

Emerging Technology Assessment

Produced for
The Company

An analysis of emerging
energy-efficient computer architectures

moonbeam



May 2022

moonbeam

Table of Contents

- I. Executive Summary
- II. Advances in Neuromorphic Computing
 - 1. Technology Categorization
 - 2. Advances in Mechanisms, Architectures, and Device Configuration – with related feature specifications
 - 3. Key Players
 - 4. Target Applications
 - 5. Requirements and Challenges
- III. Advances in In-memory Computing
 - 1. Technology Categorization
 - 2. Advances in Mechanisms, Architectures, and Device Configuration – with related feature specifications
 - 3. Key Players
 - 4. Target Applications
 - 5. Requirements and Challenges
- IV. Advances in Reservoir Computing
 - 1. Technology Categorization
 - 2. Advances in Mechanisms, Architectures, and Device Configuration – with related feature specifications
 - 3. Key Players
 - 4. Target Applications
 - 5. Requirements and Challenges
- V. Other Innovations in Novel Compute Technology Material and Architecture
 - 1. Technology Categorization
 - 2. Advances in Mechanisms, Architectures, and Device Configuration – with related feature specifications
 - 3. Key Players
 - 4. Target Applications
 - 5. Requirements and Challenges
- VI. Federal Government Analysis
 - 1. US Federal Spending
 - 2. Notable US Government Research Programs
- VII. Sources and Definitions

Section I

Executive Summary

After considering several novel in-memory computing architectures, Neuromorphic Computing offers the most promise.

The development of novel functional materials and devices incorporated into unique architectures will allow a revolutionary technological leap toward the implementation of a fully “neuromorphic” computer.

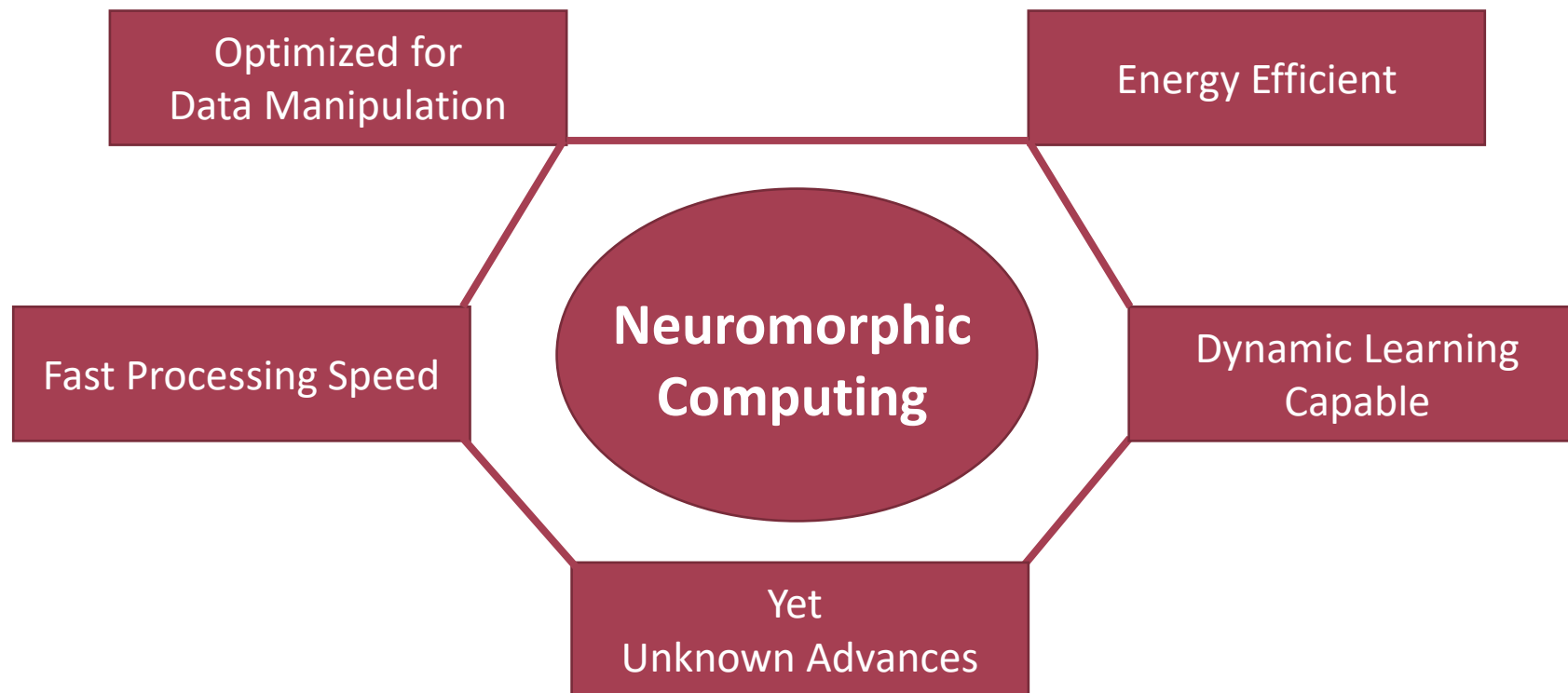
DOE Round Table Consensus, 2015

Neuromorphic Computing

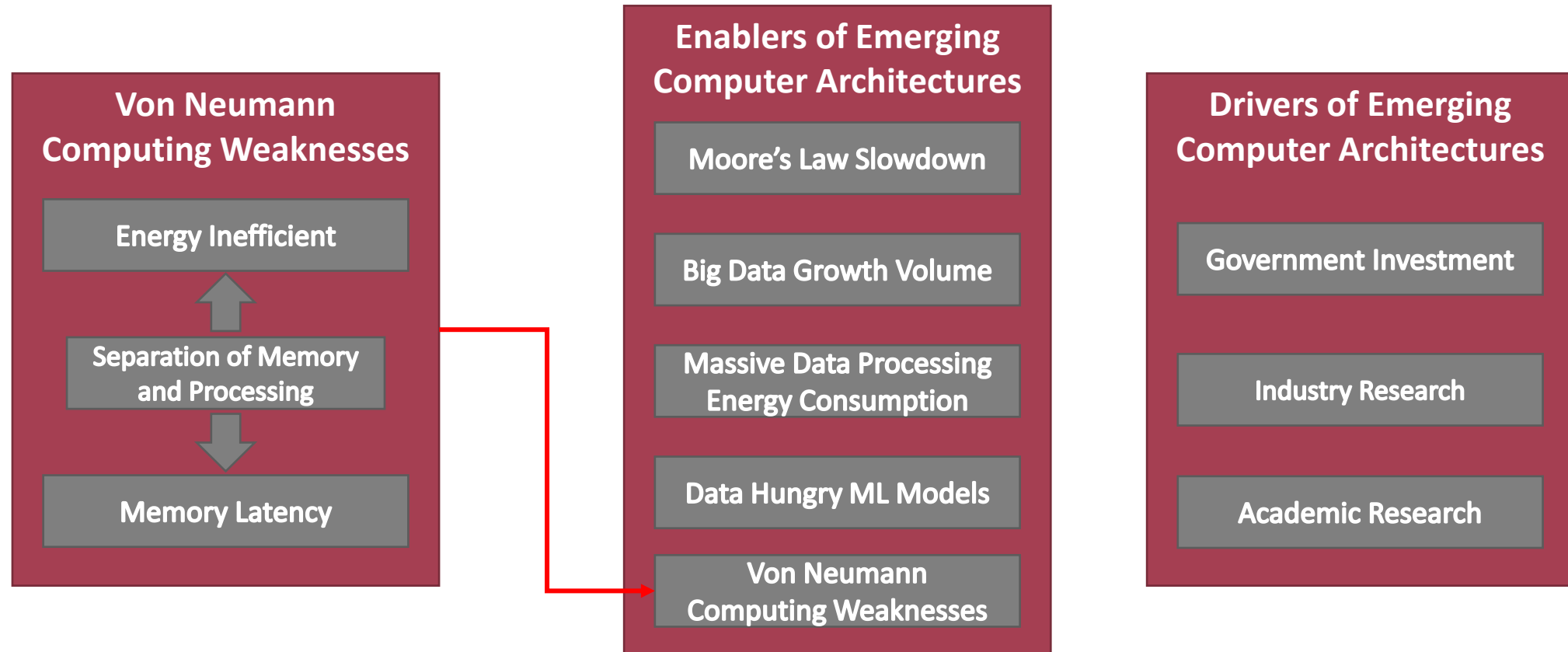
is the most promising novel computer architecture to solve emergent issues earlier identified and the only one with an outstanding need for exotic material substrates. Hence, our report focuses heavily on neuromorphic computing, and we recommend that *The Company* and Hiroshima University aim their efforts towards neuromorphic computing.

Other Technologies Evaluated:

- Reservoir Computing
- Advanced In-Memory Computing
- Near Memory Computing

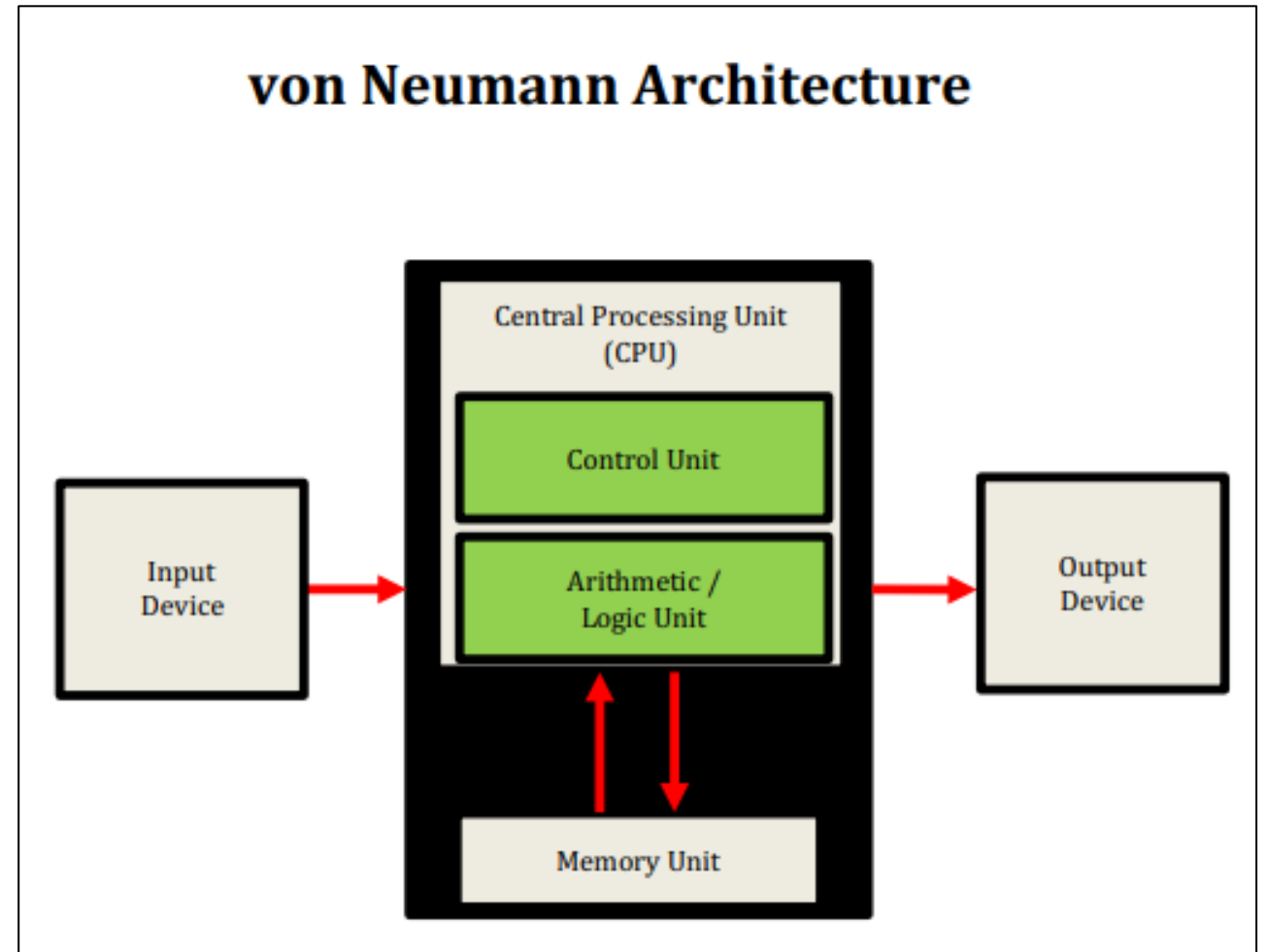
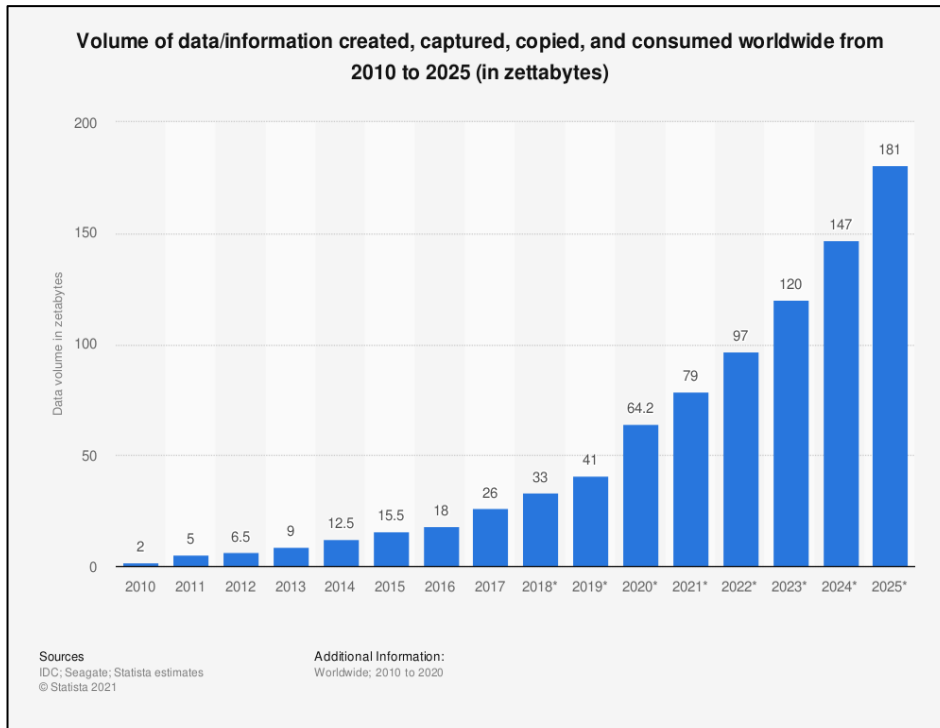


Need for new computer architectures that are energy-efficient and optimized for data processing and ML



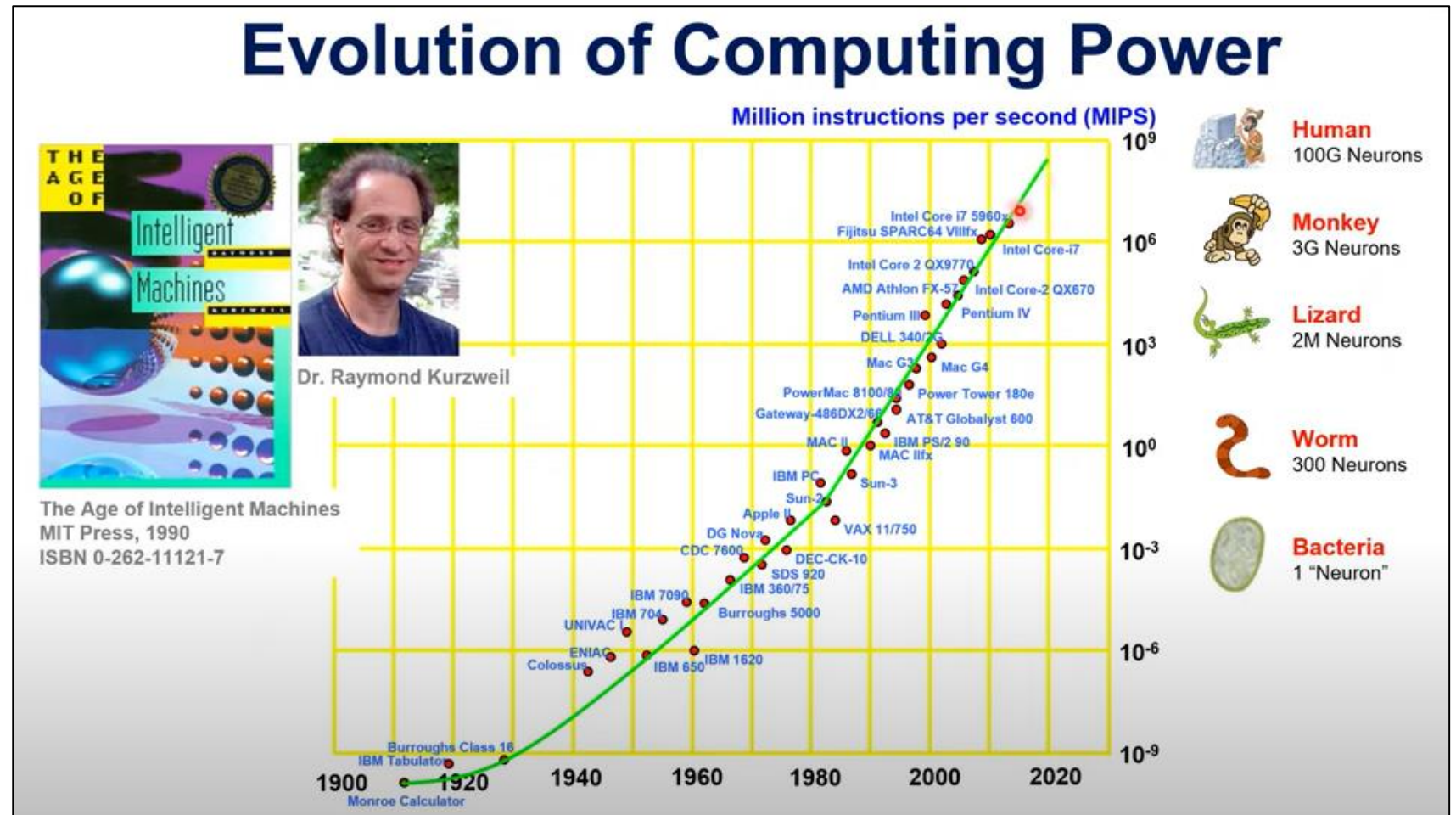
Conventional Computer Architectures: Not equipped for Data Revolution

- Traditional computer architectures separate memory and processing functions, requiring data shuttling between memory and CPU
- “...memory bandwidth and memory energy have come to dominate computation bandwidth and energy” (In-/Near-Memory Computing, page 1)



Computing Power Advances and new Data-Hungry Machine-Learning Models Result in Massive Energy Consumption

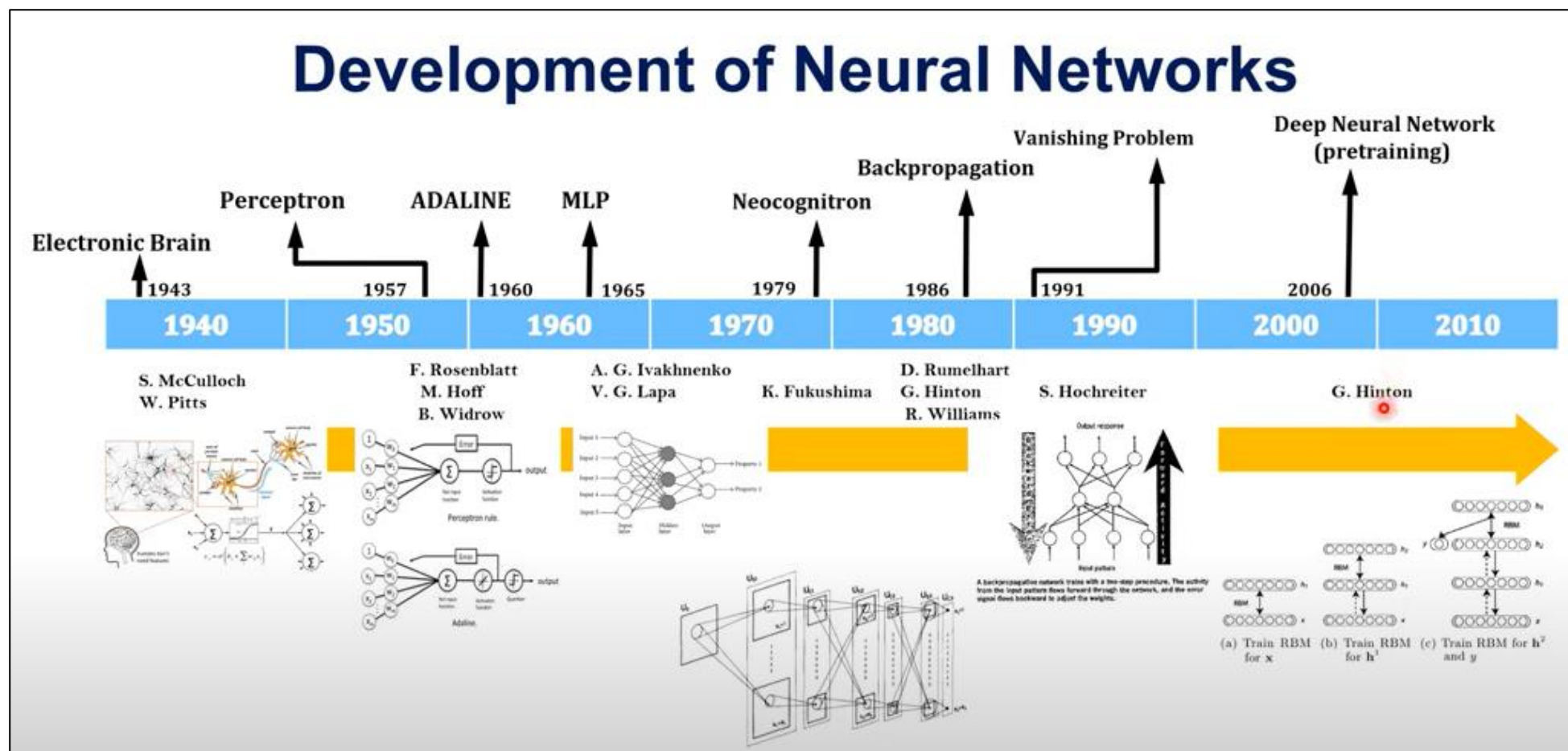
- “Currently, about 5-15% of the world’s energy is spent in some form of data manipulation, transmission, or processing” (DOE, 2015).
- “Data centers consume up to 1.5% of all the world’s electricity” (j.glanz)
- “Google’s data centers draw almost 260 MW of power...more than Salt Lake City” (j.glanz)



Computing Power Advances and new Data-Hungry Machine-Learning Models

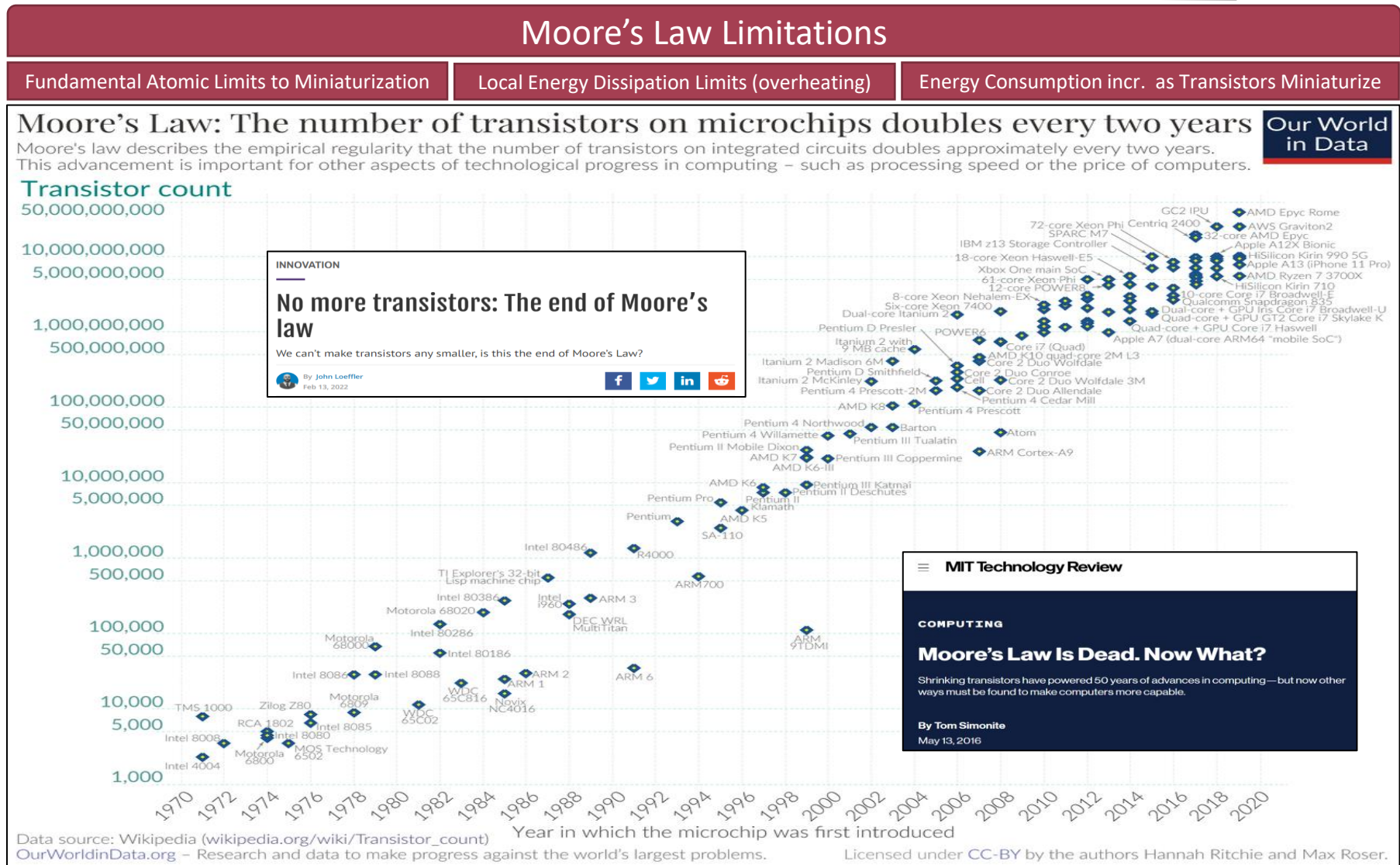
Result in Massive Energy Consumption

- “Currently, about 5-15% of the world’s energy is spent in some form of data manipulation, transmission, or processing” (DOE, 2015).
- “Data centers consume up to 1.5% of all the world’s electricity” (j.glanz)
- “Google’s data centers draw almost 260 MW of power...more than Salt Lake City” (j.glanz)



Moore's Law Decline Demands Computer Architecture Innovation

- Gordon Moore**, co-founder of Intel, observed in 1965 that the number of transistors per integrated circuit doubles approximately every 18 months. The number of transistors IC is proportional to computing power. Thus, Moore's law is an observation of the exponential growth of computing power as driven by the densification of transistors.
- Microprocessor architects** report that semiconductor advancement has slowed industry-wide since around 2010, below the pace predicted by Moore's law. Between 2019 and 2021, the highest commercially available chip transistor count increased from 39.54 billion to 57 billion or 44% - less than half of the doubling predicted by Moore's Law. Similarly, since the mid-2010s, the increase in top supercomputer performance has slowed substantially.



Section II

Advances in Neuromorphic Computing

1. Technology Categorization
2. Major Developments for this Technology
3. Advances in Mechanisms, Architectures, and Device
4. Configuration – with related feature specifications
5. Key Players
6. Target Applications
7. Requirements and Challenges

Neuromorphic Computing

Biology-informed novel computation models

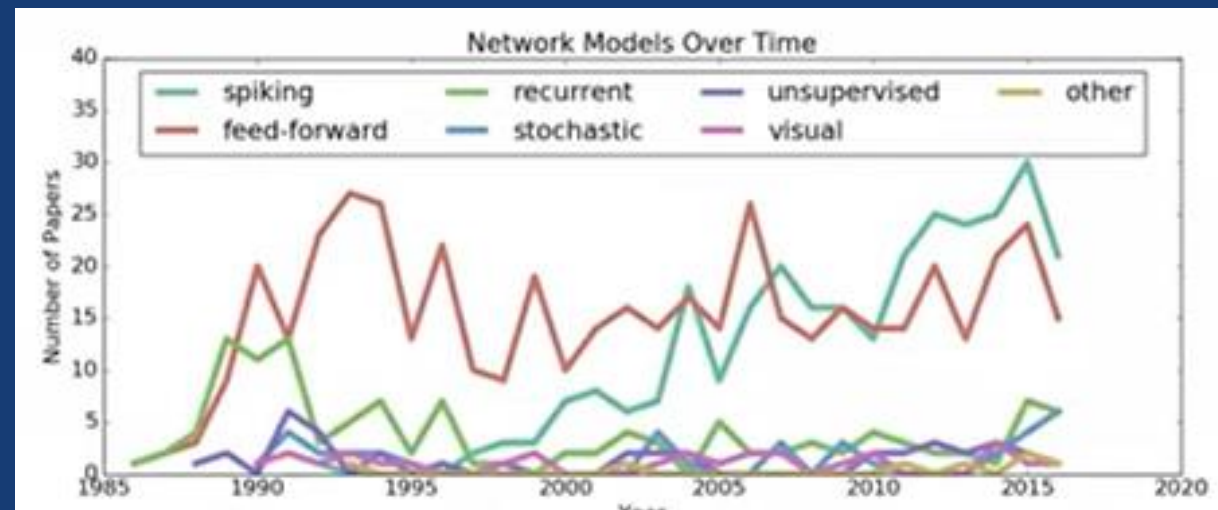
Definition

Neuromorphic computing is a special sub-category of in-memory computing that implements neural network architectures in computer hardware. Neural networks are inspired by the biological brain and have been used in software for complex pattern recognition tasks, and predictive modeling. Many distinct permutations of neural networks exist, including artificial neural networks (ANN) and spiking neural networks (SNN). SNN are more faithful approximations of how neuronal connections are updated. The remarkable achievements of deep learning (artificial neural networks) have contributed to the ascent of big data, as these algorithms require huge volumes of input data for acceptable accuracy/results. The mismatch between neural network software and Von Neumann hardware exacerbates the memory bottleneck and creates significant costs in energy and time. Direct implementations of neural networks in hardware may address these issues and augur a new age of computing for big data.

Brain:
15–30B neurons
Extremely complex
4km/mm³
35w

Realize Human Brain Advantages

Relative to AI, the human brain is remarkably accurate at minimal training pattern recognition, has good fault tolerance and remarkable energy efficiency. Neuromorphic computing aims to realize these benefits in computer hardware.



On-Chip v. Off-Chip Learning

On-chip learning allows neuromorphic chips to directly learn and train, which offers online, continuous calibration (important for edge-devices). Off-chip learning neuromorphic chips are trained elsewhere, and then downloaded to the neuromorphic chip to perform inference.

Synapse and Neuron Material

The most mature neuromorphic chips are developed using digital circuits and CMOS technologies. However, the memory latency and low-density of these products raises the possibility for the use of emerging nanotechnologies as synapse and/or neuronal material in future neuromorphic chips.

Commercial and Research Interest in Neuromorphic Computing

Rising Rapidly

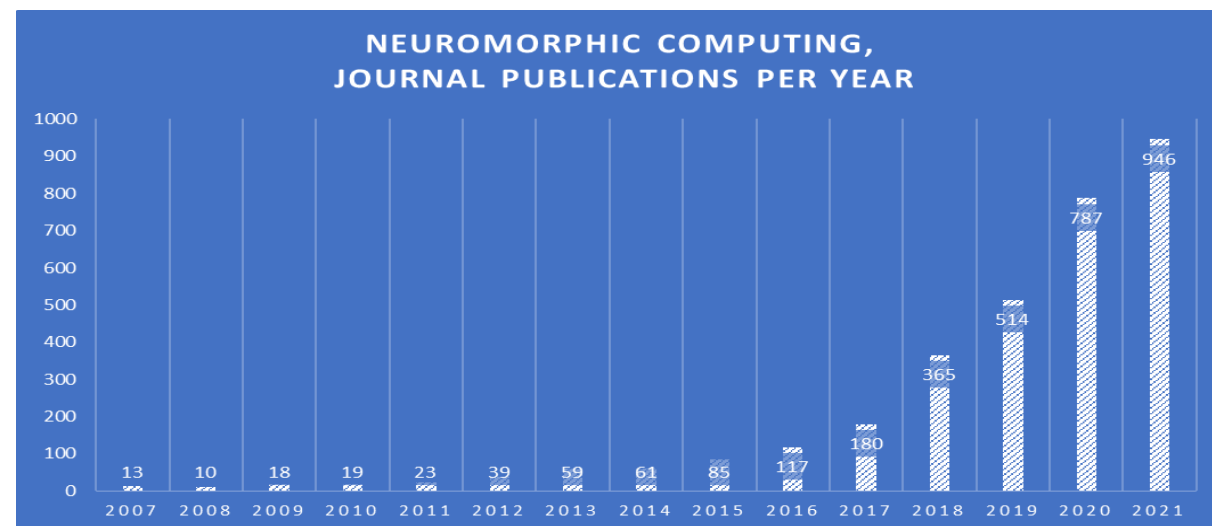
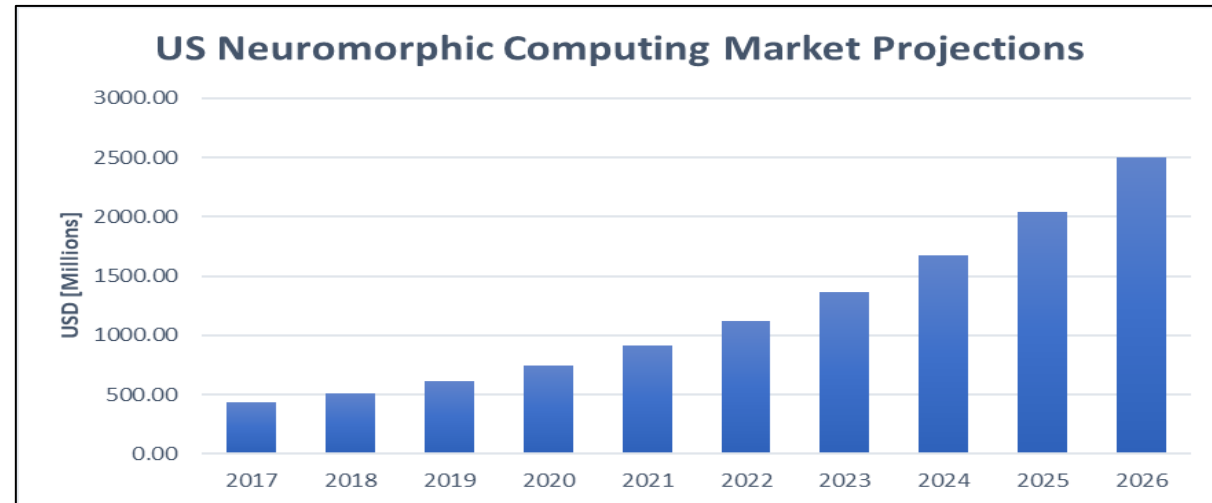
Analysis

North America is the largest regional market for neuromorphic computing and is host to many of the major players and leading research institutions for neuromorphic computing. IBM's TrueNorth chip and Intel's Loihi were pioneering examples of CMOS based neuromorphic architectures. Many of the major North American private companies with neuromorphic activities are profiled later in this report [IBM, Intel, Brainchip Holdings, Hewlett Packard]. Other significant North American companies driving North American neuromorphic growth are HRL Laboratories, Qualcomm, Numenta, General Vision, Applied Brain Research, Knowm Inc., and Vicarious.

Another key enabler aside from North America's rich tech ecosystem is substantial academic innovation. MIT research produced a "brains-on-a-chip" neuromorphic project in October 2018. Canada's University of Waterloo created the "world's largest functional artificial brain" through project SPAUN. Many other North American Universities and Research Labs are driving improvements in neuromorphic computing.

The market growth projections included at right should be interpreted as very rough estimates. Neuromorphic computing is fundamentally an extremely disruptive technology, hence, difficult to forecast. Market growth is highly dependent on innovations in technology. Should neuromorphic computing's potential be realized, the market projections at right will far underestimate impending growth.

Sources: [4]; [12]



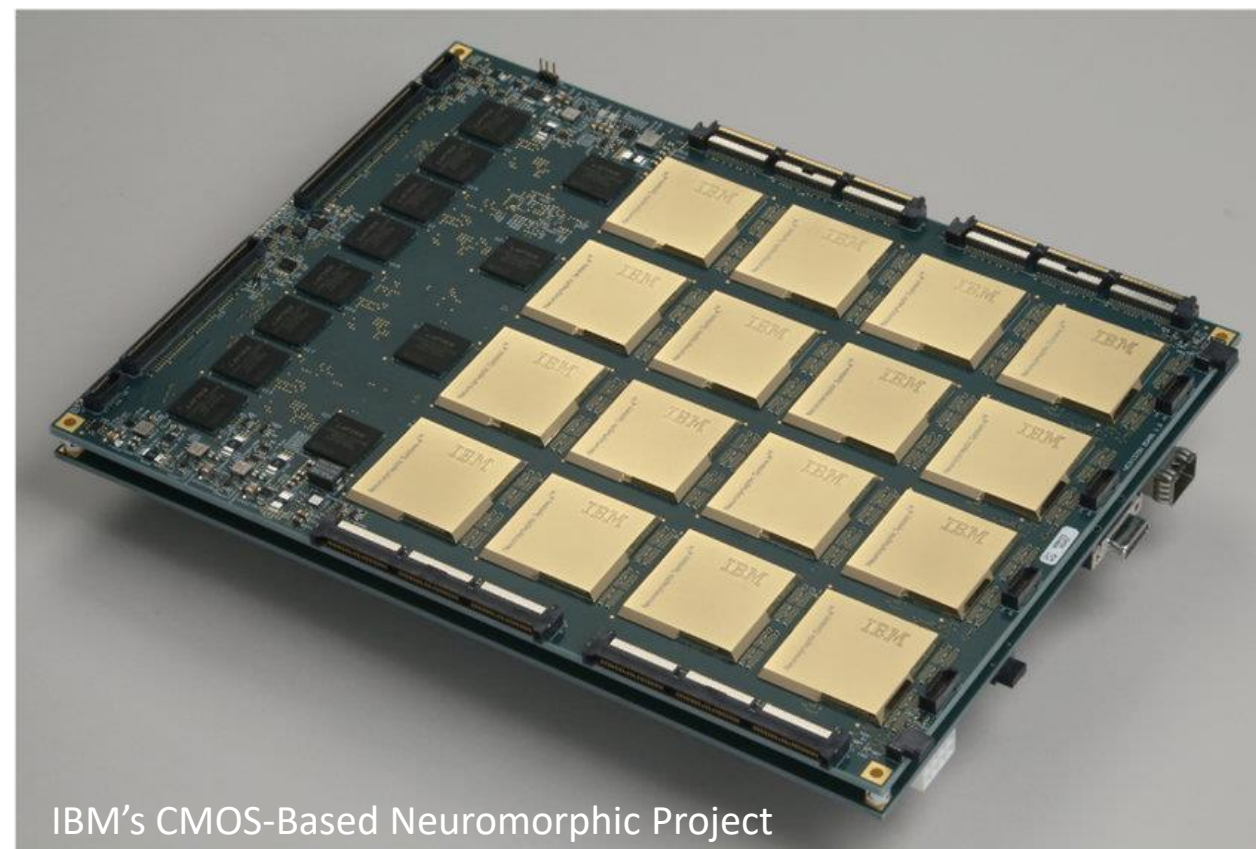
Significant need for Emerging Nanotechnologies

To function as synapses in Neuromorphic Hardware

Synapse Material Discussion

Neuromorphic hardware is composed of connective synapses and neurons. Although Neuromorphic computing architectures can be entirely constructed from existing CMOS technology (see IBM and Intel products, TrueNorth and Loihi respectively), the large scale of implemented chips inhibits commercial use-cases. Current research attempts to find substitute materials to use for synapses with CMOS neurons to reduce scale. Indeed, a long-term research ambition is to eliminate CMOS technology in neuron construction as well. Long theorized memristors (materials which change their electrical resistance in response to input charge) have been discovered. Memristors' non-volatility and high-density make them ideal candidates for implementing neuromorphic computing architectures (initially as synapses). Memristor based memory is termed resistive random-access memory (ReRAM). Synapse material candidates include phase-change memory (PCM), ferroelectric devices, valence change memory, 2D materials, organic materials, and Spintronics.

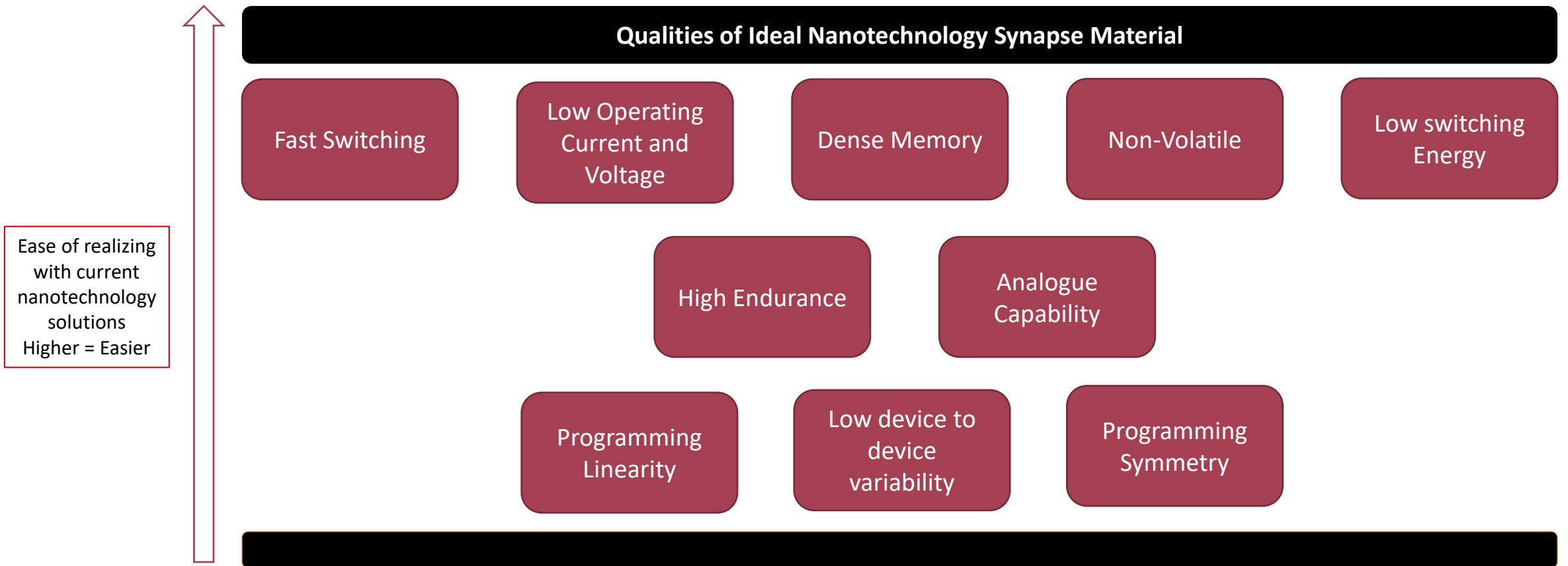
Sources: [2]



IBM's CMOS-Based Neuromorphic Project

The Missing Ideal Synapse

No Candidate Synapse Material Fulfills all Requirements of an Ideal Memristor



See appendix Definitions slide for elaboration of several synapse “qualities”

Sources: [1]; [2]; [8]; [9]

The Missing Ideal Synapse

No Candidate Synapse Material Fulfills all Requirements of an Ideal Memristor

Summary of desirable performance metrics for synaptic devices

Performance metrics	Desired targets
Device dimension	<10 nm
Multi-level states number	>100* (with linear and symmetric update)
Energy consumption	<10 fJ/programming pulse
Dynamic range (on/off ratio)	>100*
Retention	>10 years* (for inference)
Endurance	>10 ⁹ updates* (for online training)

Note: * these numbers are application dependent.

The Following Emerging Nanotechnologies

Being Considered for Employment as Synapses in Neuromorphic Hardware

1. Phase Change Memory Devices
2. Ferroelectric Devices
3. Valence Change Memory
4. Electrochemical Metallization Cells
5. 2D Materials
6. Organic Materials
7. Spintronics

Each material is analyzed in the following report section

General Materials Analysis

Material	Cell Area (F ²)	Voltage (V)	Read Time	Write Time	Write energy (J/bit)	Retention	Endurance	Multibit	Non-Volatility	Source(s)
SRAM	100	<1	~1 ns	~1 ns	~1 fJ	N/A	> 10 ¹⁶	No	No	Ref #25
DRAM	6	<1	~10 ns	~10 ns	~10 fJ	~64 ms	> 10 ¹⁶	No	No	Ref #25
Phase Change Memory Devices	4-20	<3	<10 ns	~50 ns	~10 pJ	>10 y	10 ⁸ - 10 ¹⁵	Yes	Yes	Ref #3, p. 55
Ferroelectric Devices	12 – 22	1 – 3	20 – 40 ns	10 – 60 ns	–	10 y	10 ¹⁴ - 10 ¹⁵	–	Yes	Ref #25, p.3
Valence Change Memory	>4	<2	~5 ns	~5 ns	–	>10 y	~10 ¹²	Yes	Yes	Ref #25, Ref #27, Ref #28
Electrochemical Metallization Cells	–	~7	~1 ns	~1 ns	–	-	~10 ⁶	Yes	Yes	Ref #25, Ref #2
STT – MRAM (Spintronics)	6 – 50	<2	<10 ns	<10 ns	~0.1 fJ	>10 y	<10 ¹⁵	No	Yes	Ref #3, p. 55

Phase-Change Memory Devices

Description

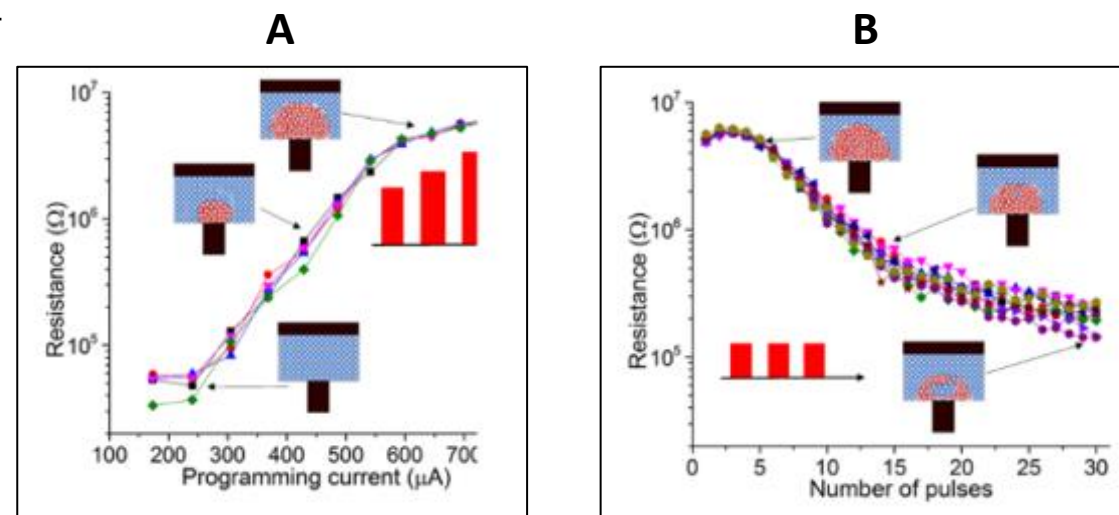
Phase-Change memory (PCM) devices are composed of two exterior electrodes and an internal chalcogenide material that can be switched between amorphous and crystalline states through joule heating, changing the resistance of the device. PCM devices are viable for binary memory storage, encoding information as either High-Resistance or Low-Resistance. Furthermore, PCM binary memory allows in-memory logical operations through imputed voltage, multi-device resistance interactions, and output current sensing. Such in-memory logic is suited for database query and hyper-dimensional computing.

PCM devices also have analogue storage capability (figure A) dependent on programming current. This feature can be exploited for matrix-vector multiply (MVM) operations in $O(1)$ time complexity – as opposed to $O(n^2)$ time complexity for MVM ran in software built on Von Neumann architecture. The application of a trained deep neural network to new data is composed of MVMs. Thus, Analogue PCM is very suited for DNN inference using off-chip trained models with excellent energy efficiency and computational speed.

An additional desirable property of PCM devices is their accumulative behavior (figure B). Training neural networks requires that synaptic weights be continually modified through multiple iterations. That PCM devices are accumulative makes possible DNN training on-chip for spiking neural networks and other deep learning algorithms.

Note: Technical parameters of state-of-the-art PCM devices is appended at end of discussion on neuromorphic materials in table.

Sources: [2]; [8]; [9]



Developmental Challenges

- PCM binary storage has an undesirably wide distribution of high-resistance values, complicating logical evaluations.
- Analogue PCM memory suffers from drift/noise as function of time, reducing precision of MVM
- behavior of PCM devices are highly non-linear, stochastic – complicating DNN training as synaptic weight updates can't be precisely controlled per algorithmic requirements
- Necessary access devices to accommodate high-current requirements of PCM increase cell sizes
- Scaling technology in large-scale arrays is not currently achievable at cost-effective values
- Set/Reset are asymmetric, and have markedly different energy requirements (Reset requires high-current)

Advances

- The unstable electrical properties of amorphous PCM devices cause read “noise” of $1/f$. Recent research has attached a projection layer, “an electrically conducting material”, to the PCM device. This device configuration results in read current largely bypassing the electrically unstable amorphized material flowing towards the projection layer, for more accurate reads reducing “noise”. Empirical research has found that Projected PCM composites significantly reduce both noise and drift “by at least an order of magnitude”
- Source: “State dependence and temporal evolution of resistance in projected phase change memory”, Nature Journal 2020

Ferroelectric Devices

Description

Ferroelectric devices were discovered in 1920 and are defined by switchable electrical polarization. Ferroelectric Random-Access-Memory (FeRAM) stores information as a polarization state. A type of FeRAM is the ferroelectric capacitor (FeCAP). To write to a FeCAP an electric field is applied to the ferroelectric layer to change the polarity of the cell. To read the information stored in a FeCAP, the device transistor applies an electric field to a particular state (e.g. 0). If the cell was already storing 0, no output pulse will occur. If the cell was previously storing a 1, a small output pulse of current will be detected. FeCAP based memory is destructive, as reading overwrites previously stored information. Thus, FeCAP based memory requires cells to be re-written when read from. FeRAM is most easily used as binary storage, however, certain crystalline structures can switch polarization direction by other than 180°, rendering analogue storage possible (albeit challenging).

“The ferroelectric field effect transistors (FeFET) features a ferroelectric capacitor as gate insulator, modulating the transistor’s threshold voltage that can be sensed non-destructively by measuring the drain-source current. Perovskite based FeFET memory arrays with up to 64kBit have been demonstrated. “

Another ferroelectric device for in-memory computer architecture is the ferroelectric tunneling junction (FTJ, see figure A). The FTJ contains a ferroelectric layer between two electrodes. The output current is dependent on the polarization of the ferroelectric layer and can be detected by running electric currents too small to adjust polarity through the device. As polarity is not adjusted, read operations are non-destructive.

Sources: [2];

Developmental Challenges

FeCAP:

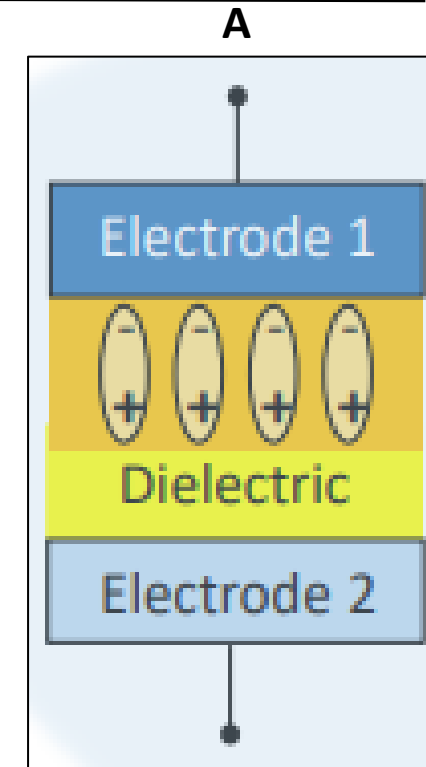
- Large capacitors lower memory density
- Necessary perovskite materials are difficult to manufacture as thin-film layers

FeFET:

- Limited scalability
- Poor data retention

FTJ:

- Manufacturing thin ferroelectric layers results in device defects: formation of interfacial dead layers; increased leakage



Advances

Ferroelectricity was discovered as an attribute of hafnium oxide in 2008. Hafnium oxide is CMOS compatible, and therefore suited for hybrid CMOS neural chip configurations. Research is being performed to create circuit architectures that rectify ferroelectric non-idealities. Too, large scale demonstrations are in the works to determine deficiencies and issues with ferroelectric based synapse construction.

Valence Change Memory

Description

Valence Change Memory (VCM) devices adjust their resistance/conductance in response to electrical pulses by exploiting oxygen ion migration effects. VCM devices are capable of binary and analogue information encoding and have been proven in demonstrations to be capable of implementing several online learning rules (spike-time dependent plasticity, and voltage threshold-based plasticity).

Filamentary VCMs are the most mature and studied type of VCM device. These devices form or destroy conductive filaments (CF) that change the conductivity of the VCM device as a binary storage method. Or, the diameter of the CF can be modulated for analogue storage.

Interfacial VCMs' conductance "scales with the junction area of the device, and the mechanism is related to a homogeneous oxygen ion movement through the oxides." Interfacial VCMs, like filamentary VCMs, are two-terminal devices.

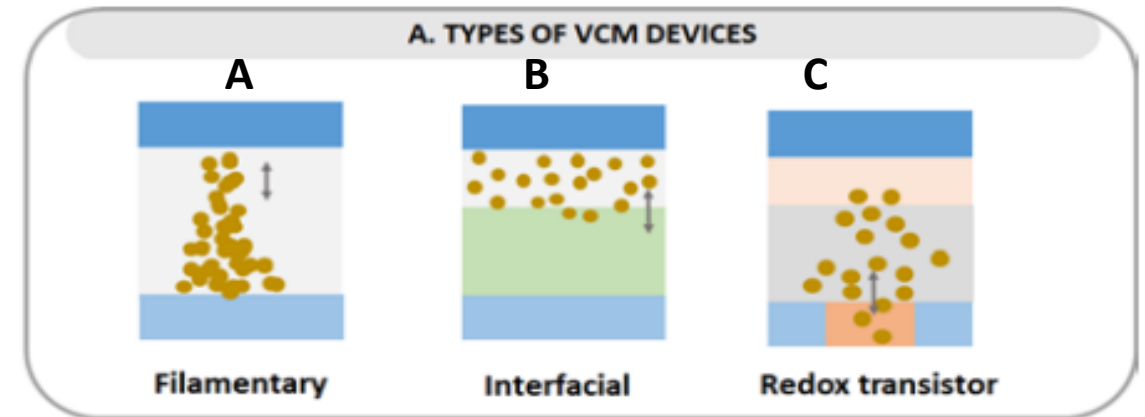
Redox Transistor VCMs are three terminal devices which adjust their resistance by control of the oxygen vacancy (reduced oxygen concentration) in the transistor channel.

Sources: [2]; [27]

Advances

Filamentary: To reduce program/read disturbs, proposed advances include the use of fast pulses for cell updates or the application of thermal engineering to device stacks to increase transition time.

Interfacial/Redox: The main research pathway is the study of interfacial and redox VCM devices at the multi device level, i.e., stack level. The outcome of these studies will be better improvements to modelling that will allow neuromorphic chip creation.



Developmental Challenges

Filamentary Devices:

- Device variability
- Device stochasticity
- Program/Read disturbs
- Low resistance levels

Interfacial Devices:

- Lower retention relative to Filamentary Devices
- Not yet scaled to nm sizes
- No simulation models currently developed

Redox-based VCM transistors

- Little studies – only demonstrated at level of single device
- Statistical data based on robust empirical studies not collected
- Lack of understanding of switching mechanism and attendant models

Electrochemical Metallization Cells

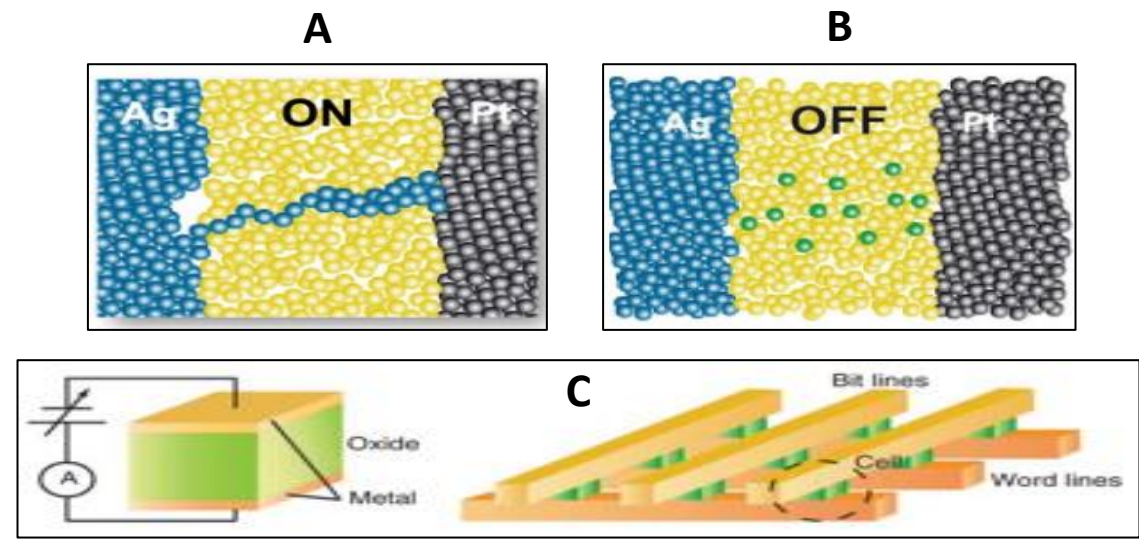
Description

Electrochemical Metallization Cells, or programmable metallization cells (PMC), are composed of two electrodes bounding a thin layer of ion transporting material. The electrodes are electrochemically active. Applied voltage creates a metallic filament, enhancing conductivity. Voltage of the opposite polarity will destroy the metallic filament, reducing conductivity. These two states (low resistance and high resistance) are used for binary storage. Visual A shows an example of low-resistance state and visual B shows an example of high-resistance state. Too, the metallic filament may have moderated width to enable analogue storage.

The electrochemically active electrode is usually composed of Ag, Cu, Fe, or Ni with the other electrode made of Pt, Ru, Pd, TiN, or W. Conductive non-metals are also explored as electrodes, including TiN, graphene, carbon nanotubes, and conductive oxides (ITO, SrRuO₃). Ag and Cu are common electrode types as they have excellent ionic mobility properties. The internal switching layer is usually a thin-film insulator or semiconductor.

Advantages of PMC include low voltage “(~ 0.2 V to ~ 1 V) and currents (from nA to μA range).” PCM devices can be fabricated from a wide range of materials and are operable in challenging conditions. Also, PCM devices are small, reducible to nanoscale sizes.

Sources: [2]; [13]



Developmental Challenges

The miniscule scale of PMC devices combined with significant current densities generates harsh, non-equilibrium device dynamics, compromising accuracy and device operation. ECMs suffer from variability in switching voltages, currents and resistive states. ECMs also experience fluctuations and drift with ongoing usage.

Advances

Scientific research is attempting to better understand the nanoscale processes and interactions of PMC cells. Other research is attempting to optimize material selection by the considerations of how internal cell materials interact. There remain device/circuit issues which are also receiving research attention.

2D Materials

Description

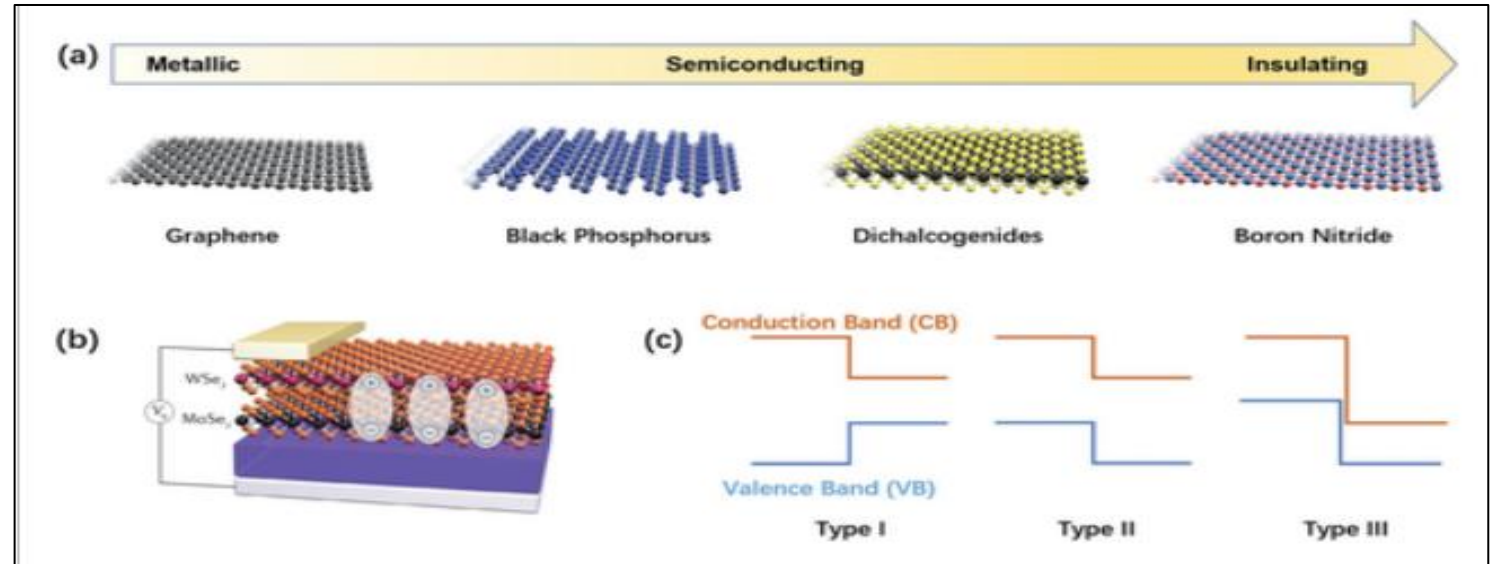
Two-dimensional atomic crystals can be aggregated into multi-layer heterostructures called van der Waals (vdW). These vdW structures can be employed as memristive devices through several possible switching devices (ways of changing resistance: conductive filament, charging-discharging grain boundary migration, ionic intercalation, and lattice phase transition). 2D materials each possess unique properties that may be ideal for certain use-cases. The fabrication of vdW heterostructures is adjustable by layer types, number of sizes, and order of layers. Many distinct vdW heterostructures can be created, each with different properties.

A graphene/MoS₂-xO_x/graphene vdW heterostructure exhibited positive memristive attributes including endurance of 10⁷ cycles, and stable performance in 340 Celsius.

The diversity of arrangements possible by 2D materials drives research to enlarge the known family of suitable 2D materials for in-memory/in-sensor computing. Another ambition of 2D material research is to reduce thickness of the internal layer to monolayer scale, allowing conductive-point resistive switching as opposed to larger conductive filament switching.

Sources: [2]

A



Developmental Challenges

Fabricated memory arrays of the larger sizes needed for in-sensor computing are beyond current fabrication capabilities. Larger synthesized areas results in greater numbers of local defects, compromising performance. Also, 2D metal, insulator, metal (MIM) arrangements for in-memory computing are limited to endurances of 10⁶.

Advances

New research of 2D materials and van der Waal aggregations is reviewing their properties of spin-orbit torque and ferroelectric polarization for in-memory neuromorphic computing. 2D materials are also being developed for use as optoelectronic synaptic devices where inputs are light stimulus, possibly opening a research pathway to the replication of human-type vision.

Organic Materials

Description

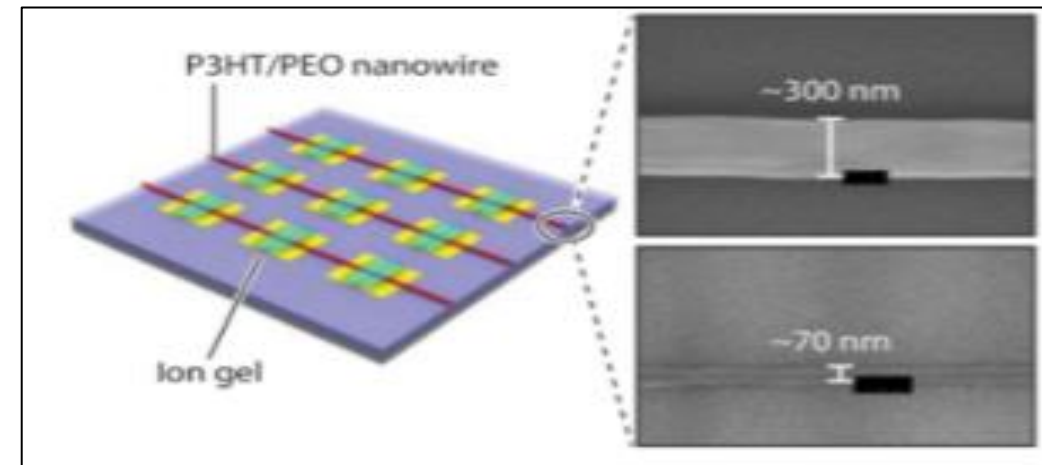
Organic semiconductors are thought to have high-potential for neuromorphic uses due to low-cost manufacturing, variety of switching mechanisms, low-power consumption, and most notably biocompatibility. Organic materials' biocompatibility opens up a range of exciting possibilities including "biointerfaces, brain-machine interfaces, and biology-inspired prosthetics."

Organic materials can replicate synaptic behaviors by filament forming, charge trapping, redox activity, and ion migration. Proof-of-concept demonstrations have shown organic synapses to be capable of replicating "synaptic weight as electrical resistance, excitatory postsynaptic potential, global connectivity, and pulse shaping"

Organic-electronics is a relatively new category of electronics, and lags its inorganic counterpart in speed, density, stability/retention, and integrability. Recent publications have shown improvements in all of these areas, yet even state-of-the-art purely experimental organic electronics has significant progress before it will have the ideal qualities for neuromorphic artificial synapses.

Source: [2]; [6]

A



Developmental Challenges

Organic semiconductor technology is new relative to inorganic alternatives and hasn't benefitted from years of effort towards realizing spatial reductions in size. As such, the current size of organic transistors is at the micrometer level, where artificial synapses should be less than 100 nms (1/10 of a single micrometer).

Advances

For biointerfaces where achievements in scaling are less important, organic artificial synapses are thought to be promising. Researchers are now envisioning demonstrations of biointegrated, high-performance computing accoutrements. Gumyusenge et. al. expect that organic based neuromorphic bio-devices will "revolutionize areas such as healthcare, entertainment, and smart textiles, to name a few."

Spintronics

Description

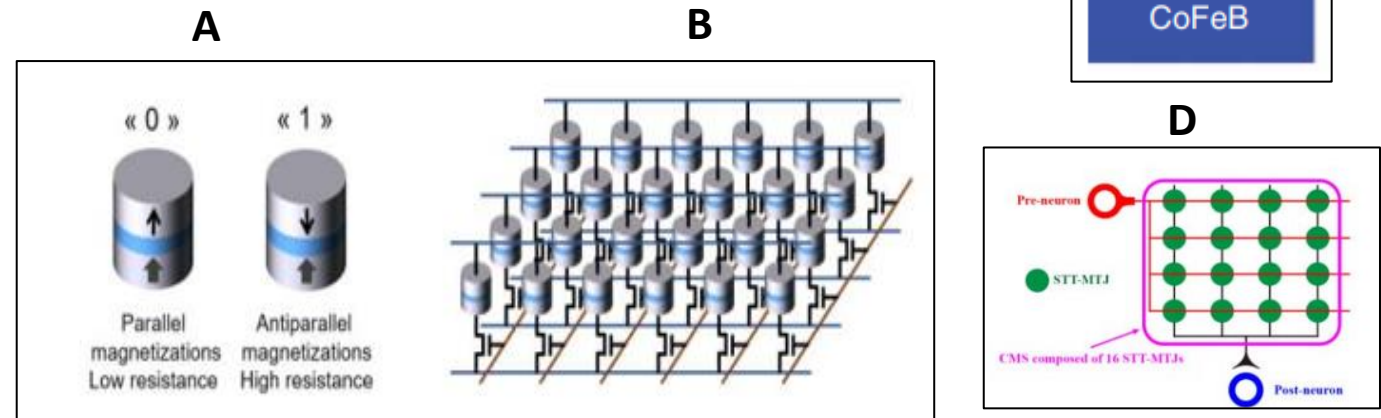
Spintronics are nanoelectronics/nano-magnets with adjustable spin and charge. Spintronics can be utilized as synapses in neuromorphic chip architecture in two ways.

- First, spintronic devices have a static magneto resistive effect; Changes to the magnetic field of spintronic devices affect resistance. The tuning/changing of the magnetic field can be employed to update synaptic weights. Magnetic tunneling junctions, spintronic memristors, and skyrmion proposals are generally based on the magneto resistive effect of spintronics.
- Spintronic's have several dynamic properties that can be used for neuromorphic implementations. "...The stochastic switch of magnetic tunneling junction, the breathing mode of skyrmion size oscillation..." (Ref #13, page 47). These dynamic properties can be used for multi-level analog neuromorphic computing.

Source: [2]; [5]; [13]

Magnetic Tunnel Junctions (MTJs)

MTJs are composed of two ferromagnetic layers containing a thin oxide layer (figure C). One of the ferromagnetic layers has a fixed direction of magnetization. The second ferromagnetic layer's magnetization is arrayed in parallel or antiparallel to the fixed layer (see figure A). Parallel composition corresponds to low resistance, and antiparallel high resistance for binary memory. Multiple binary MTJs can be combined in a network to emulate a single analog synapse (see figure D).



Developmental Challenges

- Spintronics have low resistance ratios between high resistance and low resistance when used for binary storage of between 2 to 3: other memristors have such ratios in excess of 1000
- Large write currents are needed
- Poor write endurance (number of write operations a memristor can sustain)
- STT-MTJ has stochastic switching patterns

Advances

The novel three-terminal spin orbit torque magnetic tunnel junction (SOT-MTJ) stacks a metal layer on top of the fixed magnetic ferroelectric layer. This novel configuration achieves lower write energy dissipation, and higher write energy endurance.

Neuromorphic Hardware Implementation

Either Artificial Networks or as Spiking Neural Networks

The following four slides introduce and assess each distinct neuromorphic implementation

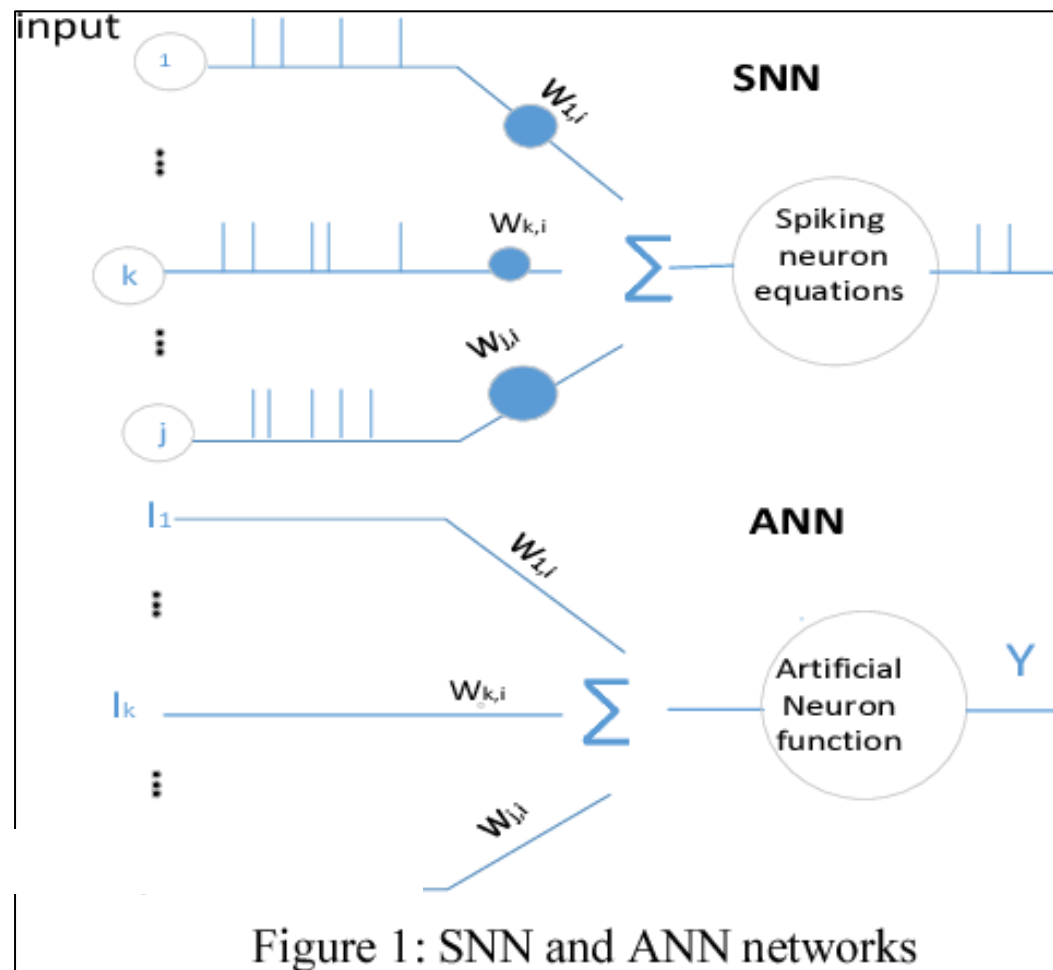


Figure 1: SNN and ANN networks

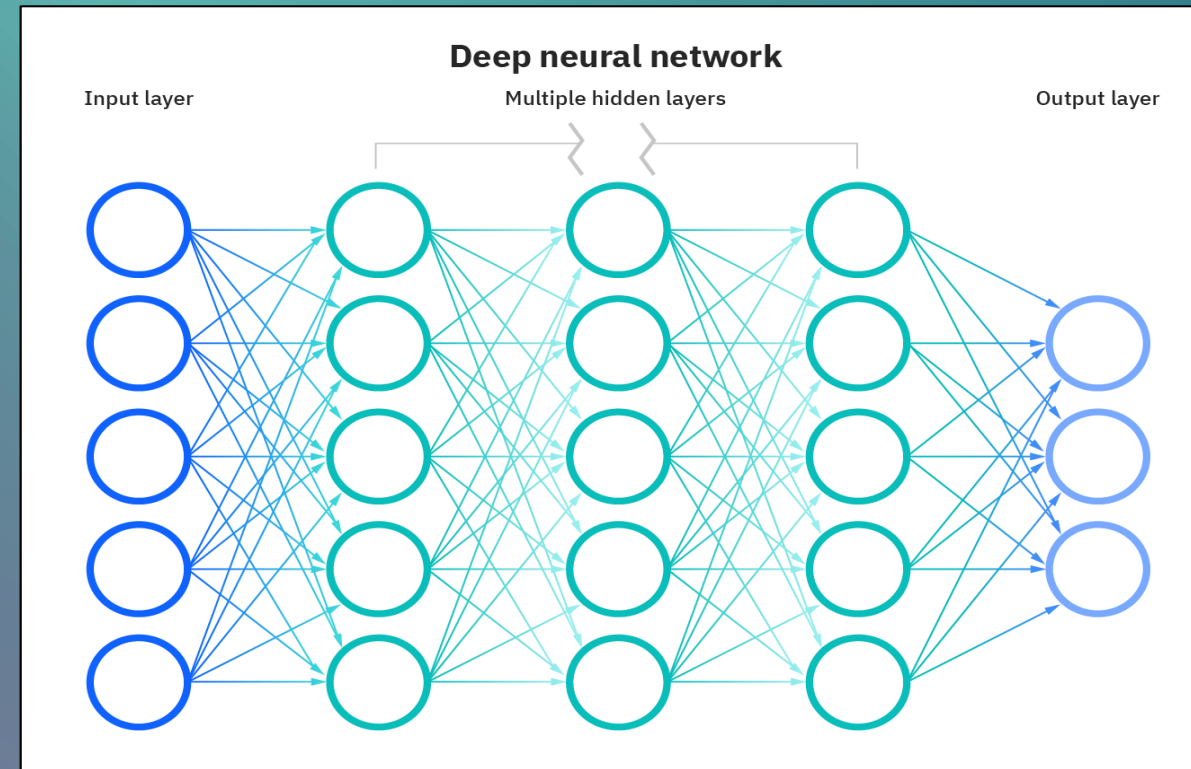
Although traditional artificial neural networks are analyzed in this report, most experts strongly believe that the future of neuromorphic computing will be realized through Spiking Neural Networks (SNNs).

Nearly all recent commercial prototypes are based on SNNs, and SNNs are lauded for their potential energy-efficiency and performance relative to traditional artificial neural networks.

Artificial Neural Networks - Introduction

Description

- Artificial Neural Networks (ANNs) are a simplified form of biological neural learning patterns. These neural networks are arrangements of neurons connected by synapses. The neurons contain activation functions which modulate the inputted values from synapses. ANNs are generally made of multiple layers, with increases in layers generally enhancing the accuracy of deployed models. Thus, the data hungry ANNs are partially responsible for the movement towards more and more data collection.
- Learning is accomplished by the application of a neural network with specified adjustable parameters to training data. The cost-function, representing model error, is progressively minimized by gradient-descent based back propagation through adjusting synapse weights. Then, a user will select a chosen model for deployment.
- Inference is the application of a trained ANN model to new data. The new data is fed into the trained model, and the trained model will make a prediction on the new data.
- The bulk of computation in training and inference is costly matrix vector multiplications. Non-Volatile-Materials are currently explored for use as an accelerator to perform matrix multiplications, speeding up deep learning.



ANN Hardware

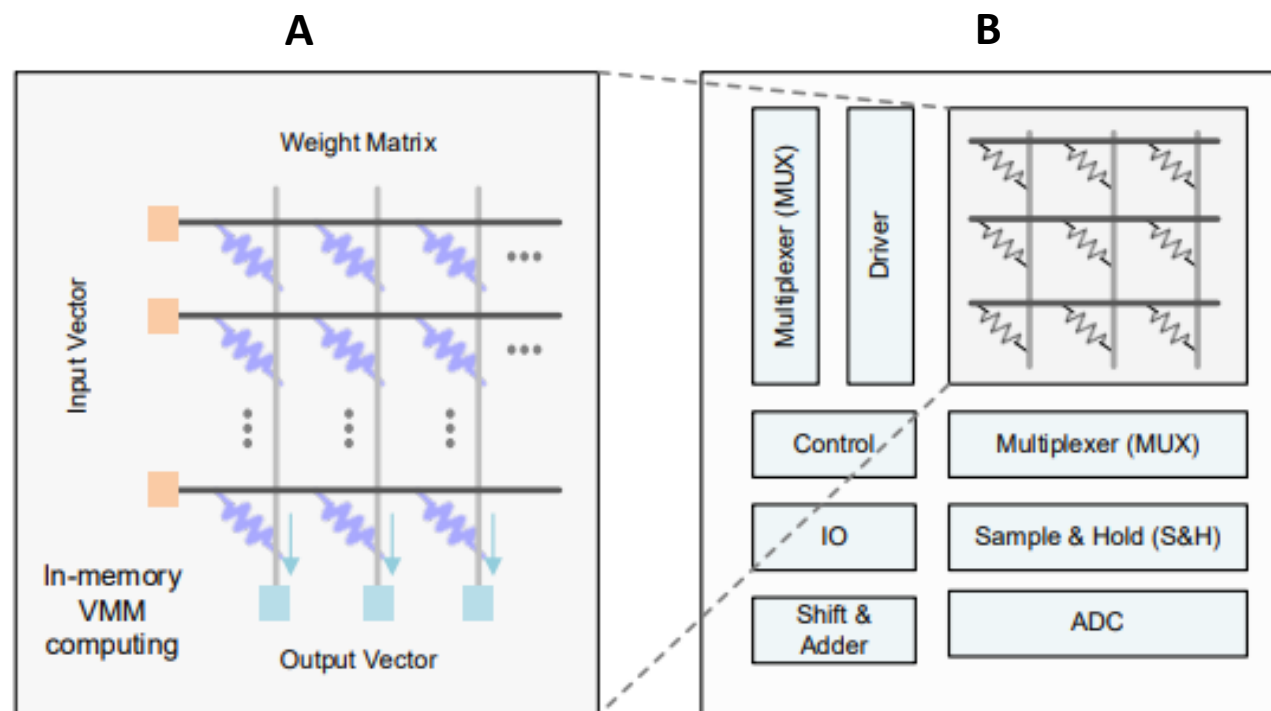
Description

The discovery of memristors in 2008 opened the possibility of memristor-based accelerators for artificial neural networks. Vector Matrix Multiplication (VMM) accounts for the bulk of computation during ANN learning and inference. Memristor Crossbars have been demonstrated to perform VMM in $O(1)$ compared to $O(n^2)$ where n is number of rows and columns (in the case of a square matrix).

For VMM the resistances of the memristors are set to match the values of the matrix to be used in the multiplication. The input vector is “multiplied” by the crossbar matrix through inputting voltage into the Crossbar structure. The interactions between the input voltage and memristors produce an output current, which is translated through the application of Ohm’s Law and Kirchoff’s law into an output vector. Each computation occurs through the electrical interactions at the data point, which are computed in parallel.

Crossbars can be configured for either analogue or binary storage. Analogue crossbars require Digital Analog Converters (DACs) to modify digital inputs to analogue and Analog Digital Converters (ADCs) to modify analogue outputs to digital, consistent with modern computer data structures.

Sources: [2]; [14]



Developmental Challenges

- The non-idealities of current non-volatile-material (NVM) memory candidates pose difficulties for the accurate programming of NVM synapse weights.
- With continued inference, the conductance values of NVM synapses will evidence drift and degrade accuracy
- Efforts to maintain 16 bit fixed-point precision strain NVMs with requiring 2^{16} unique resistance values
- Inefficiency of peripheral circuitry, especially analog-digital-converters (see figure B)

ANN Hardware

Description

The discovery of memristors in 2008 opened the possibility of memristor-based accelerators for artificial neural networks. Vector Matrix Multiplication (VMM) accounts for the bulk of computation during ANN learning and inference. Memristor Crossbars have been demonstrated to perform VMM in $O(1)$ compared to $O(n^2)$ where n is number of rows and columns (in the case of a square matrix).

For VMM the resistances of the memristors are set to match the values of the matrix to be used in the multiplication. The input vector is “multiplied” by the crossbar matrix through inputting voltage into the Crossbar structure. The interactions between the input voltage and memristors produce an output current, which is translated through the application of Ohm’s Law and Kirchoff’s law into an output vector. Each computation occurs through the electrical interactions at the data point, which are computed in parallel.

Crossbars can be configured for either analogue or binary storage. Analogue crossbars require Digital Analog Converters (DACs) to modify digital inputs to analogue and Analog Digital Converters (ADCs) to modify analogue outputs to digital, consistent with modern computer data structures.

Sources: [2]; [14]

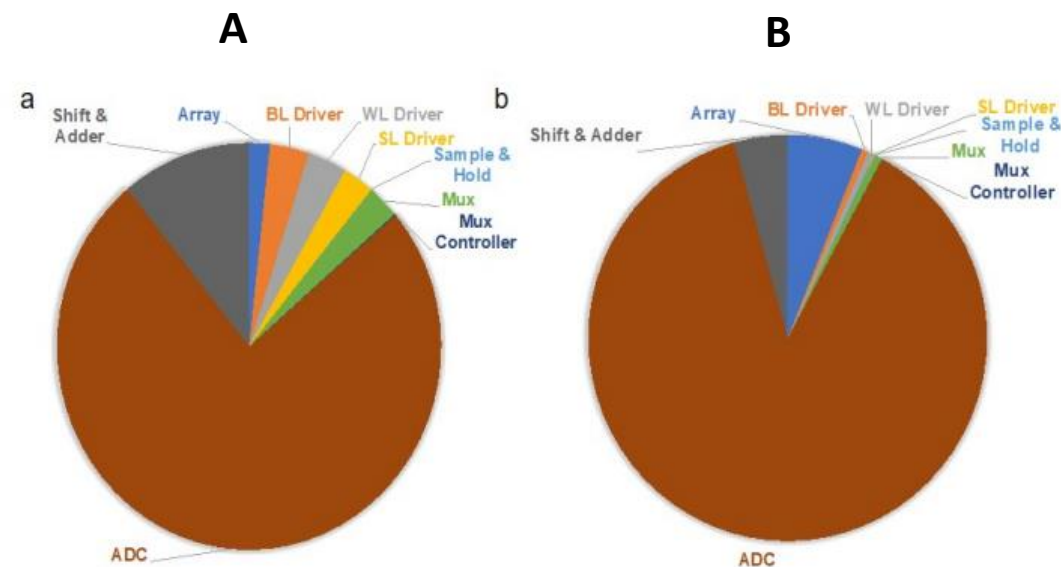


Figure 2. The breakdown of area and power consumption in a macro-circuitry instance [11]. (a) Area overhead. (b) Power overhead.

Advances

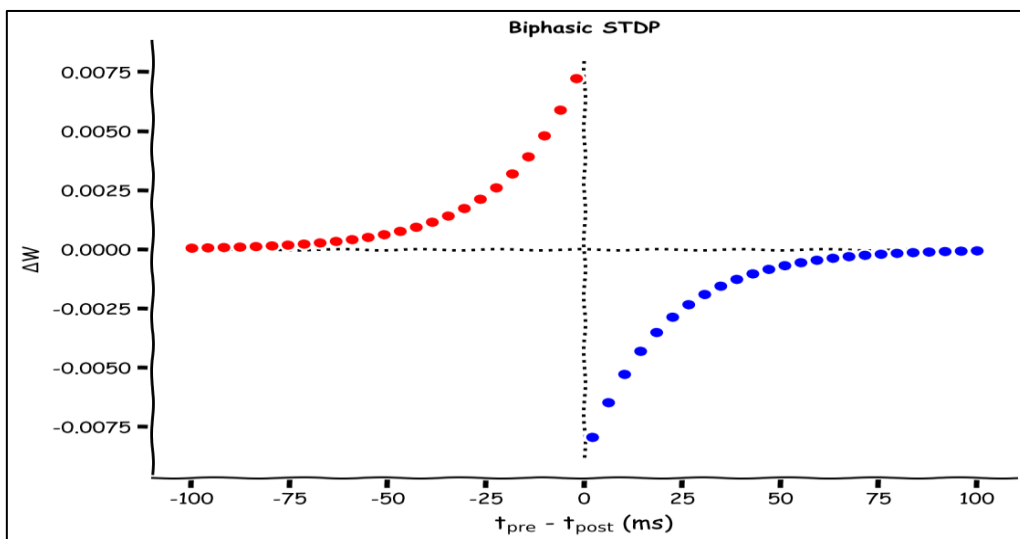
- On-chip learning and closed-loop conductance monitoring have been implemented to account for issues with NVM synapse resistance modulation
- On-chip calibration has been introduced to monitor and correct drift
- Multiple memristors are coupled together to store a single synaptic weight
- Binarized, high-accuracy networks are in development to exploit NVM’s excellent binary storage capabilities

Spiking Neural Networks - Introduction

Description

Spiking Neural Networks (SNNs) are a form of machine learning composed of synapses and neurons which can be used for supervised and unsupervised learning. Researcher interest in SNNs is driven by their more faithful replication of neural biological behavior with many neuromorphic researchers believing that large SNN models will outperform ANNs. For a comparison of SNNs to ANNs see figure B

SNNs are defined by neural networks where individual neurons communicate through spikes/impulses via synapses. The synapses which causatively contribute to neuronal spikes are modified to have higher weights, with non-causative synapses experiencing weight reductions (see figure A). Neurons fire when their membrane potential reaches a certain threshold.
Sources: [2]; [14]



B

Type of Neural Network	Neuronal Activity	Neuronal Memory	Synapse Adjustment	Function of Time
ANN	Real-valued activation functions which condition input values received from synapse(s). Common activation functions are ReLU, ELU, and the Sigmoid Function	Usually, neurons do not have memory	Synapses are usually adjusted based on stochastic backpropagation to minimize a cost function when applied to training data	No
SNN	Neuron's have adjustable states (membrane potential) and modulate received synapse spikes according to state, passing spikes on to other neurons when membrane potential is sufficiently high	The Neuron's membrane potential is memory and encodes information of past frequency of activations	Synapses are potentiated (given greater weight) if their spike preceded a neuronal spike: if their spike assisted in developing the neuronal spike. Synapses experience depression if their pulse follows the neuronal spike.	Yes, neuronal firing is temporal/ event-based

Spiking Neural Networks – Hardware

Description

Nanotechnologies are appealing for neuromorphic computing for their high-density, and excellent power efficiency: Too, their combination of computation and memory is a more faithful approximation of biological neural performance than digital circuit, CMOS SNN hardware.

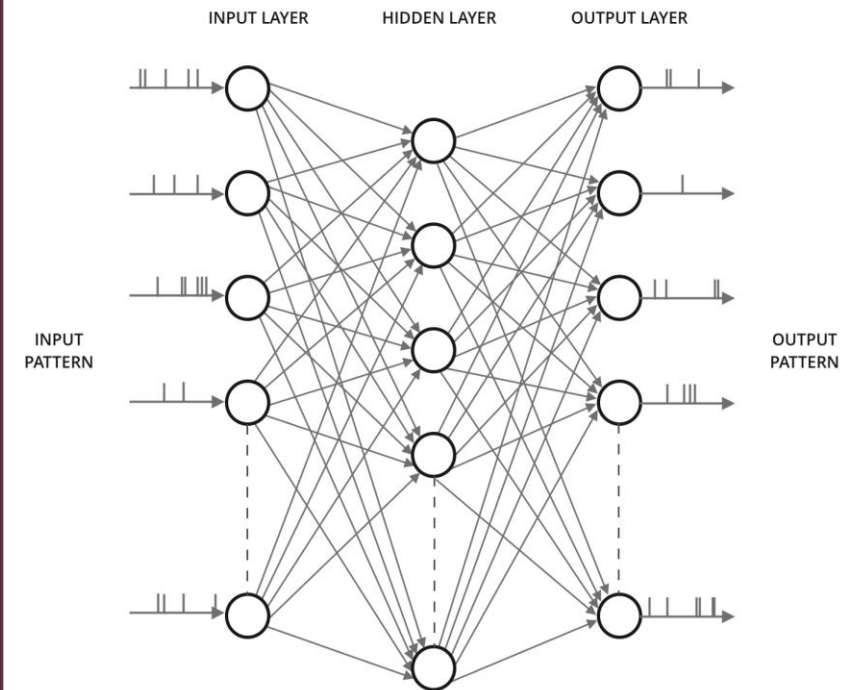
SNN nanotechnology hardware are implemented as “large crossbar arrays of synapse circuits that represent at the same time the site of memory and of computation. The synapses in each row of these arrays are connected to Integrate-and-Fire (I&F) soma circuits, located on the side of the array. The soma circuits sum spatially all the weighted currents produced by the synapses, integrate them over time, and produce an output pulse (spikes) when the integrated signal crosses a set threshold. In turn the synapses are typically stimulated with input spikes, and convert the digital pulse into a weighted analog current. Depending on the complexity of the synapse and soma circuits, it is possible to design systems that can exhibit complex temporal dynamics...or to implement adaptive and learning mechanisms that can be used to ‘train’ the network to carry out specific tasks.”

Focus on Nanotechnologies

There exist several successful developments of SNN hardware using conventional CMOS technology (TrueNorth, Loihi, BrainScaleS, etc.). These case-studies will be profiled elsewhere. These hardware implementations of SNNs have two weaknesses stemming from their storage of synaptic weights in DRAM/SRAM:

1. Accessing and updating synaptic weights has a non-insignificant time cost
2. The current DRAM/SRAM-CMOS integration density inhibits scalability of neuromorphic systems
 - A. Achieving a neuromorphic computer with the number of synapses and neurons of the human brain would take up an entire room

As such, this section will focus on SNNs based on dense emerging nanotechnologies.



Challenges

- Local-learning rules of SNNs aren't yet shown to match the high-accuracy of ANNs backpropagation
- Though SNNs are generally robust to device imperfections, current memristor's substantial non-idealities still compromise performance
- Online learning demands memristors with varied time scales

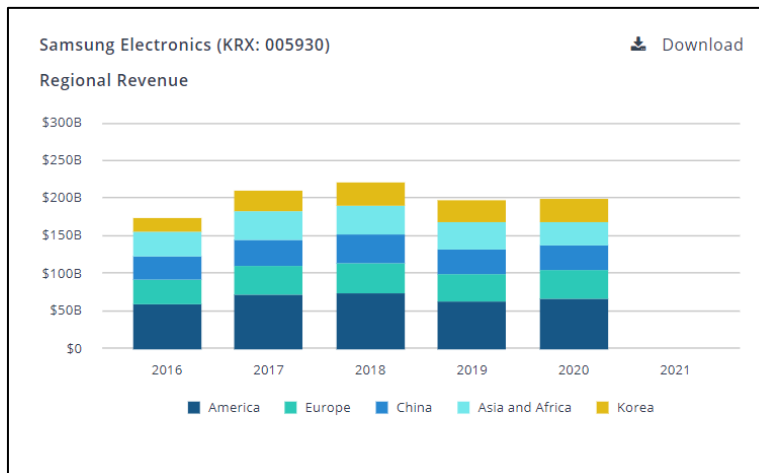
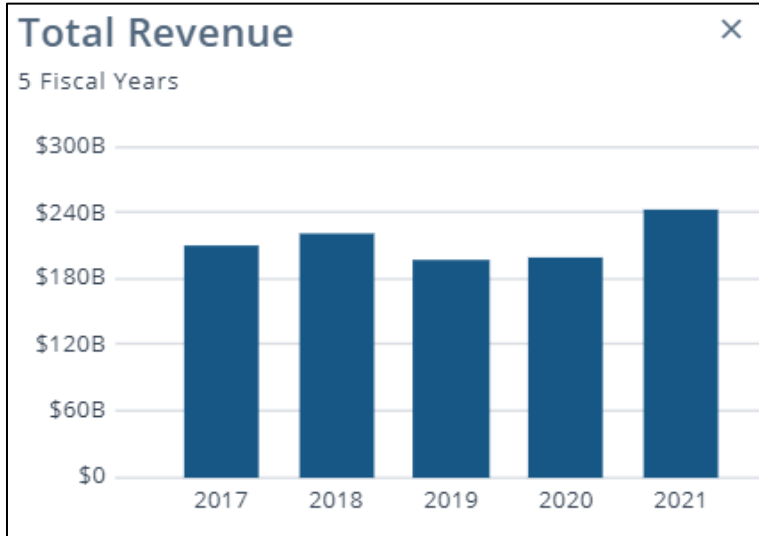
Key Players: Samsung

Research Analysis

The Samsung Advanced Institute of Technology (SAIT) is actively researching innovations in hardware to empower machine learning with enhanced power efficiency. Their research areas include near/in memory computing, asynchronous spiking neural networks, brain-inspired learning and inference algorithms, low-power mixed signal computing architectures, and new synaptic memories.

SAIT constructed the first in-memory hardware based on Magnetoresistive Random Access Memory. MRAM was previously considered an unattractive material for its low resistances. Samsung researchers adjusted in-memory computer architectures from 'current-sum' to 'resistance-sum' in response to MRAM low resistances. The MRAM computing chip was ran AI computing with 98% accuracy for classifying hand-written digits and 93% accuracy for facial recognition.

Sources: [4]; [20]; Pitchbook Data

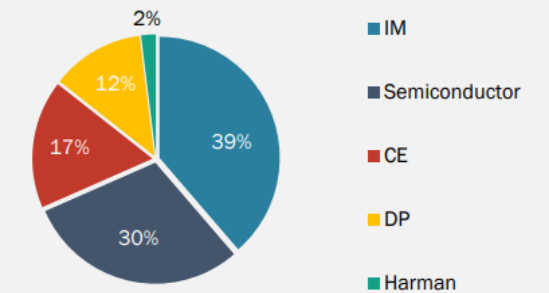


SAIT Neuromorphic Luminaries



Dr. Donhee Ham, SAIT Fellow & Harvard Professor | Dr. Seungchul Jung, SAIT Staff Researcher | Dr. Sang Joon Kim, SAIT VP of Technology

BUSINESS REVENUE MIX, 2017



Abbreviations Taxonomy

IM: Information Technology and Mobile Communications, CE: Consumer Electronics, DP: Display Business

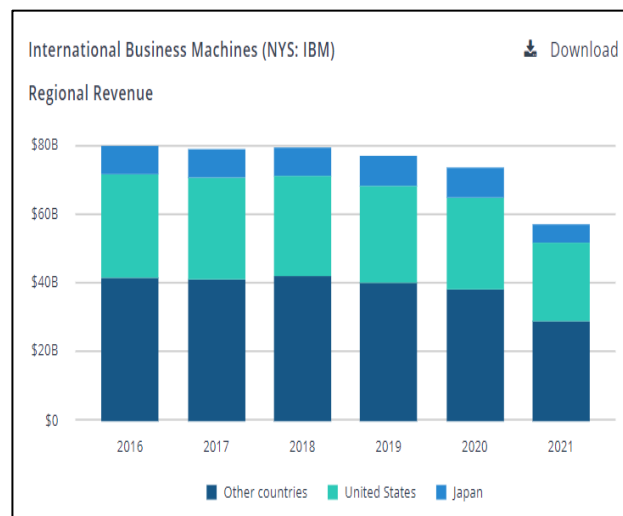
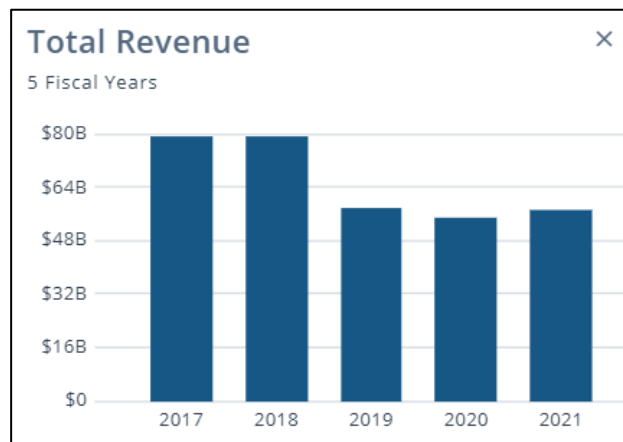
Key Players: IBM

Research Analysis


IBM was an early entrant to the field of Neuromorphic Computing, receiving funding from as early as 2008 from the DARPA SYNAPSE program to iteratively construct and improve neuromorphic chips. Early models were “Golden Gate”, and “San Francisco” (~256 neurons). IBM’s later “TrueNorth”, first created in 2014, contains 1M neurons. IBM’s 2018 NS16e-4 scale-up scale-out system scaled to 64M neurons. These neuromorphic chips are trained off-chip (on-chip training is possible and has been implemented in other models). Trained neural networks are loaded into the chip, which then performs inference. IBM has progressively been designing development / programming environments for TrueNorth (Compass Simulation Environment was the first iteration).

IBM’s TrueNorth chip is composed of CMOS technology, rendering it with a low-density and high-manufacturing cost. Exploration of non-CMOS synapse material is important to densify and reduce costs of neuromorphic chips.


IBM’s Zurich Lab is exploring many topics in neuromorphic computing in addition to IBM’s Almaden TrueNorth research initiative. According to our expert consultations, at present, IBM has discontinued its work on TrueNorth, but continues to research memristor materials at its Zurich Lab.



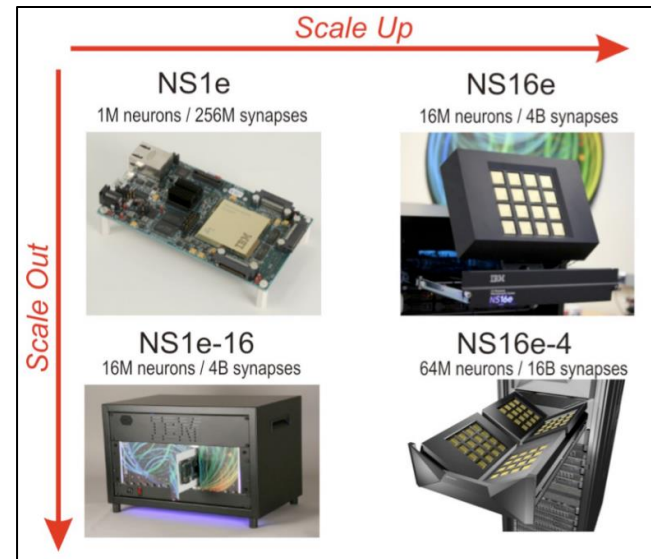
IBM Neuromorphic Luminaries



Dr. Bert Jan Offrein
Principal Research
Staff Member



Dr. Evangelous
Eleftheriou
IBM Fellow



Key Players: Intel

Research Analysis

Intel is prominent in neuromorphic computing research, having launched one of the first neuromorphic chips and founded the Neuromorphic Research Community, a network of 100+ research groups.

Intel's Loihi was created in 2017 and revamped to Loihi 2 in 2020. Loihi is based on a spiking neural network (SNN) architecture with 1M neurons and 120M synapses per chip that is constructed for on-chip learning. The 128 neuromorphic cores each contain interlinked neurons and receive spike impulses. Intel provided Nx SDK software to program Loihi through low-level programming. Intel's new offering is Lava, a higher-level programming language with enhanced accessibility that allows users to engage with Loihi 2 without detailed knowledge of the hardware implementation. Loihi 2 and Lava are made available to members of Intel's Neuromorphic Research Community to develop new use-cases and commercialize neuromorphic technologies.

Demonstrated use-cases for Loihi and Loihi 2 include:

- Robot Arm Control
- Visual-tactile sensory perception
- Odor Recognition
- Database similarity query
- Optimization problems
- Modeling Scientific Diffusion

Table 2. Comparison of Loihi to Loihi 2

Resources/Features	Loihi	Loihi 2
Process	Intel 14nm	Intel 4
Die Area	60 mm ²	31 mm ²
Core Area	0.41 mm ²	0.21 mm ²
Transistors	2.1 billion	2.3 billion
Max # Neuron Cores/Chip	128	128
Max # Processors/Chip	3	6
Max # Neurons/Chip	128,000	1 million
Max # Synapses/Chip	128 million	120 million
Memory/Neuron Core	208 KB, fixed allocation	192 KB, flexible allocation
Neuron Models	Generalized LIF	Fully programmable
Neuron State Allocation	Fixed at 24 bytes per neuron	Variable from 0 to 4096 per neuron depending on neuron model requirements
Connectivity Features	Basic compression features: <ul style="list-style-type: none"> • Variety of sparse and dense synaptic compression formats • Weight sharing of source neuron fanout lists 	In addition to the Loihi 1 features: <ul style="list-style-type: none"> • Shared synapses for convolution • Synapses generated from seed • Presynaptic weight-scaling factors • Core fan-out list compression and sharing • Broadcast of spikes at destination chip
Information Coding	Binary spike events	Graded spike events (up to 32-bit payload)
Neuron State Monitoring (for development/debug)	Requires remote pause and query of neuron memory	Neurons can transmit their state on-the-fly
Learning Architecture	Programmable rules applied to pre-, post-, and reward traces	Programmable rules applied to pre-, post-, and generalized "third-factor" traces
Spike Input	Handled by embedded processors	Hardware acceleration for spike encoding and synchronization of Loihi with external data stream
Spike Output	1,000 hardware-accelerated spike receivers per embedded processor	In addition to the Loihi 1 feature, hardware accelerated spike output per chip for reporting graded payload, timing, and source neuron
External Loihi Interfaces	Proprietary asynchronous interface	Support for standard synchronous (SPI) and asynchronous (AER) protocols, GPIO, and 1000BASE-KX, 2500BASE-KX, and 10GBase-KR Ethernet
Multi-Chip Scaling	2D tile-able chip array Single inter-chip asynchronous protocol with fixed pin-count	3D tile-able chip array Range of inter-chip asynchronous protocols with variable pipelining and pin-counts optimized for different system configurations
Timestep Synchronization	Handled by cores	Accelerated by NoC routers

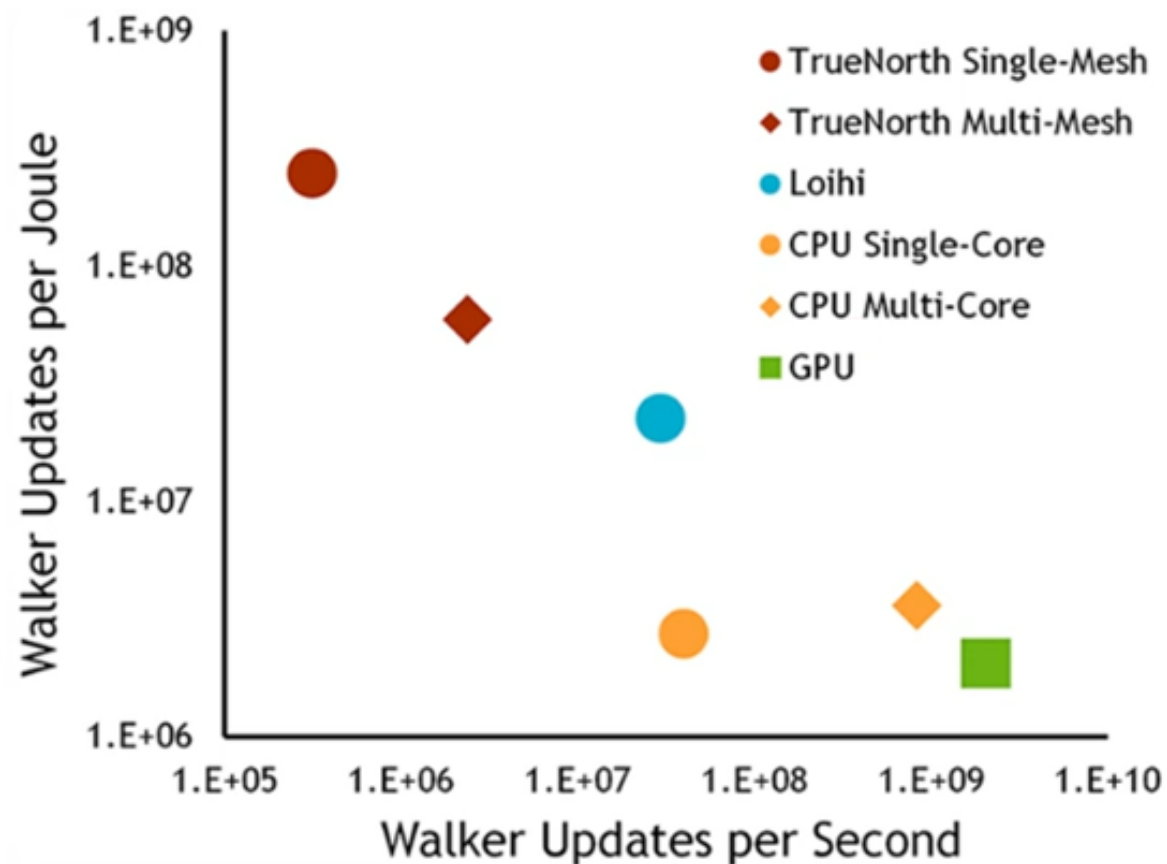
IBM and Intel Neuromorphic Processors

Demonstrate superior energy efficiency

IBM's TrueNorth was constructed mainly for energy-efficiency

Intel's Loihi (and Loihi 2) strives for a balanced mix of energy efficiency and computational speed

Loihi and TrueNorth experience improved speed/performance relative to CPU/GPUs for target applications suited for neuromorphic computing



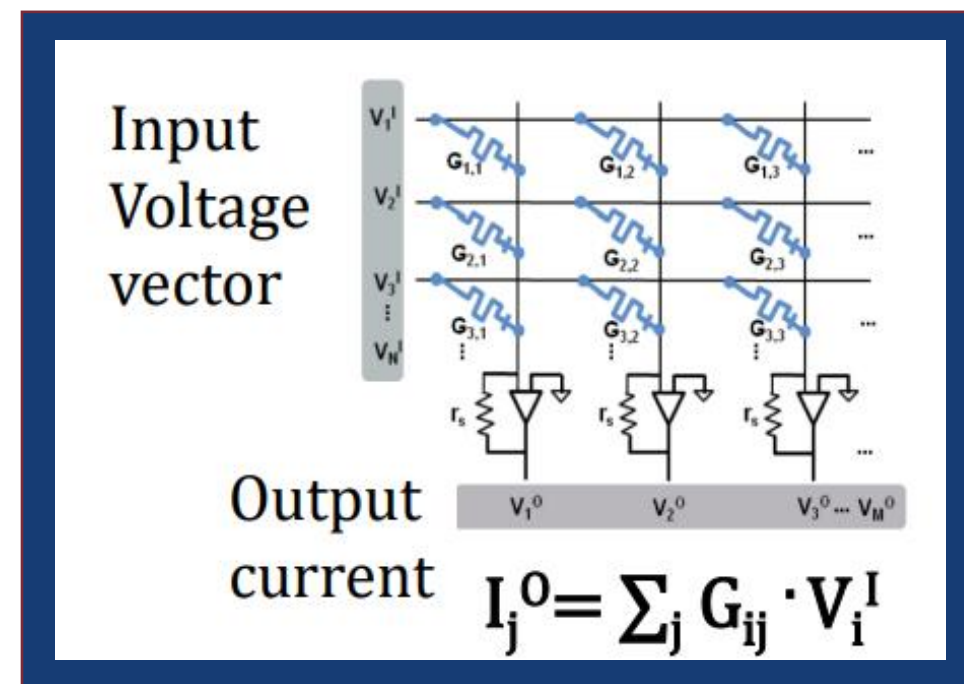
Key Players: Hewlett Packard

Research Analysis

Hewlett Packard is divided into four operating segments. The Hybrid IT segment is responsible for neuromorphic research and development. Hewlett Packard joined the Open Neural Network Exchange in 2018 to assist its ongoing research in neuromorphic computing. Hewlett Packard labs has been active in neuromorphic computing since its discovery of the memristor in 2008 during analysis of titanium dioxide.

Hewlett Packard and Dr. Miao Hu developed a Dot-Product Engine (DPE) in Hewlett Packard Labs. The DPE is an experimental product that performs matrix multiplications in parallel on memristor-based crossbars. The DPE uses the TiO_2 memristor. The DPE was able to tune each memristor to target conductance, and “do more than eight thousands of multiplications per time step in a tiny 128x64 memristor crossbar.” The project also achieved over 90% on MNIST (a popular dataset for testing deep learning algorithms). Two publications on the DPE have been published in Advanced Material and Nature Electronics.

Since its development of the DPE, Hewlett Packard has had minimal activity in the area of neuromorphics. A 2021 news release discussed the potential of neuromorphics and may signal renewed focus on neuromorphic technology.



Key Players: HRL Laboratories

Research Analysis

A 2013 HRL neuromorphic project was the creation of a continuously security authenticating edge device, for which HRL received \$2.2M in funding from the U.S. Department of Homeland Security Science and Technology Directorate.

HRL Laboratories received funding from DARPA's Foundations Required for Novel Compute (FRANC) program to develop a memristor based neuromorphic processor. Dr. Wei Yi was the Principal Investigator on the project, and the primary author of the nature journal article "Biological plausibility and stochasticity in scalable V02 active memristor neurons", which was the direct product of the HRL neuromorphics project. The HRL neuromorphic circuits contained 500 memristor synapses and 60 neurons. Moreover, the circuit "can compute convolutions for image classifications – the main workload of today's AI inference hardware", with energy efficiency "10 times better than the current state-of-the-art neuromorphic processor". Dri Yi's journal article has been cited 171 times and resulted in his elevation as an IEEE senior member.

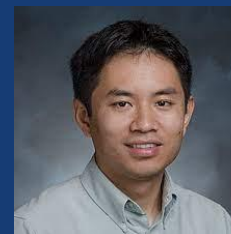
The last listed neuromorphic project in HRL's archives is dated in 2019.

Sources: [30];

HRL Neuromorphic Luminaries



Dr. Dana Wheeler
Researcher



Dr. Wei Yi
Principal Investigator



Dr. Jose Cruz-
Albrecht
Co-Principal
Investigator

HRL is notable for sponsoring
Neuromorphic Projects based on
memristor and CMOS technology.

Key Players: BrainChip Holdings

Research Analysis

Australian company BrainChip Holdings, founded in 2011, is wholly dedicated to neuromorphic computing. BrainChip holdings spent about a decade after its inception developing neuromorphic System-on-Chip (NSoC) systems. The Akida chips was the result, a neuromorphic CMOS chip designed for edge use-cases. Akida has been demonstrated to successfully complete object classification and keyword spotting, among other neural network applications. Currently, Akida chips are produced by TSMC in Taiwan.

BrainChip is presently attempting to transition from its R&D stage to the commercialization of Akida. BrainChip's website proudly declares that Akida "is the world's first commercial neuromorphic processor." As of 2021 BrainChip was seeking "Early Access Customers for further testing and verification to see how they work as part of customers' products." So, Akida is not yet at full-commercialization. Additionally, in 2022 "BrainChip licensed its Akida IP to ASIC industry heavyweights MegaChips and Renesas to help enhance and grow their technology positioning for next-generation, cloud independent AI products".

Sources: [31], [32];

A

Portfolio Breakdown ⓘ

Technology Citations

Citations are defined as the number of unique patents that have made reference to the company's patent. Citations are often used as a signal of patent value.

Most Cited Patents

All Years Last 5 Years

Patent Title	Forward Citations
1 Autonomous learning dynamic artificial neural computing device and brain inspired system	32
2 Method and a system for creating dynamic neural function libraries	30
3 Low power neuromorphic voice activation system and method	24
4 Autonomous learning dynamic artificial neural computing device and brain inspired system	22
5 Low power neuromorphic voice activation system and method	21

See Figure A above for information on BrainChip IP. Sources: Pitchbook Data;

Key Academic Research Labs

Lab Name	Description	Affiliated Faculty / Researchers
California NanoSystem's Institute	As computational tasks become increasingly difficult in a world of big data, systems to address modern challenges in collection, processing, and analysis of large datasets are increasingly necessary. Researchers at CNSI are attacking the challenges of next-generation computing by combining concepts of neuroscience and machine learning with nanoscale materials. An exemplar of this approach are James Gimzewski and Adam Stieg, who are developing complex nanoarchitectures that have structural similarity to neocortex and exhibit properties which make them an ideal platform to address the difficulty of mimicking biological neural networks in artificial computing environments	Dr. Yong Chen; Dr. James Gimzewski; Dr. Adam Stieg; Dr. Kang Wang
Carnegie Mellon: Neuromorphic Computer Architecture Lab	The Neuromorphic Computer Architecture Lab (NCAL) is a new research group in the Electrical and Computer Engineering Department at Carnegie Mellon University, led by Prof. John Paul Shen and Prof. James E. Smith to design new processor architectures that capture the capabilities and efficiencies of the brain's neocortex for energy-efficient, edge-native, on-line, sensory processing in mobile and edge devices.	Prof. John Paul Shen; Prof. James E. Smith
Institute of Neuroinformatics	The Institute of Neuroinformatics was established at the University of Zurich and ETH Zurich in 1995. The mission of the Institute is to discover the key principles by which brains work and to implement these in artificial systems that interact intelligently with the real world.	Prof. Tobi Delbruck; Prof Benjamin Grewe; Prof. Richard Hahnloser; Prof Giacomo Indiveri;
Intelligent Computing Lab - Yale	Yale's Intelligent Computing Lab is led by Prof. Priyadarshini Panda to research neuromorphic computing. The lab's research spans developing novel algorithms and new architectures (CMOS based and w/ emerging nanotechnologies). The lab has received funding from Amazon, the NSF, the Center for Brain Inspired Computing, and DARPA.	Prof. Priyadarshini Panda
Neuromorphic Artificial Intelligence Lab - RIT	Over the past decade, research in the Nu.AI lab is paving a path to revolutionize the next generation of intelligent platforms. Our short term goal is to develop lifelong learning systems that utilize minimal resources (e.g.: energy, form factor). We strive to achieve this through an interdisciplinary research approach.	Prof. Dhireesha Kudithipudi; Dr. Eric Bohannon;
Neuromorphic Computing Lab at Pennsylvania State	We believe achieving machine intelligence with brain-scale efficiency will be enabled by an end-to-end research effort ranging from sensory processing to neuromimetic hardware and associated learning methodologies. To that end, we are driven by a highly interdisciplinary perspective across the computing stack that combines knowledge from devices and circuits to machine learning and computational neuroscience.	Dr. Amit Shukla; Kezhou Yang; Sen Lu; Nafiu Islam
Salleo Research Group	The Salleo Research Group is interested in novel materials and processing techniques for large-area and flexible electronic/photonics devices. We also study defects and structure/property relations of polymeric semiconductors, as well as nano-structured and amorphous materials in thin films.	Prof. Alberto Salleo; Dr. Alexander Giovannitti; Dr. Quentin Thiburce

Key Startups

Company	Description	Year Founded	HQ	Pitchbook Valuation [USD Millions]	Number of Active Patents
Applied Brain Research	Developer of low-power edge and cloud AI applications and development tools designed to make the same.	2014	Ontario, Canada	\$.97M as of 2014	8 Active 13 Pending
FutureAI	Operator of a deep technology company intended to develop algorithms to revolutionize artificial intelligence.	NA	Washington D.C., USA	NA	NA
General Vision	Developer of embedded artificial vision silicon chips designed to permit everyday objects to have visual perception of their environment and interact with their owner.	1987	Petaluma, USA	NA	NA
GrAI Matter Labs	Developer of programmable neuromorphic computing chips designed to bring sensor analytics and machine learning to every device on the edge.	2016	Paris, France	Early stage VC deal in 2020 of \$14M for unknown ownership transfer	6 inactive 6 pending
Knowm Inc	Developer of digital computing processors designed to commercialize AHaH Computing and the neuromemristive technology stack.	2015	Santa Fe, USA	NA	NA
Numenta	Developer of a software-based neocortical theory designed to study biological learning principles. The company's platform offers an HTM (Hierarchical Temporal Memory).	2005	Redwood City, USA	\$351M as of 2021	28 active 4 pending

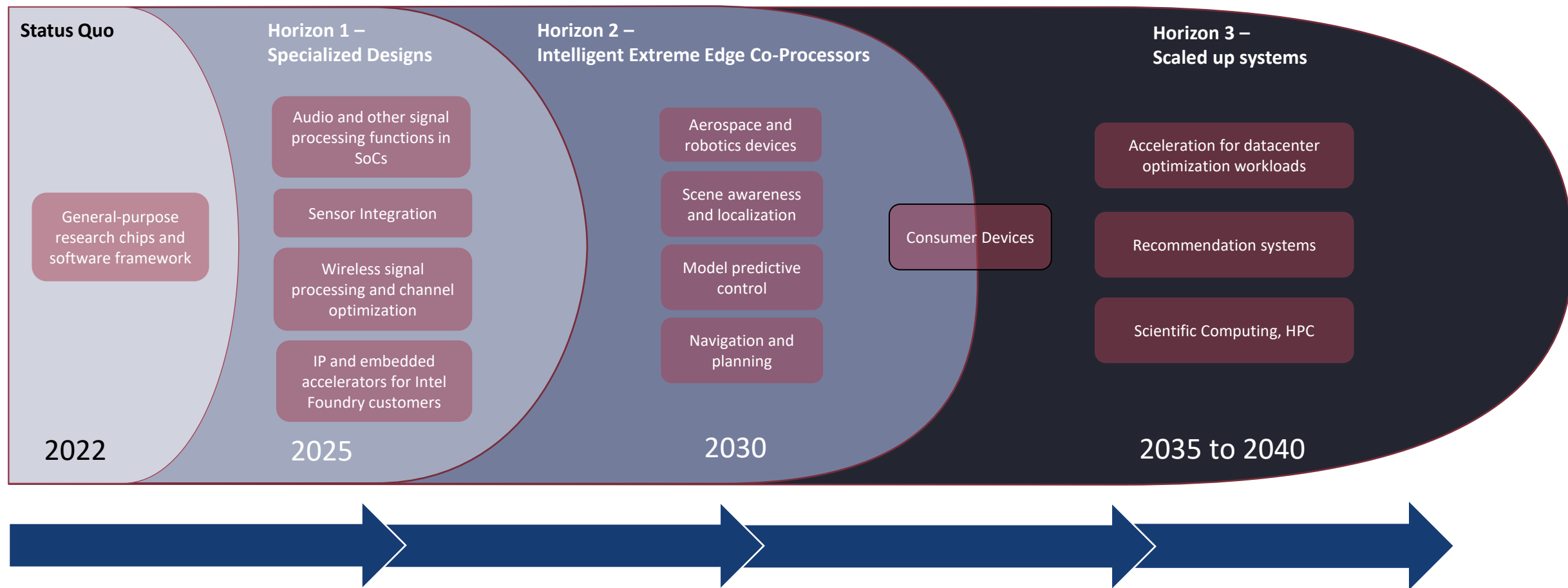
Sources: Pitchbook Data; Company Websites;

Key Startups

Company	Description	Year Founded	HQ	Pitchbook Valuation [USD Millions]	Number of Active Patents
IniLabs	Provider of neuromorphic technologies solution designed to promote neuromorphic engineering for the worldwide community of technology developers.	2009	Zurich, Switzerland	NA	NA
Innatera Nanosystems	Manufacturer of ultra-efficient neuromorphic processors intended to mimic the brain's mechanisms for processing sensory data.	2018	Rijswijk, Netherlands	NA	1 Pending
Koniku	Developer of an organic neurocomputation platform designed to interface with the real world through real biological neurons.	NA	San Rafael, USA	\$256M	1 Active 5 Pending
MemComputing	Developer of quantum computing technology designed to accelerate the optimization of complex and time-consuming problems.	2016	San Diego, USA	NA	NA
SpiNNcloud Systems	Developer of real-time computing platform designed to leverage research from the human brain project.	2021	Dresden, Germany	NA	NA
SynSense	Developer of neuromorphic computing platform designed to provide a combination of ultra-low power consumption and low latency performance.	2017	Zurich, Switzerland	\$30.89M as of 2021	NA
Vicarious	Developer of an AI-based neuro and cognitive science-based robot technology designed to mimic the function of the human brain.	2010	Union City, USA	\$400M as of 2017	10 active 4 pending

Sources: Pitchbook Data; Company Websites;

Neuromorphic Computing Applications Horizon Mapping



Sources: Expert Consultations; Intel Neuromorphic Research Conference, 2022

Neuromorphic Computing Target Application Qualities

Based on Neuromorphic Computing's Contemporary Status

An Ideal Problem to be solved by
Neuromorphic Computing

Power
Constrained

Latency
Constrained

Processes
real-time
Signals

Slowly
evolving
structure

Benefit from
shallow
online
learning

Apply deep
learning for
offline
training

With continued evolution and growth, neuromorphic target applications requirements will become less restrictive

Target Application: Robotics

Description

Neuromorphic hardware is considered a key innovation for the development of intelligent robots that can operate in dynamic, uncontrolled environments. The Robotics field is an important application for neuromorphic computing as there exists substantial demand for robots capable of safely operating in a variety of conditions with minimal training.

Early proof-of-concept neuromorphic robotics demonstrations may give way to sophisticated developments built off standardized software, dynamic vision, mature neuromorphic chips, and spiking neural network architectures.

Sources: [2]; Neura Robotis Website; Intel Neuromorphic Research Conference, 2022;

Remaining Challenges

- Remarkable hardware plasticity is a requirement for neuromorphic robotics
- The vision of autonomous, quickly-adapting robots is based on a level of intelligence that will require massive neural networks in terms of # of synapses & neurons – neuromorphic robotics will require high memory/computing densities

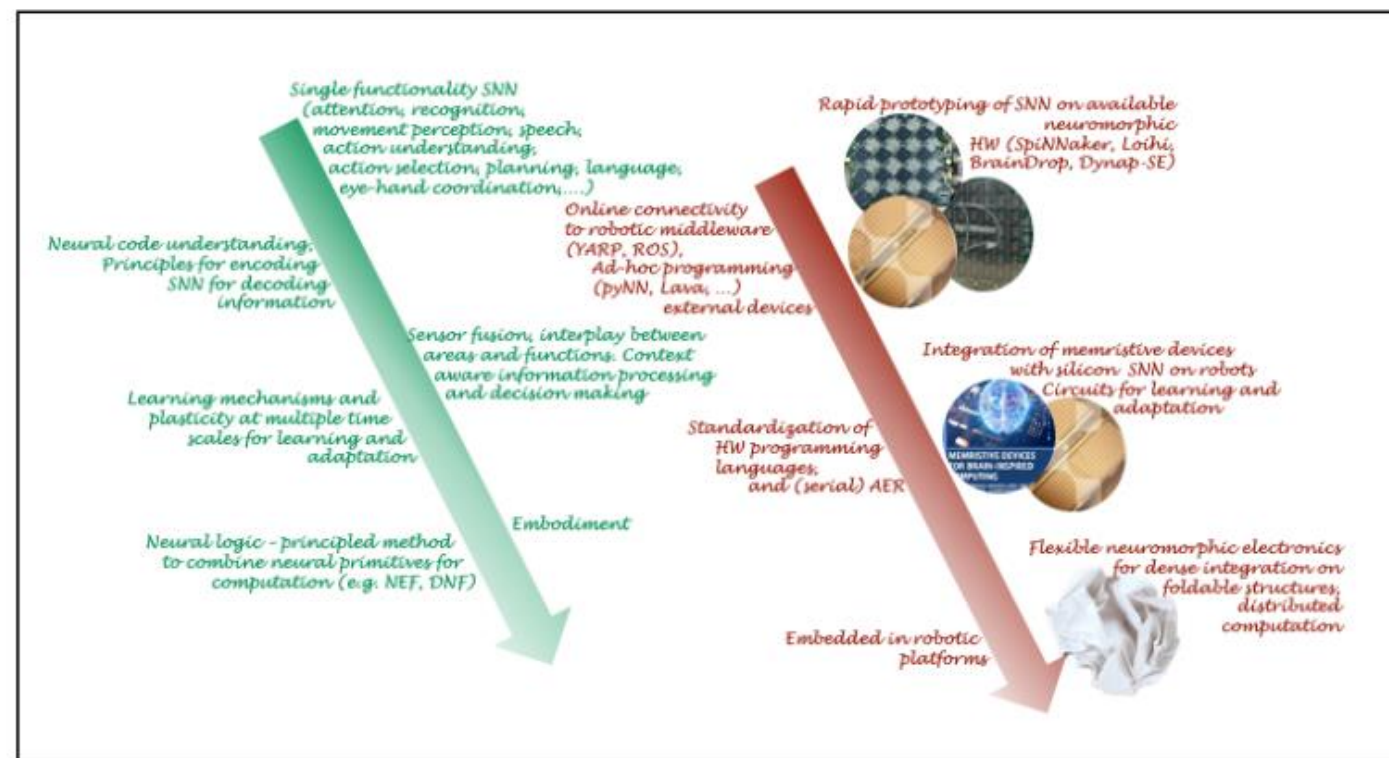


Figure 2. Timeline of a possible development roadmap. In green required theoretical advancements in order of increasing complexity. In red the technological roadmap highlighting the path for new circuit and devices development, as well as the infrastructure needed for the integration on robotic platforms.

Target Application: Robotics

Description

Neuromorphic hardware is considered a key innovation for the development of intelligent robots that can operate in dynamic, uncontrolled environments. The Robotics field is an important application for neuromorphic computing as there exists substantial demand for robots capable of safely operating in a variety of conditions with minimal training.

Early proof-of-concept neuromorphic robotics demonstrations may give way to sophisticated developments built off standardized software, dynamic vision, mature neuromorphic chips, and spiking neural network architectures.

Sources: [2]; Neura Robotis Website; Intel Neuromorphic Research Conference, 2022;

Remaining Challenges

- Remarkable hardware plasticity is a requirement for neuromorphic robotics
- The vision of autonomous, quickly-adapting robots is based on a level of intelligence that will require massive neural networks in terms of # of synapses & neurons – neuromorphic robotics will require high memory/computing densities

Research Pathways

Development of event-driven visual perception systems for dynamic robotics

Creation of integrated models of perception, cognition, and behavior based on SNNs

Merging of ANNs and SNNs to optimize SNNs for management of motor control

Neuromorphic robots capable of safely working in a range of environments

David Reger



David Reger founded Neura Robotics to make breakthrough advances in the field of cognitive, collaborative robots. David presented to Intel's Neuromorphic Research Community in April 2022: his vision for quickly adaptive, energy-efficient collaborative robots is predicated on incorporating neuromorphic computing.

Target Application: Self-driving Car

Description

The Machine Learning boom of 2015 to 2020 based on the impressive achievements of Artificial Neural Networks heralded remarkable hype for the self-driving car. The US Secretary of Transport stated in 2016 that automated self-driving would be widespread by 2021. Clearly, expectations have begun to deteriorate (see image A). The complexity of driving environments and high safety requirements of fully autonomous vehicles combine to deter their development for the foreseeable future. Too, Table 1 for measures on the Total Operations per Second to ascend to L5 based on conventional computing hardware. "One model suggests a range reduction of the order 9-20% for the control systems for Level 4 ADAS."

Enter neuromorphic computing: A key value of neuromorphic hardware is its design curated for the implementation of SNNs. Where ANNs are simplified, abstracted versions of biological neuronal information processing/storage, SNNs are more dynamical models of biological neural activity. The temporal, highly-complex, highly-dynamic nature of SNNs make them extremely difficult to build on conventional hardware. Should neuromorphic computing realize the field's potential to construct large-scale neuromorphic hardware comparable in neuronal scale to the human brain, many researchers expect a flexibility, power-efficiency, and training efficiency that will empower fully autonomous vehicles.

Sources: [2]; BMW Case Studies; Intel Neuromorphic Research Conference, 2022;

A

The more recent perception of ADAS progress can be summed up in a quote from Prof. Mary Cummings, Director of Duke University's Humans and Autonomy Laboratory [2]: "There are basically two camps. First are those who understand that full autonomy is not really achievable on any large scale, but are pretending they are still in the game to keep investors happy. Second are those who are in denial and really believe it is going to happen."

ADAS Capability	Compute Requirements
L2	2 TOPS
L3	24 TOPS
L4	320 TOPS
L5	4000+ TOPS

Table 1. Compute requirements for various ADAS levels (source: Horizon Robotics).



Dr. Mohsen Kaboli

Principal Scientist and Lead of Artificial Intelligence for the BMW group, Dr. Mohsen Kaboli presented to Intel's Neuromorphic Research Community in April 2022 to share his work on neuromorphic computing. Dr. Kaboli views neuromorphic computing as a pathway towards projecting BMW's iNEXT from level 3 to level 5 ADAS and to creating next-generation intelligent interiors.

Target Application: Edge Computing

Description

AI powered products are becoming ubiquitous: Smart TVs, refrigerators, alarm clocks, etc. are all nearly commonplace. These products collect data which delivers optimization based on AI learning. However, a significant portion of the machine learning for these edge devices is performed in data centers. This setup requires that data be transferred back and forth between collecting devices and data centers. The latency of data access is significant as data must be shuttled back and forth. Another consideration is compromised personal privacy with data stored and processed in centralized locations.

Edge AI is a solution that embeds smart processing within edge devices. However, traditional power-hungry AI algorithms often exceed the stiff resource constraints of edge devices. Neuromorphic computing is considered a leading method for achieving Edge AI as its inherent energy efficiency and latency reductions should allow energy constrained edge devices to perform smart processing.

Accenture Labs, an early exponent and research partner for Intel's neuromorphic Loihi chips, expects that early use-cases for neuromorphic edge computing will be "adaptive robots, smart vehicles, and advanced consumer interfaces." As Accenture's Lead R&D Director states, "Neuromorphic is made for the edge!"

Sources: [2]; Accenture Case Studies; NEOM Website; Intel Neuromorphic Research Conference, 2022;

A

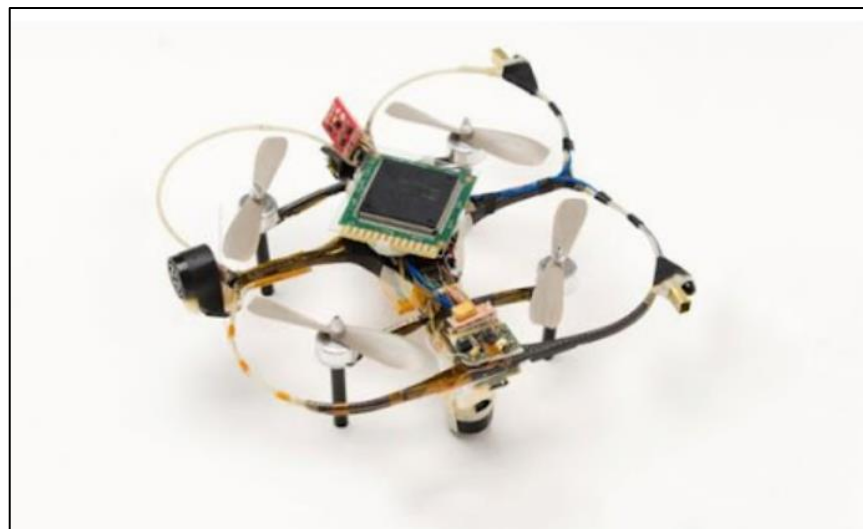


Figure A displays an MIT drone that utilized HRL's SurfRider neuromorphic chip to "learn to fly".



Noha, Al-Harthy, PhD

Dr. Al-Harthy is the Technology Lead for NEOM and has played a principal role in NEOM's plans for a smart city / cognitive city to be constructed in Saudi Arabia. NEOM is estimated to have a \$500 billion budget for the project. Dr. Al-Harthy's Neuromorphic talk presented to Intel stated that neuromorphic computing will be a key technology integrated into the urban fabric of NEOM's development.

Main Challenge by Neuromorphic Segment

Materials

Non-Ideality of Materials

Commercially ready neuromorphic chips are composed of traditional CMOS technology, limiting achievable densities. Emerging nanotechnologies have remaining challenges preventing their incorporation into novel neuromorphic chips. Thus, the density of neuromorphic chips remains limited by viable materials.

Hardware

Large-Scale Heterogeneous Integration

Neuromorphic chips perform well compared to CPUs/GPUs for certain Machine Learning tasks, especially those involving recurrence. However, there is not yet a robust framework for neuromorphic chips to operate in concert with CPUs/GPUs on large machine-learning tasks. Lava is being developed for heterogeneous optimization.

Software

Infancy of Neuromorphic Software

Neuromorphic software that is high-level and accessible to the layperson is not yet developed. This inhibits the size of the neuromorphic user base, reducing growth in neuromorphic computing.

Intel's Lava is being developed with the aim for it to be compatible with multiple neuromorphic chips and for it to become increasingly high-level/accessible.

Section III

Advances in In- memory Computing

In-Memory Products Show Significant Progress Towards Overcoming Memory wall/Von Neumann bottleneck

Description

The separation of memory and computation units in computer architecture results in significant energy costs and time delays due to non-insignificant memory latency (the time it takes for processing units to access data from memory). A promising architectural reconfiguration is to co-locate memory and processing units to enable in-memory computation. In-memory computing with traditional memory substrates (SRAM, DRAM, and NAND) attempts to modify existing memory cells to perform computation with computational circuits peripheral to memory arrays or internal memory arrays modified for certain compute operations. Historically, this research trajectory was considered infeasible. However, several commercial in-memory products have been developed in the last decade showing significant progress.

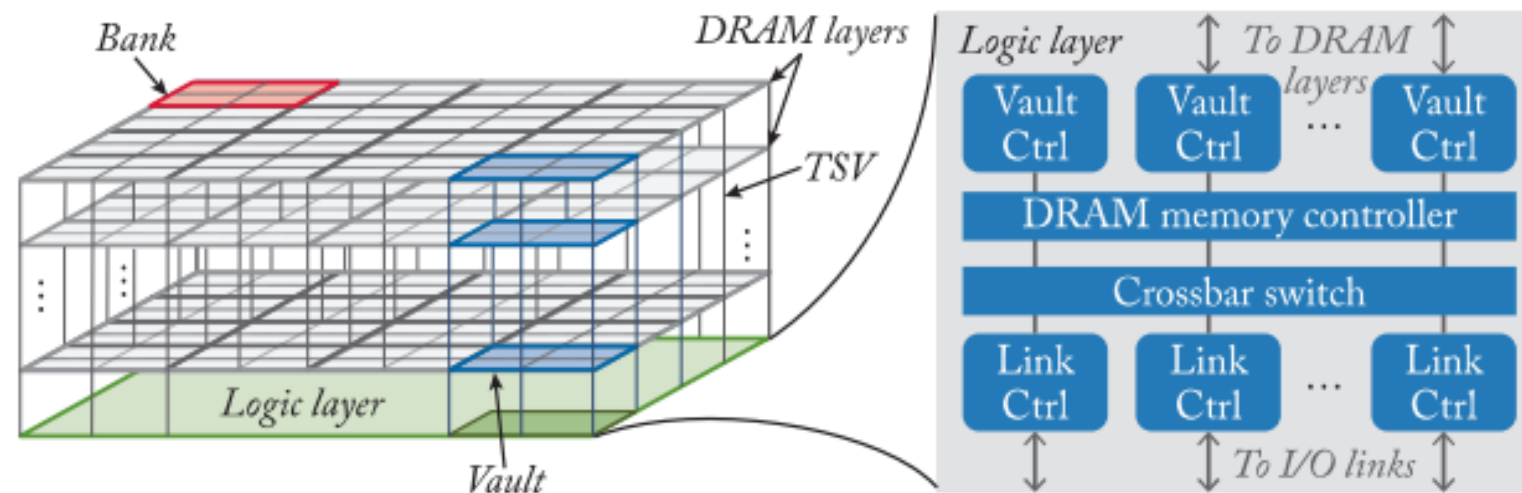
Sources: [3]

Advantages

- May have the potential to overcome the memory wall / Von Neumann bottleneck
- Enable large bandwidth access to data
- Reduces memory latency

Disadvantages

- Significant investment cost to develop new memory cell architectures
- Very low density of new cell architectures
- Additional issues unique to SRAM; DRAM; NAND



SRAM, Though more Expensive than DRAM, Potentially Addresses Data Transit Cost Associated with GPU Intensive Computing (e.g. data science, deep learning, computer vision)

Description

Static Random Access Memory (SRAM) cells are temporary storage (volatile) optimized for quick-access. SRAM has a lower density than DRAM, and higher manufacturing costs. Thus, SRAM is used for CPU cache and register memory. In-SRAM computing may be an effective replacement for dedicated accelerator chips, GPUs. The use of GPUs for co-processing incurs data movement bottlenecks as data travels via PCIe busses from CPU to GPU and vice versa. In-SRAM computing minimizes data movement and additional area required for computation (Titan XP GPU is 471 mm^2 where 9x more computing resources can be generated by enlarging SRAM by 15.8 mm^2). With slight modification of the row decoder, in-SRAM computing can support logical operations (AND, OR, NOR) through bitline sensing without overwriting data (unlike DRAM charge-sharing). These logical operations can be extended to support Bit-Serial Arithmetic for Integers and Floating-Point numbers. A key benefit of this form of in-SRAM computing is that it leaves the internal SRAM memory arrays unmodified. Other in-SRAM computing architectures (to be discussed later) alter internal SRAM memory arrays, decreasing density and oftentimes increasing access latency. Sources: [3]

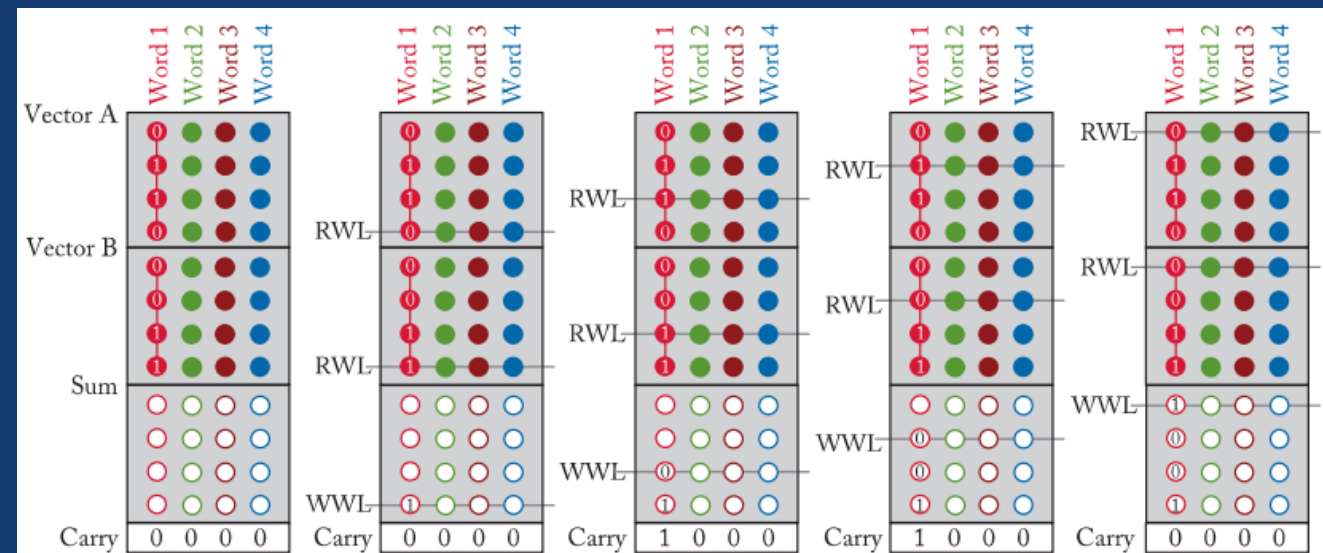


Figure 4.4: Bit-serial algorithm for in-SRAM integer addition [11].

DRAM, In-memory

Description

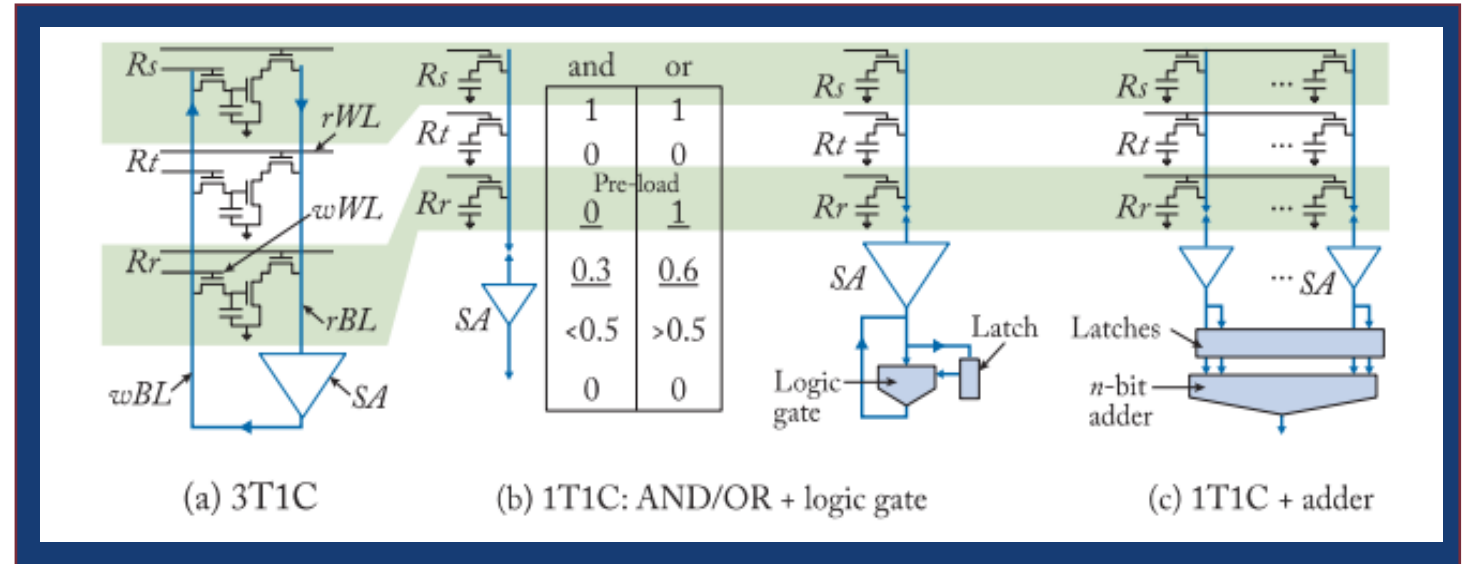
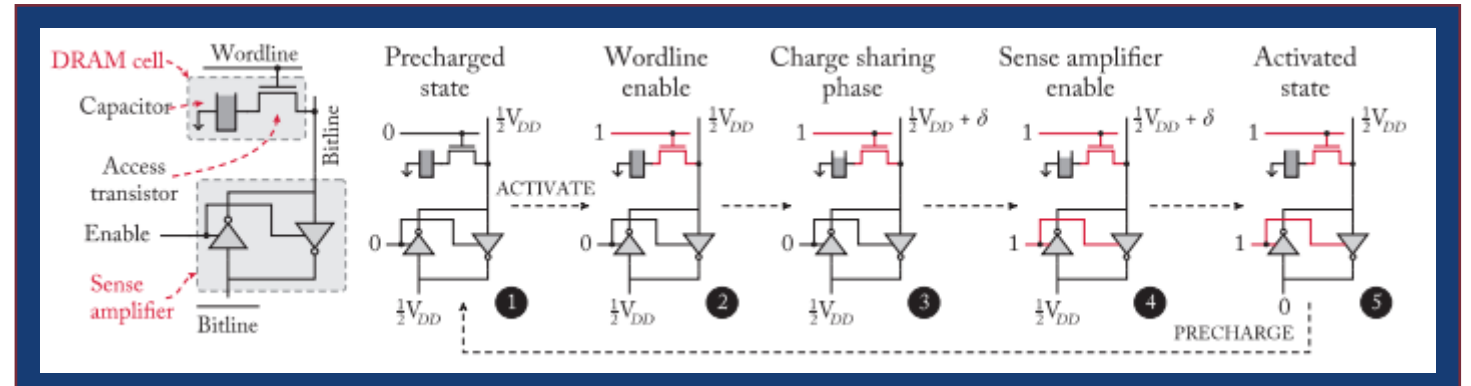
Near-memory DRAM computing is a more common novel computing approach than in-memory DRAM computing. In-memory DRAM computing faces several challenges

1. incorporation of logical elements lowers density
2. May lower computation throughput

A promising in-memory DRAM strategy is charge sharing. "Charge sharing techniques activate more than one wordline and perform bitwise operations by exploiting altered charges in capacitors connected to the same bitline." By this method, logical operations can be performed on stored memory. It should be noted that in-memory DRAM lags in maturity behind in-memory SRAM and even near-memory DRAM.

Fujiki et. al. In their excellent summary text on in-memory innovations state that despite challenges for charge sharing, they expect that "the next generation will likely see...the emergence of comput capable DRAM." However, the in-memory candidates that are likely to succeed are those which minimize changes to DRAM cell design. Conversely, "We speculate that the drastic change in cell design... may not be practical"

Sources: [3];



Major In-Memory Developments

Tend to focus on machine learning and spatial computing use cases

In-Memory SRAM for ML

Hsu et. al. built computation-in-memory (CIM) SRAM for high-speed, energy-efficient computing in 2021. Their innovation of using dynamic current-steering digital-analog-converter accomplished 2 times greater energy efficiency over previous CIM SRAMs. Hsu’s other innovation is the unified charge-processing network which boosted “area and energy efficiencies by 1.15x and 1.38x.” CIM SRAM Parameters are as follows:

- Throughput: 186.18 GOPS
- Energy Efficiency: 41.87 TOPS/W
- Area Efficiency: 3288.4 GOPS/mm²

Throughput, energy efficiency, and area efficiency represent 2.26x, 1.12x, and 2.89x improvements over previous best of class CIM SRAM demonstrations. See Table A below for a detailed comparison.]

Sources: [15]

A

	JSSC'2020 [3]	ISSCC'2020 [5]	ISSCC'2020 [6]	JSSC'2020 [7]	This Work
Technology	55 nm	28 nm	28 nm	55 nm	28 nm
Supply voltage (V)	1-0.8	0.85-1	0.7-0.9	0.9	0.9
Input precision	2 / 4	4 / 8	4 / 8	2 / 8	7
Weight precision	5	4 / 8	8	2 / 8	1 / 2
Output precision	5 / 7	12 / 20	16 / 20	7 / 19	5-7
Accuracy (CIFAR-10)	90.2%-90.42%	91.5%-91.94%	91.9%-92.02%	91.2%-91.93%	87.50%-88.87%
Throughput (GOPS)	43.2-21.2	N/A	N/A	82.29-5.14	186.18-45.51
Energy efficiency (TOPS/W)	37.5-18.37	30.4-7	33.52-11.54	10.1-0.6	41.87-12.37
Area efficiency (GOPS/mm ²)	1136.7-557.8	N/A	N/A	13.8-0.9	3288.4-803.83

3D CIM SRAM

Binary Neural Network Accelerator

Low power edge devices are more often used for inference than training of neural networks. To reduce computational complexity, edge devices will oftentimes run inference on binarized neural networks. CIM SRAM can implement many of the required operations with modification through adding additional transistors. However, the reduction of area efficiency posed by transistor addition has been a deficiency in these attempts.

Choi et. al. implemented a monolithic 3D integration through monolithic inter-tier vias to create 3D CIM SRAM optimized as a binary neural network accelerator. Their model “reduces the average execution time and energy by 39.9% and 23.2%, respectively” compared to 2D CIM SRAMs (See Table B extracted from paper).

Sources: [16]

Terms Taxonomy

- GOPS: Giga (billion) Operations Per Second
- TOPS: Tera (Trillion) Operations Per Second

B

Table 3. Normalized results of *BNN_Accels* in terms of execution time, energy and area.

	2 Layer		4 Layer
	<i>BNN_Accel</i> (2D)	<i>BNN_Accel</i> (M3D_2L)	<i>BNN_Accel</i> (M3D_4L)
Exec. time	1.000	0.801 (19.9% faster)	0.601 (39.9% faster)
Energy	1.000	1.133 (13.3% higher)	0.768 (23.2% lower)
Area	1.000	0.759 (24.1% smaller)	0.479 (53.1% smaller)

Many Startups, many in SF Bay Area, Actively Monetizing In-Memory Innovations

Noteworthy startups listed below

Company	Description	Year Founded	HQ	Pitchbook Valuation [USD Millions]	Number of Active Patents
Analog Inference	Developer of deep sub-threshold analog in-memory computation designed to run complex networks at full resolution with ultra-low latency and no active cooling.	2018	Santa Clara, USA	\$60M as of 2021	NA
Axelera AI	The company's platform integrates a custom dataflow architecture with multicore in-memory computing, enabling users to minimize power consumption to deliver edge applications for a sustainable tomorrow.	2019	Eindhoven, Netherlands	\$18M as of 2021	NA
D-Matrix	The company's platform is based on in-memory computing techniques for the data centre and is focused on attacking the physics of memory-compute integration using mixed-signal and digital signal processing techniques	2019	Cupertino, USA	\$26M as of 2021	NA
GigaSpaces Technologies	The company's platform offers in-memory computing and operational data store technologies for real-time insight into action and transactional processing	1999	New York, USA	\$13.5 M as of 2021	1 inactive
GridGain Systems	Provider of an In-Memory computing platform intended to offer services for big data systems to increase data throughput and minimize latency.	2007	Foster City, USA	\$50M as of 2016	NA
Hazelcast	Developer of an open-source in-memory data grid platform designed to unify transactional, operational, and analytical workloads. The company's platform has installed clusters that offer operational in-memory computing, enabling companies to manage their data and distribute processing using in-memory storage	2012	San Mateo, USA	\$43M as of 2017	2 Active

Sources: Pitchbook Data; Company Websites;

In-SRAM computing is Key Enabler of AI-based Internet of Things (AIoT)

AIoT Requirements

Energy Efficient

Logic Capable

Multiply &
Accumulate
(MAC) capable

Area Efficient

Inference
Accuracy

Commentary

The extension of AI inference to resource-constrained edge devices will require improvements to edge computing architecture to improve energy efficiency while reducing required data movement. CIM SRAM computing has been demonstrated to be able to perform highly parallel logic operations and MAC operations with good energy efficiency and inference accuracy. A remaining obstacle is that SRAM memory modifications that enable in-place computing decreases the density of SRAM memory arrays. Many edge devices have severe spatial constraints. However, CIM SRAM research continues to achieve better area efficiency.

Sources: [17]



Several Design Tradeoffs

(precision v. margin and density) Inhibit In-Memory potential

Overview

The transition from Von-Neumann architectures to computation-in-memory (CIM) SRAM computing (See figure A) is promising to achieve AI-based internet of things (AIoT) by enabling inference for resource constrained edge devices. Key challenges remain, several of which are summarized below.

Sources: [17]

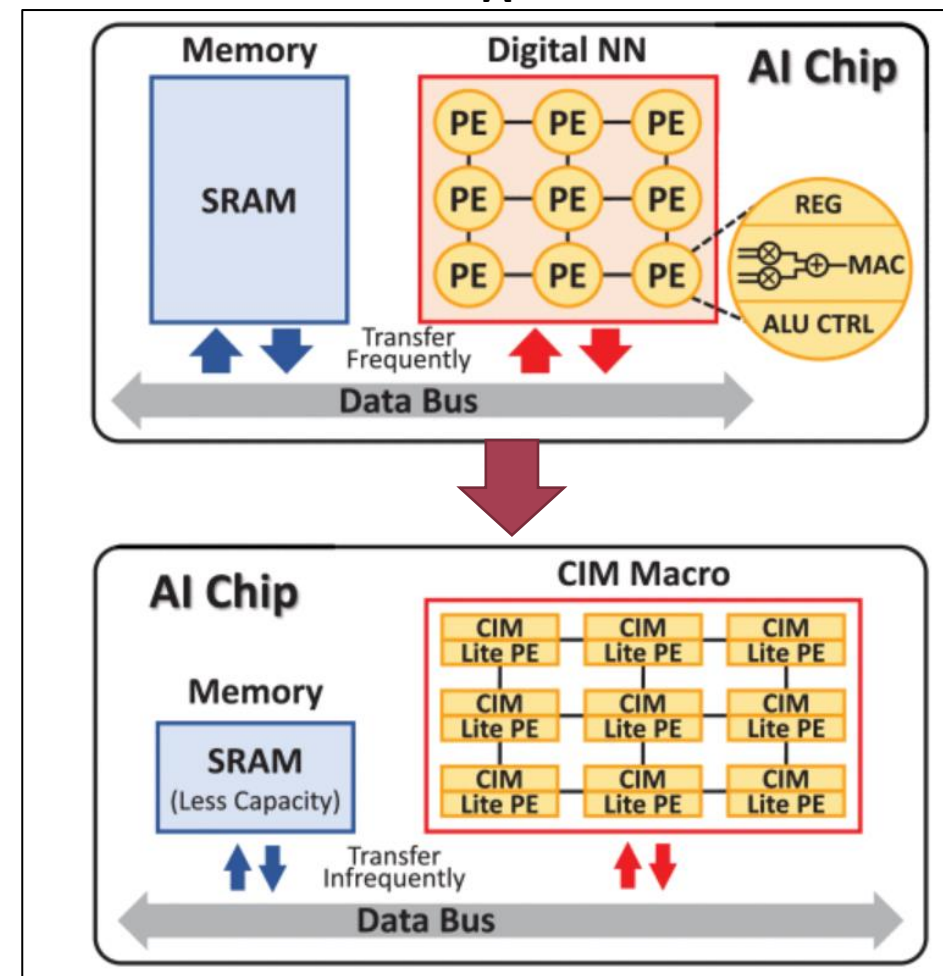
Readout precision vs. Signal Margin

As the precision of multiply and accumulate (MAC) operations is enhanced, the signal margin (difference in voltage between two consecutive MAC values) is decreased. Smaller signal margins result in greater error by sense amplifiers and decreases in system-level computing accuracy.

SRAM Densities

The modification of SRMA memories to perform in-place computations is sometimes accomplished by the imputations of transistors, lowering SRAM density / area-efficiency. Larger chip area increases costs and affects the commercial potential of the device. The tradeoff between performance and area efficiency (GOPS/mm²) is a key challenge for in-SRAM computing.

A



Section III

Advances in Reservoir Computing

Hardware-Based Reservoir Computing

A promising architecture for Machine Learning Devices

Description

Reservoir Computing is a computational framework derived from several recurrent neural network models. In reservoir computing there exists a fixed reservoir which maps data inputs into high-dimensional output. The output is trained with a simple model such as linear regression or classification. An example of a reservoir is a fixed recurrent neural network where synapse weights remain constant. As the reservoir is fixed, reservoir computing generally involves faster training than other machine-learning approaches. Reservoir Computing is most often employed for the study of dynamic time-series data.

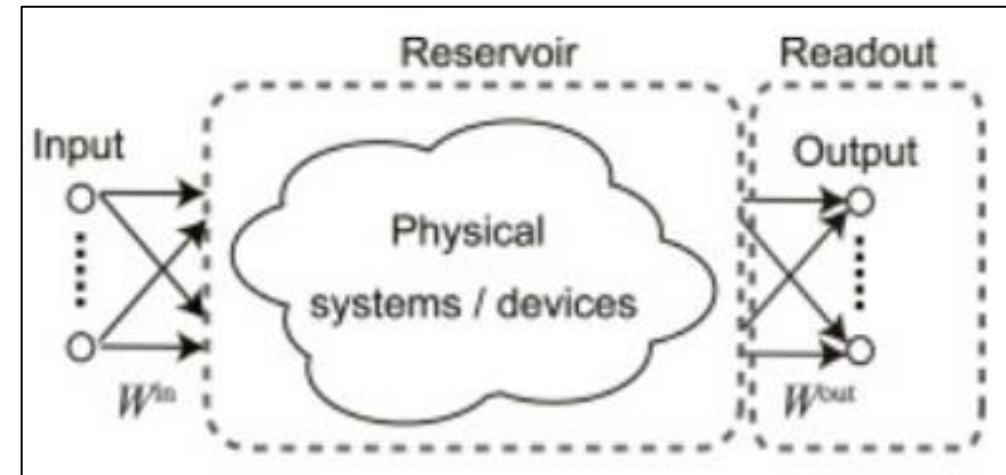
Physical Reservoir computing is possible where the reservoir can be a “complex mechanical structure” (see figure A) with two required properties:

1. Maps nonlinearly low-dimensional inputs to higher dimensions
2. Integrates information over time

The surface of water in a bucket was used by Fernando and Sojakka in 2003 as a physical reservoir for vowel classification. “The input was sound waves exciting the water and the readout was carried out through the pixelated version of video recordings of the water surface.” Reservoir computing hardware has received research interest as a novel computer architecture with low training cost.

Sources: [3]; [23]; [24]

A



Reservoir Computing Applications

- Pattern Classification
- Time Series Forecasting
- Pattern Generation
- Adaptive Filtering and Control
- System Approximation
- Short-Term Memory

Source: [23]

FPGA Reservoir Computing (RC)

Implementations Show Promise

Other Advances

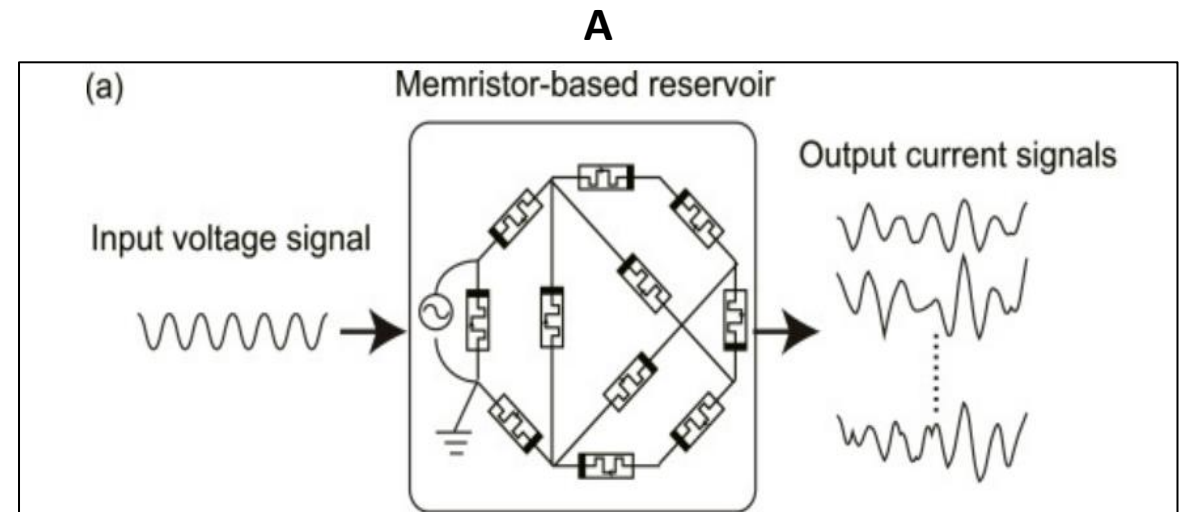
- Deep Reservoir Computing: The progressive software development of Deep Reservoir Computing wherein multiple fixed reservoirs are employed has found that this is an effective framework “for diversifying temporal representations and generating rich dynamical behavior.” Researchers are attempting to apply Deep Reservoir Computing to hardware implementations.
- Successful hardware implementations of electronic Reservoir Computing and photonic reservoir computing have been demonstrated. These are the most mature forms of RC hardware.
- Reservoir Computing is also experimented with for combination in neuromorphic spiking neural networks (see figure A)

Sources: [3];

FPGA Reservoir Computing

Field Programmable Gate Arrays (FPGAs) are often used for the implementation of Artificial Neural Networks. As such, they are readily usable to realize reservoir computing in hardware. The creation of an artificial neural network of set synaptic weights in the FPGA acts as the reservoir. Binary neurons are often chosen for their compatibility with digital logic. Empirical examinations of this layout have “...confirmed that the reservoir computer on the FPGA board has a significant advantage in the high-speed processing over the software-based reservoir implemented in a high-end laptop.

Sources: [23];

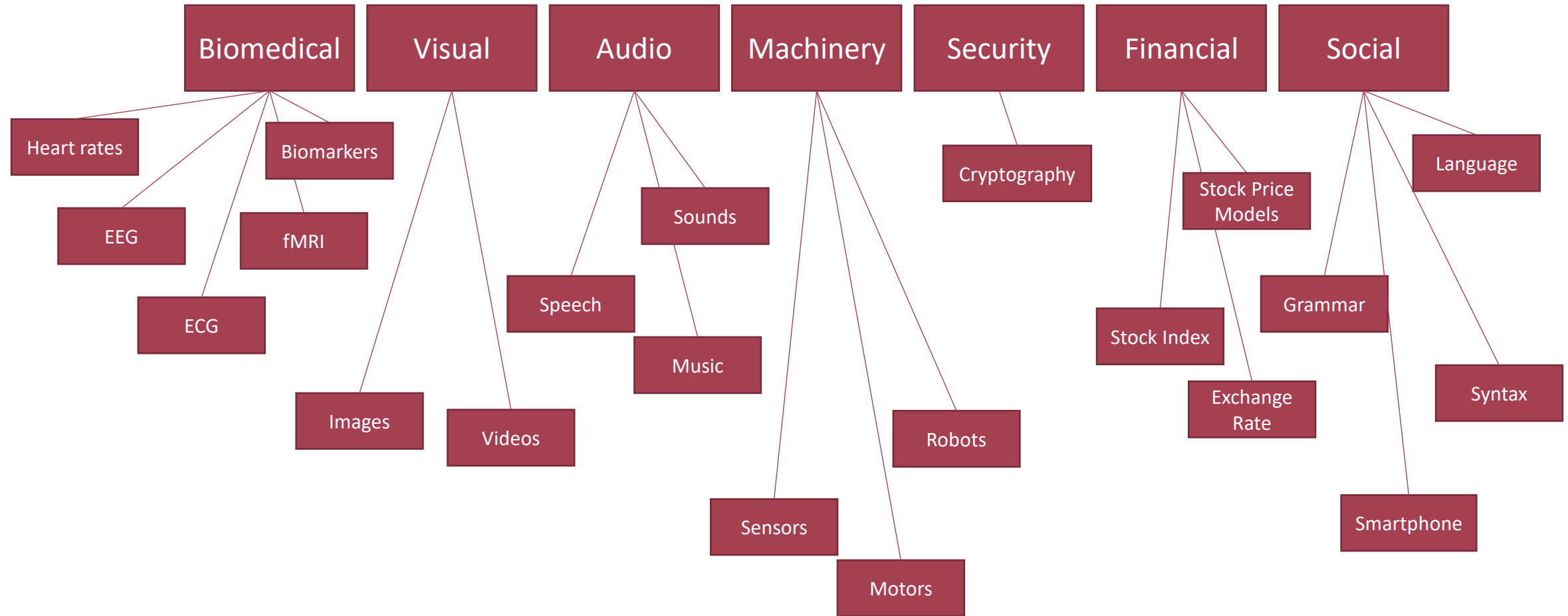


Key Academic Research Labs

Lab Name	Description	Affiliated Faculty / Researchers
Daniel P. Lathrop's Nonlinear Dynamics Lab – University of Maryland	The lab's current work with ML electronics focuses on a specific paradigm called Reservoir Computing, which excels at tasks requiring rapid inference. Its applications include prediction of complex phenomena like chaotic time series, speech recognition, and rapid image recognition/classification.	Dr. Daniel Lathrop; Dr. Katepalli Sreenivasan; Ms. Kaveri Joshi
Teuscher Laboratory – Portland State University	The mission of Teuscher Lab is to review the foundations of computer technology to help solve tomorrow's technological and societal problems. We use a radical interdisciplinary approach and apply tools from computer science, computer engineering, physics, biology, complex systems science, and cognitive science to the study and the design of next generation computing models and architectures.	Dr. Christof Teuscher; Dr. Neil Babson; Mr. Jack Cannon
Tsinghua LEMON: Laboratory of Emerging Memory and Novel Computing	An interdisciplinary Tsinghua laboratory with close research relationship with the Tsinghua Institute of Microelectronics that explores novel computing paradigms and memory devices.	Dr. He Qian; Dr. Huaqiang Wu; Dr. Ning Deng

Reservoir Computing Hardware

Ideal for applications involving dynamic, temporal data



Lagging RC Design Principles Comprehension

Constrains RC Hardware

Overview

Physical Reservoir computing aimed at hardware construction is still in its infancy. It faces numerous challenges obstructing its maturation.

Sources. The enormity of the challenges and very early research direction make reservoir computing very far from hardware commercialization. [3]

Lack of Demonstrated
Utility for Industrial
Applications

Not yet Cost-
Competitive with
Other ML Hardware

Opaque Design
Principles

Too few
Experimental Studies

Section IV

Advances in other Memory-Centric Computer Architectures

Near-Memory DRAM Computing

DRAM –Technology Categorization

Description

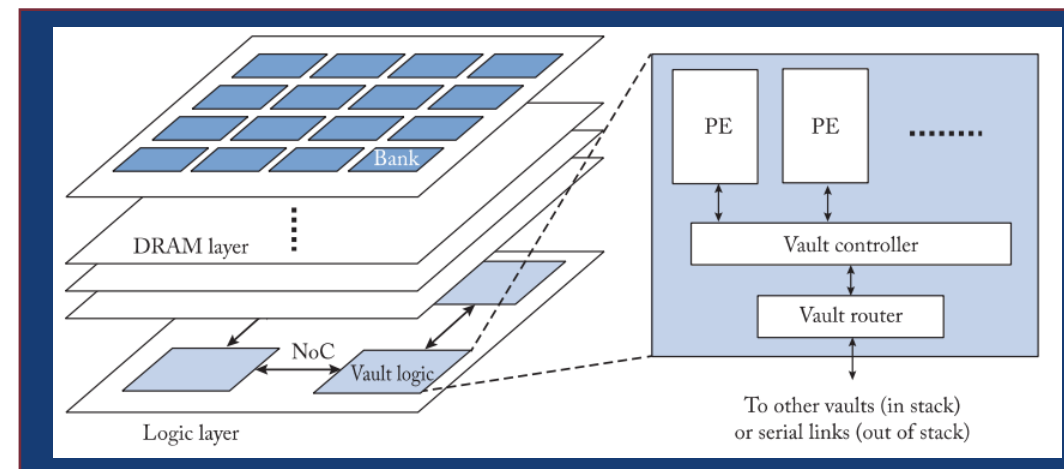
The memory latency problem can be addressed by situating processing elements closer to memory. A first attempt at this for DRAM memory was to integrate a processor and DRAM memory on a single chip to enhance memory bandwidth by on-chip data movement [see bottom right figure]. A critical failing of this approach is reduced processing performance. Rather than constructing processing elements from SRAM, on-chip integrated DRAM processors had to be built with DRAM, “a material optimized for memory cost and energy, but not logic speed”.

Improvements to Near-Memory DRAM computing have resulted in 3D integrated circuits composed of 2D DRAM memory layers with a single bottom logic layer with vertical high-speed through-silicon vias (TSV) access to memory [see top right figure]. The logic layer, separated from memory layers, can be optimized for processing performance.

Sources: [3]

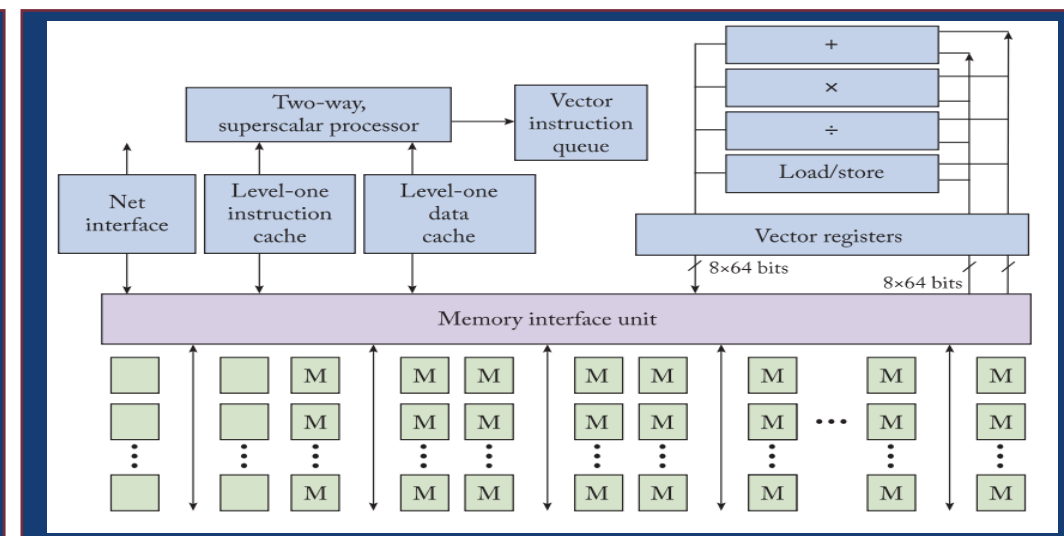
Remaining Challenges

Although reduced, data movement energy costs are still substantial in the latest versions of 3D memory cubes. HBM2’s data movement and I/O costs are 3.48 pJ/bit at a 50% toggle rate. These may be decreased by reducing distance between processing elements and memory.



Commercial Development

- Hybrid Memory Cube (HMC), Samsung/Micron
- High-Bandwidth Memory (HBM), Samsung/AMD/SK Hynix
- PIM Chips, UPMEM



Major Near-Memory DRAM Developments

Samsung's FPGA AxDIMM

Samsung's AxDIMM (not yet commercially available, see figure A) is a flexible near-memory FPGA based on DRAM internal memory. Near-memory processing elements can perform elementwise summation among other operations, minimizing data transfer costs.

Notably, the AxDIMM has two modes: non-acceleration and acceleration. In non-acceleration, host CPU can access all DRAM memory cells. In acceleration mode, the near-memory processing elements receive lookup and pooling instructions from the host CPU to reduce memory latency.

Ke et. al. tested the AxDIMM by developing a custom software stack enabling ML use-cases. They found that the "AxDIMM accelerates the execution of a broad class of recommendation models and provides up to 1.89x speedup and 31.6% memory energy savings."

Source: [18]

CGRA DRAM Accelerator

A 2015 IEEE publication explored the use of through-silicon-vias (TSVs) to construct 3D DRAM integrated memory arrays with coarse-grain reconfigurable accelerators as the local, near-memory processing layer (see figure B). Previous 3D DRAM near-memory arrays alternated memory and logic layers, with the effect of reduced capacity and poor thermal control. The single CGRA logic layer minimizes reductions in memory capacity and adverse thermal impacts.

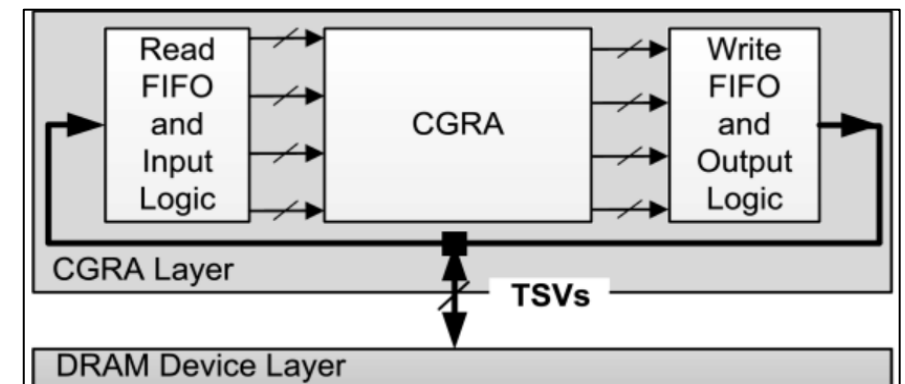
The empirical analysis of the DRAM Accelerator found that it "can reduce the energy consumption to transfer data across the memory hierarchy by 66-95 percent while achieving speedups of up to 18x over a commodity processor."

Source: [19]

A



B



Samsung and Micron lead development of Near-DRAM computing solutions

Samsung

Samsung has long been a leader in developing near-memory processing capabilities for DRAM. Specifically, Samsung partnered with AMD to develop the high-bandwidth memory (HBM, see figure A) 3D DRAM memory cube.

HBM is composed of heterogeneous 2D DRAM memory layers with a bottom logic layer. The separation of logic and memory layers allows the logic layer to be built from non-DRAM materials, which have better performance. Data communication is managed by through-silicon vias (TSVs). The HBM communicates with the host CPU “through silicon interposers with parallel inks” to optimize compatibility with highly-parallelized GPUs.

In 2021 Samsung announced that a future iteration of the HBM will have a “DRAM-optimized AI engine inside each memory bank...enabling parallel processing and minimizing data movement.”

Sources: [3]; [20]

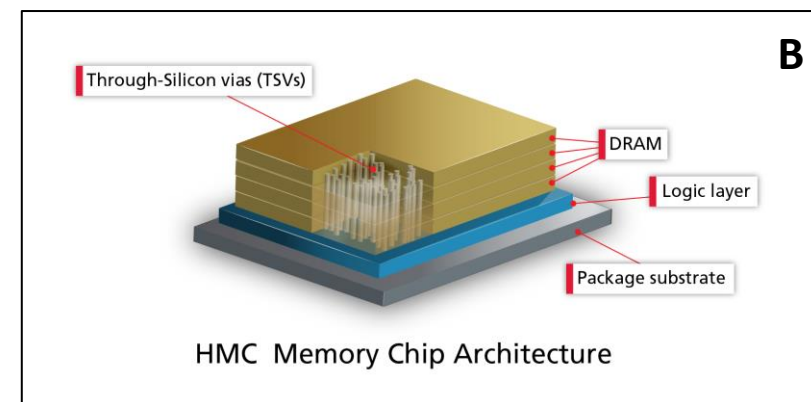
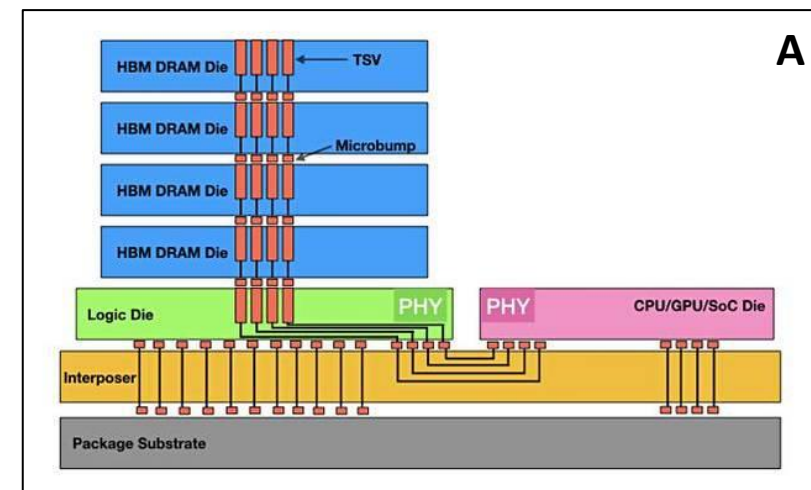
Micron

Micron developed the Hybrid Memory Cube (HMC, see figure B) in collaboration with Samsung, Open-Silicon, ARM, HP, and Microsoft. The HMC is similar to Samsung’s HBM. A key difference is how the 3D memory cube communicates with host CPUs. Micron’s HMC uses packet-based serial links to interface with host CPUs. Micron’s product is more effective at communication with CPUs, where Samsung’s HBM is suited for interacting with GPUs.

In 2018 Micron ceased focusing on HMC improvements. Rather, Micron is collaborating with Samsung on the HBM and other high-performance memory technologies.

Micron has recently partnered with the Pacific Northwest National Laboratory on a near-memory computing project to generate “specialized programming frameworks targeting Micron near memory design.” This collaboration also seeks to optimize the collaborative workflow of parallel accelerators.

Sources: [3], [21]



Near-DRAM Computing is ideal for ML Applications

Near-DRAM ML Computing

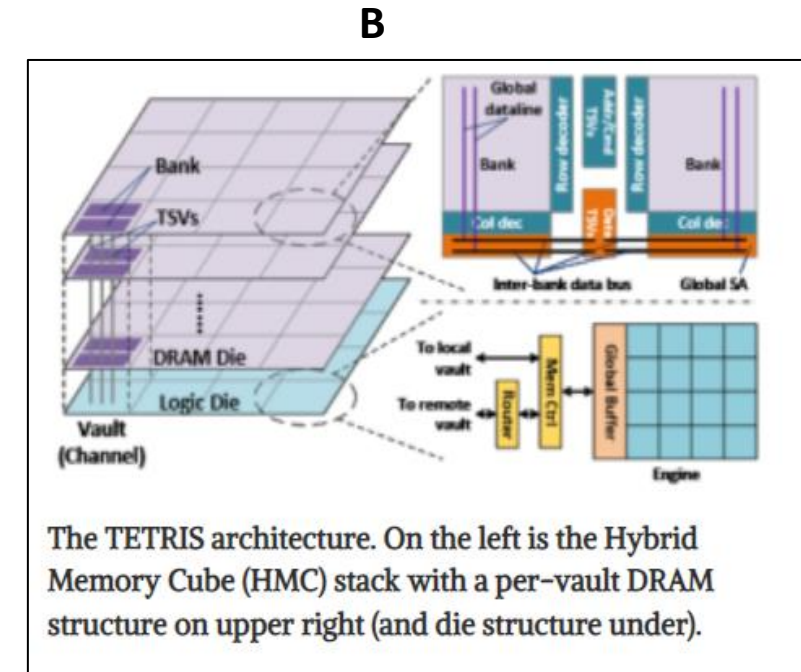
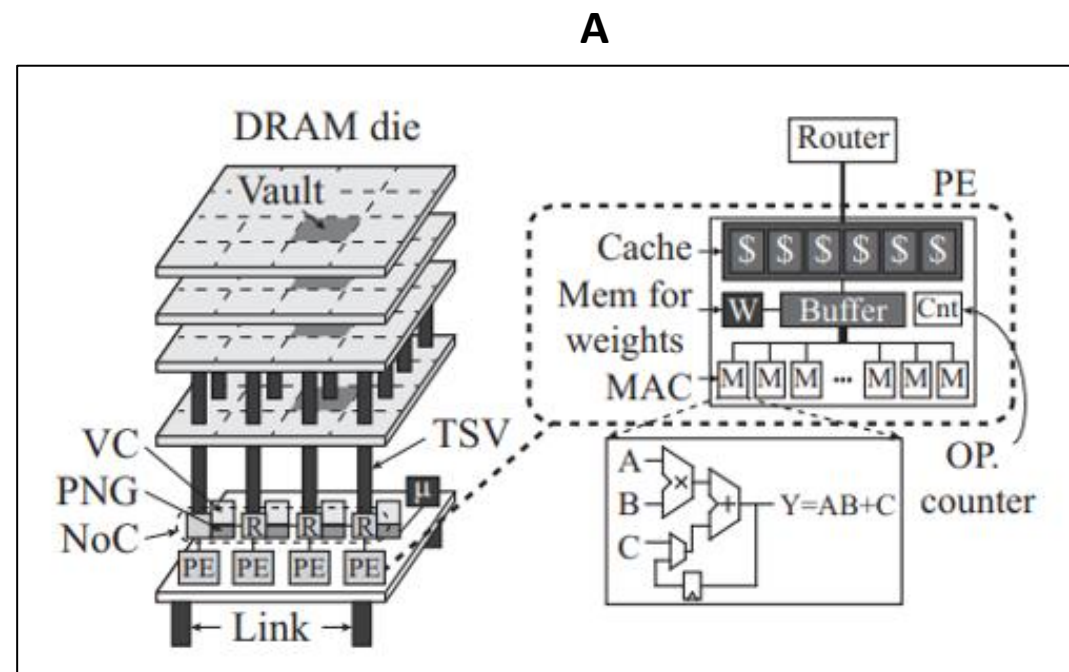
The Neurocube is a representative example of near-DRAM computing that targets ML applications. The 3D memory structure is Micron's Hybrid Memory Cube (see figure A). However, the processing elements are composed of multiple multiply accumulator (MAC) units generally involved in ML matrix operations. A similar ML aimed near-DRAM architecture titled Tetris (see figure B) was proposed by Stanford to extend the concept of Neurocube by optimizing data storage through a "hybrid work partitioning scheme" meant to reduce data access/mobility costs. Sources: [3]; [22]

Potential Benefits of Near-DRAM ML

Energy Efficient

Reduced Data Movement

Reduced Data Latency



The TETRIS architecture. On the left is the Hybrid Memory Cube (HMC) stack with a per-vault DRAM structure on upper right (and die structure under).

Near-DRAM Computing Still Suffers From Substantial Data Movement Energy Costs

Overview

Near-DRAM computing does not fully deliver on its promise of increased energy efficiency due to several outstanding challenges and tradeoffs that are profiled below. Nonetheless, Samsung, Micron and UPMEM all view the design philosophy as containing sufficient promise to warrant continued research and development.

Sources: [3]

Interconnect Energy Consumption

The 3D near-DRAM computing chips consume large amounts of energy while transferring data from memory to I/O.

Energy Savings Tradeoff

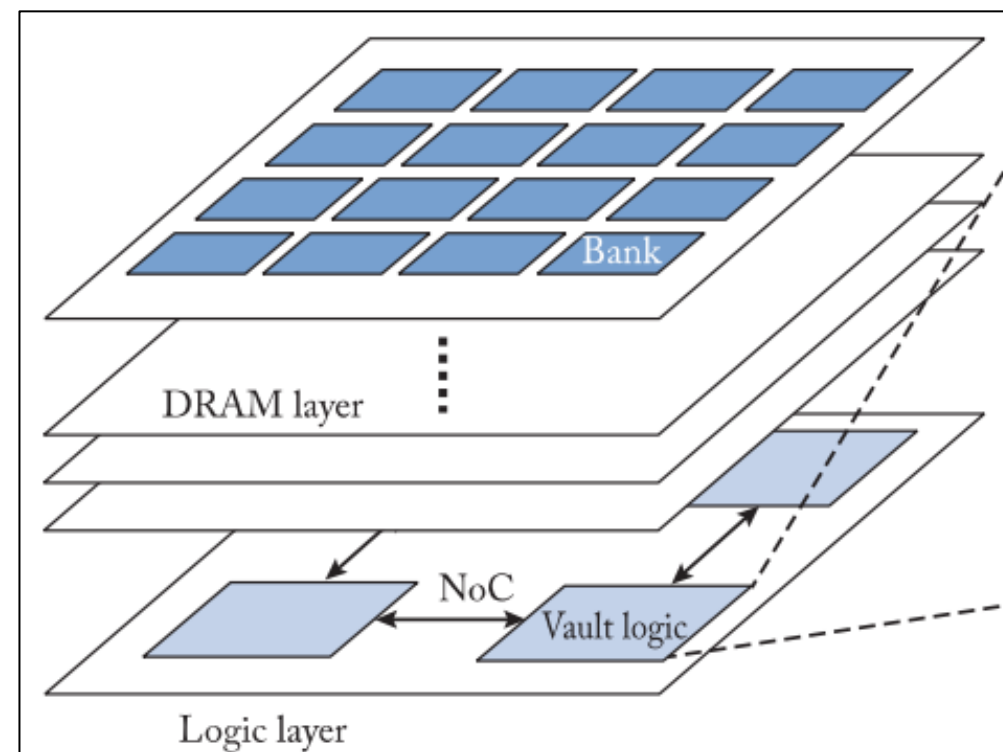
Energy efficiency is enhanced as logic is more closely coupled with memory. However, this generally involves tradeoffs of memory density and process technology.

Row Accesses Energy Consumption

The current versions of 3D near-DRAM computing products fail to reduce number of row accesses, each of which has an associated energy cost.

Latency Tradeoff

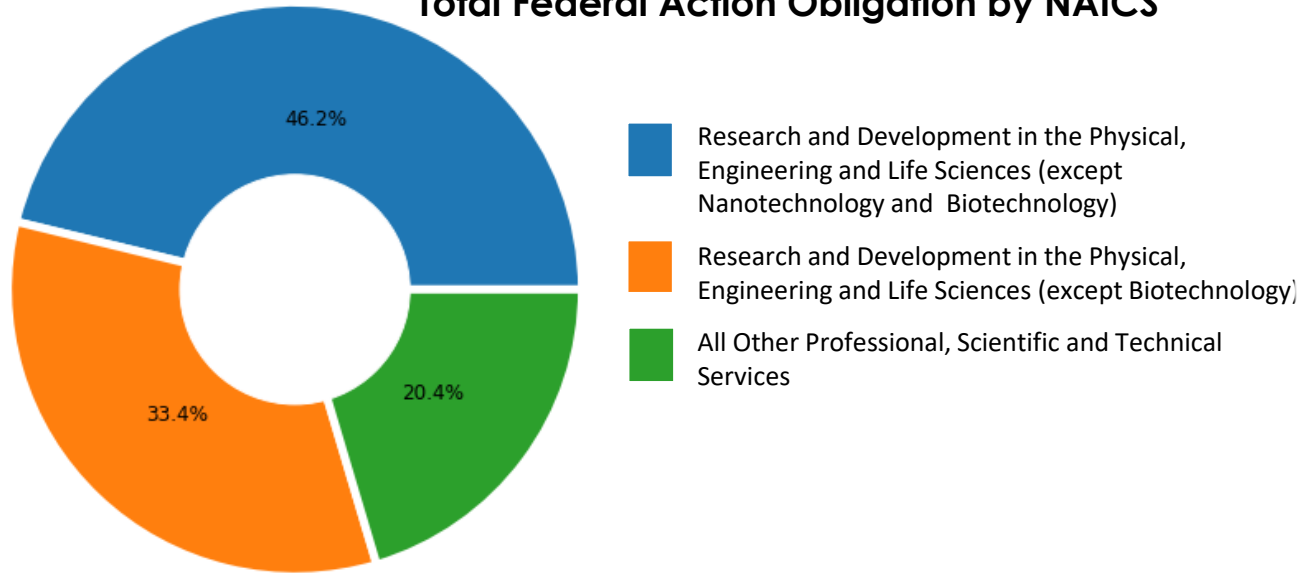
Near-memory DRAM computing solutions pose a tradeoff between reducing memory latency, and increased bandwidth of computation throughput.



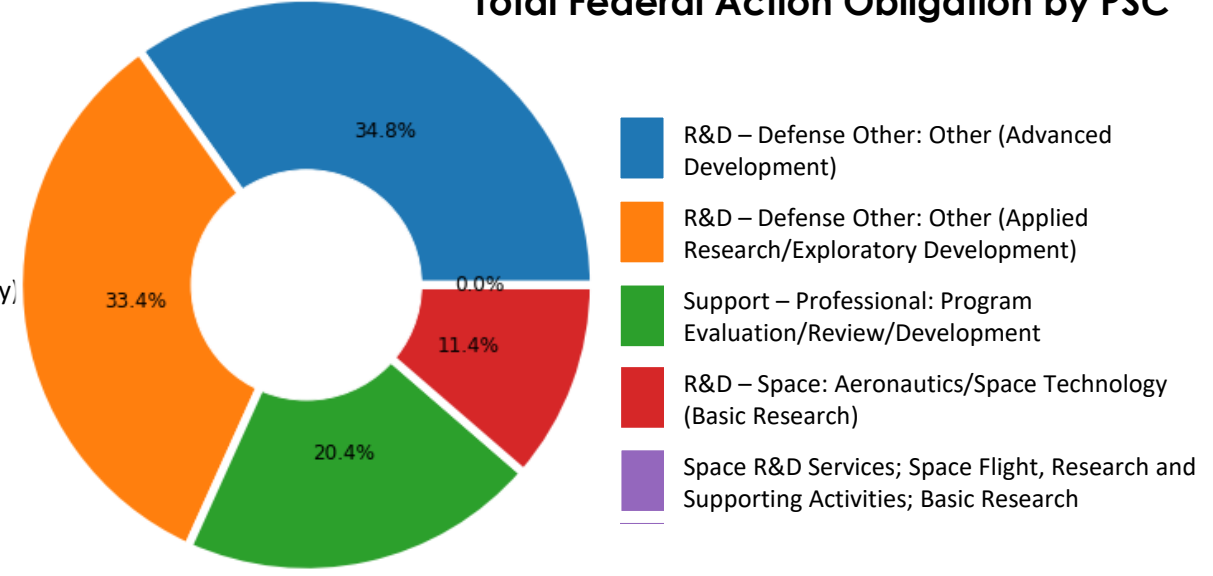
VI. Federal Government Analysis

Federal Government spent \$2.5+ million USD on Neuromorphic and In-memory Computing R&D

Total Federal Action Obligation by NAICS



Total Federal Action Obligation by PSC



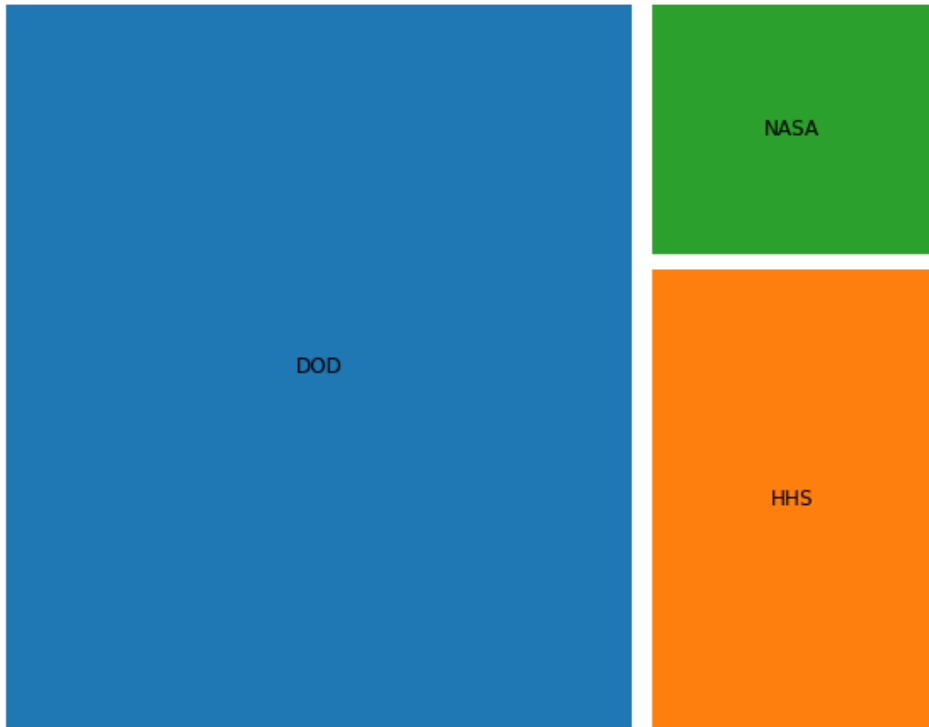
Total US Government Allocation: \$2,691,533 USD

Through 6 contracts alone, the Federal Government spent near \$ 2.7 million USD in basic and advanced research and development of Neuromorphic and In-memory Computing over the last 8 years, mainly for Defense, Life Sciences, Aeronautics and Space Flight applications.

Source: Data from Moonbeam Exchange (from 2014 to 2022), Keywords: Neuromorphic Computing, Reservoir Computing, In-Memory Computing, Near-Memory Computing

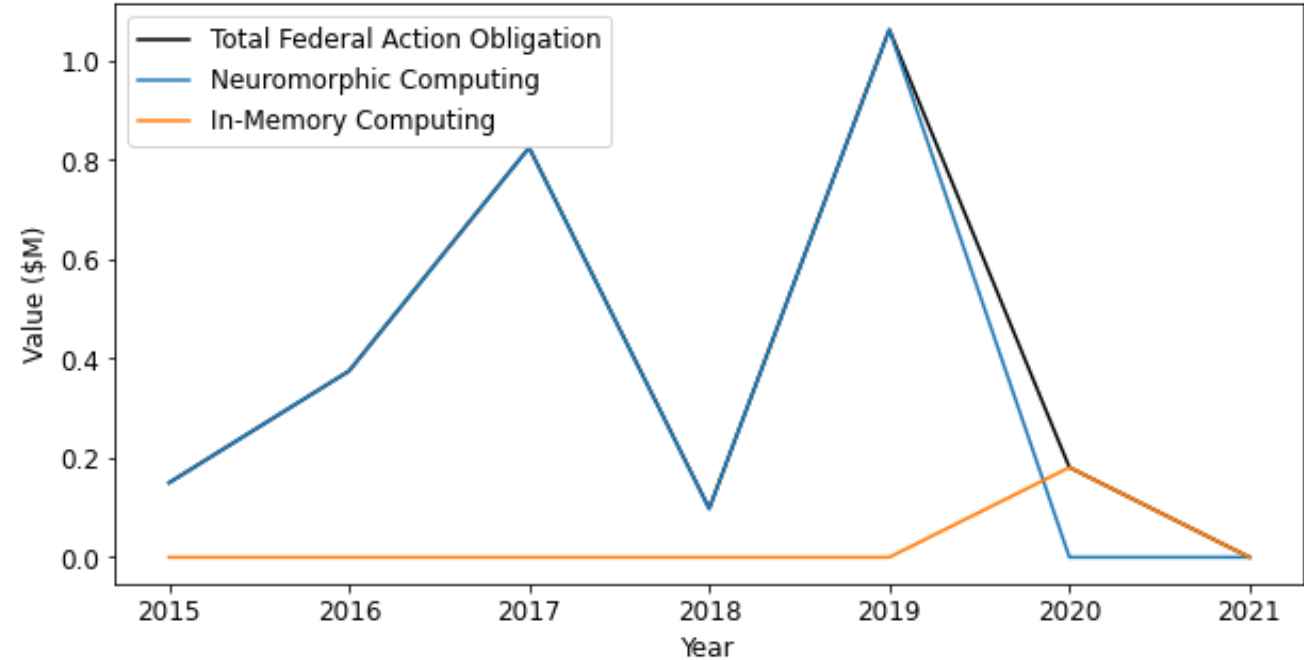
Neuromorphic Computing Investment Consistent

In-Memory Computing seemed to rise only recently



	# of contracts	Amount (in USD)
Department of Defense	2	\$1,836,847
Department of Health and Human Services (HHS)	2	\$549,105
National Aeronautics & Space Administration	2	\$305,581

Total Federal Action Obligation from 2015 to 2021 (in \$ million USD)



Neuromorphic Computing investment has steadily risen from 2015 to 2017, it dropped steeply in 2018, only to rise again and peak in 2019. On the other hand, In-Memory Computing only started to gain interest in 2020. However, since then, no more contracts were celebrated, neither for Neuromorphic nor In-Memory Computing.

Source: Data from Moonbeam Exchange (from 2014 to 2022), Keywords: Neuromorphic Computing, Reservoir Computing, In-Memory Computing, Near-Memory Computing

These 6 Federal Government Contracts Awarded to 5 Companies

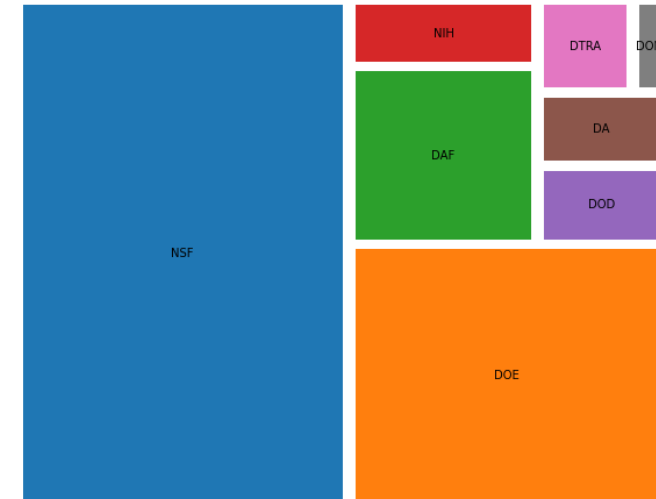
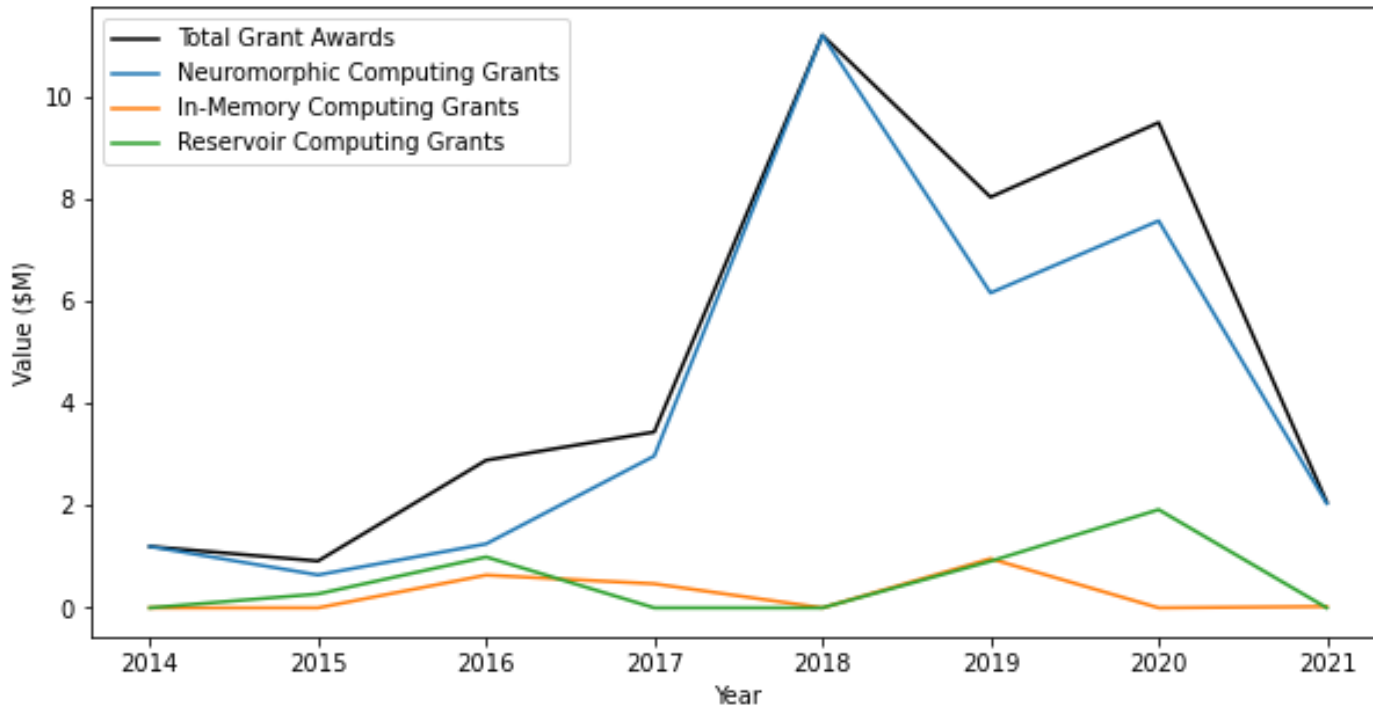
Mainly focused on R&D and producing innovative technology

	<u>Applied Research in Acoustics, LLC</u>	<u>Kalscott Engineering, Inc</u>	<u>ABT Associates, Inc.</u>	<u>Medium Technologies, Inc.</u>	<u>Avalanche Technology, Inc.</u>
Description	Brings together top-quality research scientists with systems and software engineers to solve the real-world problems and develop real-world systems.	Provides specialized, full-service research, development, test and evaluation solutions for the Aerospace, Defense and Remote Sensing industries.	Committed to employ solutions to improve quality of life and economic well-being of people worldwide.	Specializes in video distribution, visualization and Business Process Outsourcing related business solutions	Produces technology for Discrete and Embedded MRAM and Systems-On-Chip
Governmental Department	Department of Defense	Department of Defense	Department of Health and Human Services	National Aerospace & Space Administration	National Aerospace & Space Administration
Number of Contracts	1	1	2	1	1
City, State	Washington, DC	Lawrence, KS	Atlanta, GA	Santa Barbara, CA	Fremont, CA
Amount	\$937,994	\$898,853	\$549,105	\$181,000	\$124,581
Type of Computing	Neuromorphic	Neuromorphic	Neuromorphic	In-memory	Neuromorphic

(Source: Data from Moonbeam Exchange (from 2014 to 2022); Keyword: Neuromorphic Computing, Reservoir Computing, In-Memory Computing, Near-Memory Computing)

The US Government Granted \$39.2 million USD

Grant Award Amount from 2014 to 2021 (in \$ million USD)



Total Funding: \$39,201,169

of contracts

Amount (in USD)

Agency	# of contracts	Amount (in USD)
National Science Foundation	40	\$20,098,575
Department of Energy	2	\$9,869,188
Department of the Air Force	7	\$4,022,128
National Institutes of Health	4	\$1,499,873
Department of Defense	3	\$1,198,524
Department of the Army	3	\$1,113,315
Defense Threat Reduction Agency	1	\$1,041,261
Department of the Navy	3	\$358,305

As observed in Federal Contracts, Grants seemed to be mainly devoted to the R&D of Neuromorphic Computing (compared with In-Memory and Reservoir Computing, that had fewer and less significant investments). National Science Foundation (NSF) is responsible for most grants/funding.

Source: Data from Moonbeam Exchange (from 2014 to 2022), Keywords: Neuromorphic Computing, Reservoir Computing, In-Memory Computing, Near-Memory Computing

Universities seem to be the Top Governmental Contract Recipients

	University of California, San Diego	Leland Stanford Junior University	University Corporation	University of Michigan
Governmental Department	Department of Energy National Science Foundation	National Science Foundation Department of Defense	National Science Foundation	National Science Foundation
Amount	\$10,690,000	\$4,497,332	\$1,749,588	\$1,739,895
Type of Computing	Neuromorphic	Neuromorphic	Neuromorphic	In-Memory (\$1,360,352) & Neuromorphic (\$379,543)
Research Grant Titles	<ul style="list-style-type: none"> Quantum Materials For Energy Efficient Neuromorphic Computing (Q-MEEN-C) - DESC0019273 CRI: CI-NEW: Trainable Reconfigurable Development Platform For Large-scale Neuromorphic Cognitive Computing - 1823366 	<ul style="list-style-type: none"> Quantum Neuromorphic Computing And Simulation With Multimode Cavity QED - W911NF1910262 Coherent Ising Machines For Optimization, Machine Learning And Neuromorphic Computing - 1918549 	<ul style="list-style-type: none"> DMREF: Data Driven Discovery Of Conjugated Polyelectrolytes For Neuromorphic Computing - 1922042 	<ul style="list-style-type: none"> FET: MEDIUM: Memory Processing Unit (MPU) - An Efficient, Reconfigurable In-memory Computing Fabric - 1900675 SHF: SMALL: Efficient In-memory Computing Architecture Based On RRAM Crossbar Arrays - 1617315 Scaled Non-volatile Bulk Analogue Memory For Neuromorphic Computing - 2106225
Published Papers	<p>From DESC0019273:</p> <ul style="list-style-type: none"> https://arxiv.org/abs/2204.01832 <p>From 1823366:</p> <ul style="list-style-type: none"> 10.3389/fnins.2018.00583 10.3389/fnins.2019.00357 https://doi.org/10.1109/ICRC2020.2020.00013 https://doi.org/10.1145/3407197.3407209 ... 	<p>From 1918549:</p> <ul style="list-style-type: none"> https://doi.org/10.1038/s41567-021-01492-w https://doi.org/10.1038/s41586-021-04223-6 https://doi.org/10.1103/PhysRevResearch.4.013009 https://doi.org/10.1364/OPTICA.442550 https://doi.org/10.1364/OPTICA.442332 ... 	<p>From 1922042:</p> <ul style="list-style-type: none"> https://doi.org/10.1038/s41524-021-00541-5 10.1002/adma.201908120 10.1021/acs.jctc.9b01121 	<p>From 1900675:</p> <ul style="list-style-type: none"> https://doi.org/10.1109/ISCAS51556.2021.9401307 https://doi.org/10.1109/TCSI.2020.3000468 ... <p>From 1617315:</p> <ul style="list-style-type: none"> https://doi.org/10.1002/adma.201700527 https://doi.org/10.1038/s41928-019-0270-x ...

(Source: Data from Moonbeam Exchange (from 2014 to 2022); Keyword: Neuromorphic Computing, Reservoir Computing, In-Memory Computing, Near-Memory Computing)



Universities seem to be the Top Governmental Contract Recipients

	Arizona State University	Princeton University	Research Foundation University of New York	Georgia Tech Research Corporation	
Governmental Department	Defense Threat Reduction Agency National Science Foundation	Department of the Air Force National Science Foundation	Department of the Air Force National Science Foundation	National Science Foundation Department of Defense	
Amount	\$1,541,261	\$1,390,414	\$1,056,344	\$1,048,524	
Type of Computing	Neuromorphic	Reservoir (\$1,056,379) & Neuromorphic (\$334,035)	Neuromorphic	Neuromorphic	
Research Grant Titles	<ul style="list-style-type: none"> Characterizing And Modelling Radiation Effects In Neuromorphic Computing Paradigm - HDTRA11710038 FET: SMALL: 2D Material Compound Synapse Arrays For Robust In-memory And Neuromorphic Computing (2DNEURO) - 2001107 	<ul style="list-style-type: none"> E2CDA: TYPE I: Collaborative Research: Nanophotonic Neuromorphic Computing - 1740262 Reservoir Computing As A General Framework For A Comparative Study Of Classical And Quantum Information Processing - FA95502010177 	<ul style="list-style-type: none"> Fabrication Technologies For Superconducting Optoelectronic Neuromorphic Computing - FA87501910031 RUI: Structural And Compositional Modification Of Memristive Niobium Oxide Films For Neuromorphic Computing Applications - 2103185 	<ul style="list-style-type: none"> Radiation Tolerance Of New Self Healing Crystalline Memristors For Neuromorphic Computing - HDTRA11210031 	
Published Papers	<p>From 2001107:</p> <ul style="list-style-type: none"> https://doi.org/10.1088/2634-4386/ac0242 https://doi.org/10.1088/1361-6528/ac55d2 https://doi.org/10.1109/TCSI.2022.3144240 	<p>From 1740262:</p> <ul style="list-style-type: none"> 10.1063/1.5109689 10.1109/ASAP49362.2020.00028 https://doi.org/doi.org/10.1515/nanoph-2020-0172 			

(Source: Data from Moonbeam Exchange (from 2014 to 2022); Keyword: Neuromorphic Computing, Reservoir Computing, In-Memory Computing, Near-Memory Computing)

Universities seem to be the Top Governmental Contract Recipients

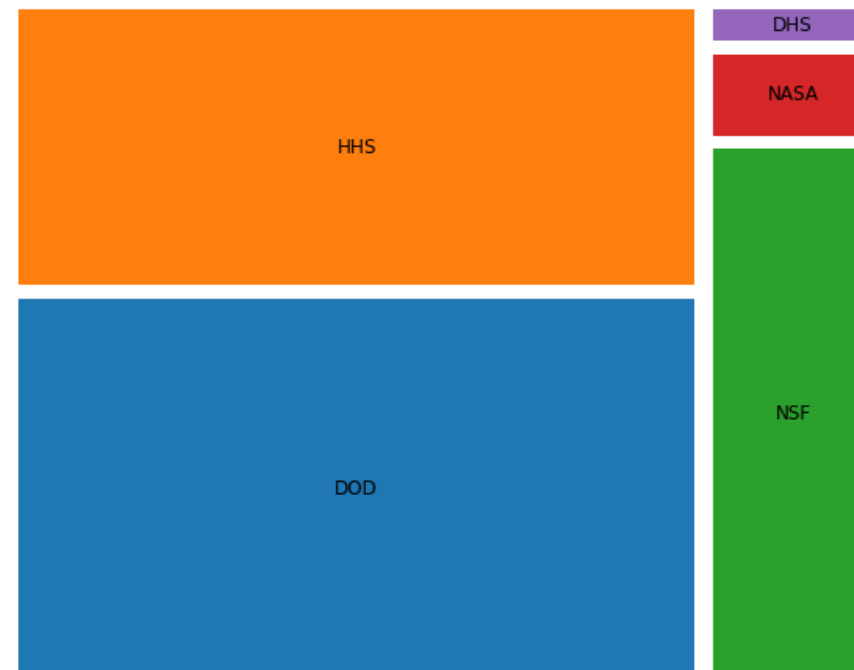
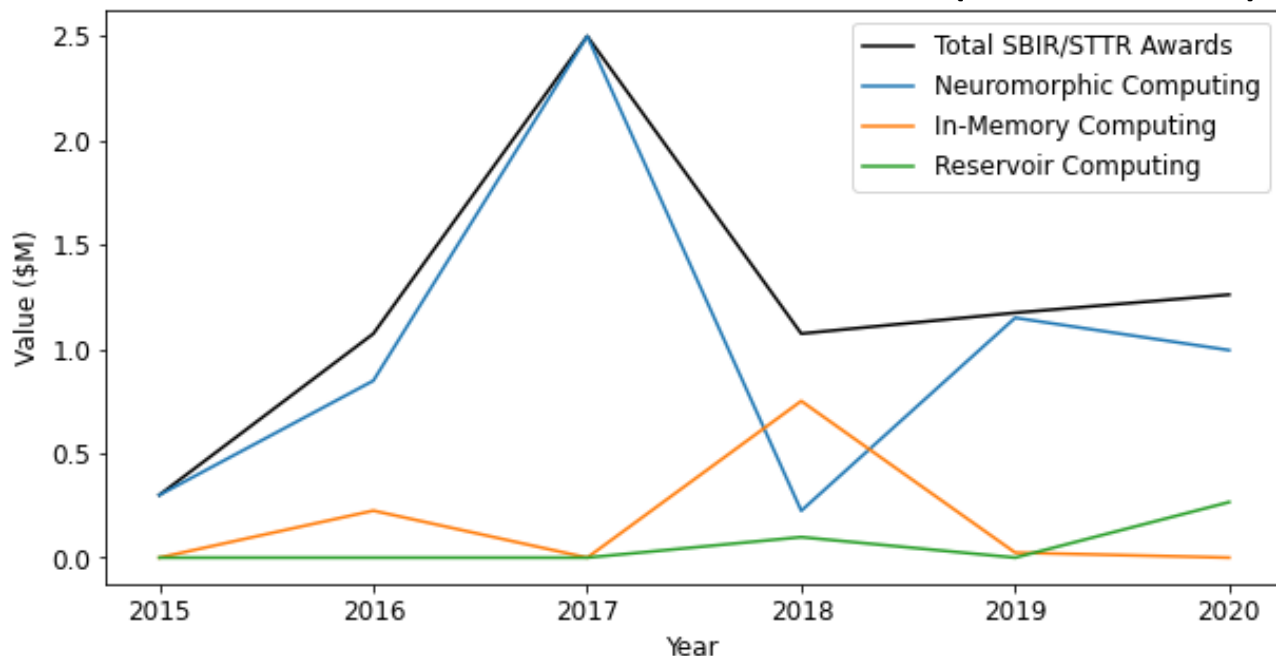
	Inheret, Inc	Rensselaer Polytechnic Institute	Virginia Polytechnic Institute & State University	Drexel University
Governmental Department	National Institutes of Health	National Science Foundation	National Science Foundation	National Science Foundation
Amount	\$1,016,123	\$1,000,000	\$978,811	\$933,107
Type of Computing	Neuromorphic	Neuromorphic	Reservoir (\$499,999) Neuromorphic (\$478,812)	Neuromorphic
Type of Research or Development	<ul style="list-style-type: none"> A SAAS Solution To Identify Patients At Increased Risk For Hereditary Disease - R42CA239842 	<ul style="list-style-type: none"> PFI:BIC: Multimodal-sensor-enabled Environments With Advanced Cognitive Computing Enabling Smart Group Meeting Facilitation Services. - 1631674 	<ul style="list-style-type: none"> RTML: SMALL: Achieving Real-time And Energy-efficient Computing For 5g Networks (ARTEN): A Deep Reservoir Computing Approach - 1937487 SPECEES: Collaborative Research: Enabling Spectrum And Energy-efficient Dynamic Spectrum Access Wireless Networks Using Neuromorphic Computing - 1811497 	<ul style="list-style-type: none"> RTML: SMALL: Design Of System Software To Facilitate Real-time Neuromorphic Computing – 1937419 CAREER: Facilitating Dependable Neuromorphic Computing: Vision, Architecture, And Impact On Programmability - 1942697
Published Papers		For 1631674: <ul style="list-style-type: none"> 10.1145/3304109.3325816 10.1145/3242969.3243022 10.1109/JSTSP.2020.2992394 ... 	For 1937487: <ul style="list-style-type: none"> https://doi.org/10.1109/INFOCOM42981.2021.9488865 https://doi.org/10.1109/INFOCOM42981.2021.9488764 10.1145/3392717.3392749 ... For 1811497: <ul style="list-style-type: none"> https://doi.org/10.1109/JIOT.2021.3052691 https://doi.org/10.1109/TWC.2021.3051317 https://doi.org/10.1109/TNNLS.2020.3029711 ... 	For 1937419: <ul style="list-style-type: none"> https://doi.org/10.1109/LES.2020.3025873 https://doi.org/10.1145/3457388.3458664 https://doi.org/10.1007/s11265-020-01573-8 ... For 1942697: <ul style="list-style-type: none"> https://doi.org/10.1109/VTS50974.2021.9441050 https://doi.org/10.1145/3394885.3431529 https://doi.org/10.1145/3457388.3458664 ...

(Source: Data from Moonbeam Exchange (from 2014 to 2022); Keyword: Neuromorphic Computing, Reservoir Computing, In-Memory Computing, Near-Memory Computing)



Total SBIR/STTR given between 2016 to 2020

SBIR/STTR Award Amount from 2015 to 2020 (in \$ million USD)



As for SBIR (Small Business Innovation Research) and STTR (Science Technology Transfer), 17 were given over the last 5 years. They were mainly for Neuromorphic Computing, but In-Memory gathered some interest in 2018 and Reservoir Computing gaining traction in 2020.

Department of Defense and Health and Human Services are the top contributors for these programs.

Total Funding: \$7,374,565

	# of contracts	Amount (in USD)
Department of Defense	9	\$3,372,445
Health and Human Services	2	\$2,493,533
National Science Foundation	3	\$1,199,586
National Aeronautics and Space Administration	2	\$210,670
Department of Homeland Security	1	\$98,331

Source: Data from Moonbeam Exchange (from 2014 to 2022), Keywords: Neuromorphic Computing, Reservoir Computing, In-Memory Computing, Near-Memory Computing

The top beneficiaries were companies that develop work in Health sector (e.g. melanoma detection or identification of increased risk for hereditary disease) or Defense sector (e.g. cognitive computing applications or navigations systems)

	Vignet, Inc	Applied Research in Acoustics, Inc.	Decibel Research, Inc.	Inheret, Inc	Green Mountain Semiconductor, Inc.	Kalscott Engineering, Inc.	Mentium Technologies, Inc.
Governmental Department	Health and Human Services	Department of Defense	Department of Defense	Health and Human Services	National Science Foundation	Department of Defense	National Science Foundation
Type of Award	SBIR Phase II	SBIR Phase I & II	STTR Phase II	STTR Phase II	SBIR Phase I & II	SBIR Phase I & II	SBIR Phase I
Amount	\$1,499,971	\$1,087,899	\$997,449	\$993,562	\$975,000	\$898,853	\$224,586
Type of Computing	Neuromorphic	Neuromorphic	Neuromorphic	Neuromorphic	In-memory	Neuromorphic	Neuromorphic
Grant Titles	<ul style="list-style-type: none"> Melanoma Early Detection RCT With Smartphones Cognitive Computing And Family Social Support 	<ul style="list-style-type: none"> Machine Interface for Contracting Assistance (MICA) Cognitive Computing Application for Defense Contracting 	<ul style="list-style-type: none"> RF-IR Data Fusion 	<ul style="list-style-type: none"> InheRET: A SaaS solution to identify patients at increased risk for hereditary disease 	<ul style="list-style-type: none"> SBIR Phase I: Ultra-High Speed In-Memory Searchable Dynamic Random Access Memory SBIR Phase II: In-Memory Artificial Neural Network 	<ul style="list-style-type: none"> Cognitive Computing Application for Defense Contracting (SBIR Phase I & II) 	<ul style="list-style-type: none"> SBIR Phase I: Addressing the memory bottleneck in deep neural networks in cloud platforms

(Source: Data from Moonbeam Exchange (from 2014 to 2022); Keyword: Neuromorphic Computing, Reservoir Computing, In-Memory Computing, Near-Memory Computing)



The top beneficiaries were companies that develop work in Health sector (e.g. melanoma detection or identification of increased risk for hereditary disease) or Defense sector (e.g. cognitive computing applications or navigations systems)

	Potomac Research, LLC	Rescon Technologies, LLC	Avalanche Technology, Inc.	Datanova Scientific, LLC	Warrant Technologies, LLC	Stalwart Technologies, Inc.
Governmental Department	Department of Defense	Department of Defense	National Aeronautics & Space Administration	Department of Homeland Security	National Aeronautics & Space Administration	Department of Defense
Type of Award	SBIR Phase I	STTR Phase I	SBIR Phase I	SBIR Phase I	SBIR Phase I	SBIR Phase I
Amount	\$218,291	\$145,953	\$124,581	\$98,331	\$86,089	\$24,000
Type of Computing	Reservoir	Reservoir	Neuromorphic	Neuromorphic	Neuromorphic	In-memory
Type of Research or Development	<ul style="list-style-type: none"> Machine Learning Enabled Near-Real-Time Situational Response for Mechanical Systems Near-Term Forecasting of Nonstationary Dynamic Processes 	<ul style="list-style-type: none"> Data fusion for sUAS navigation systems using reservoir computing 	<ul style="list-style-type: none"> pMTJ STT-MRAM based Chiplets for Neuromorphic Computing 	<ul style="list-style-type: none"> RIDER on the Storm: A Cognitive Cloud for Resilience Assessment 	<ul style="list-style-type: none"> Neuroevolution of Electronic Liquid State Machines 	<ul style="list-style-type: none"> Open Call for Innovative Defense-Related Dual-Purpose Technologies/Solutions with a Clear Air Force Stakeholder Need

(Source: Data from Moonbeam Exchange (from 2014 to 2022); Keyword: Neuromorphic Computing, Reservoir Computing, In-Memory Computing, Near-Memory Computing)

SyNAPSE Program

- *Systems of Neuromorphic Adaptive Plastic Scalable Electronics* (SyNAPSE) is a DARPA program that started in 2008 with the aim to develop electronic neuromorphic machine technology, in an attempt to build a new kind of cognitive computer with form, function, and architecture similar to the mammalian brain.
- The initial phase of the SyNAPSE program developed nanometer scale electronic synaptic components capable of adapting the connection strength between two neurons in a manner analogous to that seen in biological systems (Hebbian learning), and simulated the utility of these synaptic components in core microcircuits that support the overall system architecture.
- The program will also focus on hardware development through the stages of microcircuit development, fabrication process development, single chip system development, and multi-chip system development. In support of these hardware developments, the program seeks to develop increasingly capable architecture and design tools, very large-scale computer simulations of the neuromorphic electronic systems to inform the designers and validate the hardware prior to fabrication, and virtual environments for training and testing the simulated and hardware neuromorphic systems.

Source: <https://www.darpa.mil/program/systems-of-neuromorphic-adaptive-plastic-scalable-electronics>

SyNAPSE Program is mainly lead by two teams

IBM Team

led by Dharmendra Modha

IBM Research

Rajagopal Ananthanarayanan, Leland Chang, Daniel Friedman, Christoph Hagleitner, Bulent Kurdi, Chung Lam, Paul Maglio, Dharmendra Modha, Stuart Parkin, Bipin Rajendran, Raghavendra Singh

Stanford University

Brian A. Wandell, H.-S. Philip Wong

Cornell University

Rajit Manohar

Columbia University Medical Center

Stefano Fusi

University of Wisconsin–Madison

Giulio Tononi

University of California, Merced

Christopher Kello

iniLabs GmbH

Tobi Delbruck

1st Phase (2008) - \$4.9 million

2nd Phase - \$16.1 million

3rd Phase (2011) - \$21 million

HRL Team

led by Narayan Srinivasa

HRL Laboratories

Narayan Srinivasa, Jose Cruz-Albrecht, Dana Wheeler, Tahir Hussain, Sri Satyanarayana, Tim Derosier, Youngkwan Cho, Corey Thibeault, Michael O' Brien, Michael Yung, Karl Dockendorf, Vincent De Sapio, Qin Jiang, Suhas Chelian

Boston University

Massimiliano Versace, Stephen Grossberg, Gail Carpenter, Yongqiang Cao, Praveen Pilly

Neurosciences Institute

Gerald Edelman, Einar Gall, Jason Fleischer

George Mason University

Giorgio Ascoli, Alexei Samsonovich

Portland State University

Christof Teuscher

Stanford University

Mark Schnitzer

University of Michigan

Wei Lu

Georgia Institute of Technology

Jennifer Hasler

University of California, Irvine

Jeff Krichmar

Set Corporation

Chris Long

1st Phase (2008) - \$5.9 million

2nd Phase - \$10.7

3rd Phase (2011) - \$17.9 million

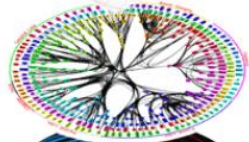
Several publications have been produced through the SyNAPSE Program

Published papers highlights:



Supercomputer Simulations

Preissl et al., Compass: A scalable simulator for an architecture for Cognitive Computing



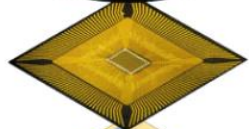
Neuroscience Data

Modha and Singh, Network architecture of the long-distance pathways in the macaque brain



Simulation with 100 trillion synapses

Wong et al., 10¹⁴



Neurosynaptic Core

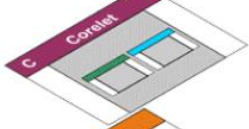
Architecture: A Network of Neurosynaptic Cores, Neuron Model

Cassidy et al., Cognitive Computing Building Block: A Versatile and Efficient Digital Neuron Model for Neurosynaptic Cores



Programming Model, End-to-end Cognitive Ecosystem

Amir et al., Cognitive Computing Programming Paradigm: A Corelet Language for Composing Networks of Neurosynaptic Cores



Algorithms and Applications

Esser et al., Cognitive Computing Systems: Algorithms and Applications for Networks of Neurosynaptic Cores



Conceptual Models of Cognitive Systems

Shaw et al., Cognitive Computing Commercialization: Boundary Objects For Communication

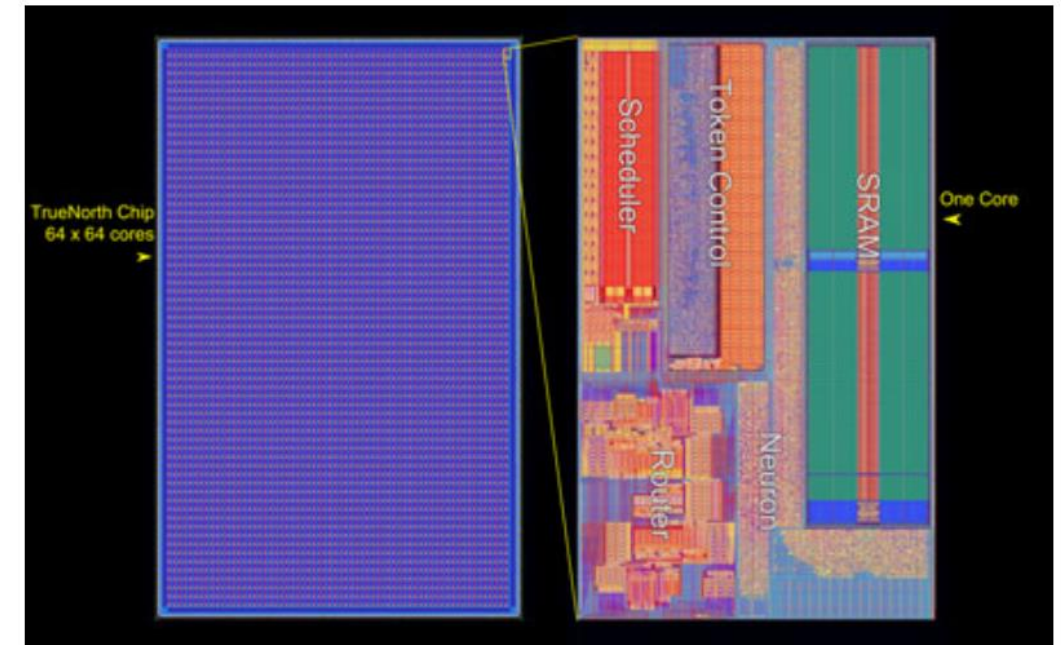


Adapted from <https://modha.org/2013/08/darpa-synapse-phase-3-a-new-foundation-to-program-synapse-chips/>

Several products have also been produced through the SyNAPSE Program

Published product highlights:

- clockless operation (event-driven), consumes 70 mW during real-time operation, power density of 20 mW/cm²
- manufactured in Samsung's 28 nm process technology, 5.4 billion transistors
- one million neurons and 256 million synapses networked into 4096 neurosynaptic cores by a 2D array, all programmable
- each core module integrates memory, computation, and communication, and operates in an event-driven, parallel, and fault-tolerant fashion



Esser et al., 2016. Convolutional networks for fast, energy-efficient neuromorphic computing. PNAS

Adapted from <https://www.nextbigfuture.com/2016/09/ibm-neuromorphic-chip-hits-darpa.html> and <https://research.ibm.com/blog>

“Nanotechnology-Inspired Grand Challenge For Future Computing” Program

- This program started in 2016 and is a shared vision of several collaborating Agencies: Department of Energy (DOE), National Science Foundation (NSF), Department of Defense (DOD), National Institute of Standards and Technology (NIST), Intelligence Community (IC)
- The Grand Challenge addresses three Administration priorities—the National Nanotechnology Initiative (NNI), the National Strategic Computing Initiative (NSCI), and the Brain Research through Advancing Innovative Neurotechnologies (BRAIN) Initiative to **create a new type of computer that can proactively interpret and learn from data, solve unfamiliar problems using what it has learned, and operate with the energy efficiency of the human brain.**
- A successful result of this Grand Challenge may be the identification of application areas (that could be Grand Challenges themselves) that represent new approaches to computing, and then demonstrating the approach’s effectiveness through a physical device technology with scalable manufacturing methods, a compatible computer architecture, and demonstrations of applications performance and capabilities.

Source: https://www.nano.gov/sites/default/files/pub_resource/federal-vision-for-nanotech-inspired-future-computing-grand-challenge.pdf

“Nanotechnology-Inspired Grand Challenge For Future Computing” Program

- Achieving this Grand Challenge would lead to many game-changing capabilities, addressing the following technology priorities shared by multiple Federal agencies:
 - Intelligent big data sensors that act autonomously and are programmable via the network for increased flexibility, and that support communication with other networked nodes
 - Machine intelligence for scientific discovery enabled by rapid extreme-scale data analysis, capable of understanding and making sense of results and thereby accelerating innovation
 - Online machine learning, including one-shot learning, and new methods and techniques to deal with high-dimensional and unlabelled data sets
 - Cybersecurity systems that can prevent (or minimize) unauthorized access, identify anomalous behavior, ensure data and software code integrity, and provide contextual analysis for adversary intent or situational awareness
 - Technology that enables trusted and secure operation of complex platforms, energy, or weapons systems that require software (or combination of multiple codes) so complicated that it exceeds a human’s ability to write and verify the software and its performance
 - Emerging computing architecture platforms, neuromorphic or quantum or others, that significantly accelerate algorithm performance, concurrency, and performance execution while maintaining and/or reducing energy consumption by over six orders of magnitude compared to today’s state-of-the-art systems
 - Autonomous or semi-autonomous platforms supporting the observe-orient-decide-act (OODA) process for both military and civilian purposes, such as transportation, medicine, scientific discovery, exploration, and disaster response

Program has multiple research and development focus areas, with specific long term goals

Research and Development Focus Areas

The research and development needed to achieve the Grand Challenge can be categorized into the following seven focus areas:

1. Materials
2. Devices and Interconnects
3. Computing Architectures
4. Brain-Inspired Approaches
5. Fabrication/Manufacturing
6. Software, Modeling, and Simulation
7. Applications

3. Computing Architectures long term goals:

- 5-year goal: Enable large-scale design, modeling, characterization, and verification of future computing architectures in both digital and analog domains. Leverage advances in high-performance computing platforms to enable parallel, high-concurrency, and large-scale simulations beyond exascale performance. This will enable the hybridization and interfacing of current digital computing with quantum- or biology-inspired computing approaches that require analog and other novel interfaces.
- 10-year goal: Be able to predict the performance of new architectures incorporating new material systems and physical nonlinear phenomena.
- 15-year goal: Be able to predict the design and characterization of computing architectures based on user applications needs. These results should enable ready-to-fabricate designs and specifications.

This program has multiple research and development focus areas, with specific long term goals

Research and Development Focus Areas

The research and development needed to achieve the Grand Challenge can be categorized into the following seven focus areas:

1. Materials
2. Devices and Interconnects
3. Computing Architectures
4. Brain-Inspired Approaches
5. Fabrication/Manufacturing
6. Software, Modeling, and Simulation
7. Applications

4. Brain-Inspired Approaches long term goals:

- 5-year goal: Translate knowledge from biology, neuroscience, materials science, physics, and engineering into useable information for computing system designers.
- 10-year goal: Identify and reverse engineer biological or neuro-inspired computing architectures, and translate results into models and systems that can be prototyped.
- 15-year goal: Enable large-scale design, development, and simulation tools and environments able to run at exascale computing performance levels or beyond. The results should enable development, testing, and verification of applications, and be able to output designs that can be prototyped in hardware.

Section IV

Sources and Definitions

Experts Consulted in Support of Report



Dr. Asim Iqbal

Dr. Asim completed a Masters in Neural Systems and Computation at the Institute of Neuroinformatics followed by a PhD in Computational Neuroscience and Machine Learning at the Institut für Hirnforschung (HiFo) and Zentrum für. Dr. Asim's experience includes a 2016 research project running classification on IBM's TrueNorth, and software development for a spike-based neuromorphic retina sensor. He is currently working as the lead Machine Learning Scientist at Cajal Neuroscience.



Dr. Sumit Bam Shrestha

Dr. Shrestha completed his PhD in Spiking Neural Networks at Nanyang Technological University of Singapore. He has been a Research Scientist at Intel since 2020, and works on Spiking Neural Networks, Neuromorphic Chips, and Dynamic Vision Sensors.



Alpha Renner

Alpha is a PhD student at the Institute of Neuroinformatics. His work focuses on the simulation and emulation of spiking neural networks to explore memory and perception in a dynamical system. Alpha works closely with Intel's Loihi team to refine and innovate neuromorphics using the Loihi. He presented his work on implementing a backpropagation algorithm on Loihi at Intel's 2022 Conference on Neuromorphics .

Note: These figures were consulted in the crafting of this research report. However, any errors or deficiencies within the report must be credited to the makers of the report alone.

Research References [Ref #]

1. Cao, Guiming, et al. "2D material based synaptic devices for neuromorphic computing." *Advanced Functional Materials* 31.4 (2021): 2005443.
2. Christensen, Dennis Valbjørn, et al. "2022 roadmap on neuromorphic computing and engineering." *Neuromorphic Computing and Engineering* (2022).
3. Fujiki, Daichi, et al. "In-/near-memory Computing." *Synthesis Lectures on Computer Architecture* 16.2 (2021): 1-140.
4. *Global neuromorphic computing market*. Global Neuromorphic Computing Market Size and Growth Forecast 2026. (n.d.). from <https://www.bccresearch.com/partners/verified-market-research/global-neuromorphic-computing-market.html>
5. Grollier, Julie, et al. "Neuromorphic spintronics." *Nature electronics* 3.7 (2020): 360-370.
6. Gumyusenge, Aristide, et al. "Materials Strategies for Organic Neuromorphic Devices." *Annual Review of Materials Research* 51 (2021): 47-71.
7. Loeffler, J. (2022, February 14). *No more transistors: The end of Moore's law*. Interesting Engineering. from <https://interestingengineering.com/transistors-moores-law#:~:text=Moore's%20Law%20is%20Dead!&text=In%20the%20end%2C%20Moore's%20Law,we%20could%20accomplish%20that%20task>.
8. Miranda, Enrique, and Jordi Suñé. "Memristors for neuromorphic circuits and artificial intelligence applications." *Materials* 13.4 (2020): 938.
9. Nandakumar, S. R., et al. "A Phase-Change Memory Model for Neuromorphic Computing." *Journal of Applied Physics*, vol. 124, no. 15, AMER INST PHYSICS, 2018, p. 152135, <https://doi.org/10.1063/1.5042408>.
10. *Neuromorphic computing report - U.S. doe office of science ...* (n.d.). from https://science.osti.gov/-/media/ascr/pdf/programdocuments/docs/Neuromorphic-Computing-Report_FNLBLP.pdf
11. Simonite, T. (2020, April 2). *Moore's law is dead. now what?* MIT Technology Review. from <https://www.technologyreview.com/2016/05/13/245938/moores-law-is-dead-now-what#:~:text=Moore's%20Law%20is%20named%20after,with%20Intel%20leading%20the%20charge>.
12. *Trusted publisher-independent citation database*. Web of Science Group. (2022, January 3). from <https://clarivate.com/webofsciencegroup/solutions/web-of-science/>
13. Wan, & Shi, Yi. (2022). *Neuromorphic devices for brain-inspired computing: artificial intelligence, perception and robotics*. Wiley-VCH
14. Zheng, Nan, and Pinaki Mazumder. *Learning in Energy-Efficient Neuromorphic Computing: Algorithm and Architecture Co-Design*. John Wiley & Sons, 2019.

Research References [Ref #], continued

15. Hsu, Ying-Tuan, et al. "A High-Throughput Energy-Area-Efficient Computing-in-Memory SRAM Using Unified Charge-Processing Network." *IEEE Solid-State Circuits Letters*, vol. 4, IEEE, 2021, pp. 146–49, <https://doi.org/10.1109/LSSC.2021.3103759>
16. Choi, Jeong Hwan, et al. "A System-Level Exploration of Binary Neural Network Accelerators with Monolithic 3D Based Compute-in-Memory SRAM." *Electronics (Basel)*, vol. 10, no. 5, MDPI AG, 2021, p. 623, <https://doi.org/10.3390/electronics10050623>
17. C. -J. Jhang, C. -X. Xue, J. -M. Hung, F. -C. Chang and M. -F. Chang, "Challenges and Trends of SRAM-Based Computing-In-Memory for AI Edge Devices," in *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 68, no. 5, pp. 1773-1786, May 2021, doi: 10.1109/TCSI.2021.3064189.
18. Ke, Liu, et al. "Near-Memory Processing in Action: Accelerating Personalized Recommendation With AxDIMM." *IEEE MICRO*, vol. 42, no. 1, IEEE, 2022, pp. 116–27, <https://doi.org/10.1109/MM.2021.3097700>
19. Farmahini-Farahani, Amin, et al. "DRAMA: An Architecture for Accelerated Processing Near Memory." *IEEE Computer Architecture Letters*, vol. 14, no. 1, IEEE, 2015, pp. 26–29, <https://doi.org/10.1109/LCA.2014.2333735>
20. *Samsung develops industry's first high bandwidth memory with AI processing power.* – Samsung Global Newsroom. (n.d.). Retrieved March 29, 2022, from <https://news.samsung.com/global/samsung-develops-industrys-first-high-bandwidth-memory-with-ai-processing-power>
21. Russell, J. (2022, March 14). *PNNL, Micron work on new memory architecture for blended HPC/AI workflows.* HPCwire. Retrieved March 29, 2022, from <https://www.hpcwire.com/2022/03/09/pnnl-micron-work-on-new-memory-architecture-for-blended-hpc-ai-workflows/>
22. Kim, Duckhwan, et al. "Neurocube." *Computer Architecture News*, vol. 44, no. 3, 2016, pp. 380–92, <https://doi.org/10.1145/3007787.3001178>
23. anaka, Gouhei, et al. "Recent Advances in Physical Reservoir Computing: A Review." *Neural Networks*, vol. 115, Elsevier Ltd, 2019, pp. 100–23, <https://doi.org/10.1016/j.neunet.2019.03.005>
24. Nakajima, Kohei, and Ingo Fischer. *Reservoir Computing*. Springer Singapore Pte. Limited, 2021

Research References [Ref #], continued

25. Hsu, Ying-Tuan, et al. "A High-Throughput Energy-Area-Efficient Computing-in-Memory SRAM Using Unified Charge-Processing Network." *IEEE Solid-State Circuits Letters*, vol. 4, IEEE, 2021, pp. 146–49, <https://doi.org/10.1109/LSSC.2021.3103759>
26. Yin, Lei, et al. "Emerging 2D Memory Devices for In-Memory Computing." *Advanced Materials (Weinheim)*, vol. 33, no. 29, Wiley Subscription Services, Inc, 2021, p. 2007081–n/a, <https://doi.org/10.1002/adma.202007081>
27. Celano, Umberto, et al. "Scalability of valence change memory: From devices to tip-induced filaments." *AIP Advances* 6.8 (2016): 085009.
28. Lim, Ee Wah, and Razali Ismail. "Conduction mechanism of valence change resistive switching memory: a survey." *Electronics* 4.3 (2015): 586-613.
29. Schuman, C.D., Kulkarni, S.R., Parsa, M. et al. Opportunities for neuromorphic computing algorithms and applications. *Nat Comput Sci* 2, 10–19 (2022). <https://doi.org/10.1038/s43588-021-00184-y>
30. Yi, W., Tsang, K.K., Lam, S.K. et al. Biological plausibility and stochasticity in scalable VO₂ active memristor neurons. *Nat Commun* 9, 4661 (2018). <https://doi.org/10.1038/s41467-018-07052-w>
31. *BrainChip completes testing production version of the Akida Chip*. Design And Reuse. (n.d.). Retrieved May 12, 2022, from <https://www.design-reuse.com/news/50911/brainchip-akida-neuromorphic-chip-in-production.html>
32. *BrainChip reflects on a successful 2021, with move to market readiness behind next-generation edge-based AI Solutions*. Design And Reuse. (n.d.). Retrieved May 12, 2022, from <https://www.design-reuse.com/news/51300/brainchip-2021-achievements.html>

Definitions

- GOPS: Giga (billion) Operations Per Second
- TOPS: Tera (Trillion) Operations Per Second
- Fast Switching: Quick transfer from high-resistance states to low-resistance states for binary storage materials. Or, quick transfers from one multilevel state to another multilevel state for analogue storage materials
- Non-Volatile: Memory is retained even when disconnected from power supply
- Analogue Capability: Memory device can store more than two memory states
- Programming Symmetry: Transition from lower resistance states to higher resistance states is analogous to transition from higher resistance states to lower resistance states in terms of pulse requirements
- Programming Linearity: Memory device resistance increases/decreases linearly with uniform application of pulses
- Low device to device variability: Memory devices are sufficiently predictable in their behavior (don't vary from one to another) that neuromorphic computations are not inhibited