataset/model zoos
- Integration with the cloud
- Hardware demonstrators

