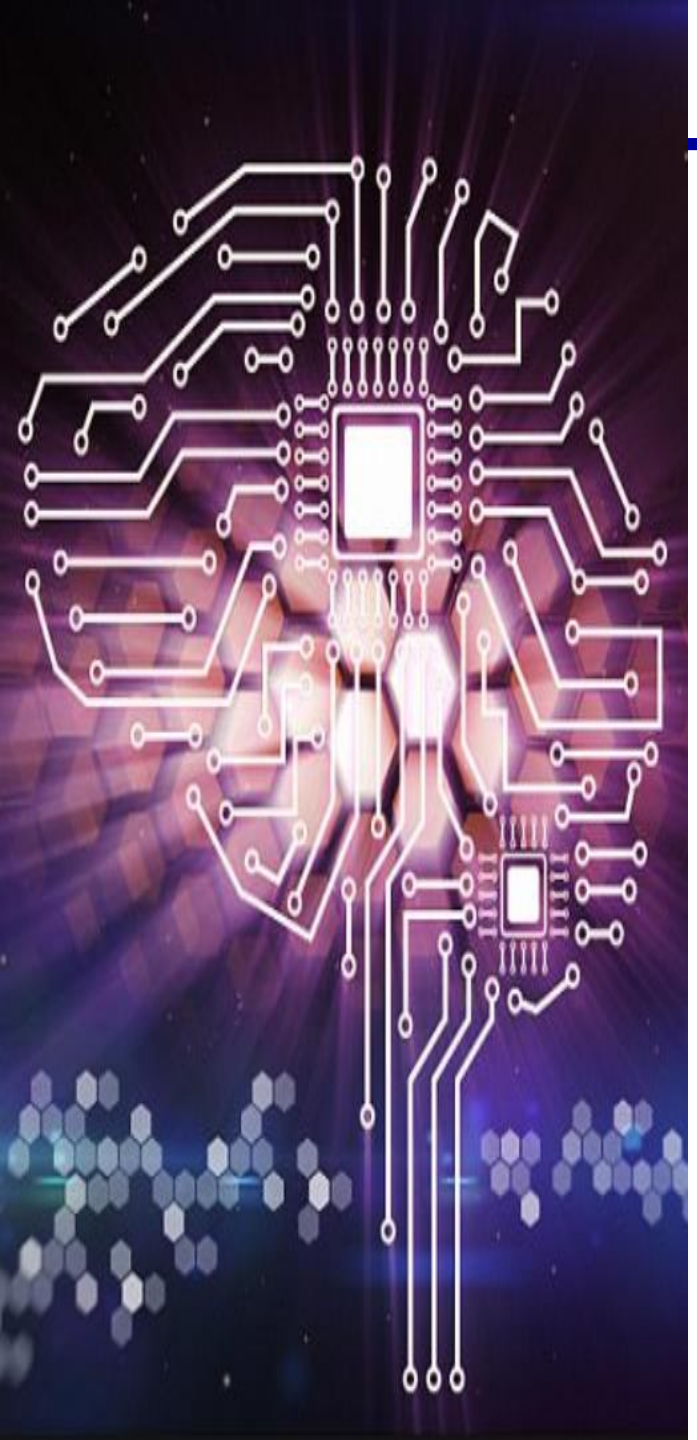
Neuromorphic Computing: Beyond-CMOS Approach to Future Computing

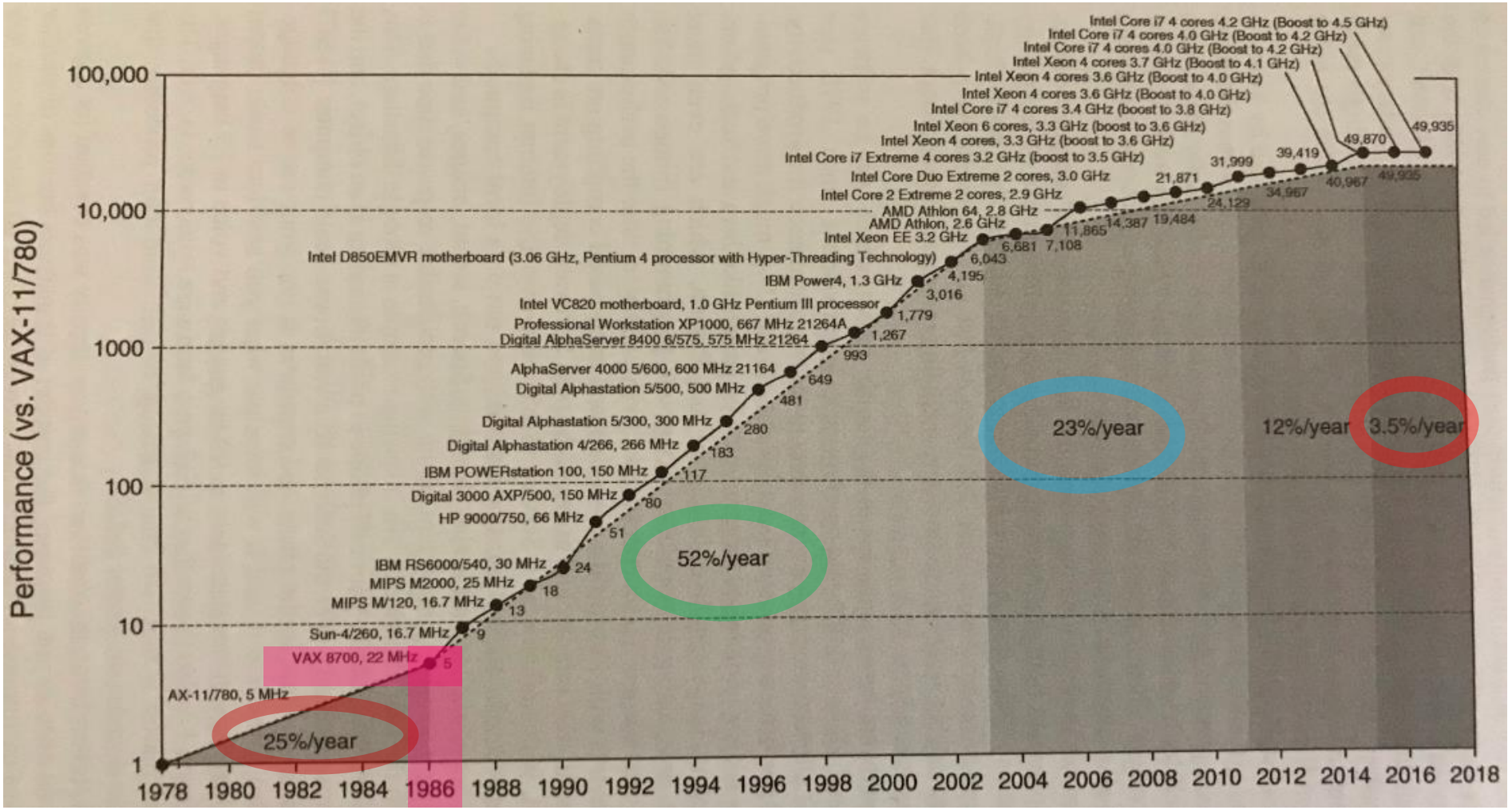
Abderazek BEN ABDALLAH
University of Aizu
Graduate School of Computer Science and Eng.
Adaptive Systems Laboratory
Aizu-Wakamatsu, Japan
Email: benab@u-aizu.ac.jp
URL: https://www.u-aizu.ac.jp/~benab/



Agenda

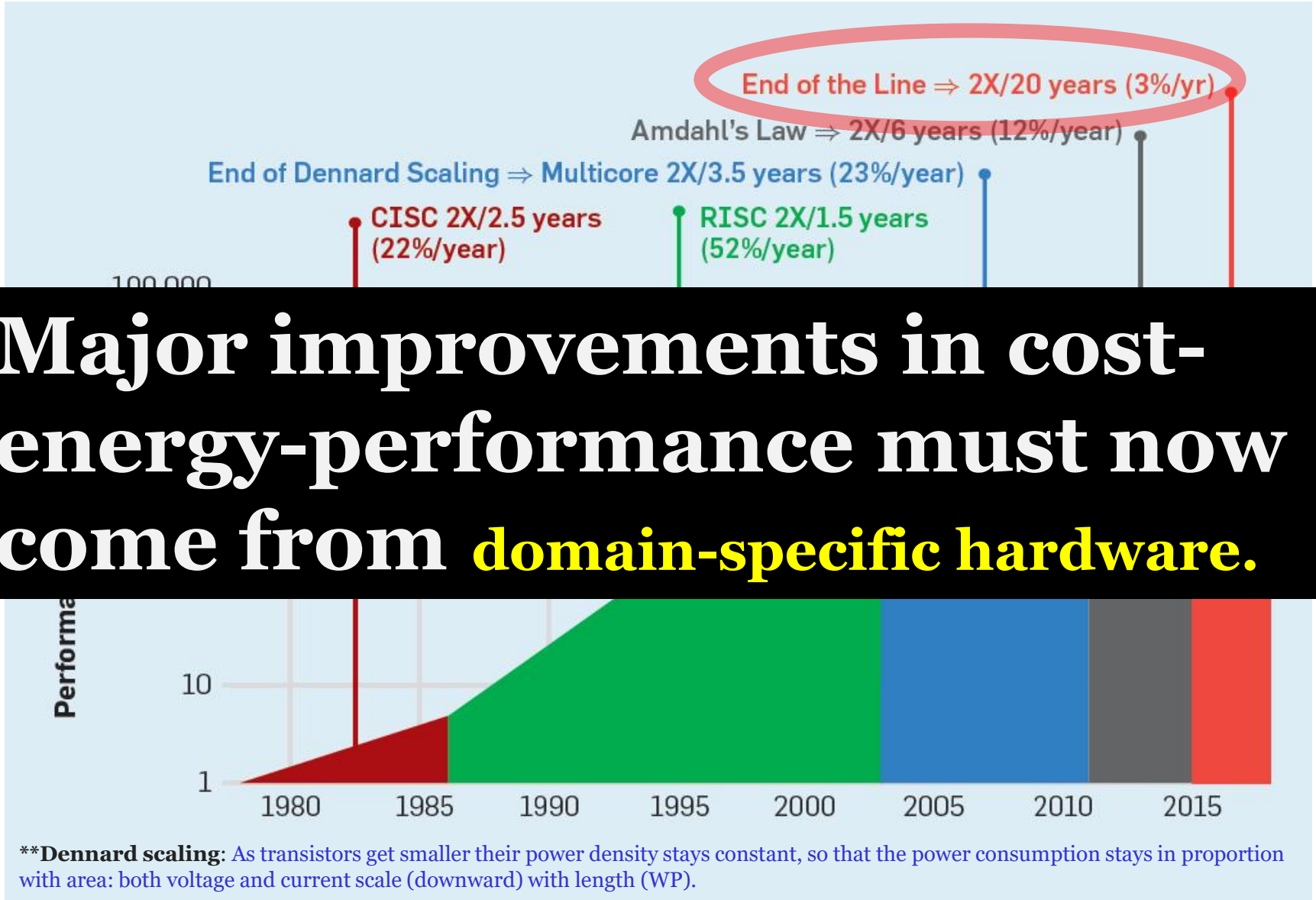
- **Fundamental Trends**
- **AI – The Emerging Industrial Revolution**
- **Neuromorphic Computing Systems**
- **Our AI-Chips**
- **Future Direction**

Moore's law is no longer providing more Compute



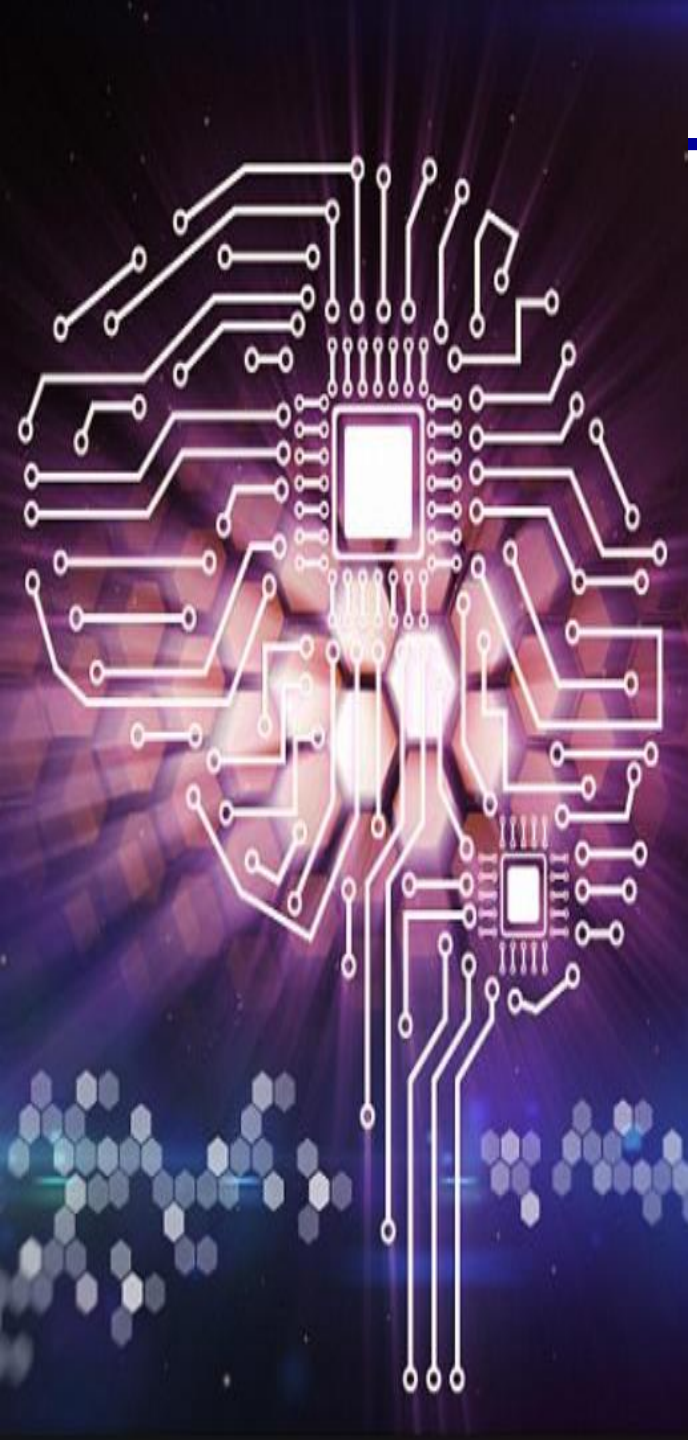
Source: Norman P. Jouppi, Cliff Young, Nishant Patil, David Patterson, Communications of the ACM, September 2018, Vol. 61 No. 9, Pages 50-59

Moore's law is no longer providing more compute



Major improvements in cost-energy-performance must now come from domain-specific hardware.

**Dennard scaling: As transistors get smaller their power density stays constant, so that the power consumption stays in proportion with area: both voltage and current scale (downward) with length (WP).



Agenda

- Fundamental Trends
- AI – The Emerging Industrial Revolution
- Neuromorphic Computing Systems
- Our AI-Chips
- Future Direction

Four factors in promoting AI/AI-HW

Image:kdnuggets.com

AI algorithms are being applied to nearly everything we do.

Strong Gov. & Industry Engagements


Image: kdnuggets.com

Growth of computational power

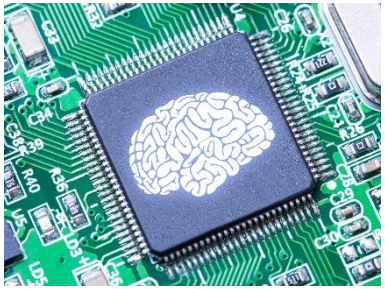


Image: spectrum.ieee.org

More compute means new solutions to previously intractable problems, i.e. GO


Image: sas.com

Larger data sets and models lead to better accuracy but also increase the computation time

AI-Chips are … everywhere

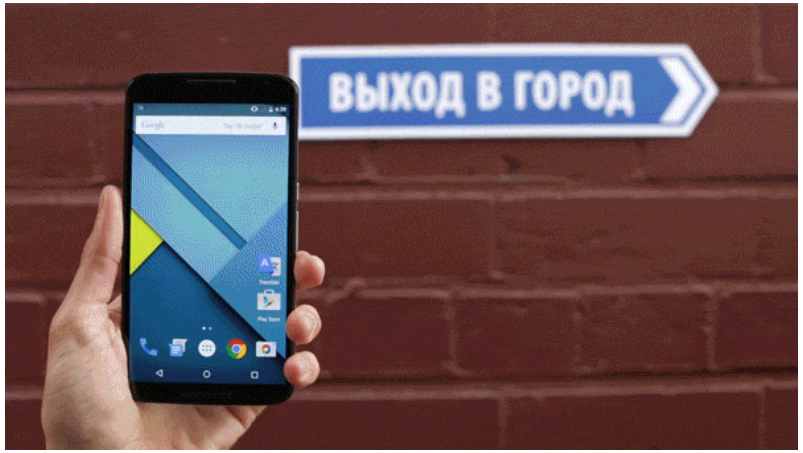
Self-driving, EV Car

Bottom Image source: edition.cnn.com

Smart Robots

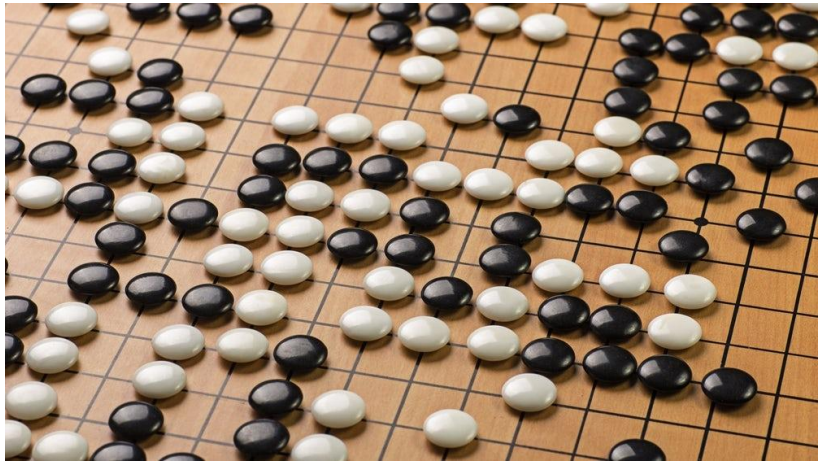
Image source: roboticsbusinessreview.com

Machine Translation



Bottom Image source: missqt.com

Gaming



Bottom Image Source: newatlas.com

AI-Chips are … everywhere

Self-driving, EV Car

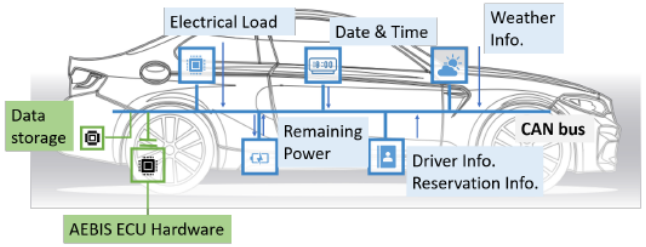


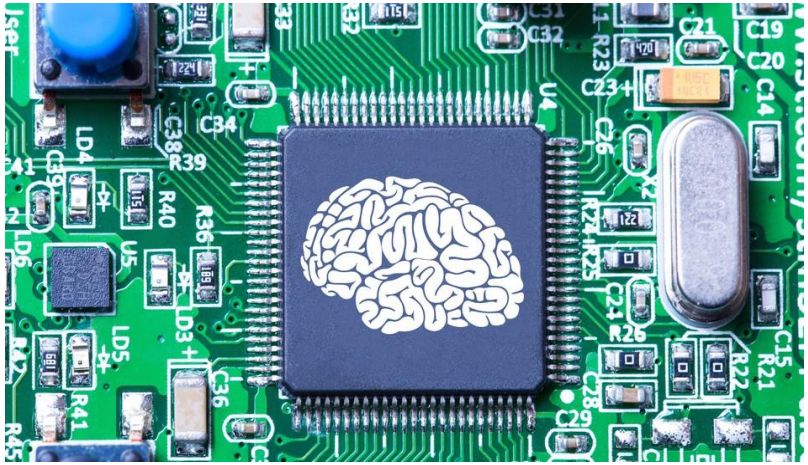
FIGURE 2. The integration of the proposed AEBIS system into the built-in Controller Area Network (CAN) of Electrical Vehicles (EVs). A CAN bus is a robust vehicle interconnect standard allowing microcontrollers and devices to communicate with each other. Each blue box indicates a built-in electronic controller unit (ECU), which shares with other ECUs its data via the CAN bus. The green box on the left shows a customized ECU for data storage, collecting and processing the data from other ECUs. The data storage ECU then transmits the data to the AEBIS ECU hardware for training and inference.

Ref. Zhishang Wang, Mark Ogbodo, Huakun Huang, Chen Qiu, Masayuki Hisada, Abderazek Ben Abdallah, "**AEBIS: AI-Enabled Blockchain-based Electric Vehicle Integration System for Power Management in Smart Grid Platform**," IEEE Access, 12/2020. DOI:10.1109/access.2020.30446

Smart Robots

Image source: roboticsbusinessreview.com

Machine Translation



Bottom Image source: missqt.com

Gaming

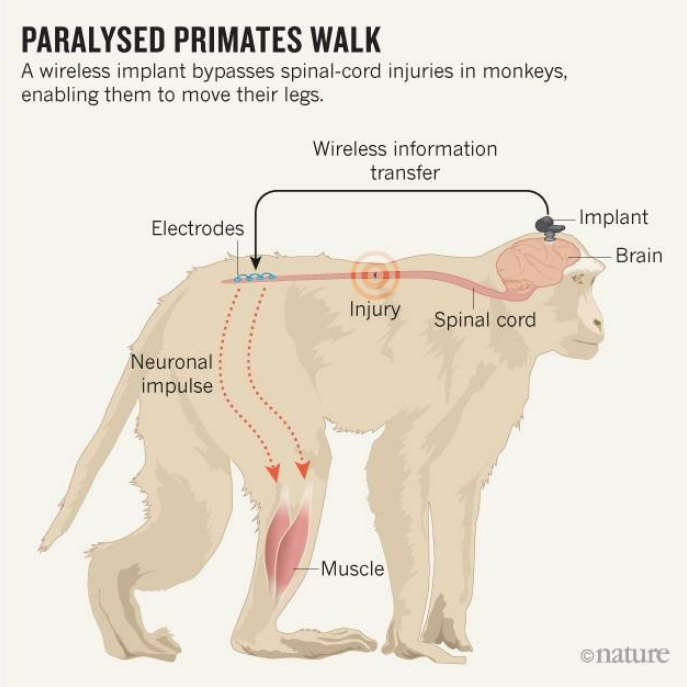
Bottom Image Source: newatlas.com

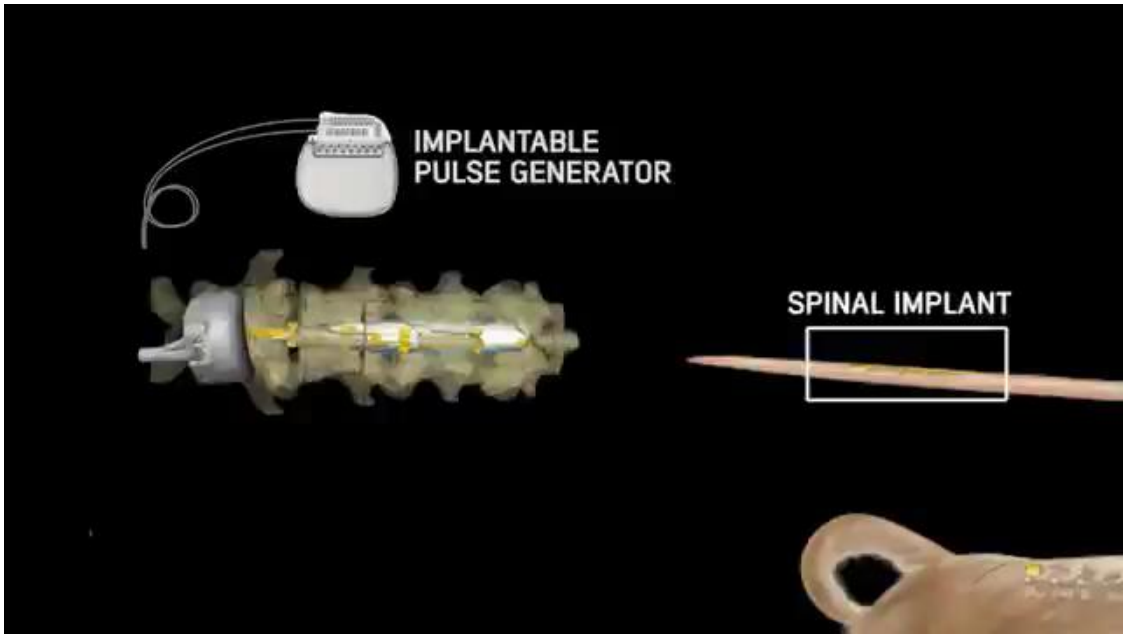
AI-Chips are … everywhere

Brain implant allows paralysed monkey to walk

There really is a kind of intelligence inside the spinal cord. We are not just talking about reflexes that automatically activate muscles. In the spinal cord there are networks of neurons able to take their own decisions

-Grégoire Courtine-
Neuroscientist, Federal Institute of Technology, Lausanne



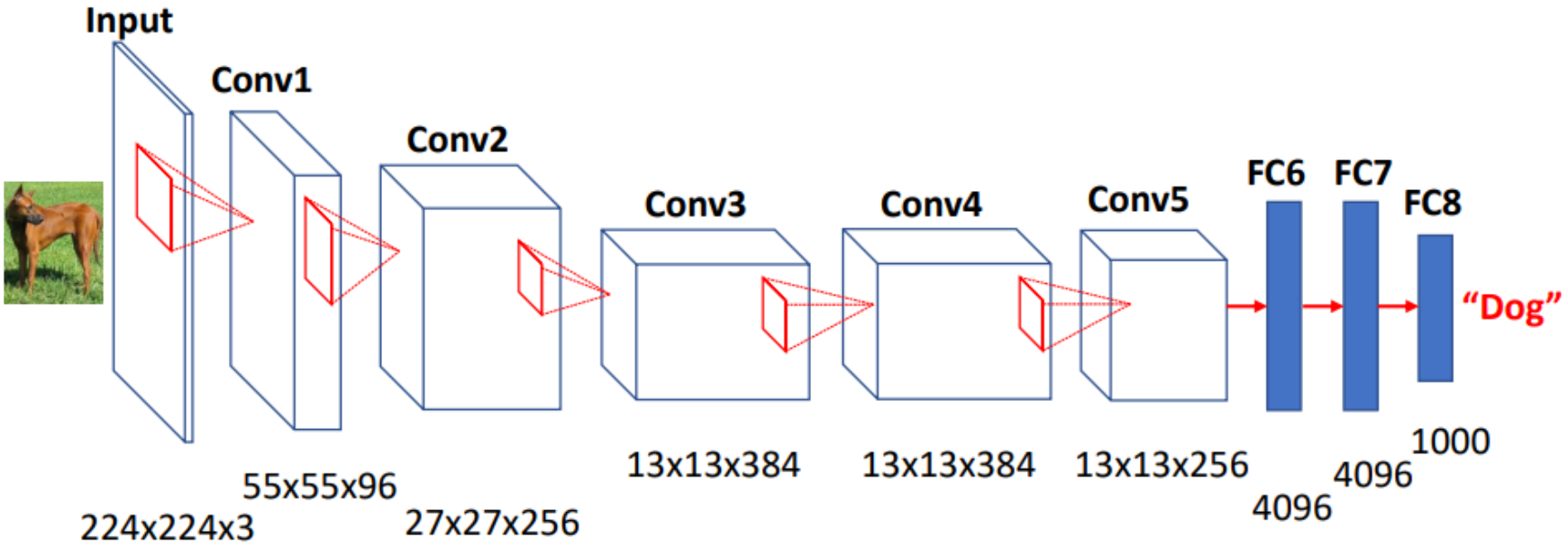


Nature volume539, pages284–288 (10 November 2016)

Deep learning requires massive compute power

- A 32-bit convolutional NN requires calculations for every floating point operation (FLOP)

- Number of FLOPS for a single inference are on the order of billions



Deep learning requires massive compute power

Metrics	LeNet-5	AlexNet	VGG-16	GoogLeNet (v1)	ResNet-50
Top-5 error	n/a	16.4	7.4	6.7	5.3
Input Size	28x28	227x227	224x224	224x224	224x224
# of CONV Layers	**2**	**5**	**16**	**21 (depth)**	**49**
Filter Sizes	5	3, 5,11	3	1, 3 , 5, 7	1, 3, 7
# of Channels	1, 6	3 - 256	3 - 512	3 - 1024	3 - 2048
# of Filters	6, 16	96 - 384	64 - 512	64 - 384	64 - 2048
Stride	1	1, 4	1	1, 2	1, 2
# of Weights	2.6k	2.3M	14.7M	6.0M	23.5M
# of MACs	283k	666M	15.3G	1.43G	3.86G
# of FC layers	**2**	**3**	**3**	**1**	**1**
# of Weights	58k	58.6M	124M	1M	2M
# of MACs	58k	58.6M	124M	1M	2M
Total Weights			**138M**	**7M**	**25.5M**
Total MACs	**341k**	**724M**	**15.5G**	**1.43G**	**3.9G**

Source: Joel Emer, ISCA Tutorial, 2017

Deep learning requires massive compute power

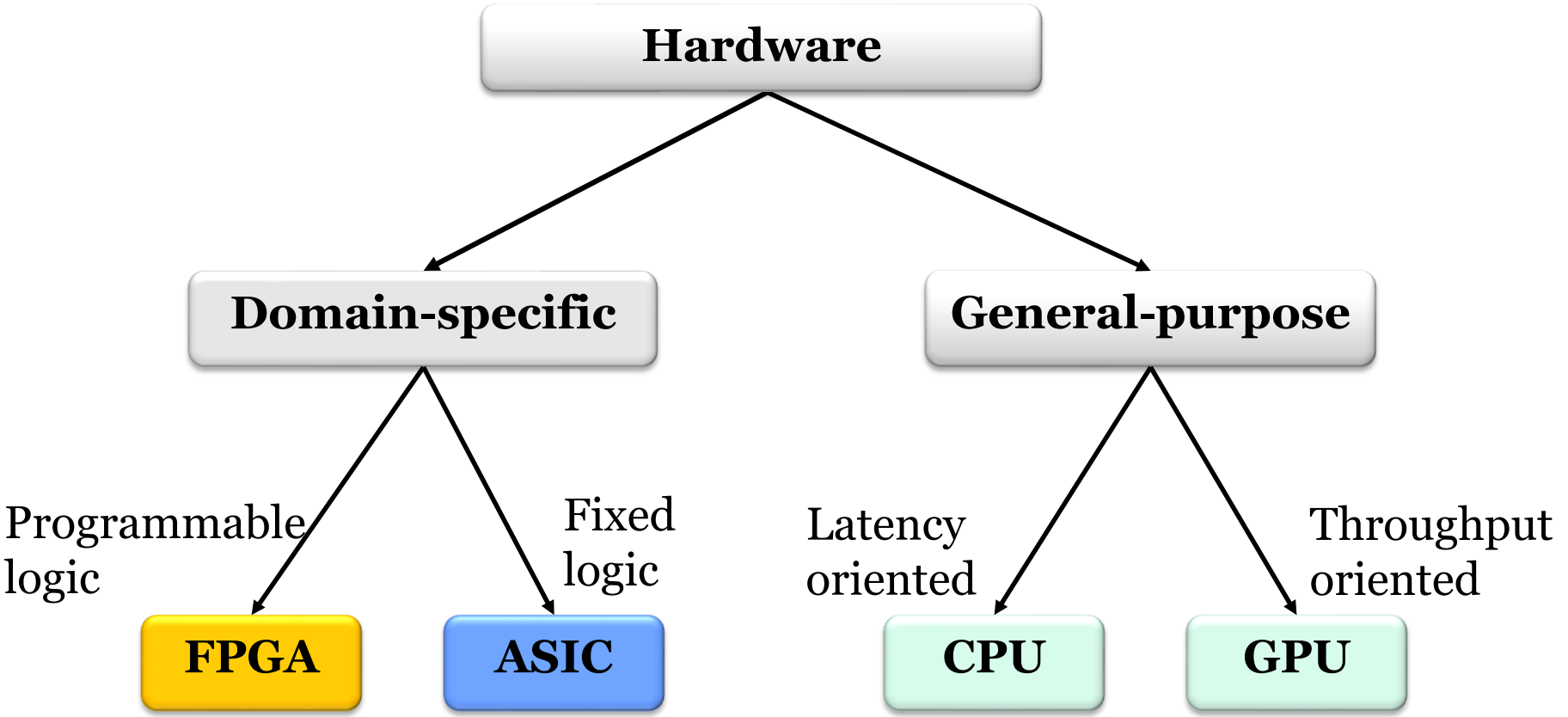
Metrics	LeNet-5	AlexNet	VGG-16	GoogLeNet (v1)	ResNet-50
Top-5 error	n/a	16.4	7.4	6.7	5.3
Input Size	28x28	227x227	224x224	224x224	224x224
# of CONV Layers	**2**	**5**	**16**	**21 (depth)**	**49**
Filter Sizes	5	3, 5,11	3	1, 3 , 5, 7	1, 3, 7
# of Channels	1, 6	3 - 256	3 - 512	3 - 1024	3 - 2048
# of Filters	6, 16	96 - 384	64 - 512	64 - 384	64 - 2048
Stride	1	1, 4	1	1, 2	1, 2
# of Weights	2.6k	2.3M	14.7M	6.0M	23.5M
# of MACs	283k	666M	15.3G	1.43G	3.86G
# of FC layers	**2**	**3**	**3**	**1**	**1**
# of Weights	58k	58.6M	124M	1M	2M
# of MACs	58k	58.6M	124M	1M	2M
Total Weights	**60k**	**61M**	**138M**	**7M**	**25.5M**
Total MACs			**15.5G**	**1.43G**	**3.9G**

Source: Joel Emer, ISCA Tutorial, 2017

What does it mean ?

End of Moore's Law + **Exponential Increase in Compute Requirements** = **Needs New Approach**

Current State of the Art in Neural Algorithms HW Computing

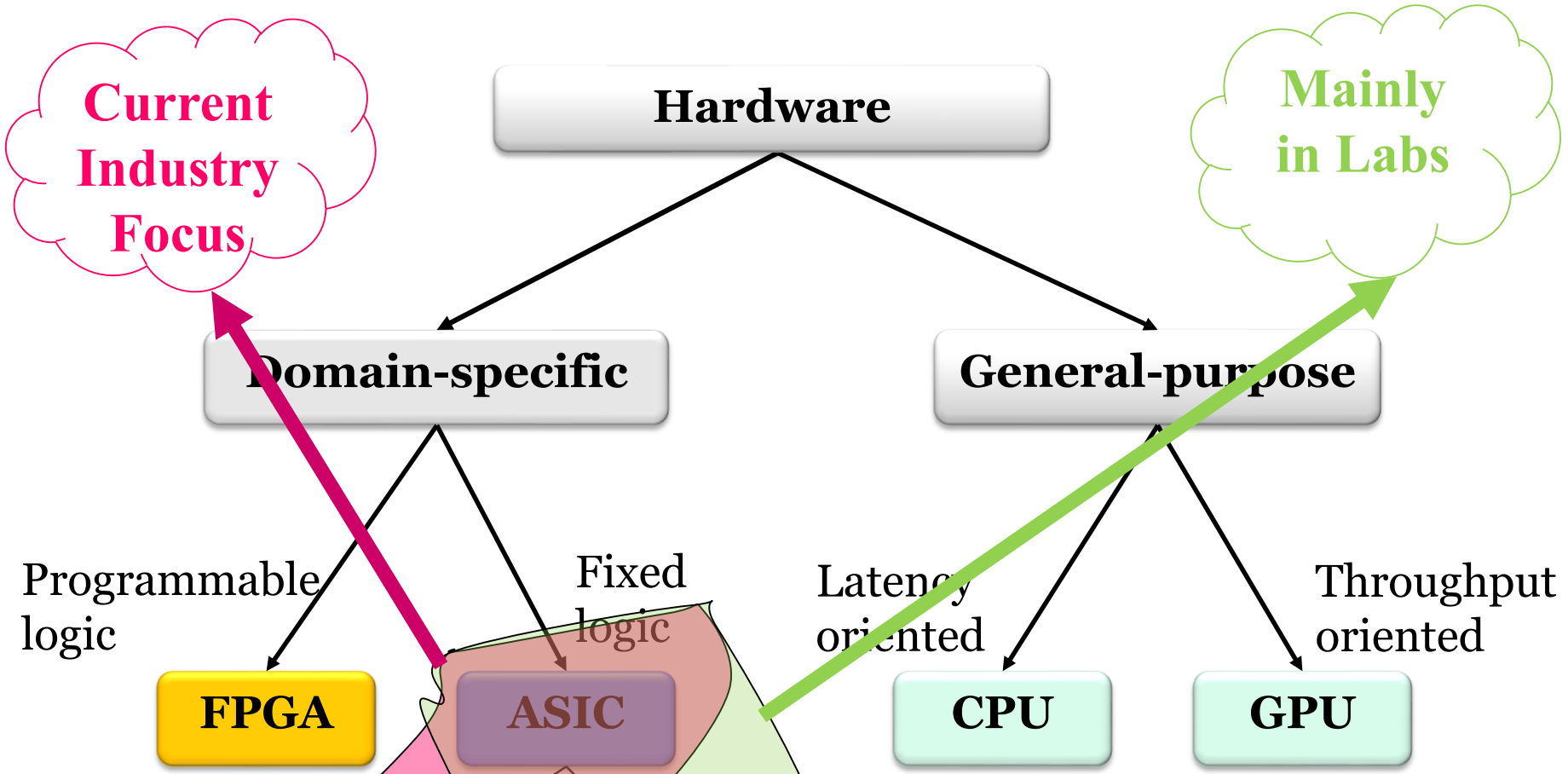


- General; requires HDL
- Moderate performance & efficiency

- Specific: executes STDP
- HP & efficiency
- Expensive, 40MB local memory Example: IBM TrueNorth

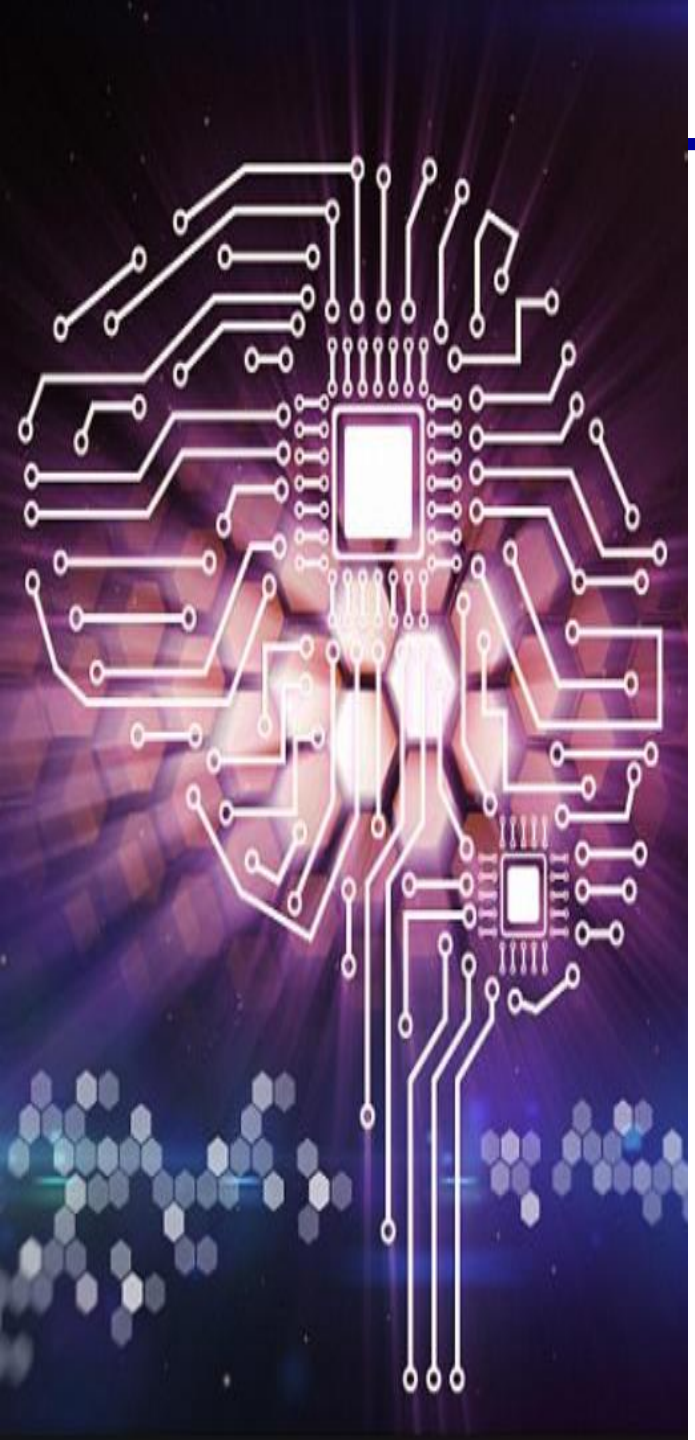
- **Most general; common programming languages**
- **Lowest power efficiency and performance**
- **Memory separate from chip**
- **Example: Google deep learning study**

Current State of the Art in Neural Algorithms HW Computing



- General; requires HDL
- Moderate performance efficiency

- Most general; common programming languages
- Lowest power efficiency and performance
- Memory separate from chip
- Example: Google deep learning study



Agenda

- Fundamental Trends
- AI – The Emerging Industrial Revolution
- Neuromorphic Computing Systems
- Our AI-Chips
- Future Direction

AI HW is inspired by Nature – Biological neuron

AI-Chips are inspired by biology

➔ parallel computation.

AI HW is inspired by Nature – Biological neuron

AI-Chips are inspired by biology

➔ parallel computation.

Latest digital DL processors: ~10TOPS/W

Synapse op. in **brain**: 0.1~1 fJ/op
1,000~10,000 TOPS/W
=1~10 POPS/W

- # of neurons: ~10^{11}
- # of synapses: ~10^{15}
- Power consumption: ~ 20 W;
- Operating frequency: 10~100 Hz

- Works in parallel: 10^6 parallelism vs. <10^1 for PC (VN)
- Faster than current computers: i.e. simulation of a **5 s** brain activity takes **~500 s** on state-of-the- art supercomputer

Different approaches of AI-Chips

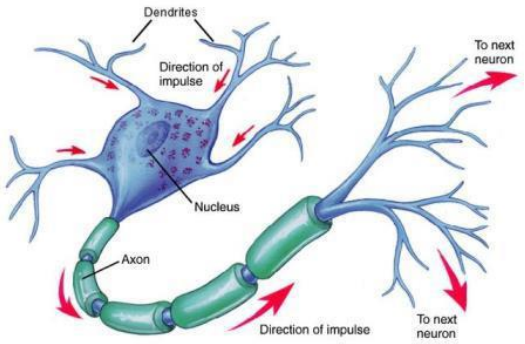
Poor/Simple

Good/Complex

Neuron	Digital, Analog. LIF. . . .	Izhikevich model	Huxley-Hodgkin model	. . .
Synapse	MAC (weighted sum) . .	Spiking STDP	. . . .	Many nonlinear properties . . . .

Generally Used in DL algorithms

Frequency	10~100 Hz (brain)

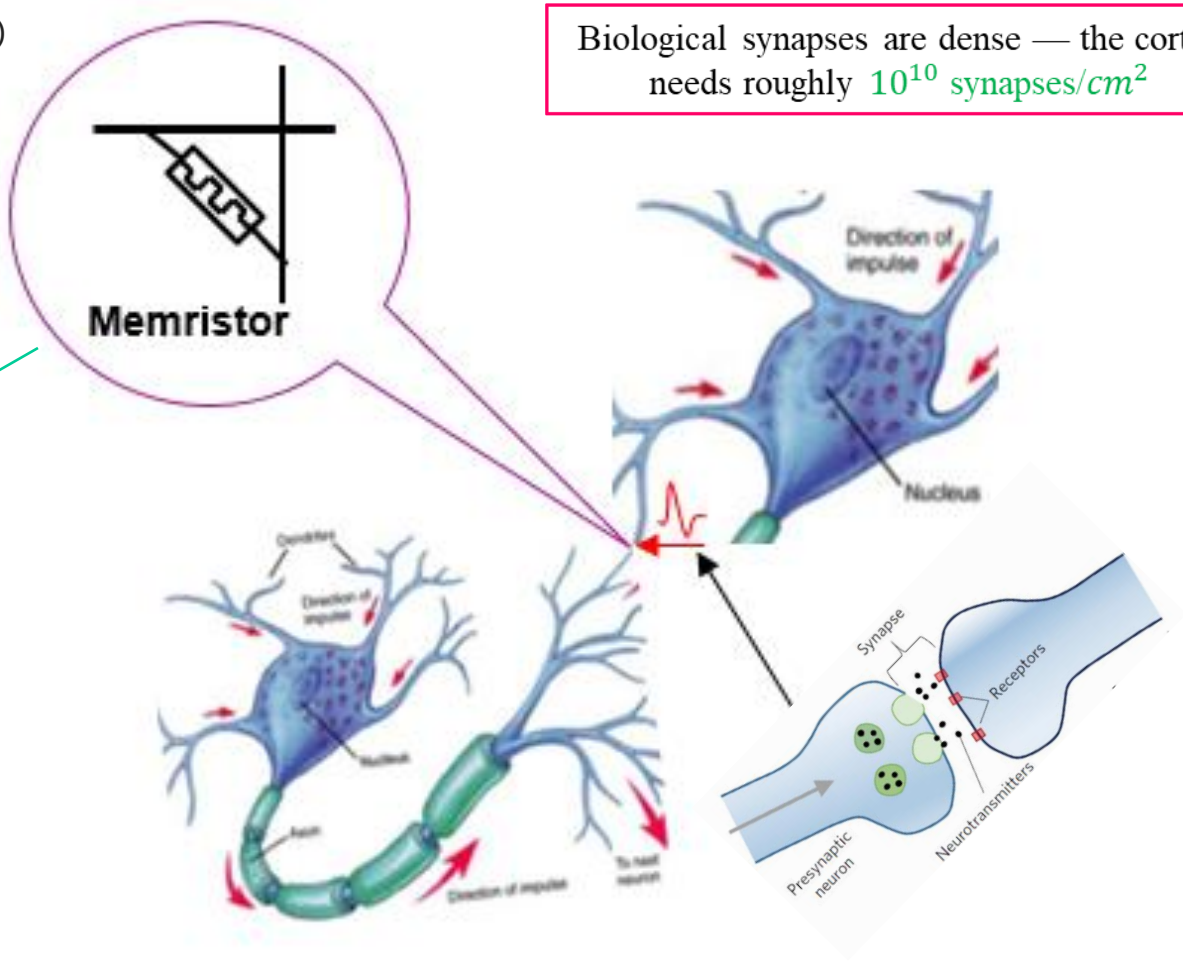


Memristor for Synapse Design

(Chua, 1971)

Biological synapses are dense — the cortex needs roughly 10^{10} synapses/cm^2

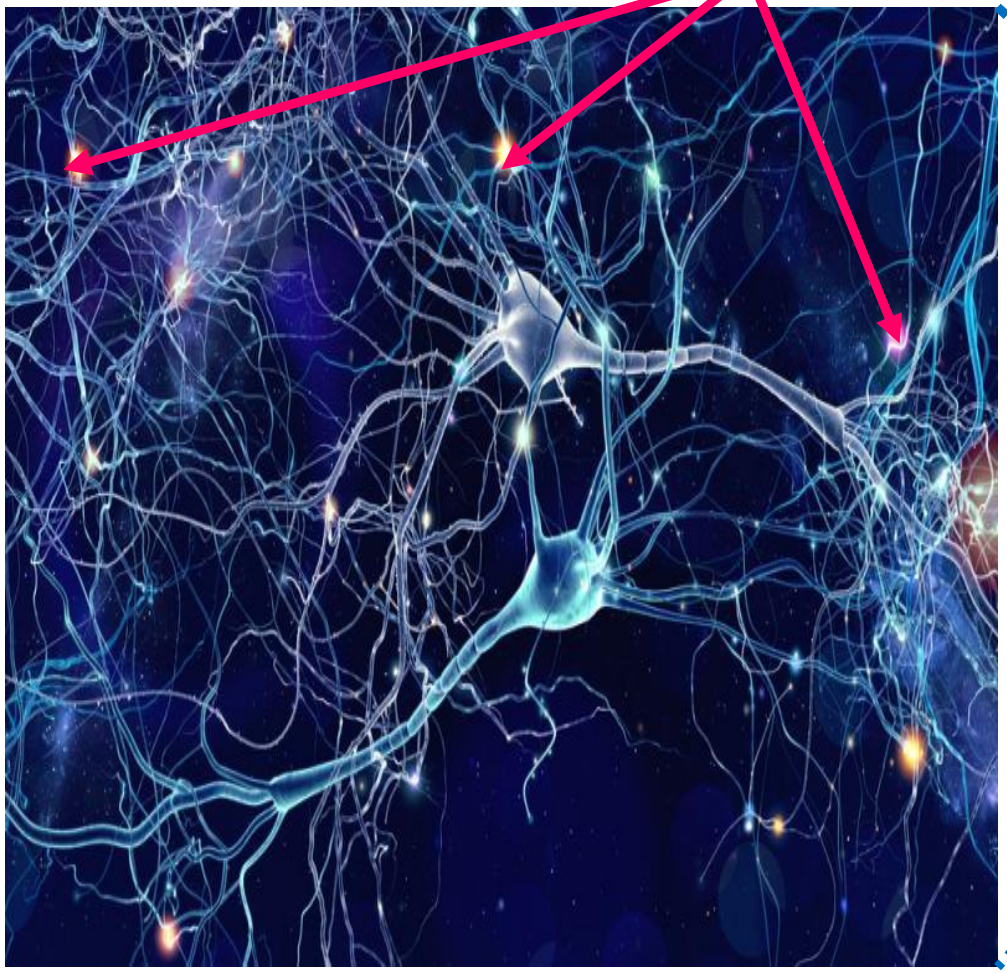
The electrical resistor is not constant but depends on the history of current that had previously flowed through the device.

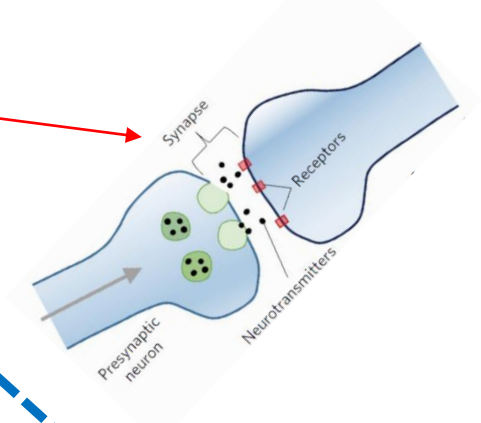


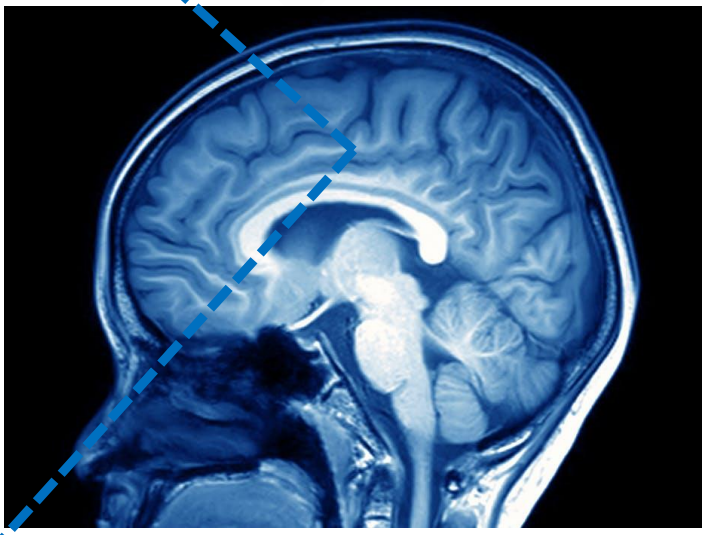
❖ Voltage **pulses** can be applied to a **memristor** to change its **resistance**, just as **spikes** can be applied to a **synapse** to change its **weight**.

How biological neurons learn?

Brain is a large network of neurons connected and communicating via synapses

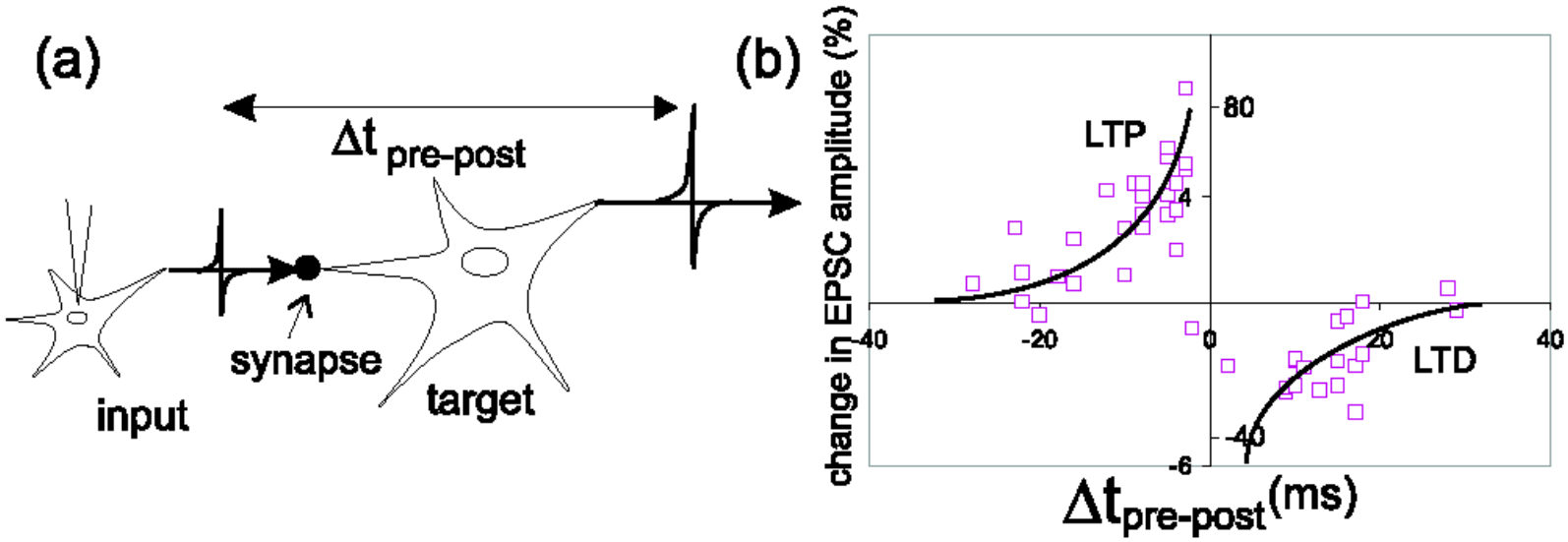






How biological neurons learn?

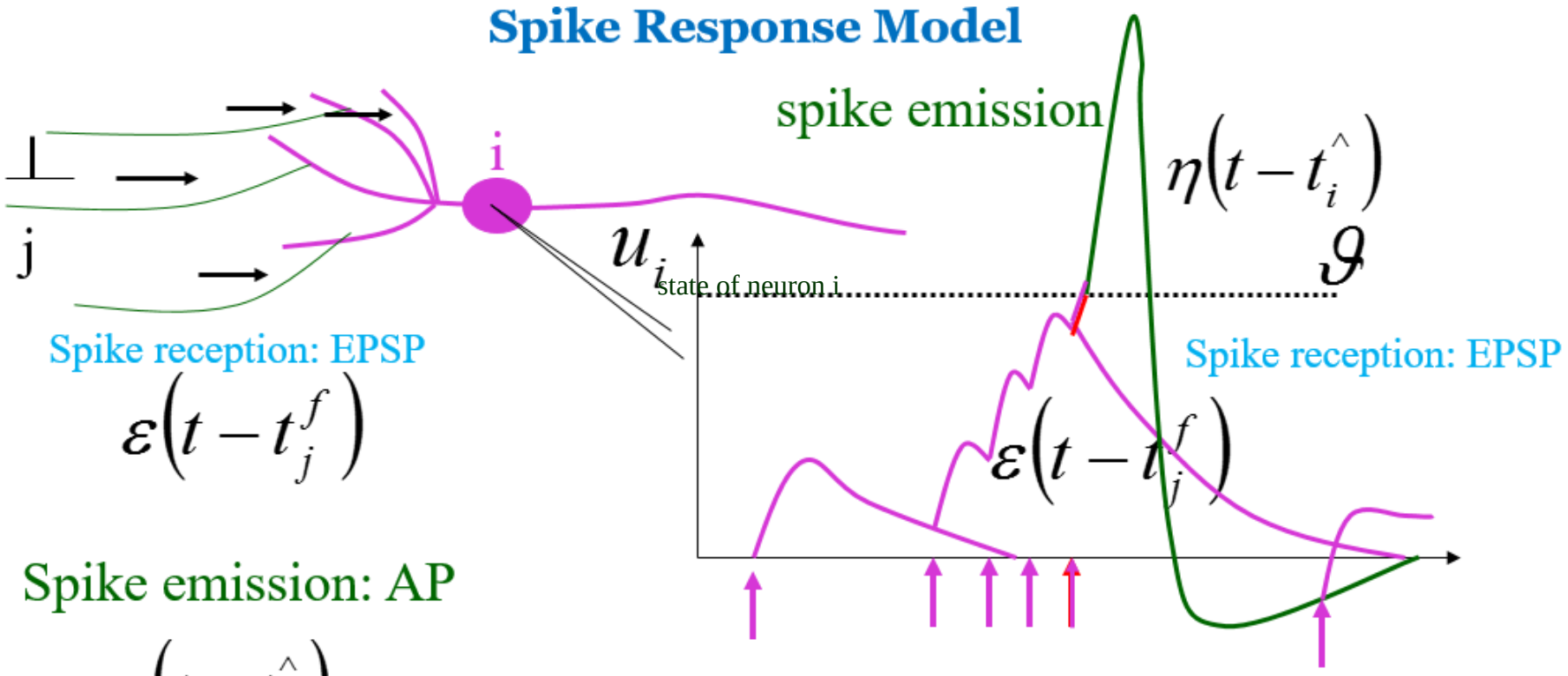
- Learning rules based on STDP specify changes in synaptic strength depending on the time interval between each pair of presynaptic and postsynaptic events.



Spike-timing-dependent plasticity (STDP)

- If the presynaptic neuron fire **before** the postsynaptic neuron within a preceding 20ms, LTP occurs
- If the presynaptic neuron fire **after** the postsynaptic neuron within the following 20ms, LTD occurs

Spiking Neuron Model



Spike emission: AP

$$\eta(t-\hat{t}_i)$$

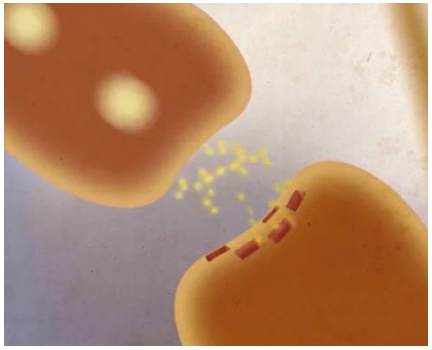
reset of the membrane potential (action potential)

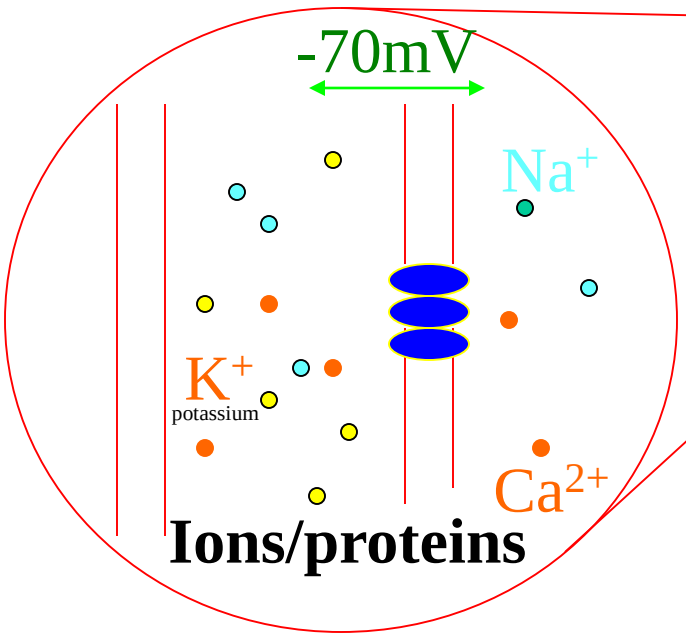
$$u_i(t) = \eta(t-\hat{t}_i) + \sum_j \sum_f w_{ij}\, \varepsilon(t-t_j^f)$$

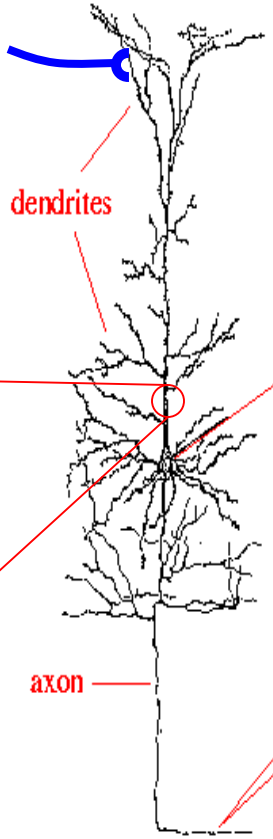
$$u_i(t) = \vartheta \Rightarrow \text{Firing:} \quad \hat{t}_i = t$$

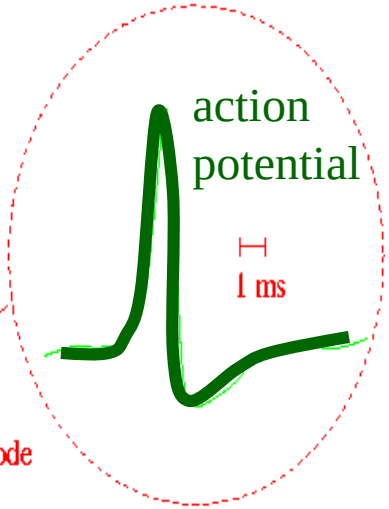
[Ref. 18]

Spiking Neuron Model- Molecular Basis

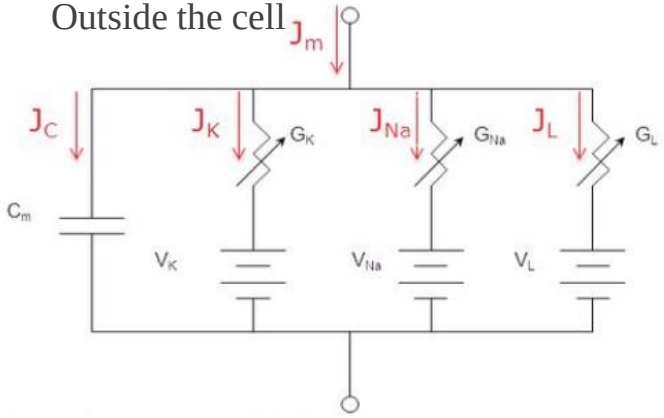






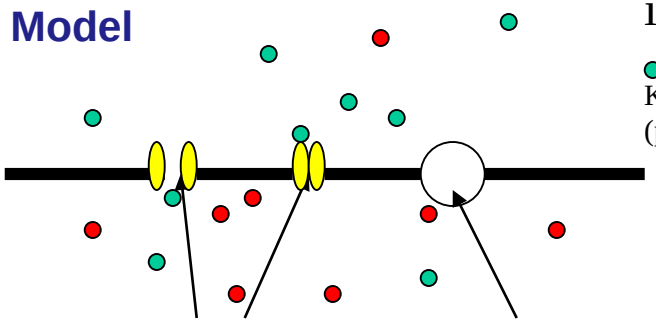


Spiking Neuron Model- Molecular Basis



Hodgkin-Huxley Model

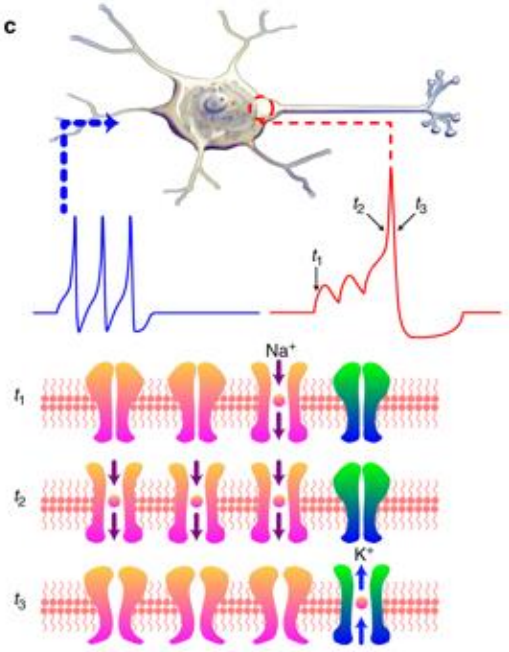
~



$$J_c = C_m \partial V_m / \partial t \qquad J_{Na^+} = G_{Na^+}\left(V_m - V_{Na^+}\right)$$

$$J_{K^+} = G_{K^+}\left(V_m - V_{K^+}\right) \qquad J_L = G_L\left(V_m - V_L\right)$$

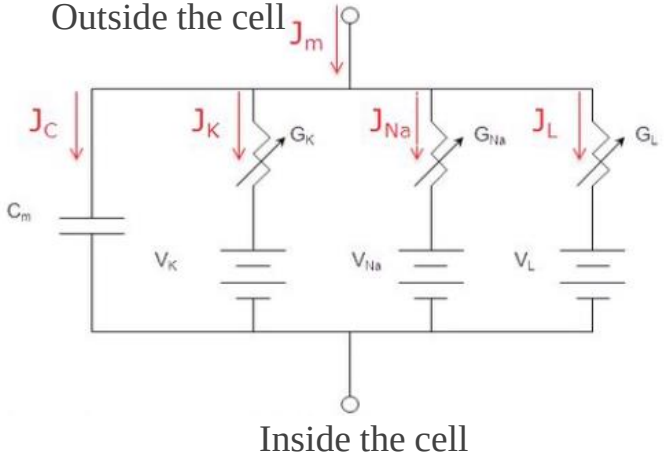
$$J_m = J_c + J_{K^+} + J_{Na^+} + J_L$$



$$J_m = C_m \partial V_m / \partial t + G_{K^+}\left(V_m - V_{K^+}\right) + G_{Na^+}\left(V_m - V_{Na^+}\right) + G_L\left(V_m - V_L\right)$$

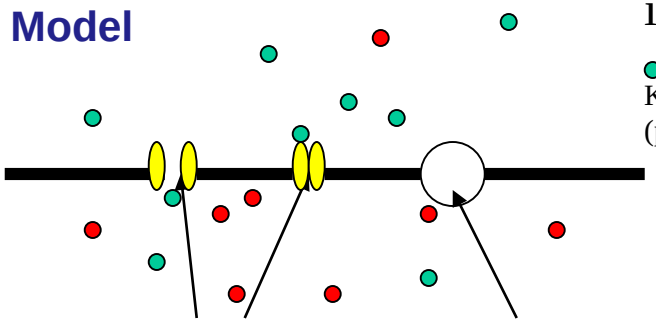
[Ref. 18]

Spiking Neuron Model- Molecular Basis



Hodgkin-Huxley Model

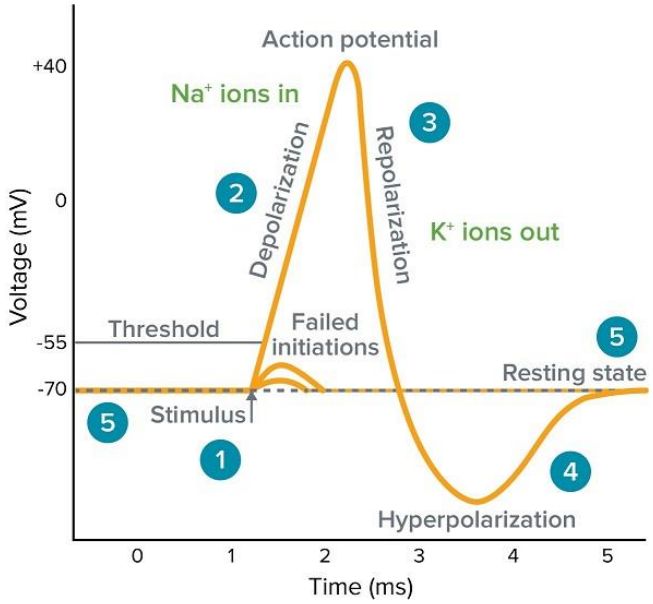
~



$$J_c = C_m \partial V_m / \partial t \qquad J_{Na^+} = G_{Na^+}\left(V_m - V_{Na^+}\right)$$

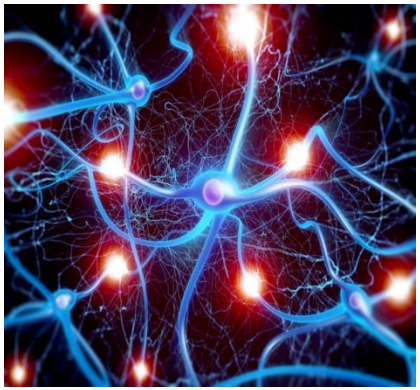
$$J_{K^+} = G_{K^+}\left(V_m - V_{K^+}\right) \qquad J_L = G_L\left(V_m - V_L\right)$$

$$J_m = J_c + J_{K^+} + J_{Na^+} + J_L$$

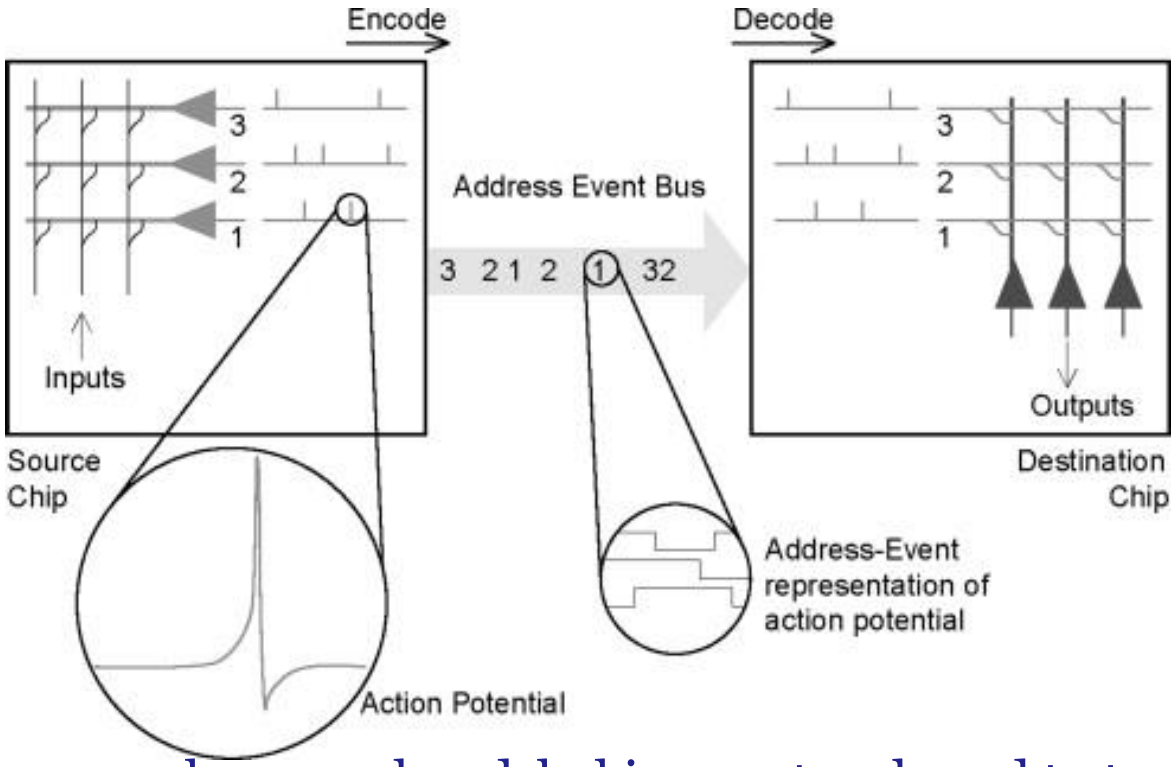


$$J_m = C_m \partial V_m / \partial t + G_{K^+}\left(V_m - V_{K^+}\right) + G_{Na^+}\left(V_m - V_{Na^+}\right) + G_L\left(V_m - V_L\right)$$

Wiring via AER (Address Event Representation)

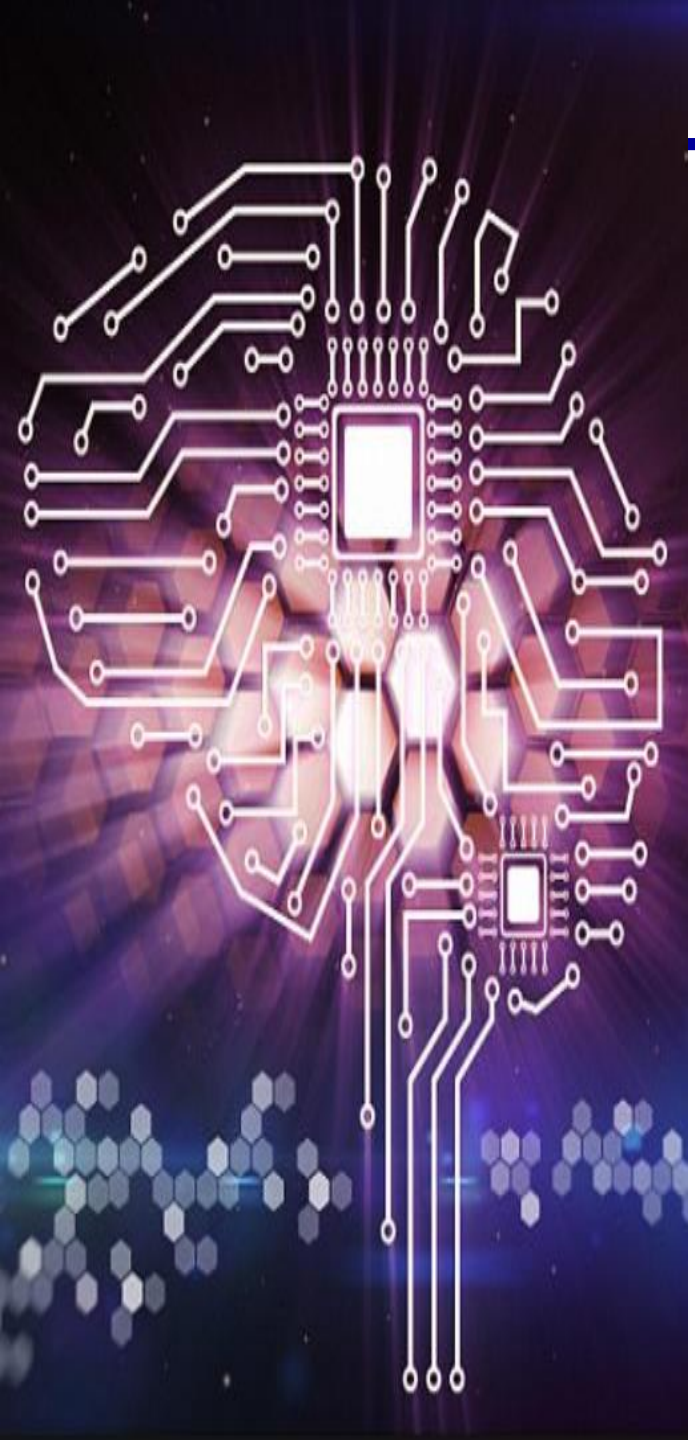


(Courtesy: iStock/Henrik5000)



Ref. 4

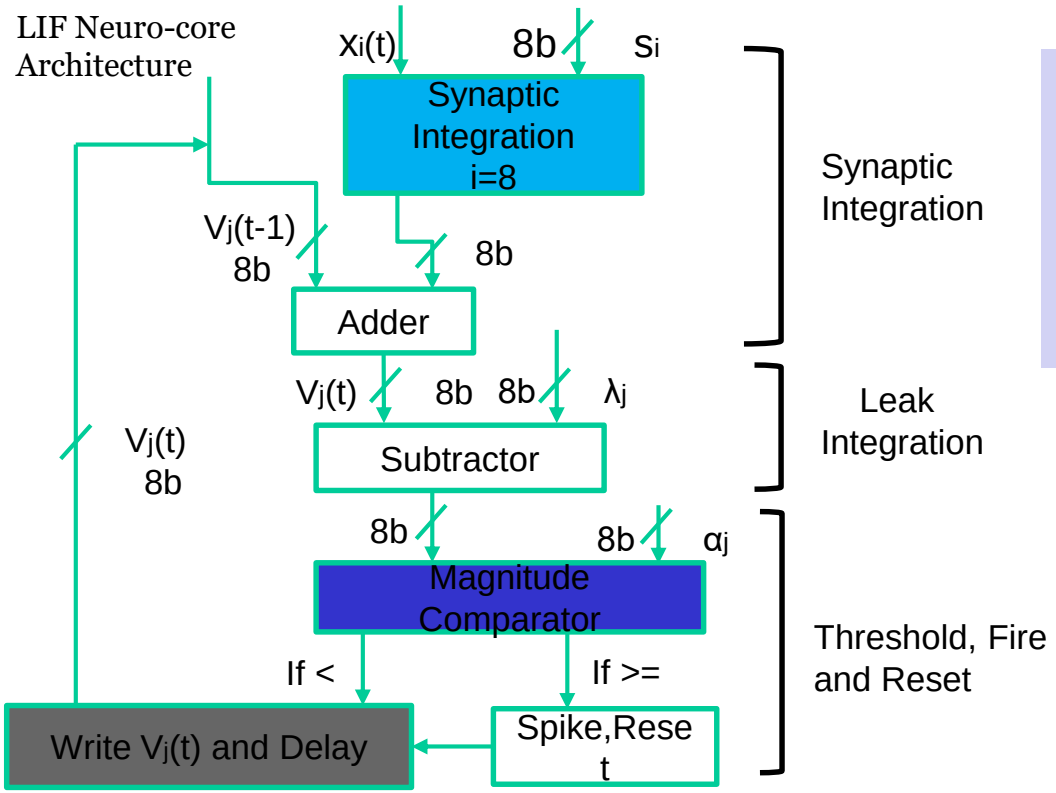
- AER is an asynchronous handshaking protocol used to transmit signals between neuromorphic systems.



Agenda

- Fundamental Trends
- AI – The Emerging Industrial Revolution
- Neuromorphic Computing Systems
- Our AI-Chips
- Future Direction

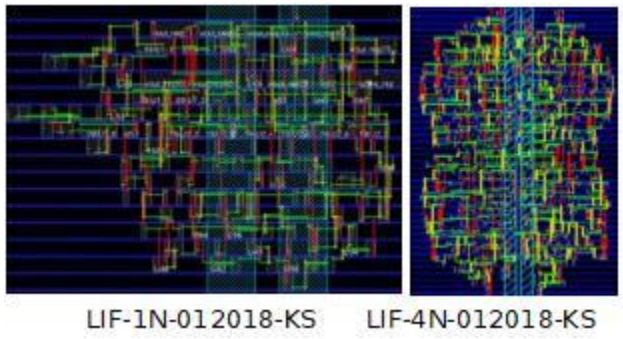
LIF Neuro-core for NASH System



- Xi(t) – Spike input to the synapse
- Si – synaptic weight
- Vj(t) – Membrane potential
- αj – Neuron threshold
- Λj – Leak value

Table 1: Area Evaluation

Item	NC-1N	NC-4N
Cell Internal Power	6.9680 µW	20.5040 µW
Net Switching Power	4.8271 µW	14.8272 µW
Total Dynamic Power	11.7950 µW	35.3312 µW
Cell Leakage Power	4.6943 µW	14.3147 µW



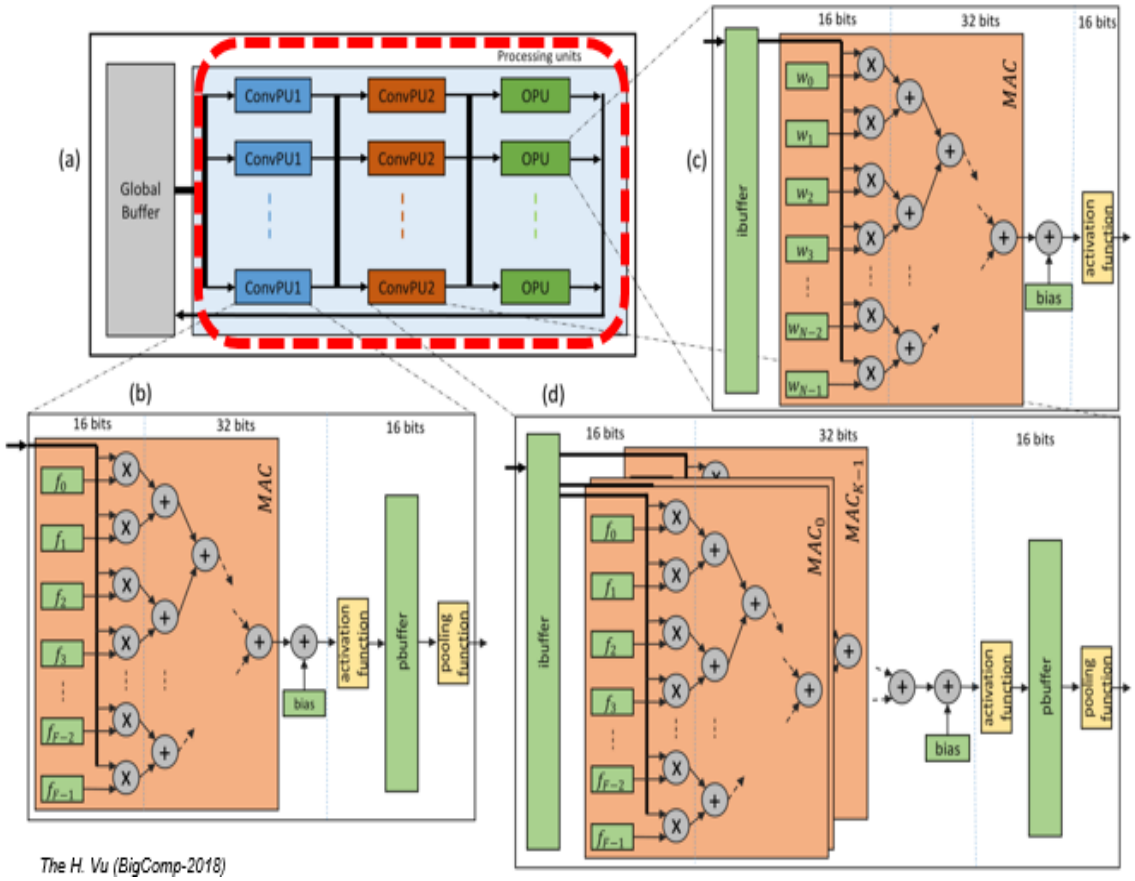
Placement of LIF-1N (Left) and LIF-4N (right)

Table 1: Power Evaluation

Item	NC-1N	NC-4N
Combinational Area	186.998 µm	562.856001 µm
Non-Comb Area	47.88002 µm	213.864000 µm
Total Cell Area	234.878002 µm	776.720001 µm

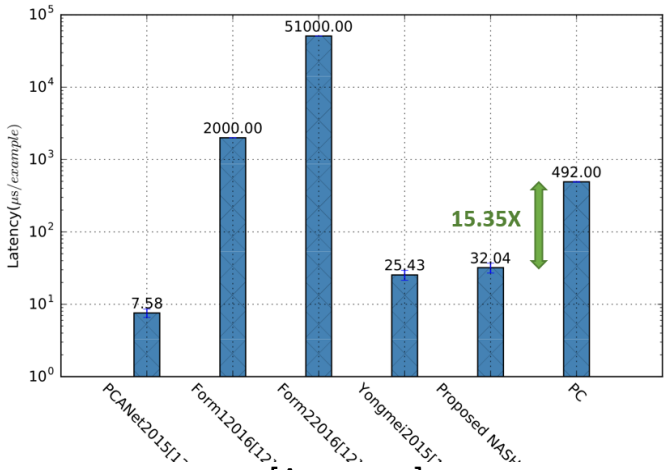
Kanta Suzuki, Yuichi Okuyama, Abderazek Ben Abdallah, "Hardware Design of a Leaky Integrate and Fire Neuron Core Towards the Design of a Low-power Neuro-inspired Spike-based Multicore SoC", Proc. Of IPSJ, 2018

Application I
Neuro-inspired Hardware System for Image Recognition

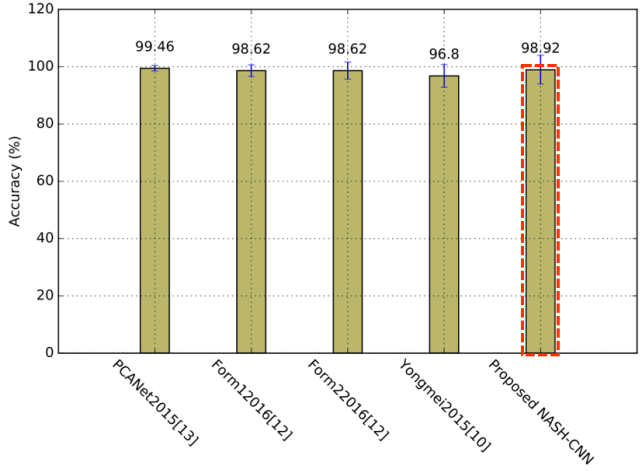


The H. Vu (BigComp-2018)

[Execution time]



[Accuracy]



The H. Vu, Ryunosuke Murakami, Yuichi Okuyama, Abderazek Ben Abdallah, "Efficient Optimization and Hardware Acceleration of CNNs towards the Design of a Scalable Neuro-inspired Architecture in Hardware", Proc. of the IEEE International Conference on Big Data and Smart Computing (BigComp-2018), January 15-18, 2018

NASH: Low-power Event-driven Adaptive Neuromorphic System for Autonomous Cognitive Behaviour

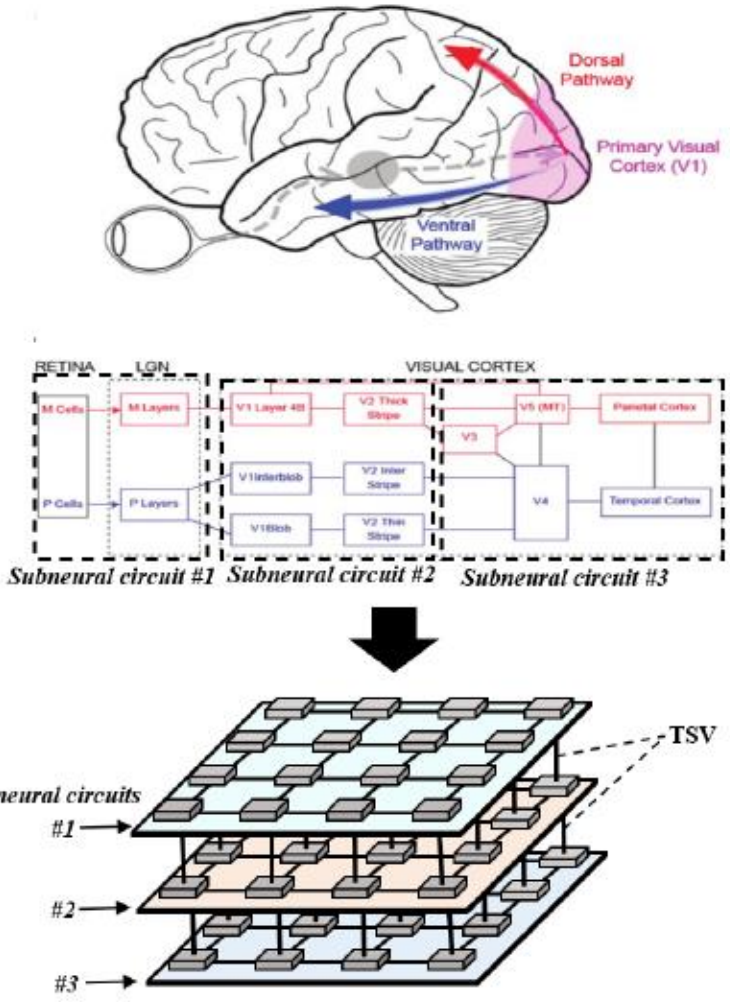
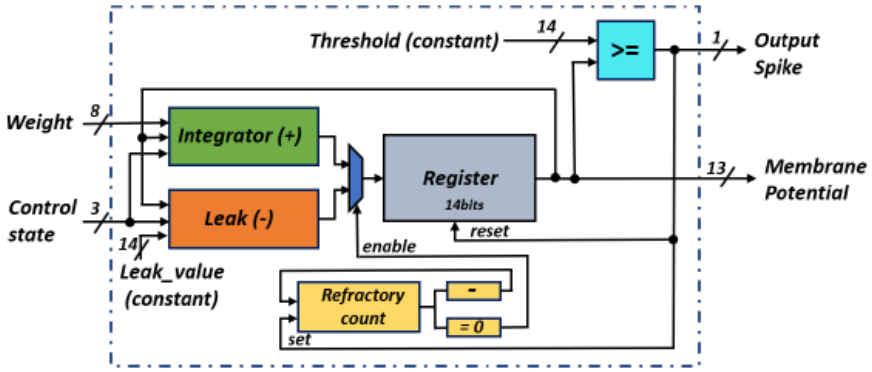
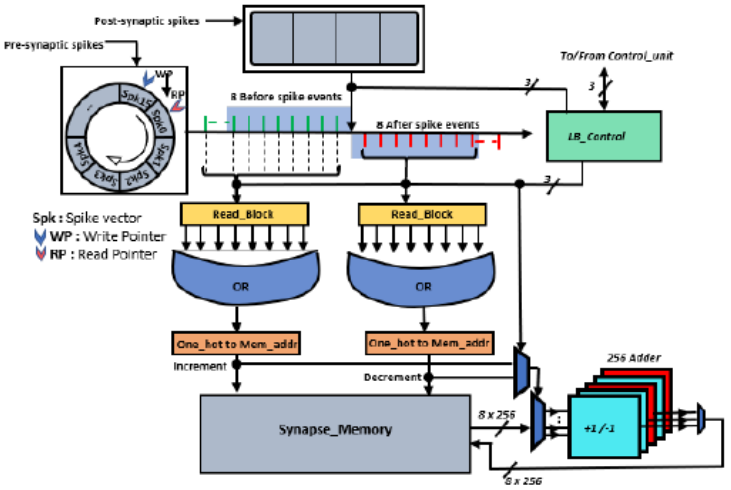


FIGURE 1: An illustration of large scale neural task mapping on a 3D neuromorphic architecture. (a) illustration of information flow path from retina to primary visual cortex (V1) in the human brain (b) Block diagram of connections among



LIF Neuron



Learning Module

NASH: Low-power Event-driven Adaptive Neuromorphic System for Autonomous Cognitive Behaviour

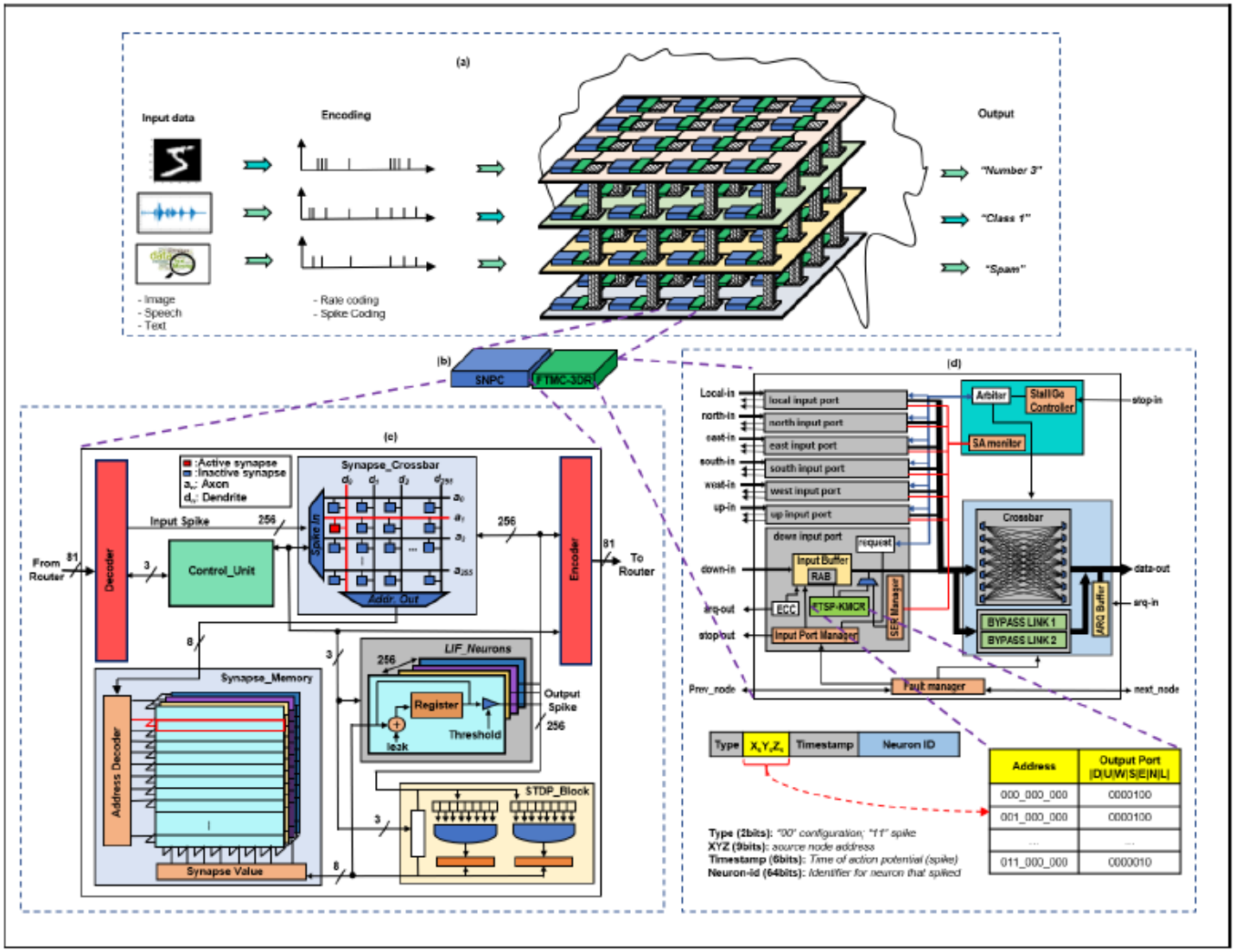


FIGURE 3: High level view of the NASH System Architecture: (a)NASH system architecture illustrated in a 4×4 topology. (b) A single node in the NASH system comprising of an SNPC and a FTMC-3DR. (c) An architecture of the SNPC comprising of the synapse memory, synapse crossbar, LIF neurons, control unit, encoder/decoder, and an STDP learning block (d) Architecture of the FTMC-3DR comprising of the input ports, crossbar, swich allocator, arbiter and flow control.

NASH: Low-power Event-driven Adaptive Neuromorphic System for Autonomous Cognitive Behaviour

TABLE 3: Hardware Complexity of the KMCR and FTSP-KMCR under the benchmarks.

System	Spike Injector [9]				SNPC (This work)	
	KMCR		FTSP-KMCR		KMCR	FTSP-KMCR
Testbench	Inv. Pen.	Wis.	Inv. Pen.	Wis.	MNIST	MNIST
Area (mm^2)	0.102	0.346	0.108	0.365	0.549	0.616
Power (mW)	10.13	34.20	10.64	35.92	562.69	647.61

TABLE 4: Comparison results between the proposed NASH and existing works.

Parameters/Systems	Loihi [45]	ODIN [38]	Seo et al [46]	This work
Benchmark	MNIST	MNIST	MNIST	MNIST
Accuracy (%)	84	85	77.2	79.4
Number of Cores	128	1	1	27
Number of Neurons / core	max. 1024	256	256	256
Neuron Model	IF	LIF and Izh.	LIF	LIF
Neuron Update	serial	serial	serial	parallel
Membrane Potential Resolution	16 bits	11-bits	8-bits	14 bits
Number of Synapses /core	114k to 1M	65K	64k	65k
Synaptic Connection	Crossbar	Crossbar	Crossbar	Crossbar
Synapse Resolution	9- to 1-bit	4-bit	1-bit	8-bits
Learning Rule	On-chip STDP	On-chip Stochastic SDSP	on-chip STDP	On-chip STDP
Memory Technology	Memristor	SRAM	SRAM	SRAM
Interconnect	2D-NoC	2D-NoC	-	3D-NoC
Implementation	Digital	Digital	Digital	Digital
Technology	14-nm FinFet	28-nm FD-SOI CMOS	45nm SOI-CMOS	45-nm NANGATE CMOS
core Area (mm^2) (excl. pad)	N/A	0.086 (ODIN)	0.8	5.03
Supply Voltage (v)	0.75	0.55 - 0.1	-	1.1
Power ($_mW$)	6.45	0.28	N/A	9.588

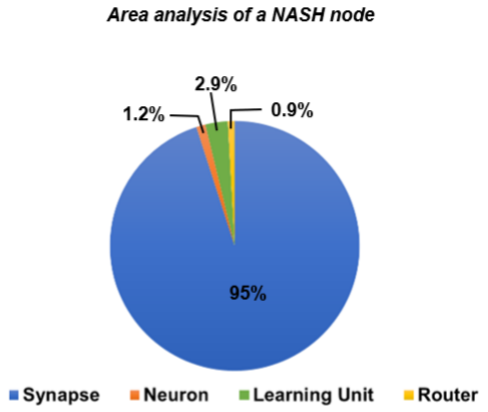


FIGURE 15: Area analysis of NASH node

References

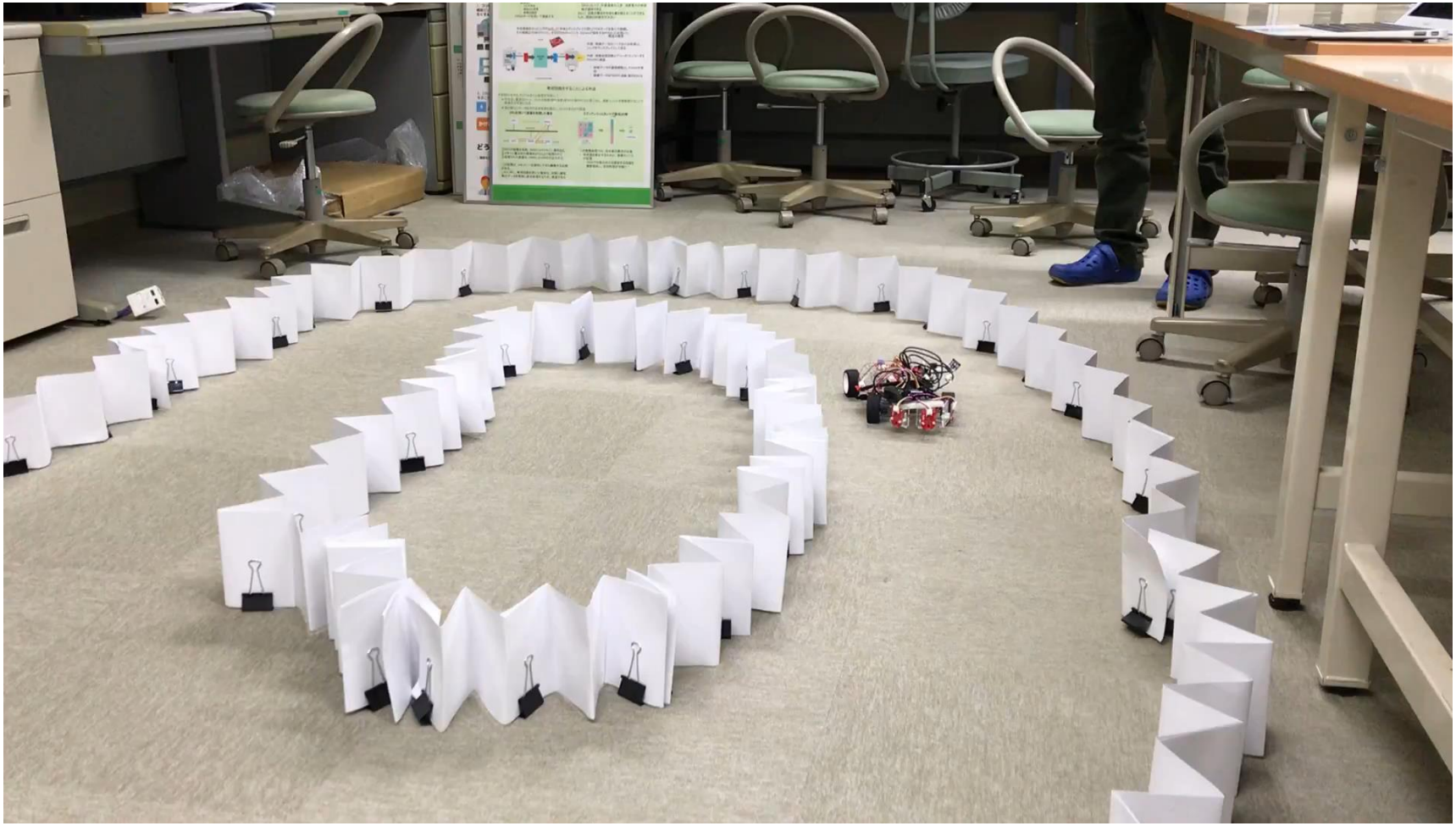
•The H. Vu,Yuichi Okuyama, Abderazek Ben Abdallah, "Comprehensive Analytic Performance Assessment and K-means based Multicast Routing Algorithms and Architecture for 3D-NoC of Spiking Neurons," ACM Journal on Emerging Technologies in Computing Systems (JETC), Special Issue on Hardware and Algorithms for Learning On-a-chip for Energy-Constrained On-Chip Machine Learning, Vol. 15, No. 4, Article 34, October 2019. doi: 10.1145/3340963

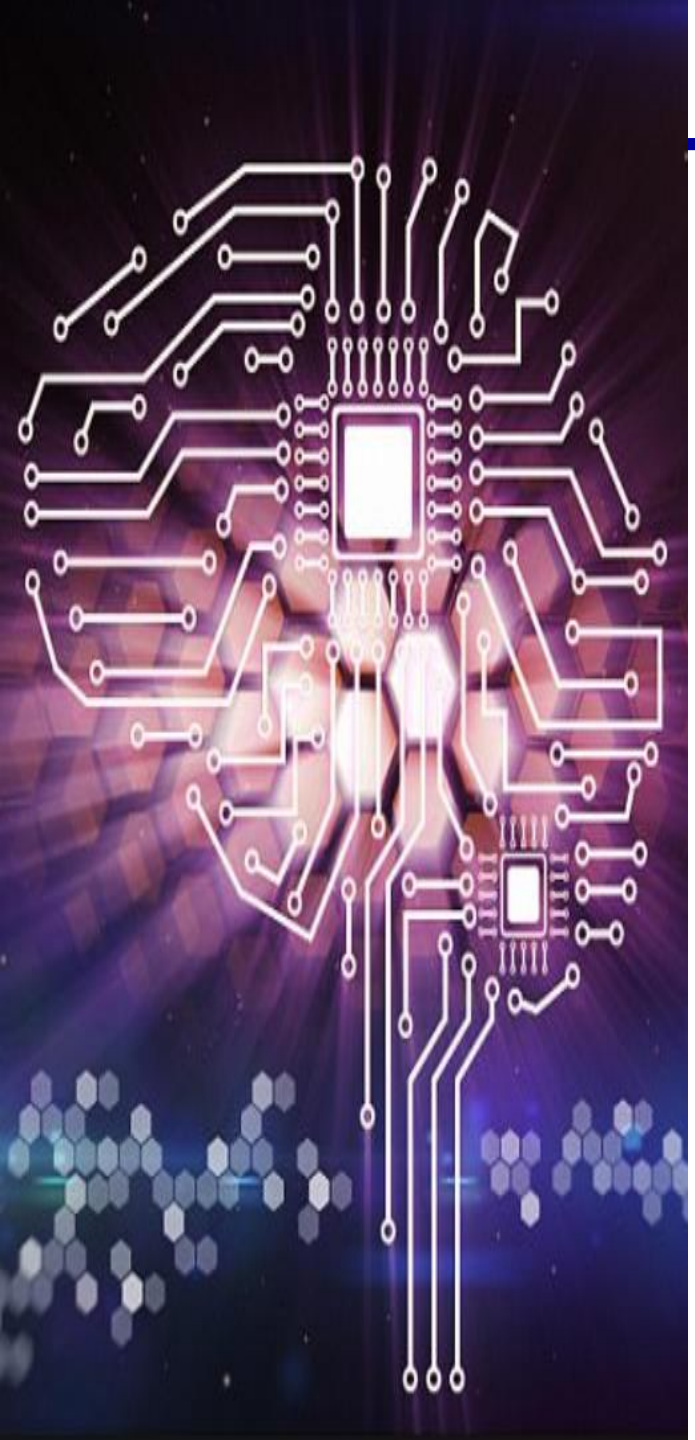
•The H. Vu, Ogbodo Mark Ikechukwu, and Abderazek Ben Abdallah, "Fault-tolerant Spike Routing Algorithm and Architecture for Three Dimensional NoC-Based Neuromorphic Systems'", IEEE Access, vol. 7, pp. 90436-90452, 2019.

•K. N. Dang, A. B. Ahmed, A. Ben Abdallah and X. Tran, "TSV-OCT: A Scalable Online Multiple-TSV Defects Localization for Real-Time 3-D-IC Systems," IEEE Transactions on Very Large Scale Integration (VLSI) Systems, vol. 28, no. 3, pp. 672-685, 3/2020. doi: 10.1109/TVLSI.2019.2948878

Demo 1

Demo 2





Agenda

- Fundamental Trends
- AI – The Emerging Industrial Revolution
- Neuromorphic Computing Systems
- Our AI-Chips
- Future Direction

Conclusions

❖ Memory access in AI-Chip is the bottleneck

Worst case: ALL memory R/W are DRAM accesses

Ex. AlexNet [NIPS 2012] has 724M MACs → 2896M DRAM accesses required

Possible HW/SW techniques to cope with the memory access problem:

❖Advanced Storage Technology

- ✓ Embedded DRAM (eDRAM) → Increase on-chip storage capacity
- ✓ 3D Stacked DRAM → Increase memory bandwidth
- ✓ Use memristors as programmable weights (resistance)

❖Reduce size of operands for storage/compute

- ✓ Floating point → Fixed point
- ✓ Bit-width reduction

❖Reduce number of operations for storage/compute

- ✓ Network Pruning; Compact Network Architectures

ACKNOWLEDGEMENT

Faculty/PI

BEN ABDALLAH Abderazek

Visiting Researchers

Dr. DANG Nam Khanh (2020, 2019)

Dr. HUANG Huakun (2020)

Doctoral Students

D2 OGBODO M. Ikechukwu

Research: NASH

D1 WANG Zhishang

Research: AEBiS

D1 FUKUCHI Tomohide

Research: NeuroSys

D1 LIANG Yuxiao

Research: T-AEBiS

D1 WILLIAMS Y. Yerima

Research: NeuroSys

D1 WANG Jiangkun

Research: AIRBiS

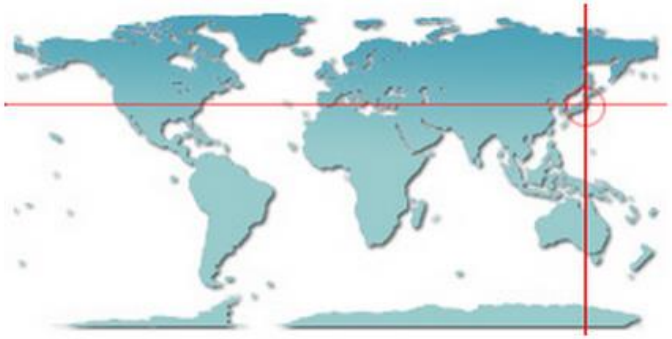
Thank you!

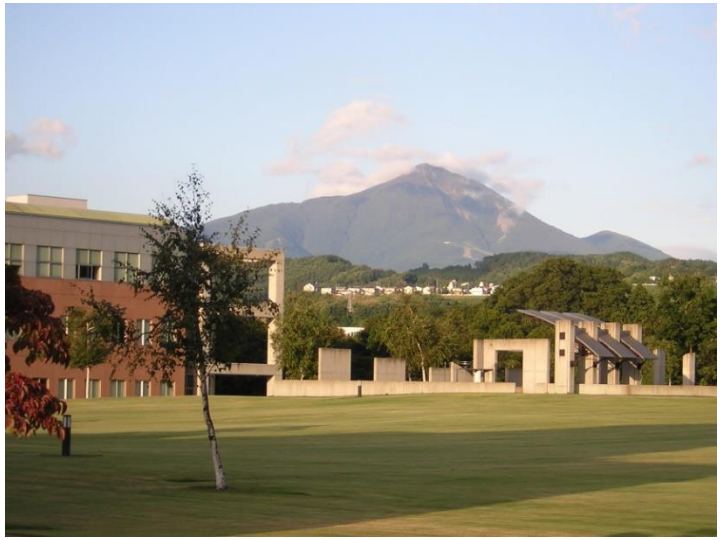
ありがとうございました

Abderazek BEN ABDALLAH

University of Aizu

benab@u-aizu.ac.jp





to Advance Knowledge for Humanity