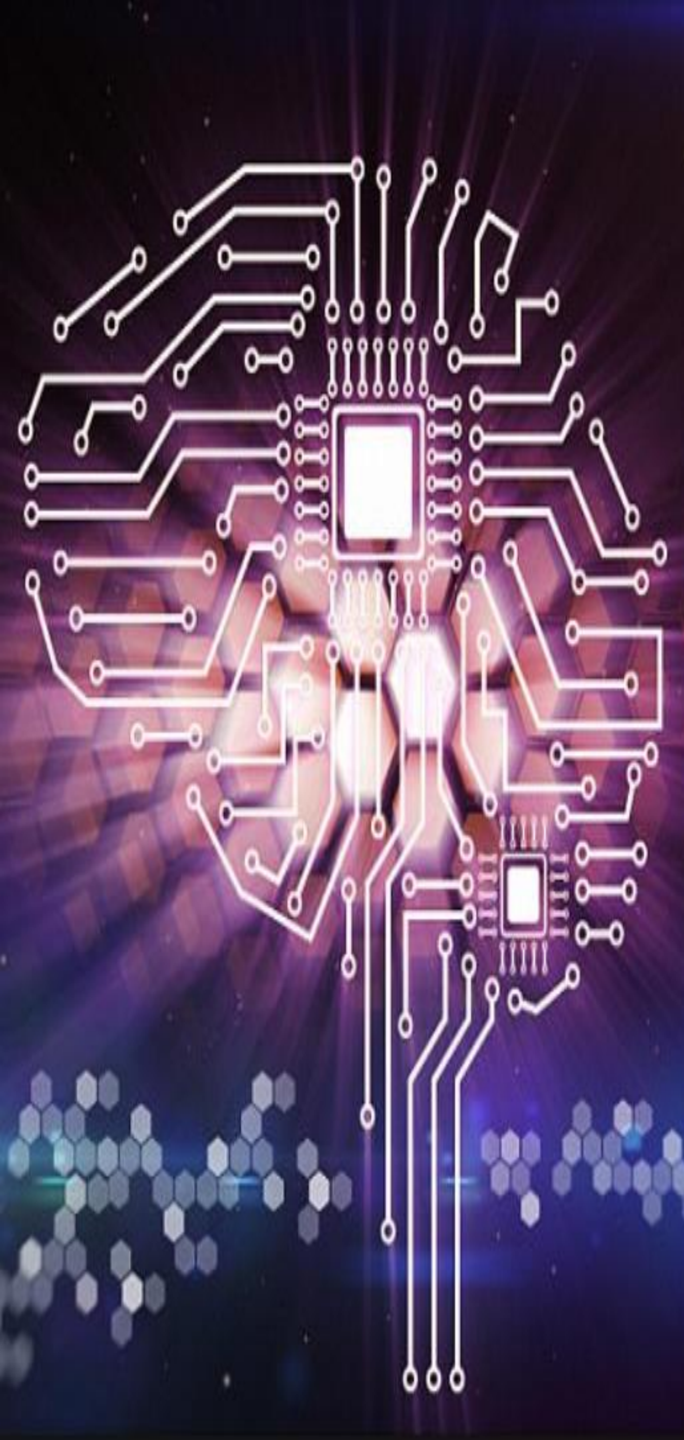
The Future of Machine Learning: Neuromorphic Chips

Abderazek Ben Abdallah
Adaptive Systems Laboratory
The University of Aizu, Aizu, Japan
Email: benab@u-aizu.ac.jp



Agenda

- Fundamental Trends
- AI – The Emerging Industrial Revolution
- AI at the Edge
- ASL Neuromorphic Chips
- Conclusions

AI-Chips are … everywhere

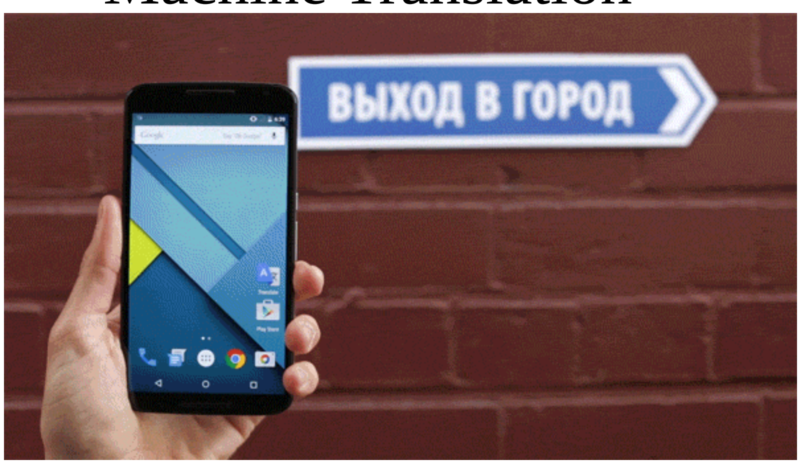
Self-driving Car

Bottom Image source: edition.cnn.com

Smart Robots

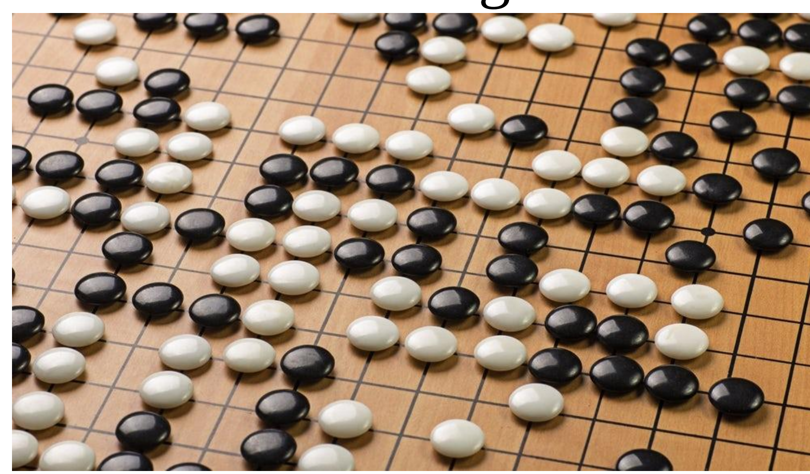
Image source: roboticsbusinessreview.com

Machine Translation



Bottom Image source: missqt.com

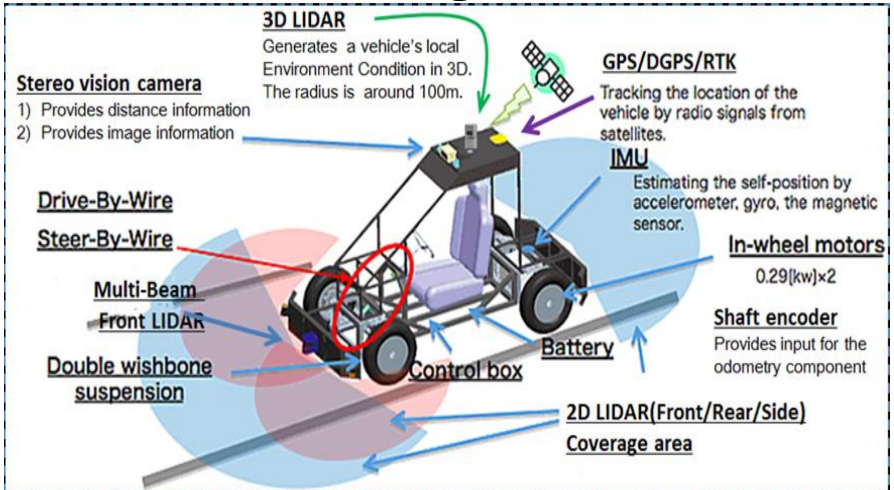
Gaming



Bottom Image Source: newatlas.com

AI-Chips are … everywhere

Self-driving Car

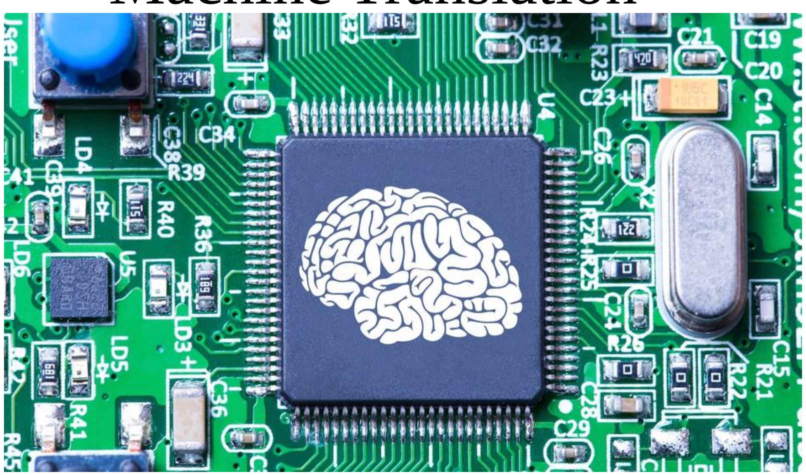


Bottom Image source: edition.cnn.com

Smart Robots

Image source: roboticsbusinessreview.com

Machine Translation



Bottom Image source: missqt.com

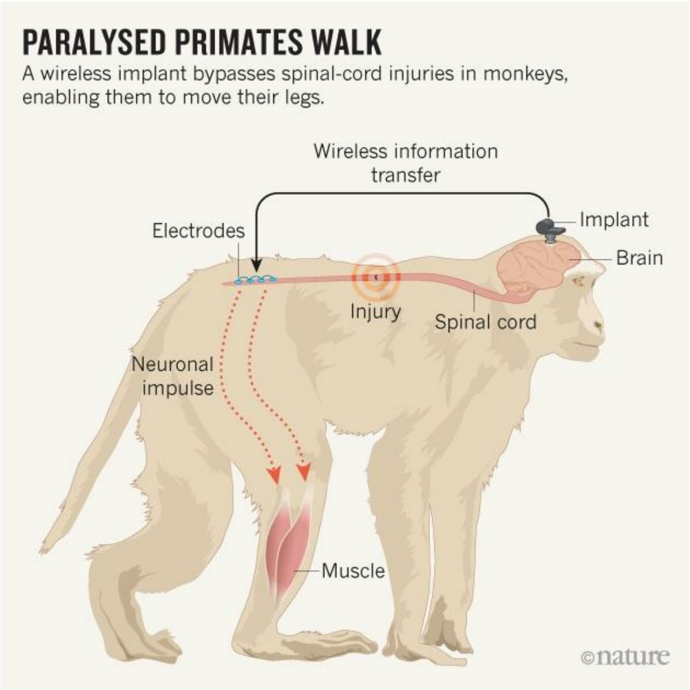
Gaming

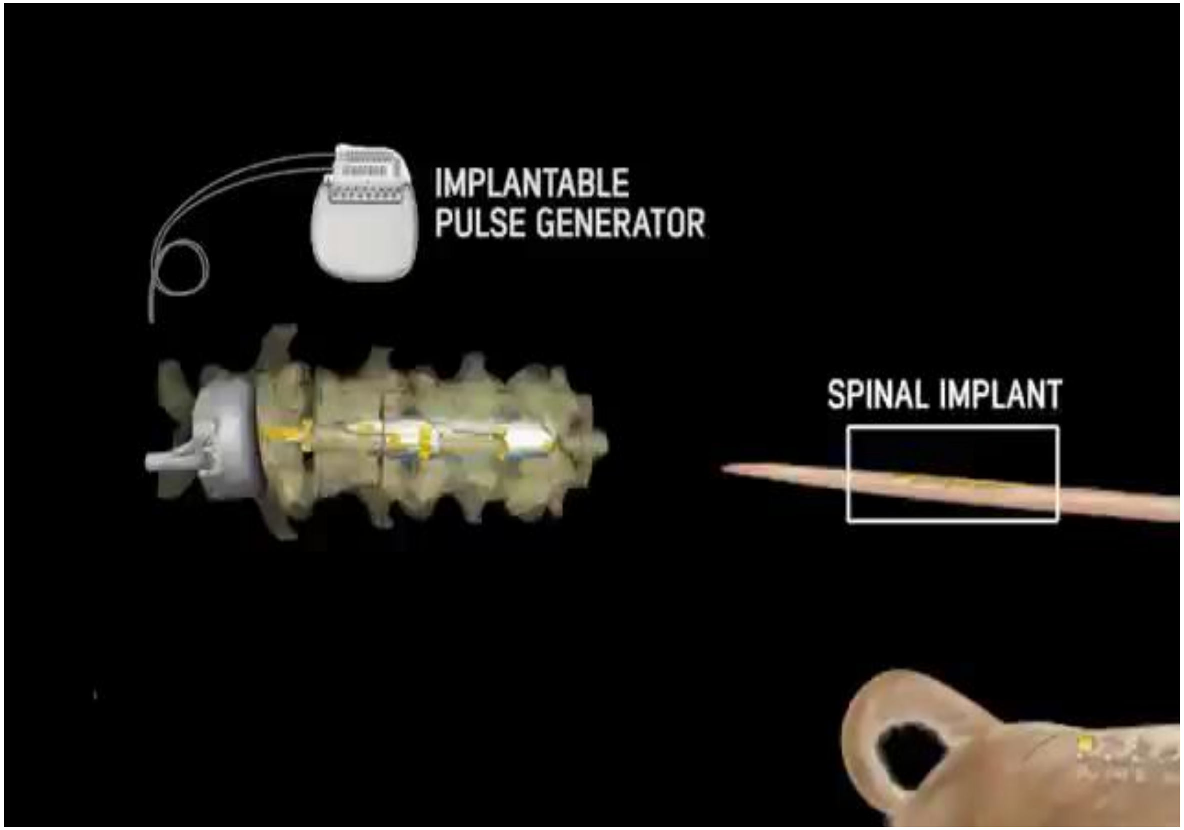
Bottom Image Source: newatlas.com

AI-Chips are … everywhere

Brain implant allows paralysed monkey to walk

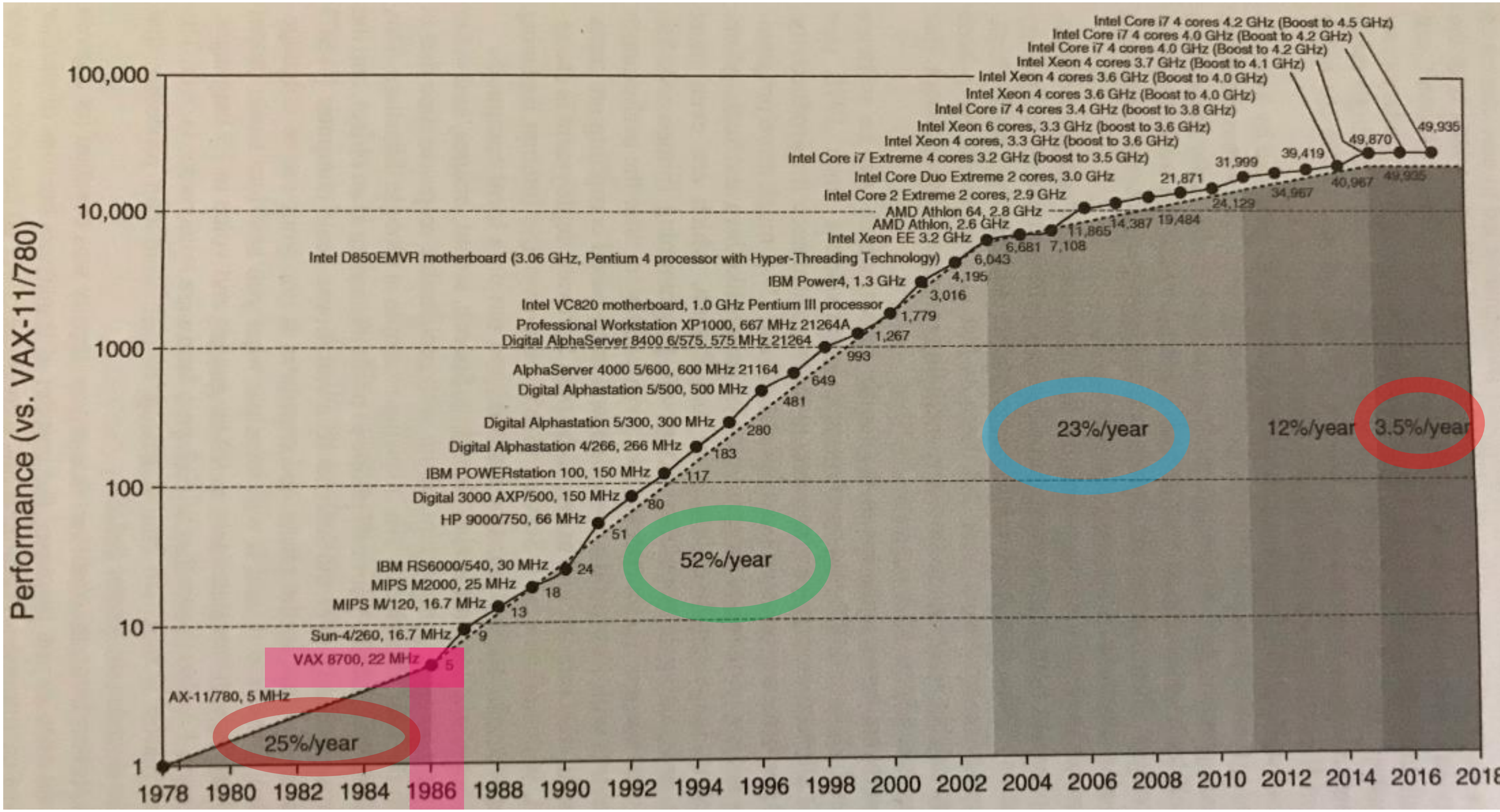
There really is a kind of intelligence inside the spinal cord. We are not just talking about reflexes that automatically activate muscles. In the spinal cord there are networks of neurons able to take their own decisions





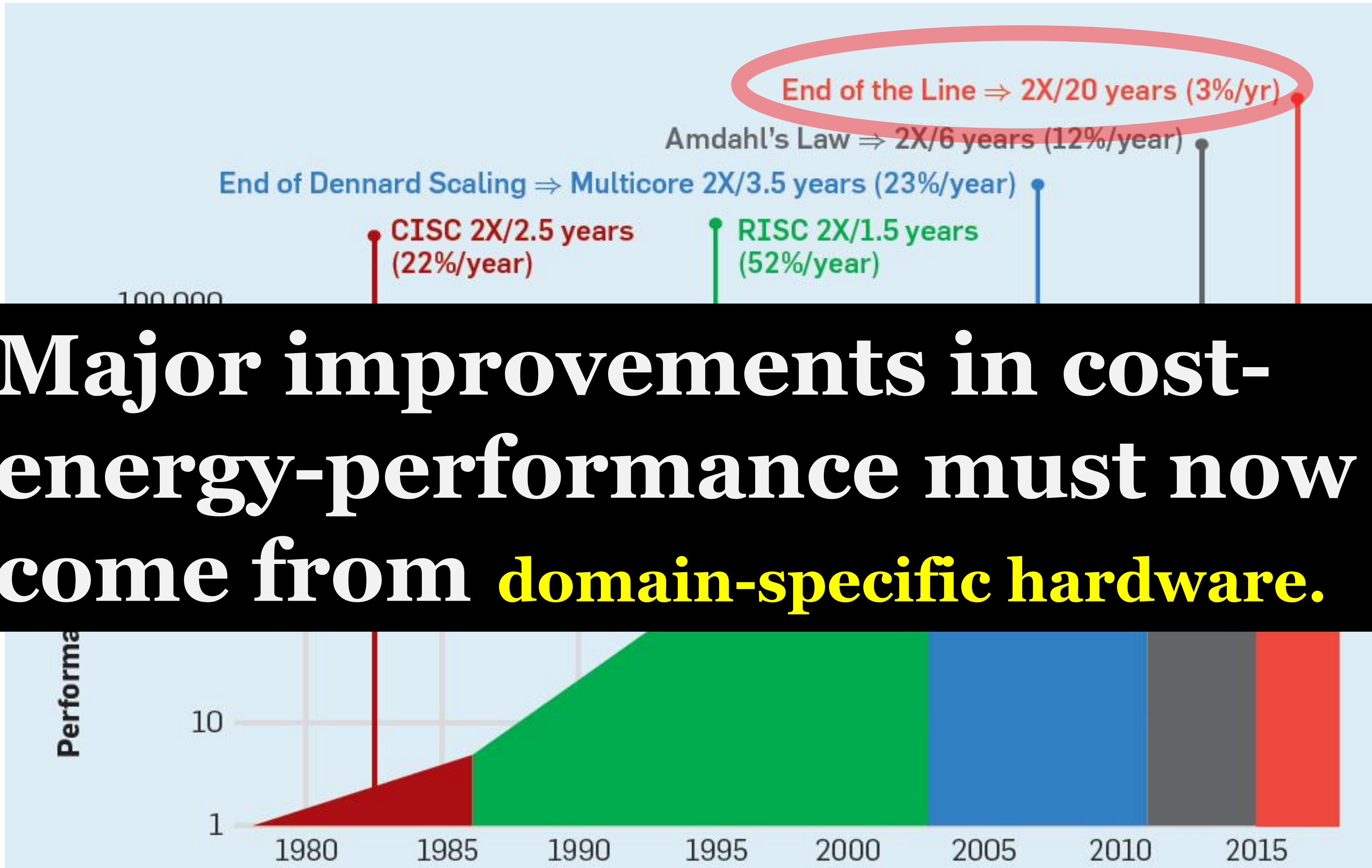
Nature volume539, pages284–288 (10 November 2016)

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



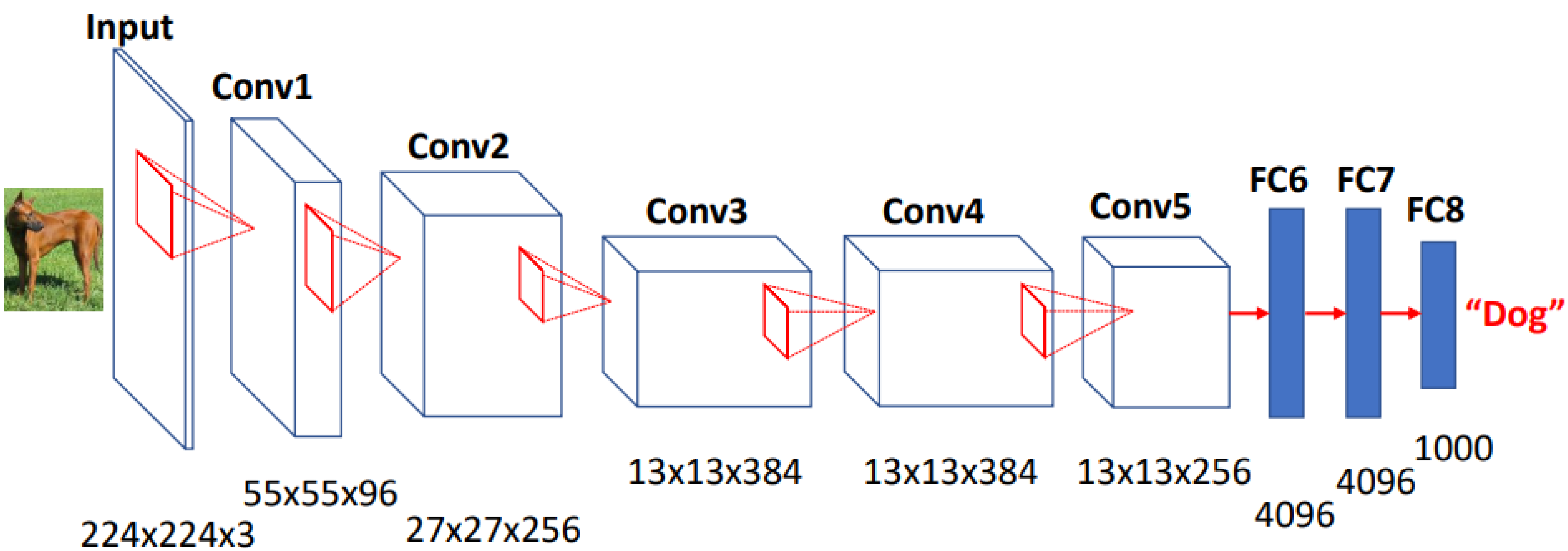
Major improvements in cost-energy-performance must now come from domain-specific hardware.

**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).

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	[illegible]	[illegible]	**138M**	**7M**	**25.5M**
Total MACs	**341k**	**724M**	**15.5G**	**1.43G**	**3.9G**

Source: Joel Emer, ISCA Tutorial, 2017

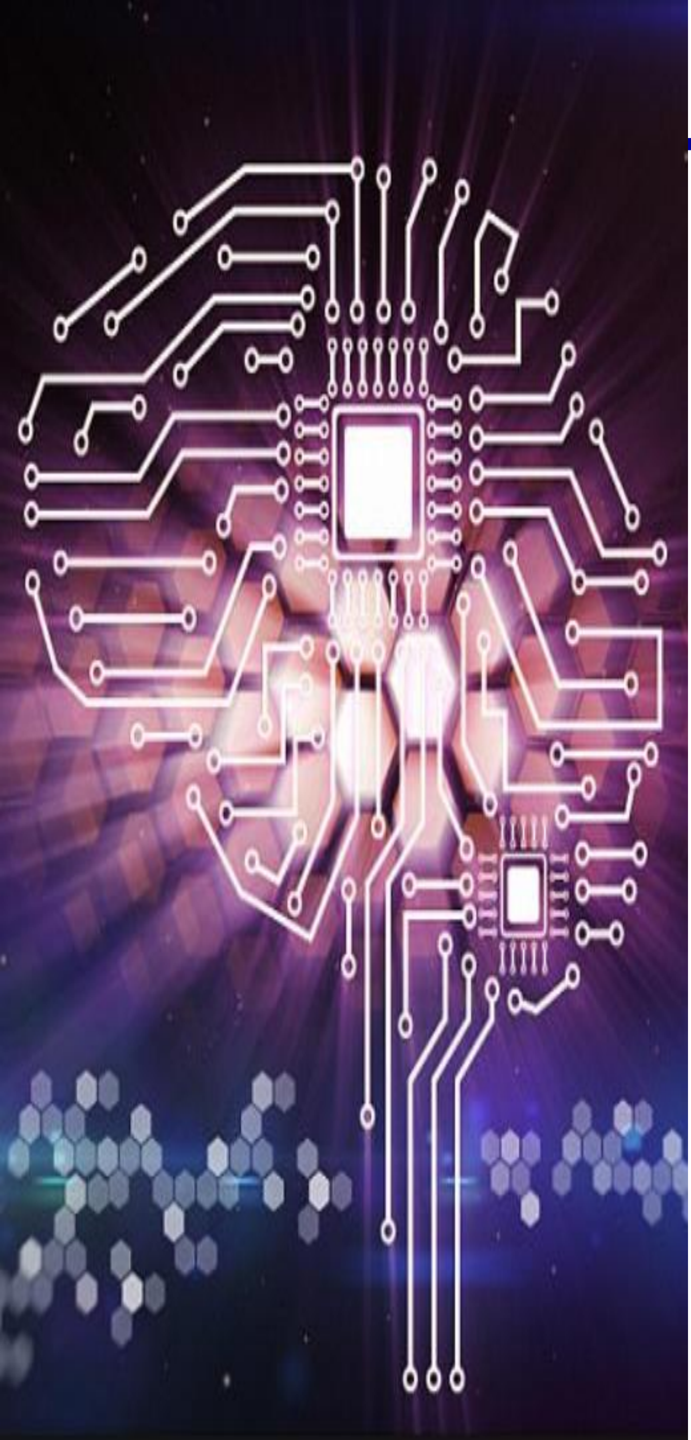
Deep learning requires massive compute power

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**



Agenda

- **Fundamental Trends**
- **AI – The Emerging Industrial Revolution**
- **AI at the Edge**
- **ASL Neuromorphic Chips**
- **Conclusions**

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

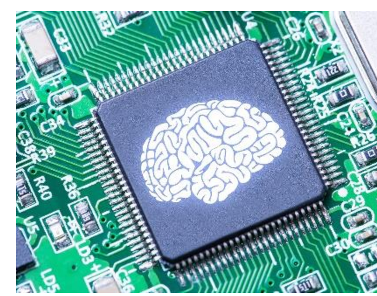


Image: spectrum.ieee.org

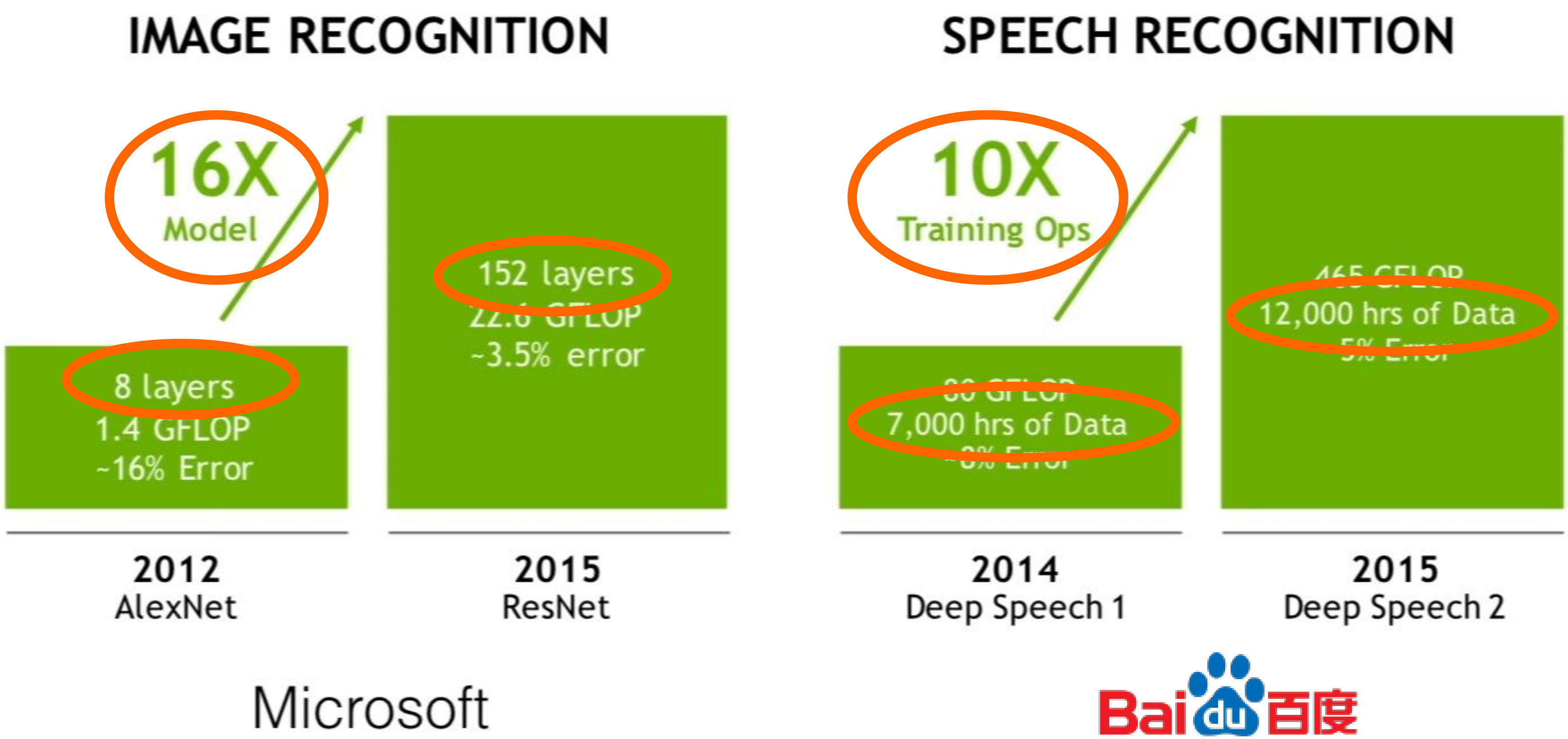
Image: sas.com

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

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

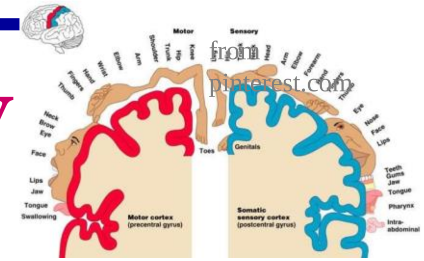
Hardware & Data Enable DNNs

AI model performance scales with dataset size and the # of model parameters, thus necessitating more compute.



Dally, NIPS'2016 workshop on Efficient Methods for Deep Neural Networks

AI HW is inspired by Nature – Biological neuron



from pinterest.com

AI-Chips are inspired by biology

→ parallel computation.

AI HW is inspired by Nature – Biological neuron

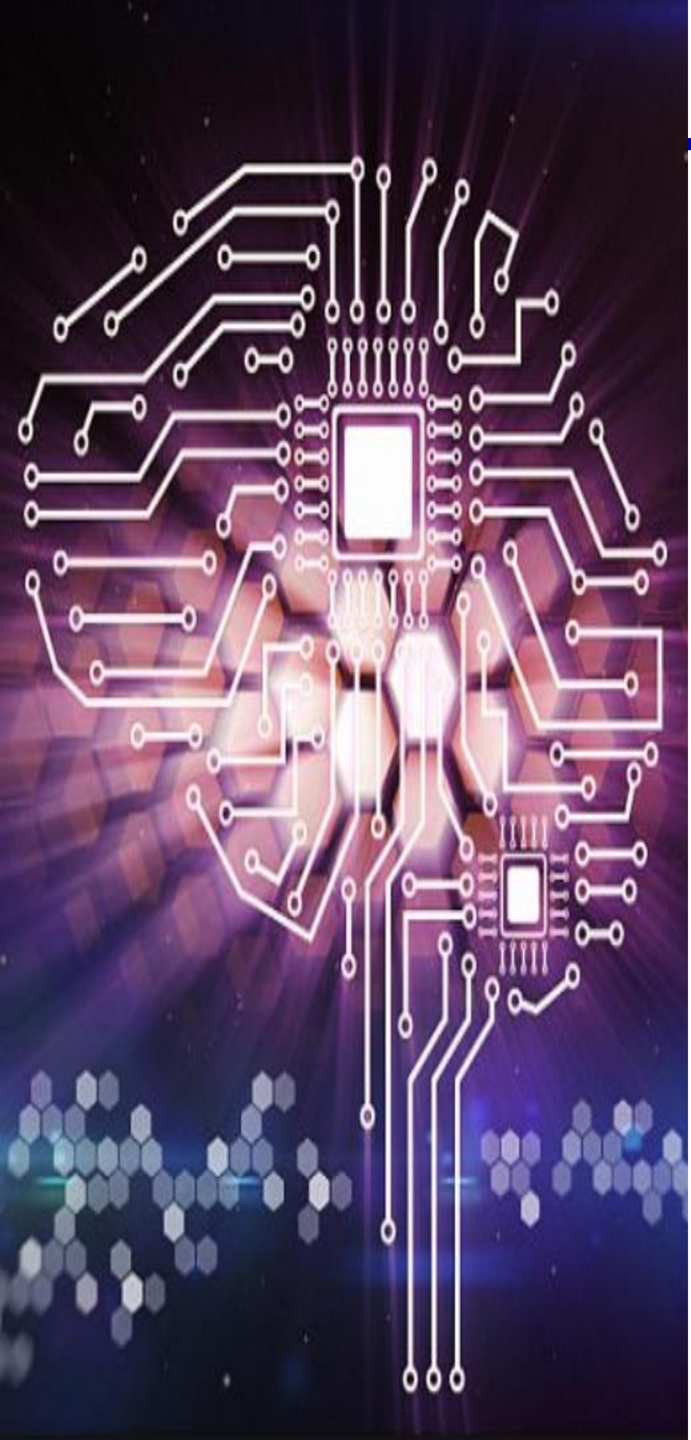
AI-Chips are inspired by biology
➔ parallel computation.

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: $\sim 10^{11}$
- # of synapses: $\sim 10^{15}$
- 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



Agenda

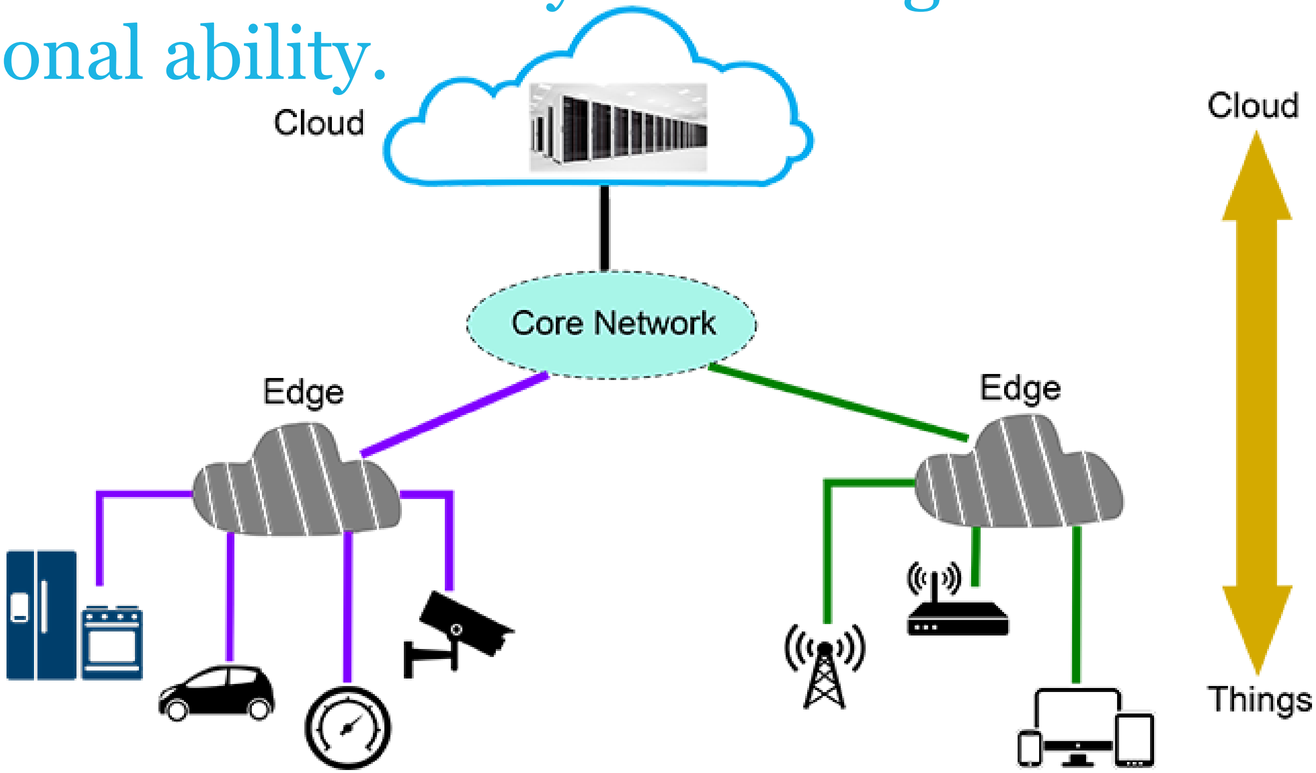
- **Fundamental Trends**
- **AI – The Emerging Industrial Revolution**
- **AI at the Edge**
- **ASL Neuromorphic Chips**
- **Conclusions**

17

AI at the Edge – High Security with Fast Computing

- The need for no latency, higher security, faster computing, and low cost would drive the adoption of devices that are able to offer AI at the **EDGE** ➔ Give devices the capability to run ML independent of the cloud by increasing their computational ability.

Delivers computing + intelligence where it is needed.



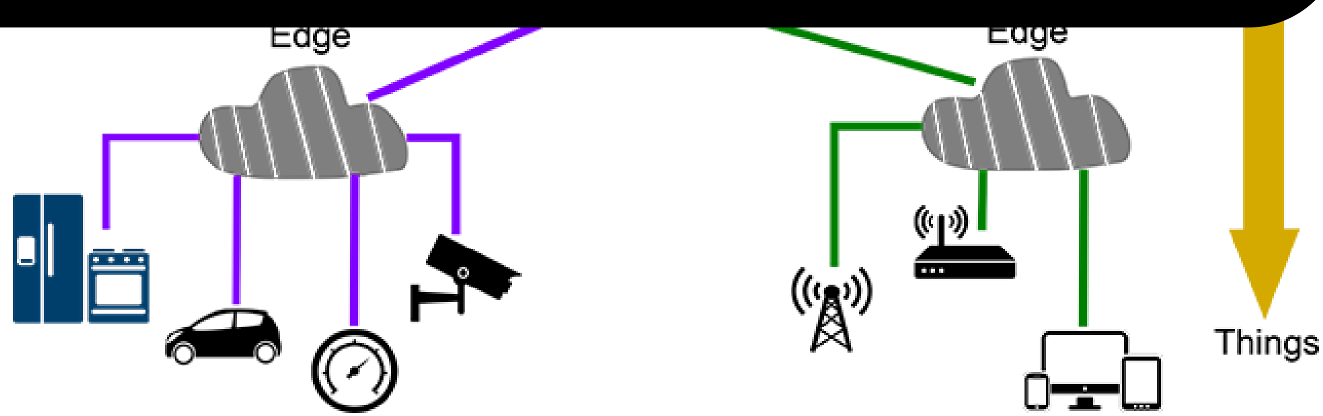
Edge devices will be equipped with special AI-chips based on FPGAs and/or ASICs

AI at the Edge – High Security with Fast Computing

- The need for no latency, higher security, faster computing, and low cost would drive the adoption of devices that are able to offer AI at the

On-device approach helps reduce latency for critical applications, lower dependence on the cloud, and better manage the massive data being generated by the IoT/Edge device.

+ **intelligence** **where it is needed.**



Edge devices will be equipped with special AI-chips based on FPGAs and/or ASICs

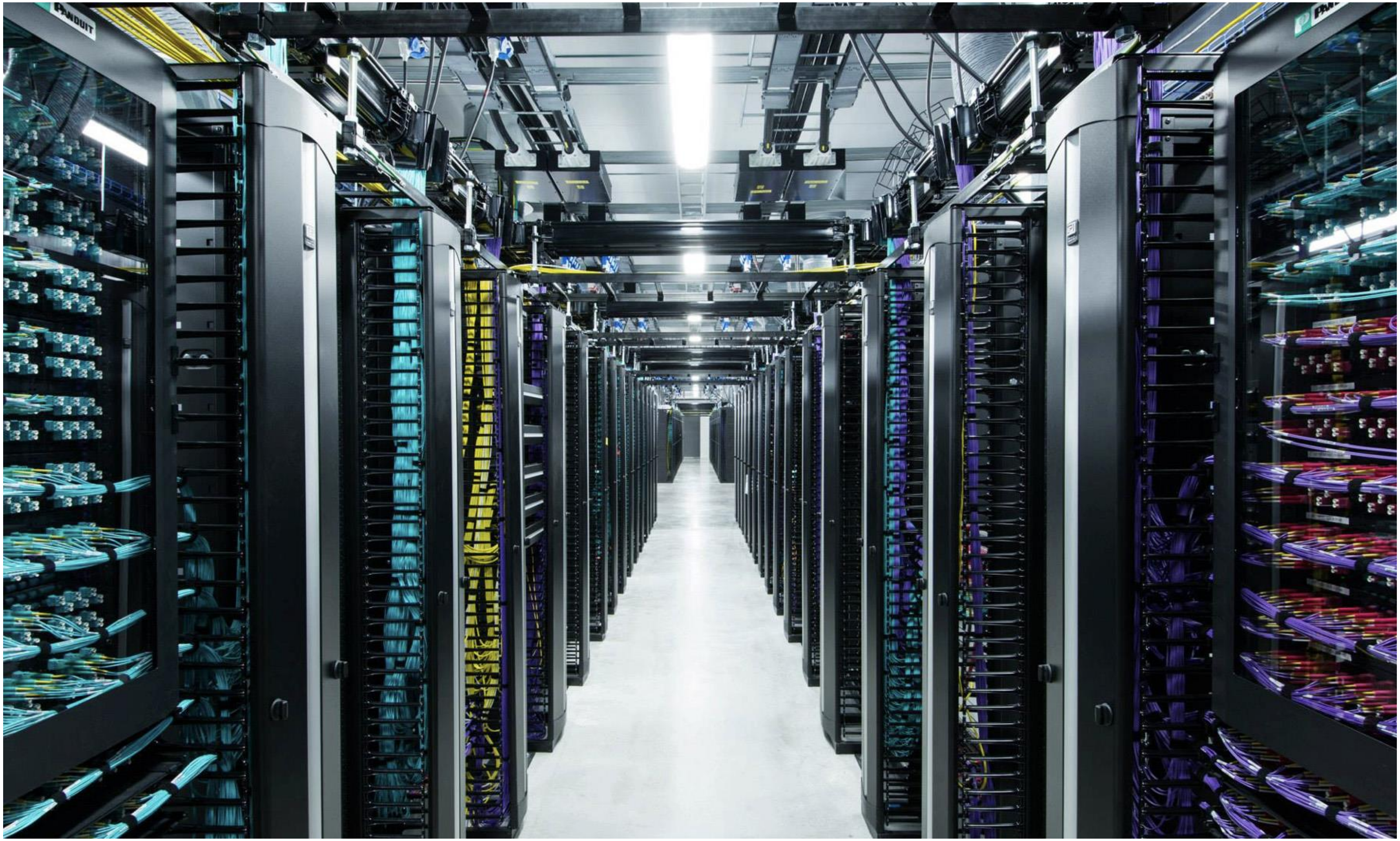
Deep learning requires massive compute power

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**
# 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**

To solve this level of computation, we need a GPU

Source: Joel Emer, ISCA Tutorial, 2017

Deep learning requires massive compute power



…most of ML models run in Data Center (Cloud)

…but there are cases where the ‘’cloud’’ cannot solve

Japanese AI security camera [theverge.com]

…but there are cases where the ‘’cloud’’ cannot solve

If the data is sent to the cloud, you cannot have RT decision.

Intel Falcon 8+ Drone transforms inspections conducted in the oil and gas industry [sustainableoilfield.com]

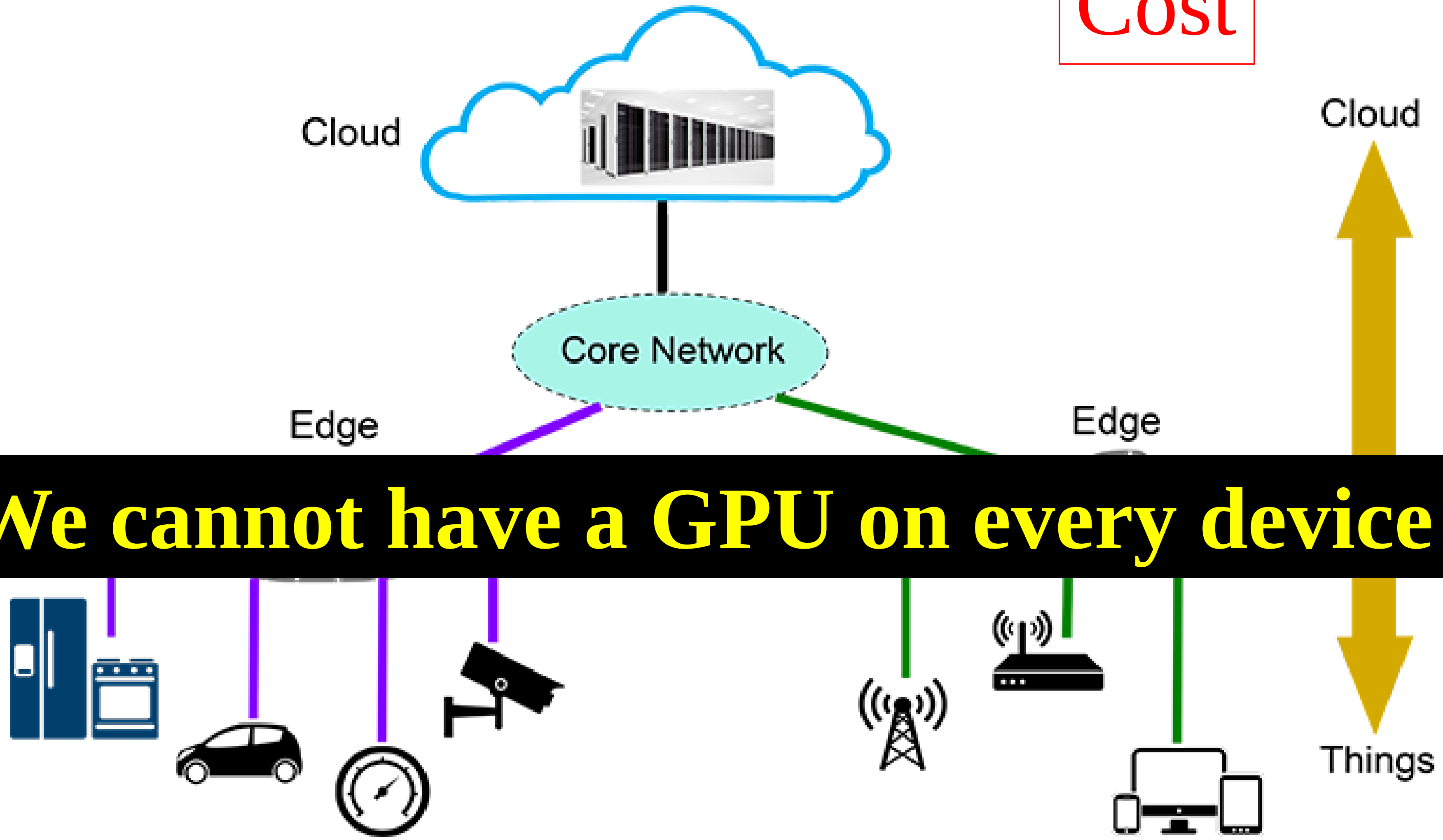
…but there are cases where the ‘’cloud’’ cannot solve

[Ref. life-of-coco.com]

Privacy

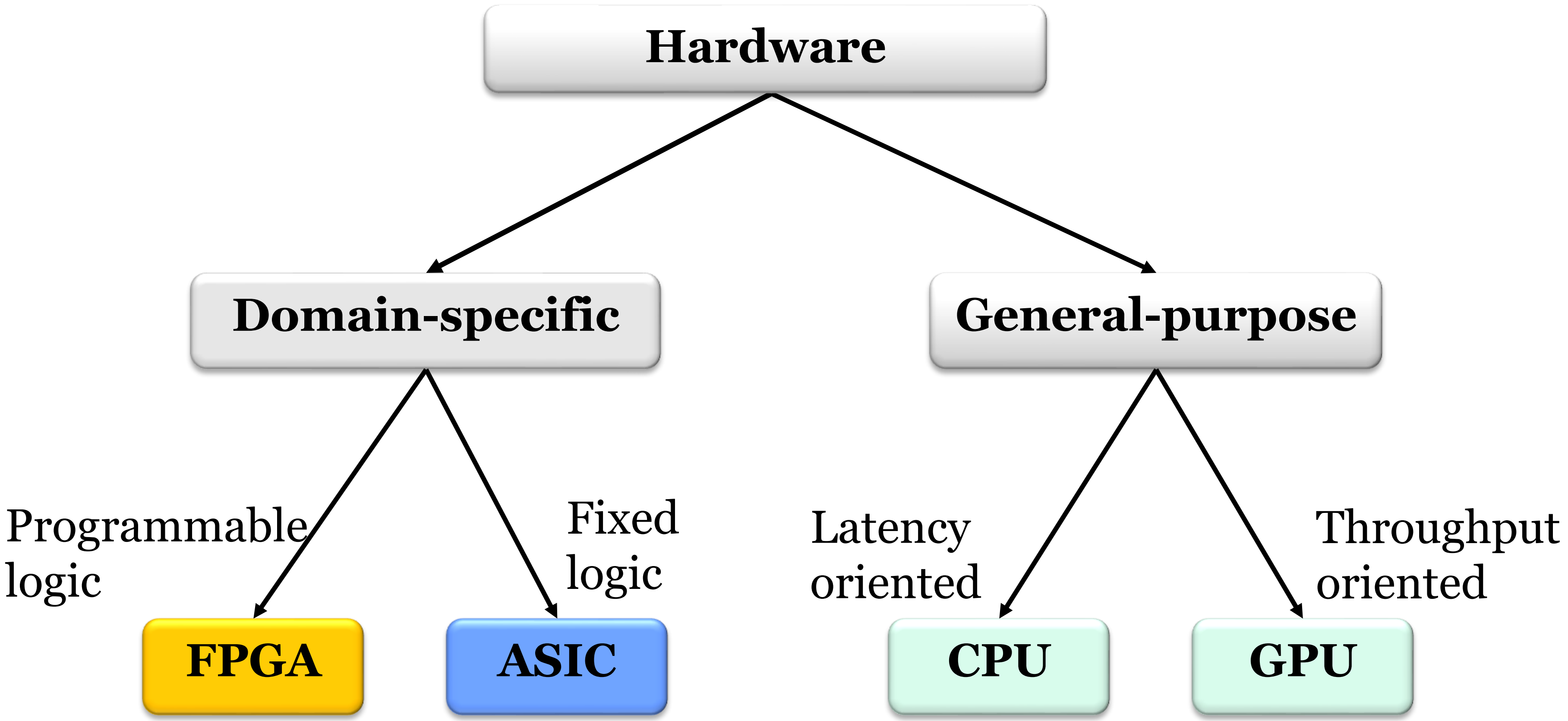
Real world product are often resource constrained

Cost



We cannot have a GPU on every device

Current State of the Art in Neural Algorithms HW Computing



FPGA

- General; requires HDL
- Moderate performance & efficiency

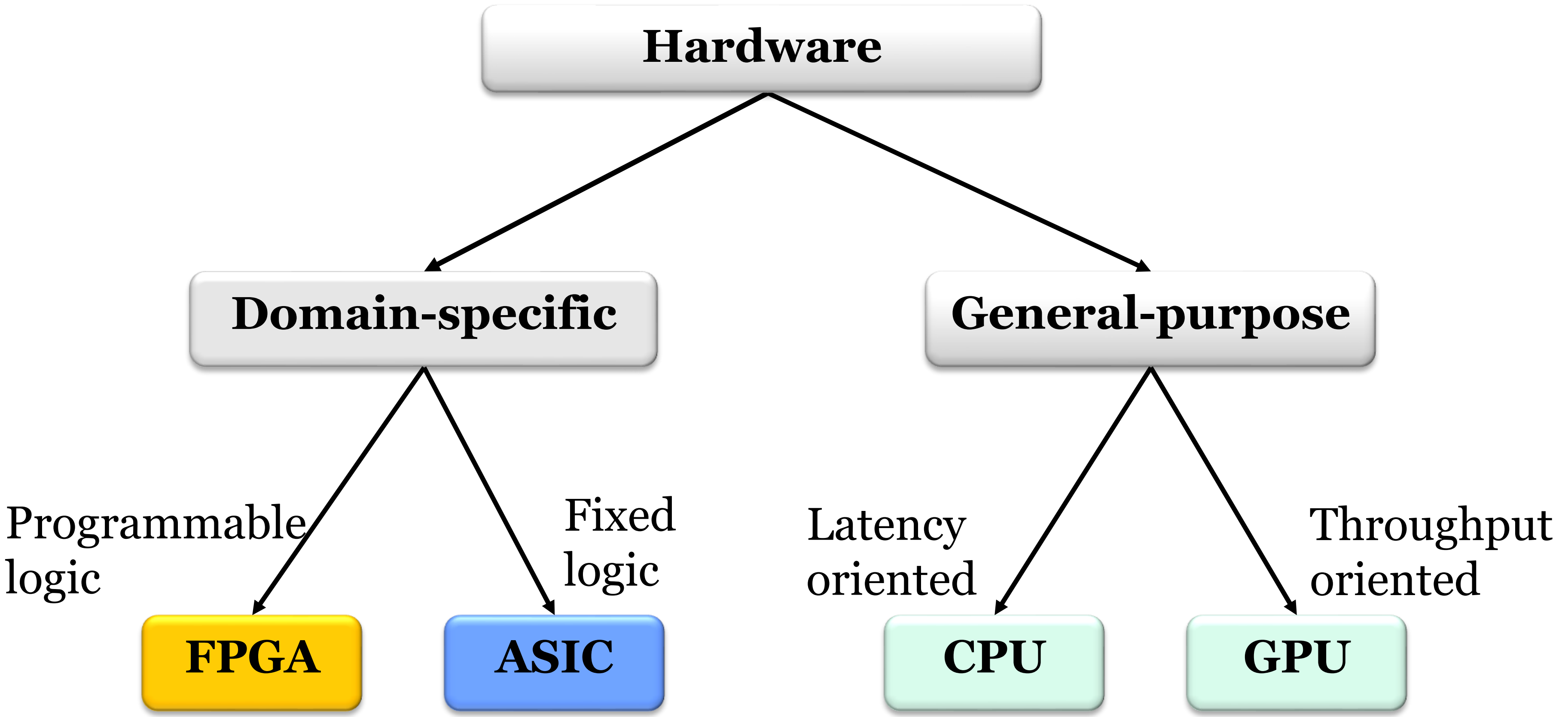
ASIC

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

CPU / GPU

- 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

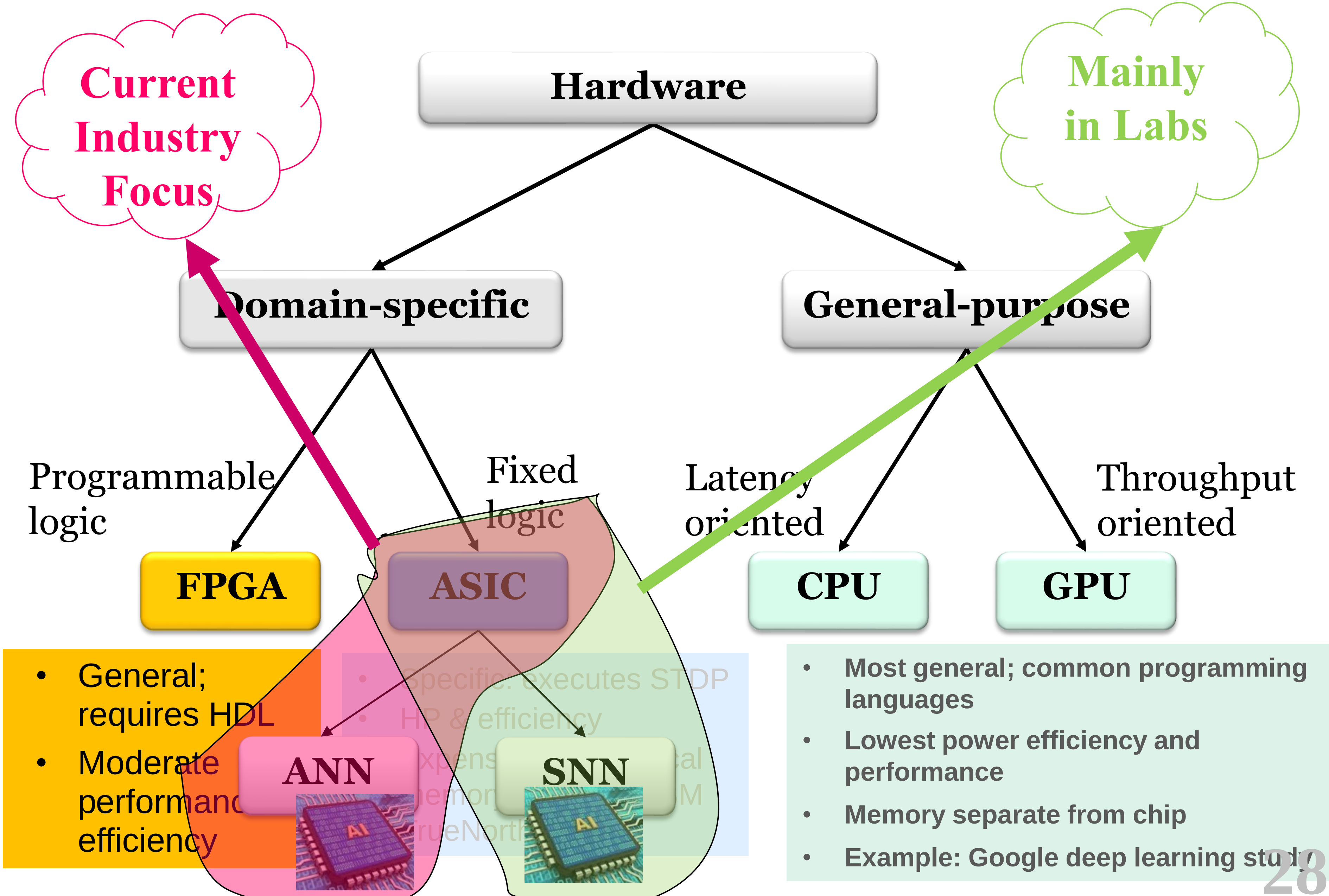


- General; requires HDL
- Moderate performance & efficiency

- 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



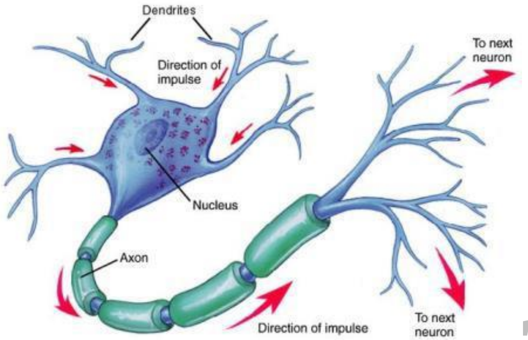
Different approaches of AI-Chips

Poor/Simple → Good/Complex

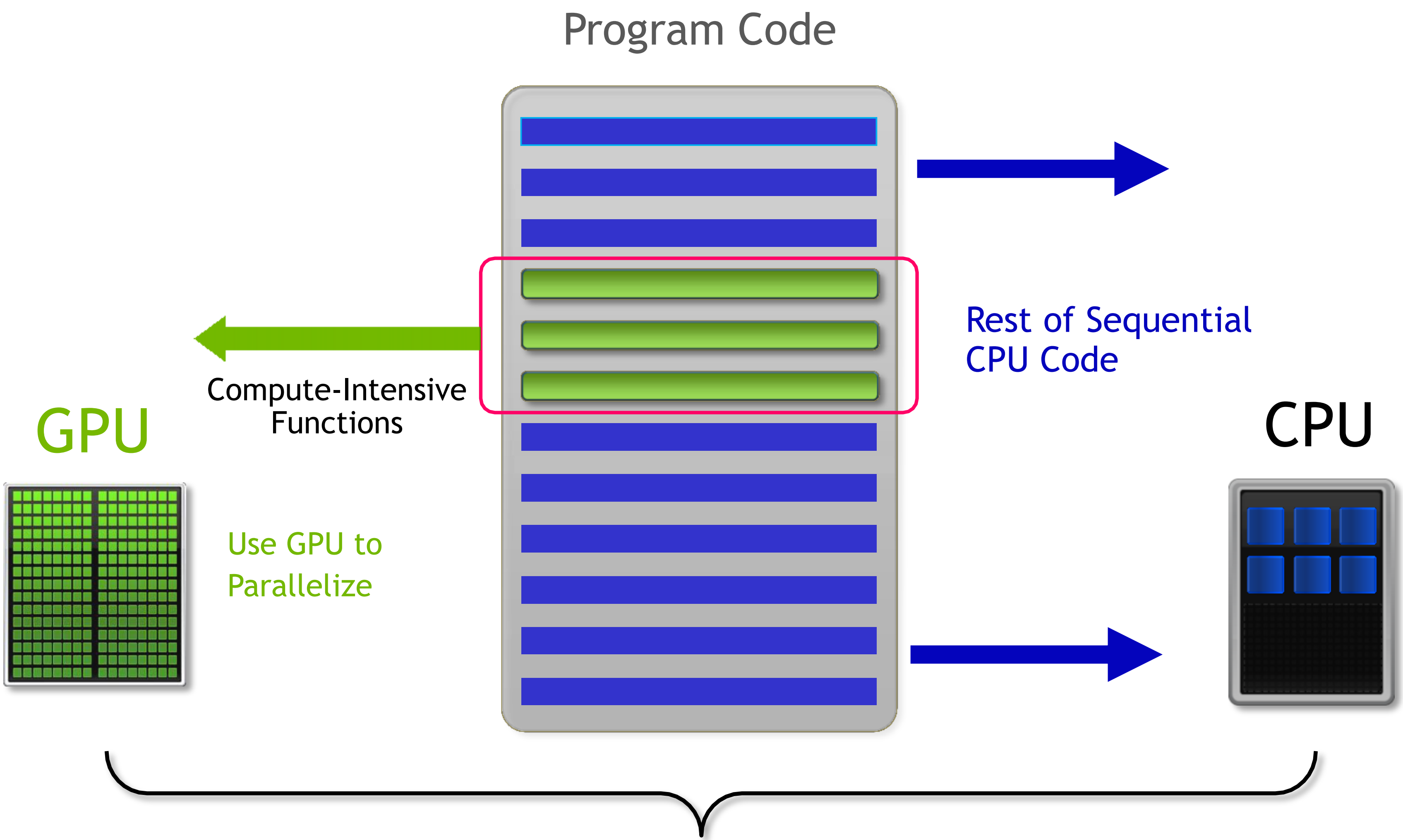
Neuron	Digital, Analog. LIF.	. . .	Izhikevich model	Huxley-Hodgkin model	. . .
Synapse	MAC (weighted sum)	. . .	Spiking STDP	. . . . Many nonlinear properties	. . . .

Generally Used in DL algorithms

Frequency	10~100 Hz (brain)

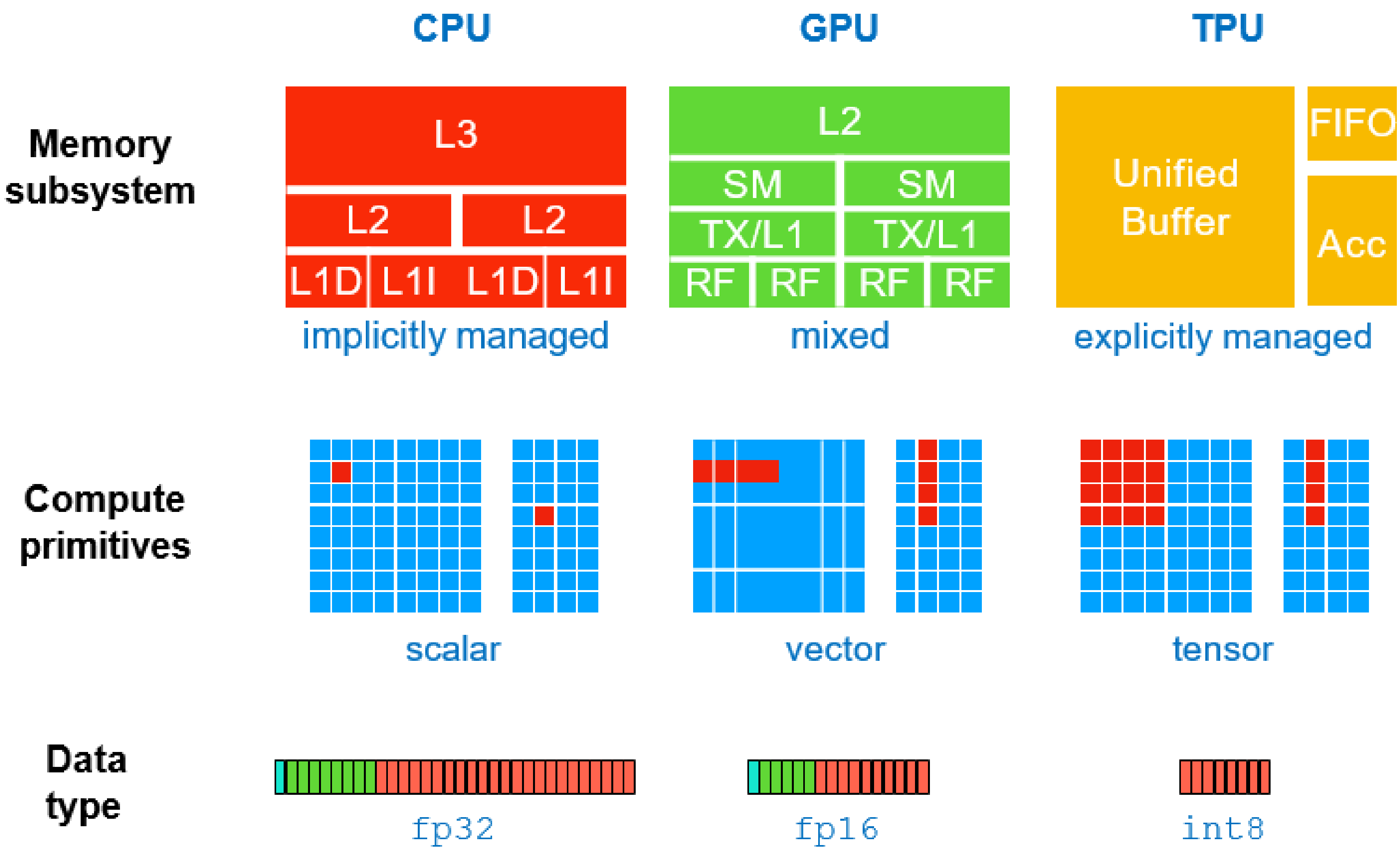


Current AI Chip = Accelerator/Co-processor



Acceleration with GPU

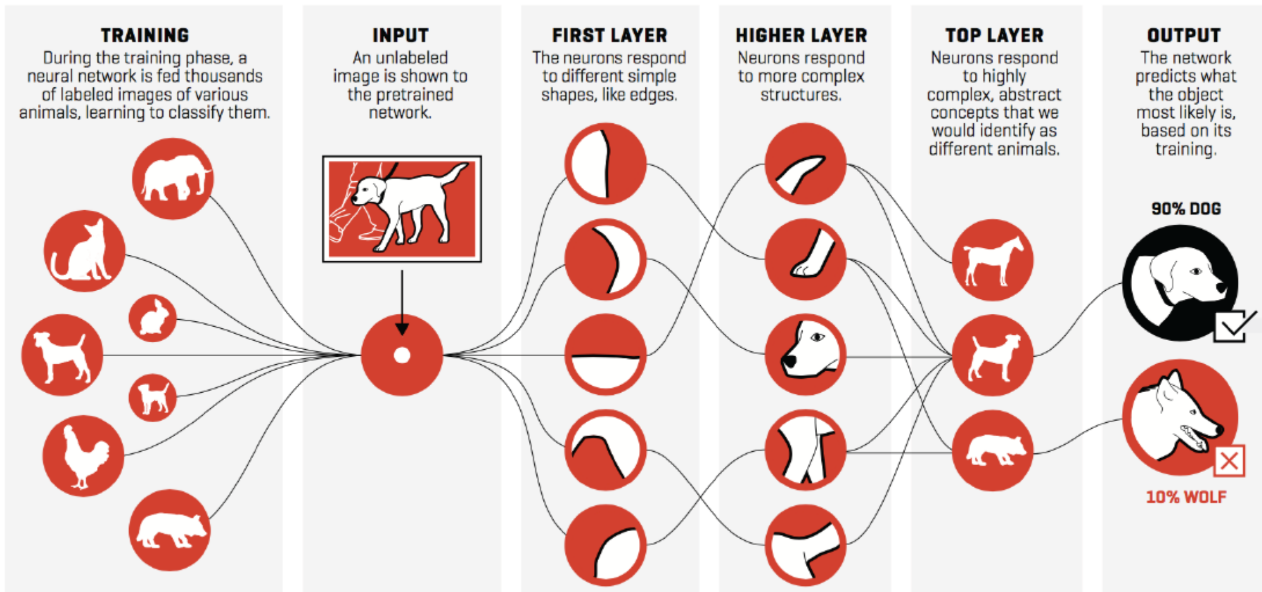
Accelerator Characteristics



Ref 3

…Deep Leering is considered as a sophisticated ''rocket'' of Machine Learning!!

Fuel = Data!

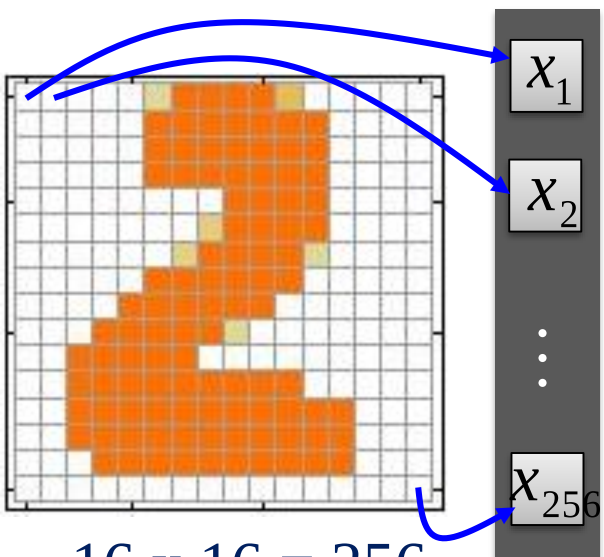


1. ''Deep Learning'' means using a neural network with <u>several layers of nodes</u> between input & output

2. the series of layers between input & output do feature identification and processing in a series of stages, just as our brains seem to.

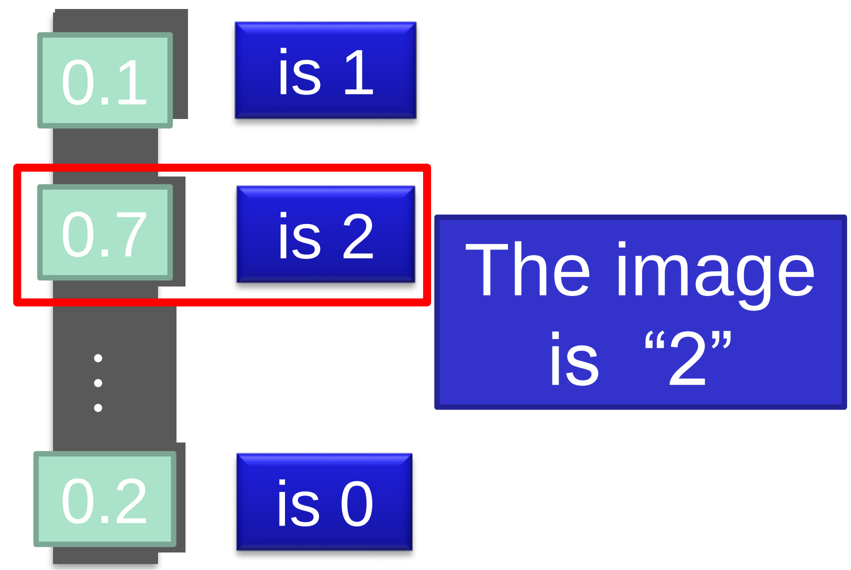
Example1: Handwriting Digit Recognition on FPGA

Input



16 x 16 = 256
Ink → 1
No ink → 0

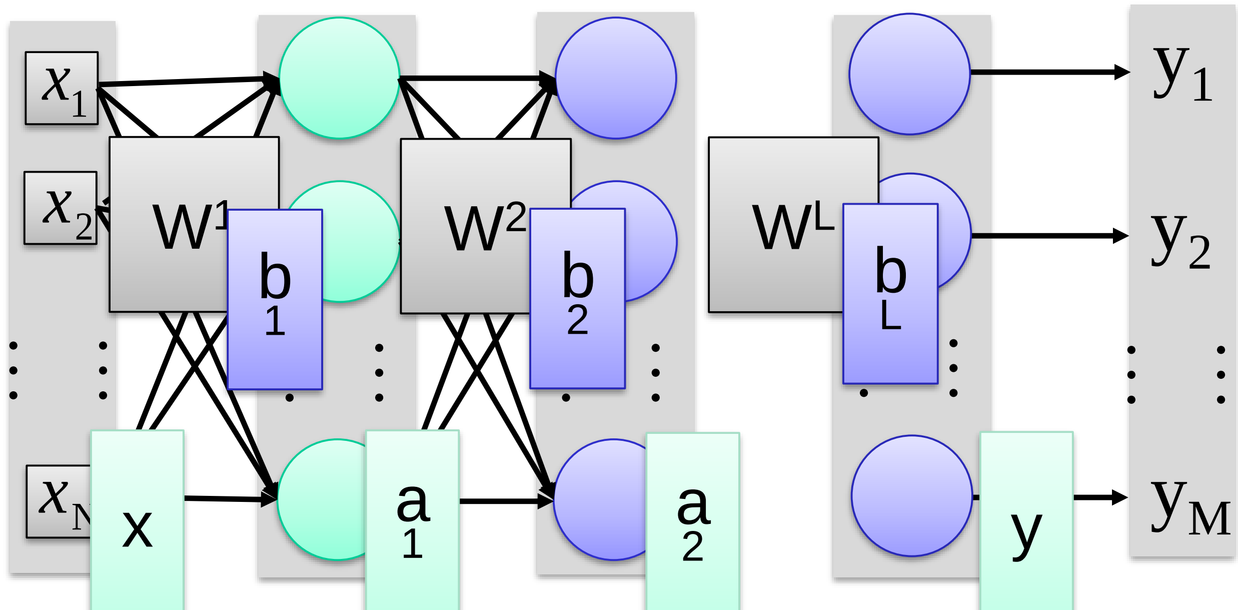
Output



Each dimension represents the confidence of a digit.

Example1: Handwriting Digit Recognition on FPGA

Conventional Artificial Neural Network

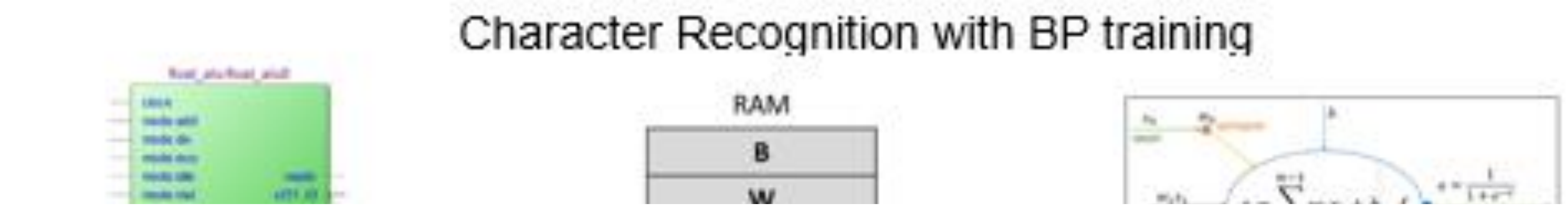


$$y = f(x)$$

Parallel computing techniques are needed to speed up **matrix operations**

$$= \sigma(W^L \quad W^2 \quad W^1 \; x \quad b_1 \quad b_2 \quad b_L)$$

Example1: Handwriting Digit Recognition on FPGA



Implementation of detecting 16 patterns from 16 inputs with BP.

Device: EP2C35F672C6
Family: Cyclone2
Synthesis: Quartus2 13.1

Table 1 : ANN Performance Evaluation

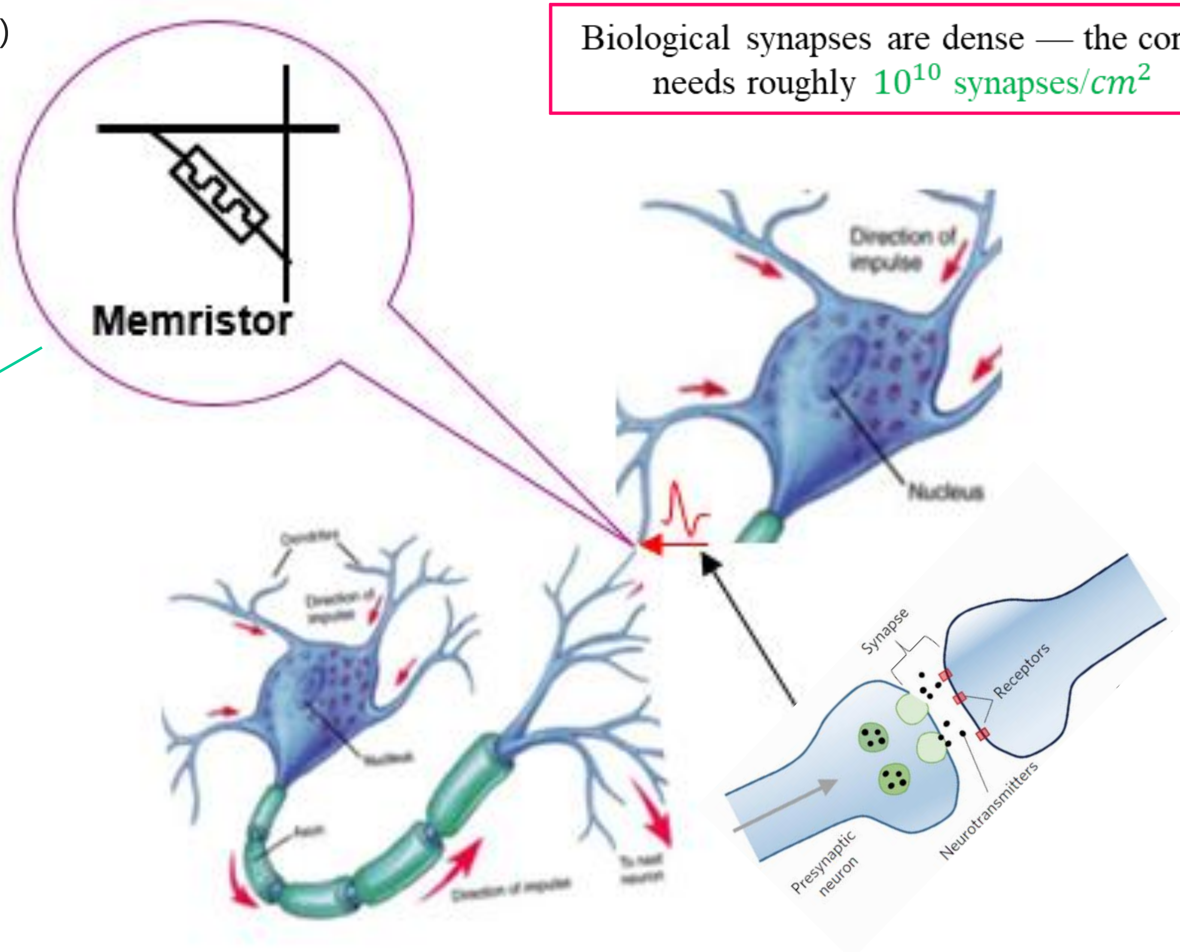
ALUs	Registers	Pins	Fmax
10,989 (33%)	5,814 (18%)	432 (89%)	76.02 MHz
Memory	**DSP Block**	**Power Consumption**	
4,956 (1%)	54 (77%)	286.84 mW	

Memristor for Synapse Design

(Chua, 1971)

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

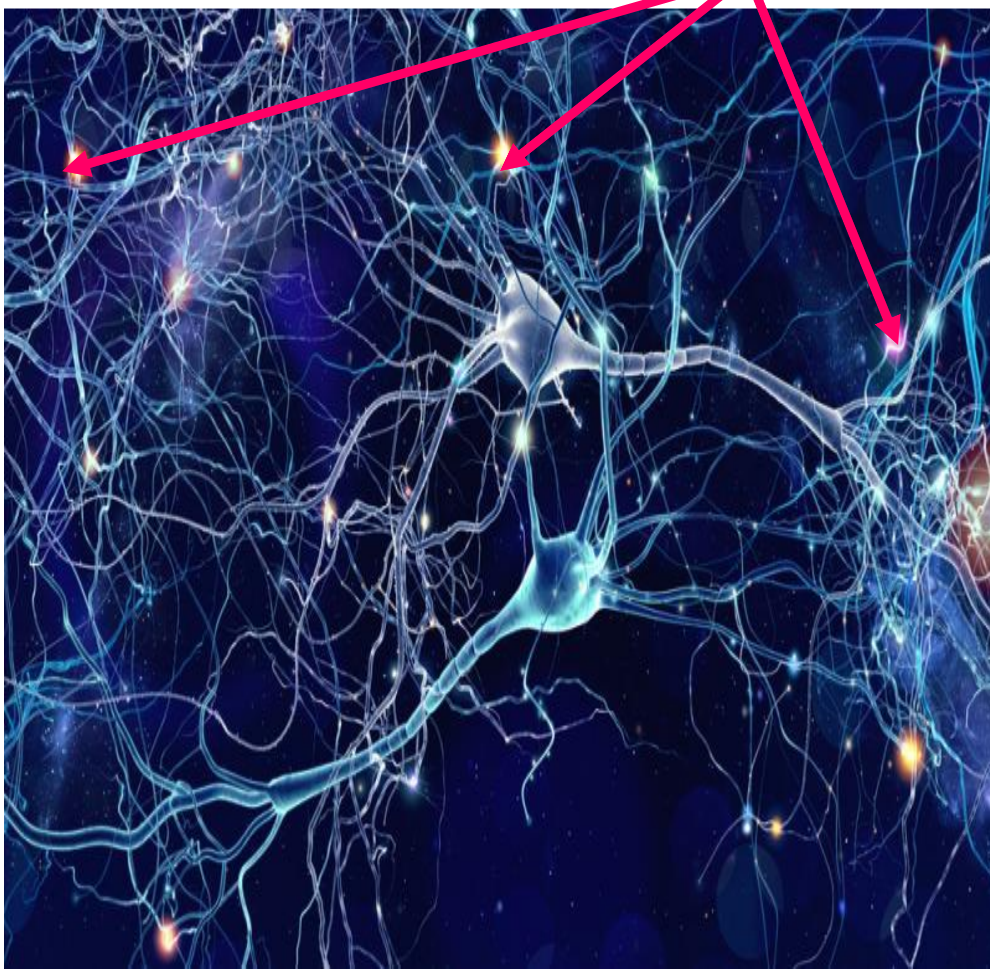
The electrical resistor is not constant but depends on the history of current that had previously flowed through the device.

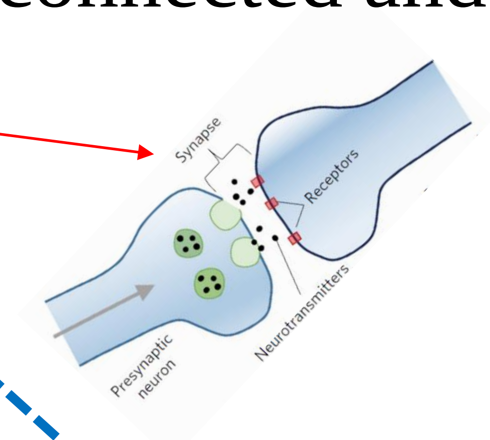


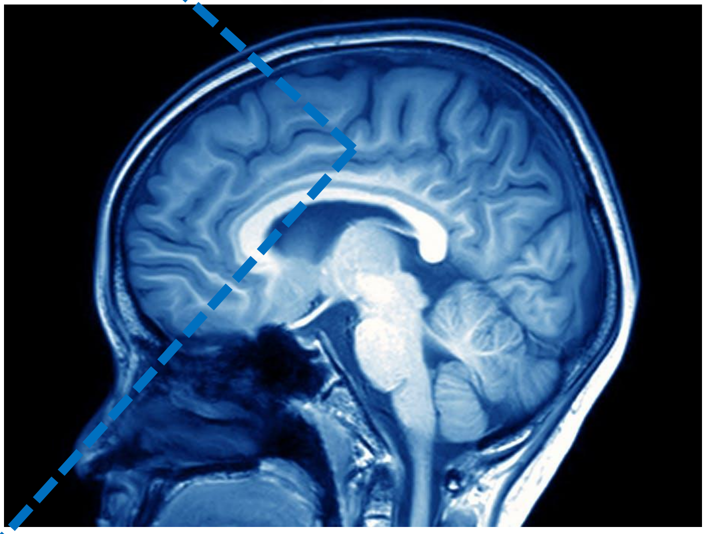
❖ Voltage **pulses** can be applied to a **memristor** to change its **resistance**, just as **spikes** can be applied to a **synapse** to change its **weight**.

How biological neurons learn?

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

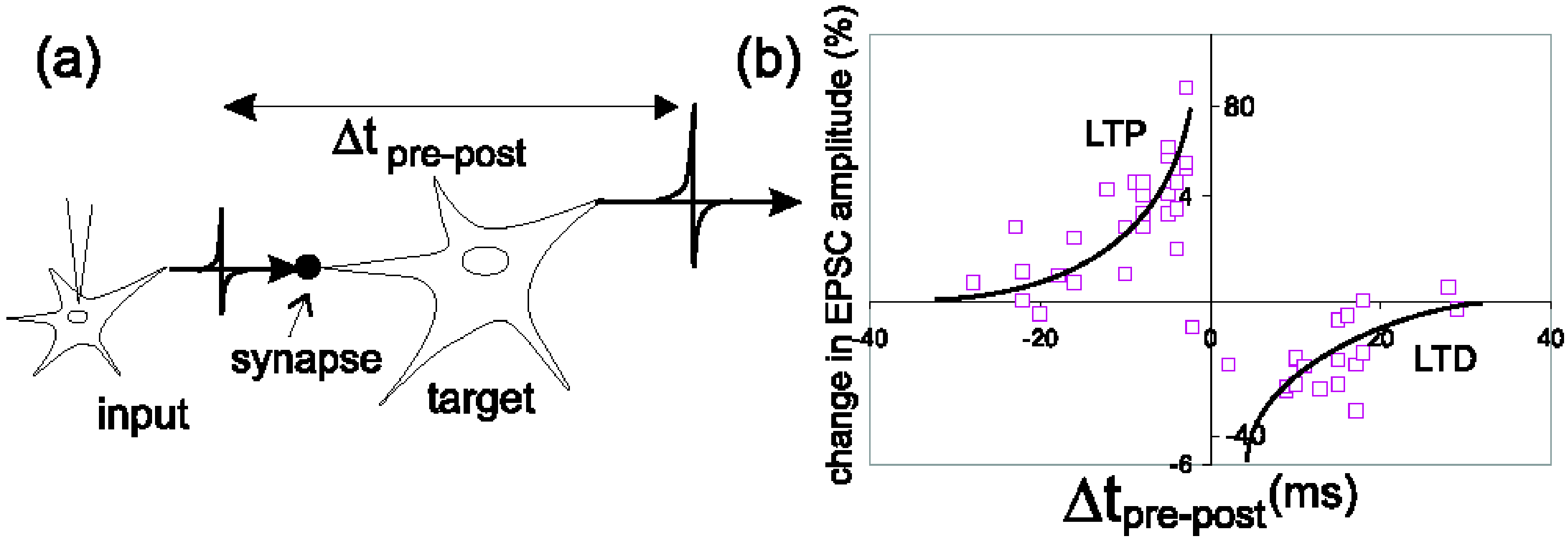






How biological neurons learn?

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



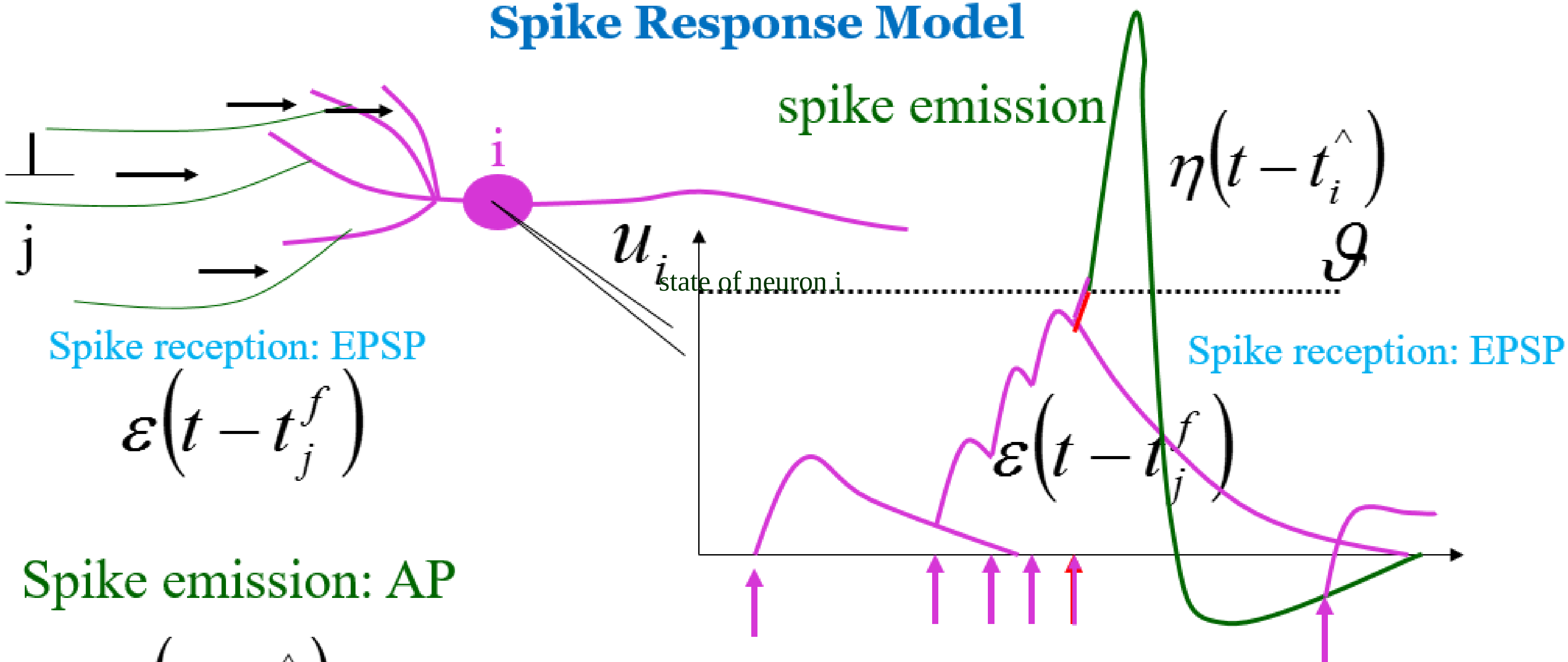
Spike-timing-dependent plasticity (STDP)

- 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 Response Model



Spike emission: AP

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

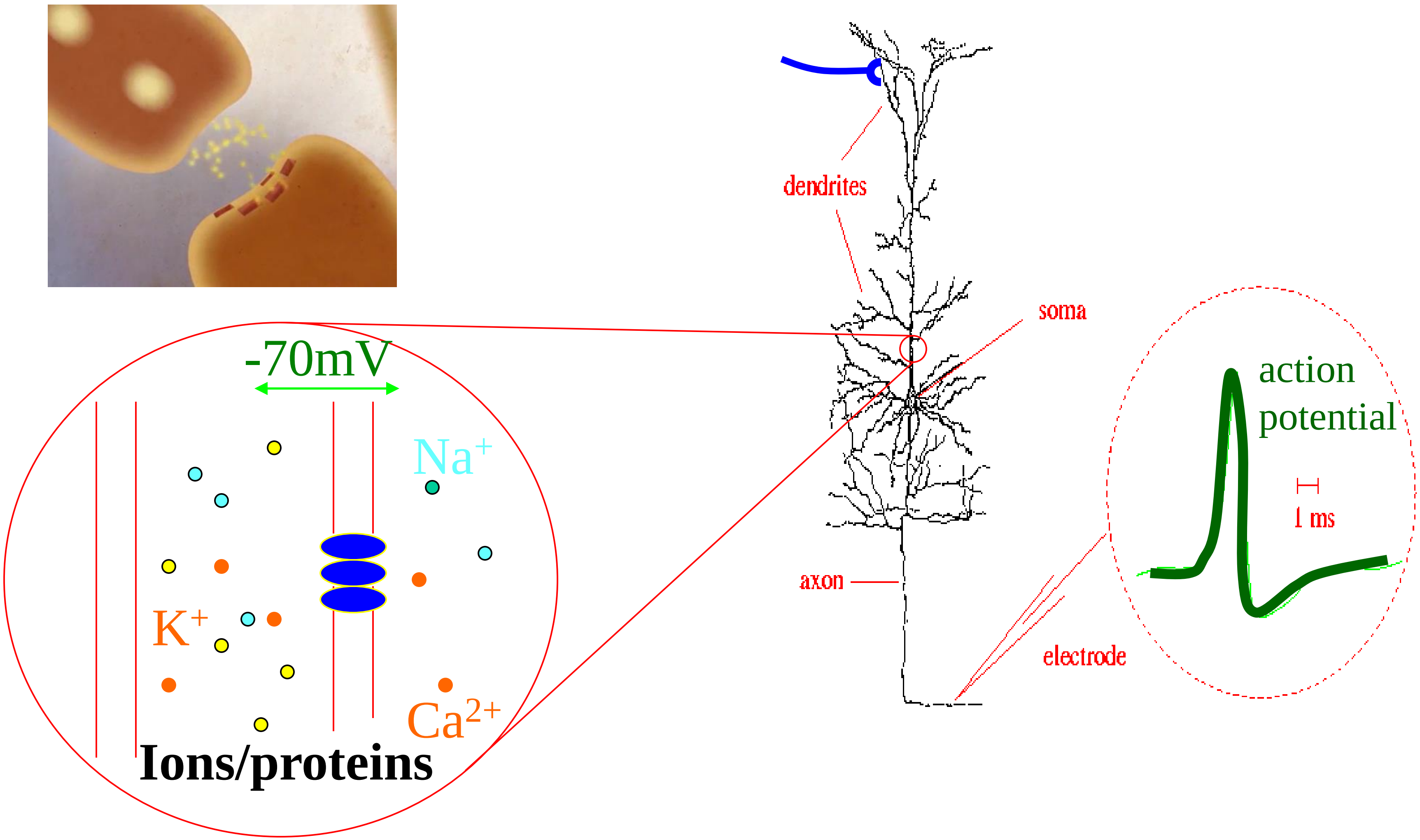
reset of the membrane potential (action potential)

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

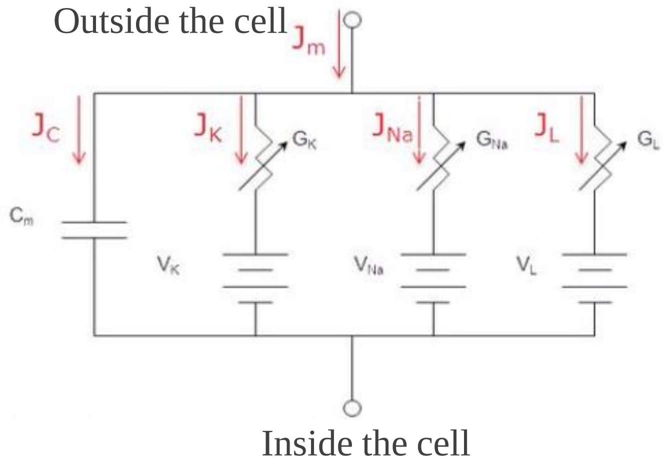
$$u_i(t) = \vartheta \Rightarrow \text{Firing: } \quad \hat{t}_i = t$$

[Ref. 18]

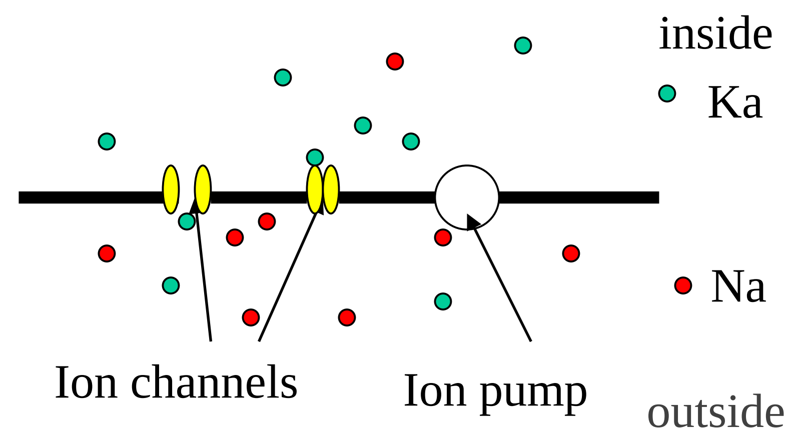
Spiking Neuron Model- Molecular Basis



Hodgkin-Huxley Model



~



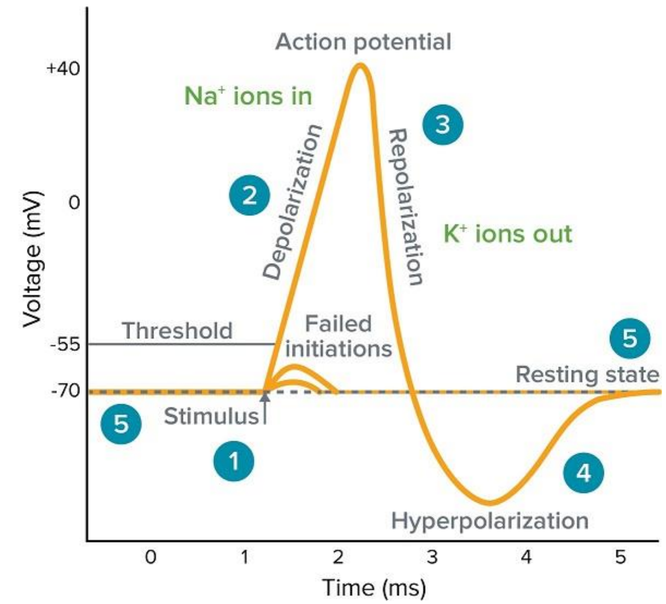
$$J_c = C_m \, \partial V_m / \partial t$$

$$J_{Na^+} = G_{Na^+}\left(V_m - V_{Na^+}\right)$$

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

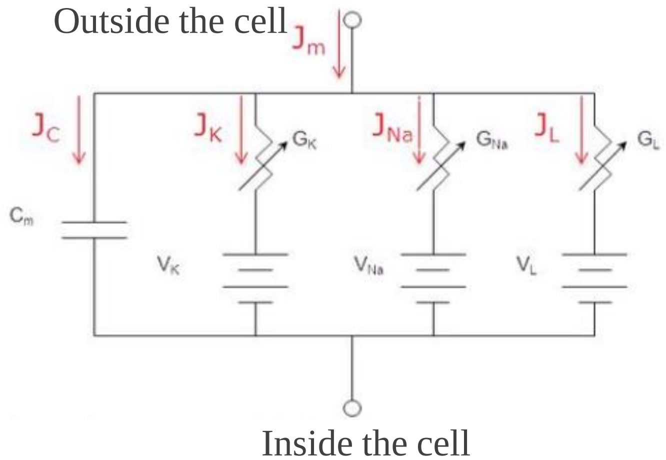
$$J_L = G_L\left(V_m - V_L\right)$$

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

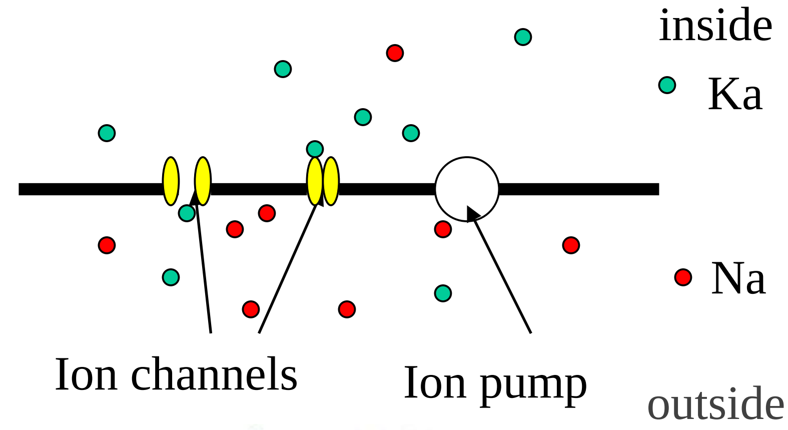


$$J_m = C_m \, \partial V_m / \partial t + G_{K^+}\left(V_m - V_{K^+}\right) + G_{Na^+}\left(V_m - V_{Na^+}\right) + G_L\left(V_m - V_L\right)$$

Hodgkin-Huxley Model



~

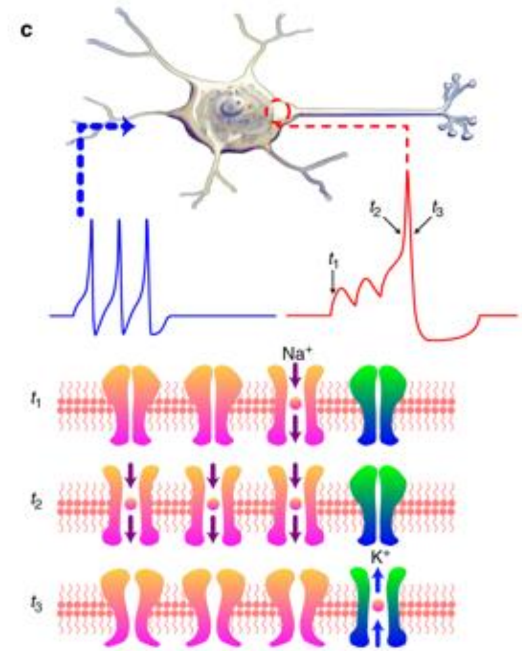


$$J_c = C_m \, \partial V_m / \partial t$$

$$J_{Na^+} = G_{Na^+}\left(V_m - V_{Na^+}\right)$$

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

$$J_L = G_L\left(V_m - V_L\right)$$

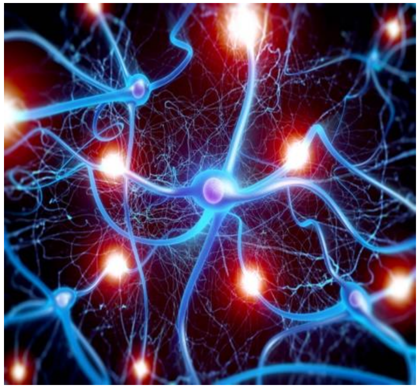


$$J_m = J_c + J_{K^+} + J_{Na^+} + J_L$$

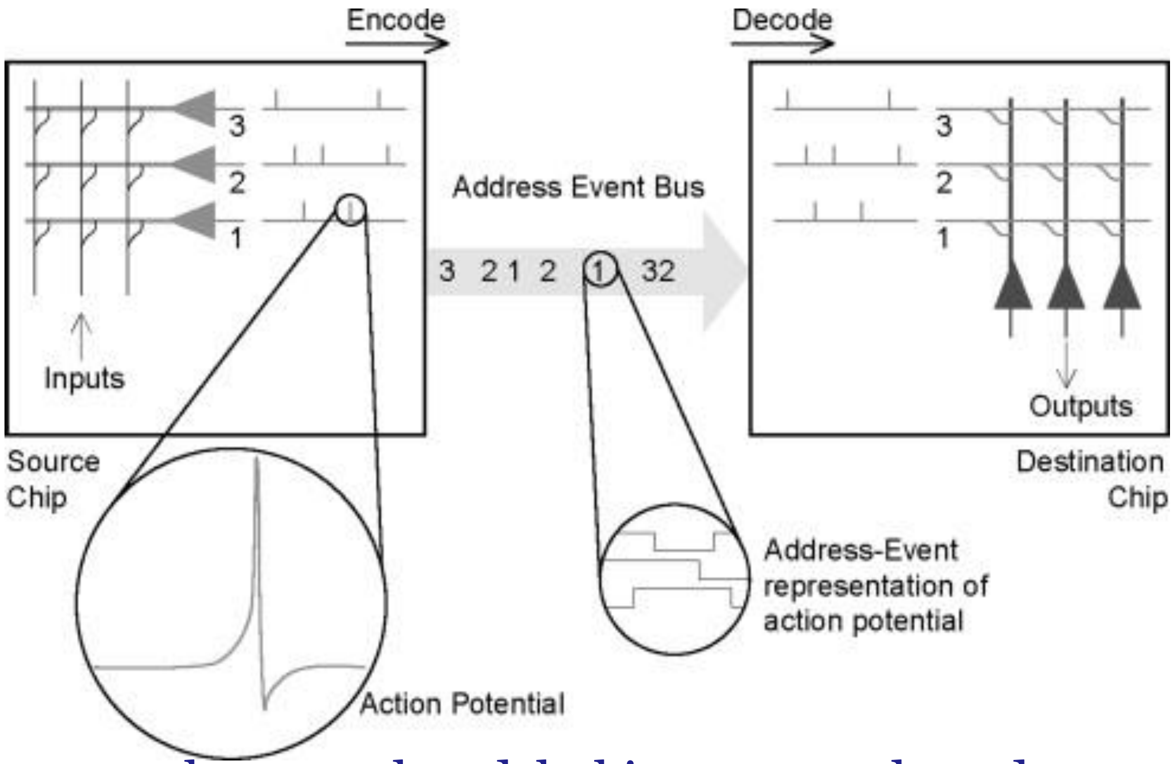
$$J_m = C_m \, \partial V_m / \partial t + G_{K^+}\left(V_m - V_{K^+}\right) + G_{Na^+}\left(V_m - V_{Na^+}\right) + G_L\left(V_m - V_L\right)$$

[Ref. 18]

Wiring via AER (Address Event Representation)



(Courtesy: iStock/Henrik5000)



Ref. 4

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

NN Training Works with Low-precision FP

fp32: Single-precision IEEE Floating Point Format

Range: (10^-45) to (10^38)

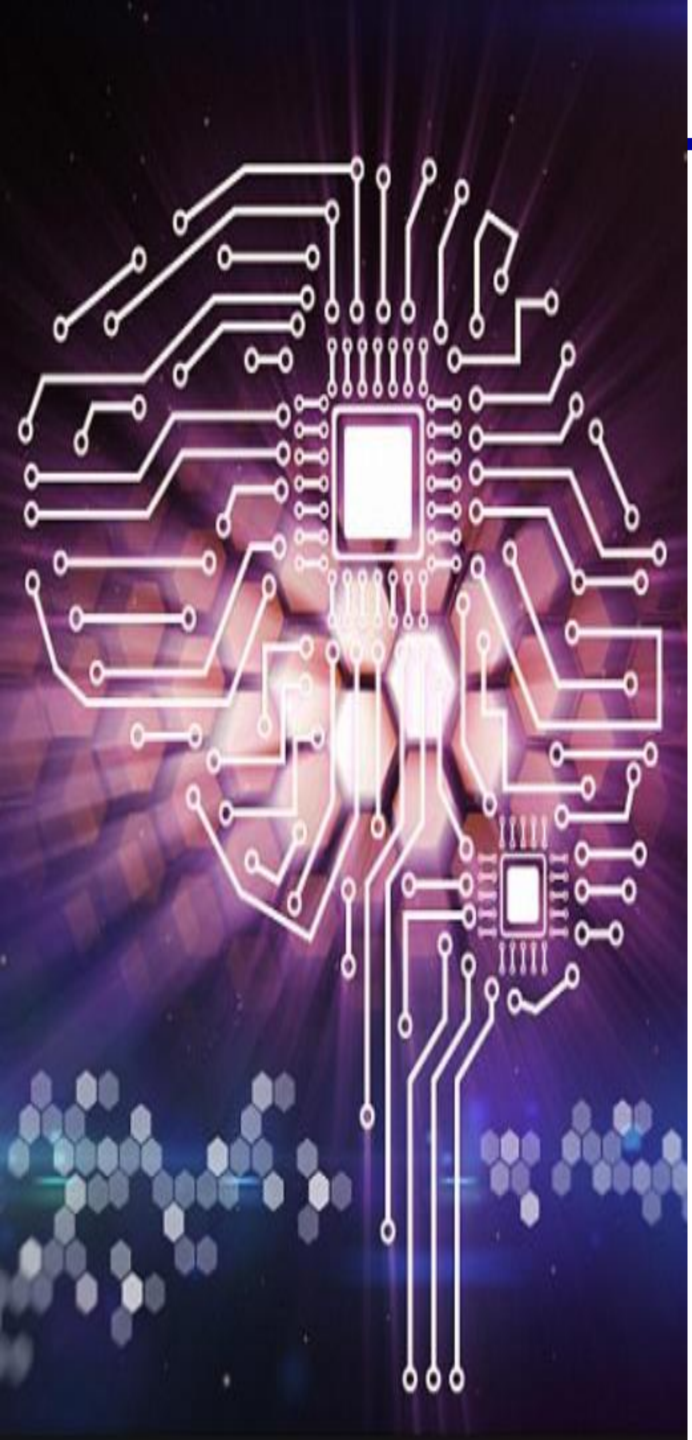
fp16: Half-precision IEEE Floating Point Format

Range: 10^-8 to 65504

bfloat16: Brain Floating Point Format

Range: (10^-45) to (10^38)

- Represent the same range of numbers of fp32 just at a much lower position.
- It turns out that we don’t need all that precision for NN training, but we do actually need all the range.

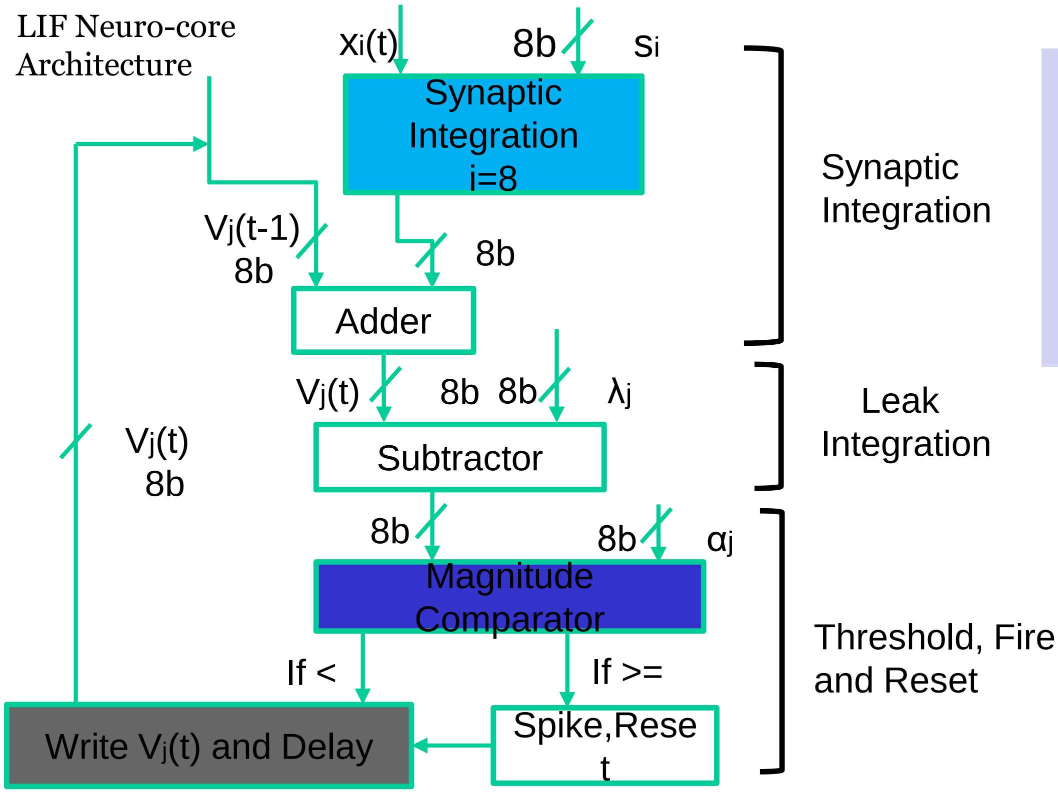


Agenda

- **Fundamental Trends**
- **AI – The Emerging Industrial Revolution**
- **AI at the Edge**
- **ASL Neuromorphic Chips**
- **Conclusions**

46

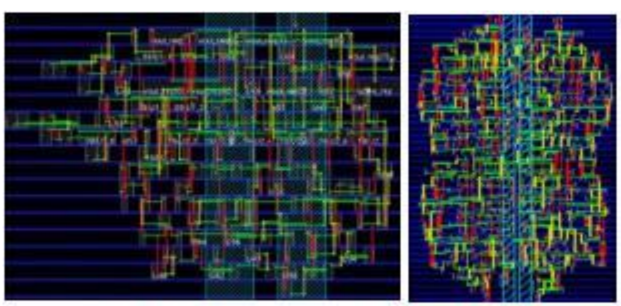
LIF Neuro-core for NASH System



- 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



LIF-1N-012018-KS LIF-4N-012018-KS

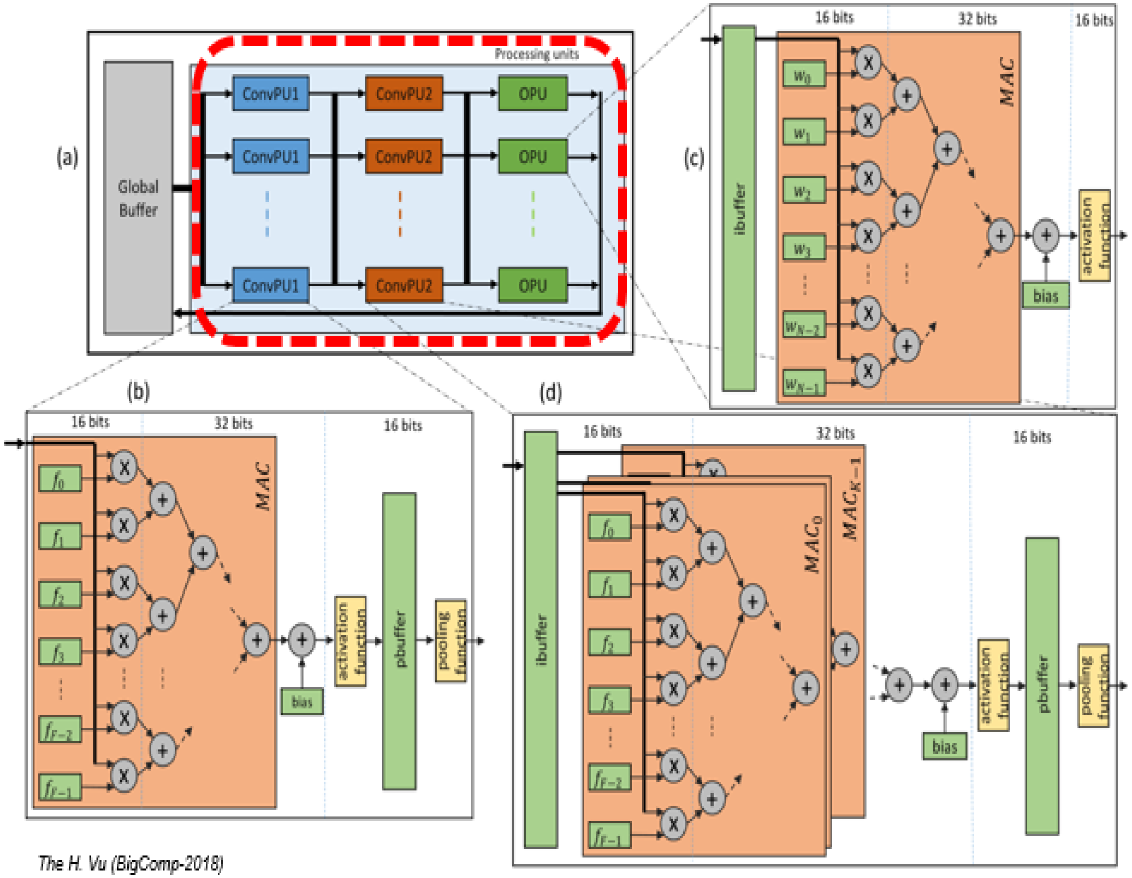
Placement of LIF-1N (Left) and LIF-4N (right)

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

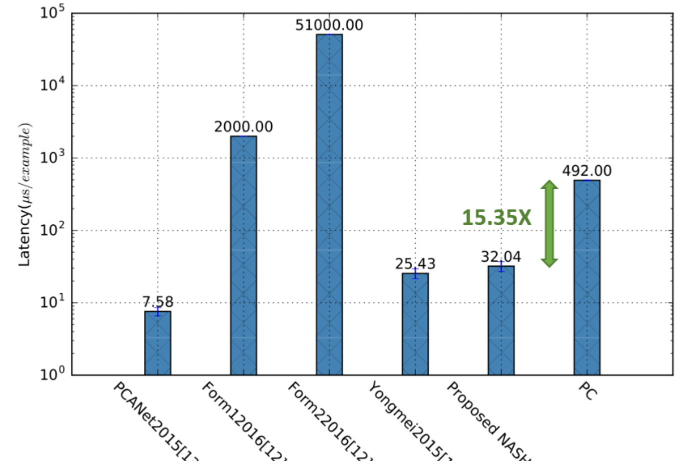
Kanta Suzuki, Yuichi Okuyama, Abderazek Ben Abdallah, "Hardware Design of a Leaky Integrate and Fire Neuron Core Towards the Design of a Low-power Neuro-inspired Spike-based Multicore SoC", Proc. Of IPSJ, 2018

Application I
Neuro-inspired Hardware System for Image Recognition

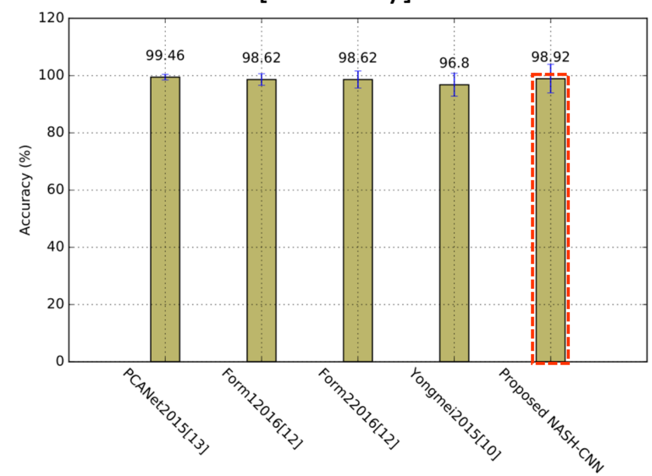


The H. Vu (BigComp-2018)

[Execution time]

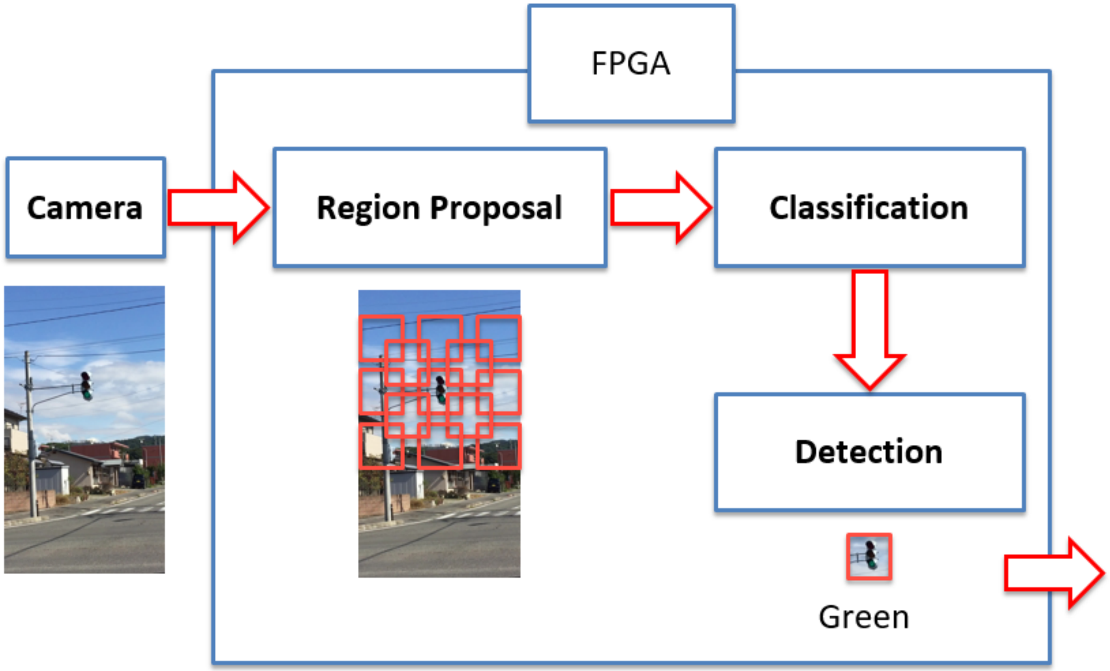


[Accuracy]

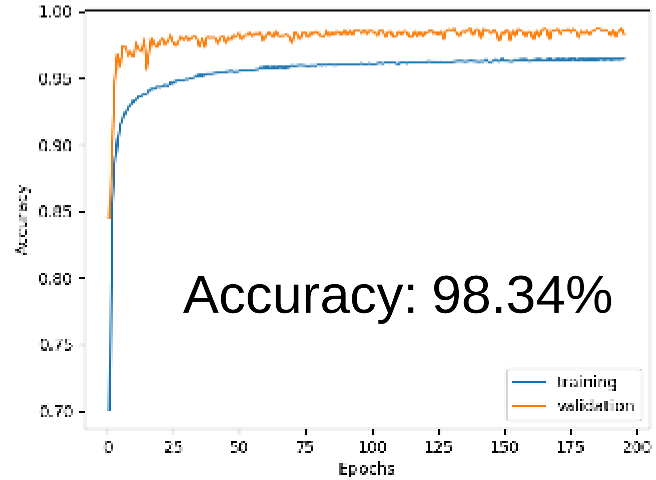


The H. Vu, Ryunosuke Murakami, Yuichi Okuyama, Abderazek Ben Abdallah, "Efficient Optimization and Hardware Acceleration of CNNs towards the Design of a Scalable Neuro-inspired Architecture in Hardware", Proc. of the IEEE International Conference on Big Data and Smart Computing (BigComp-2018), January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

Application II
Neuro-inspired Hardware System for Autonomous Vehicles



CNN



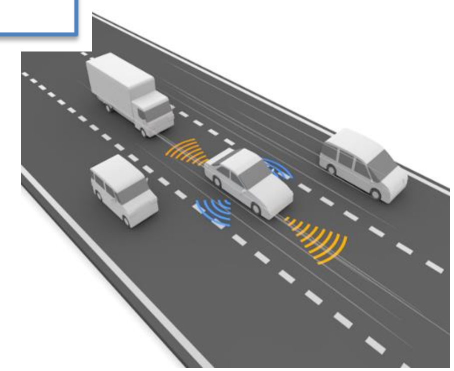
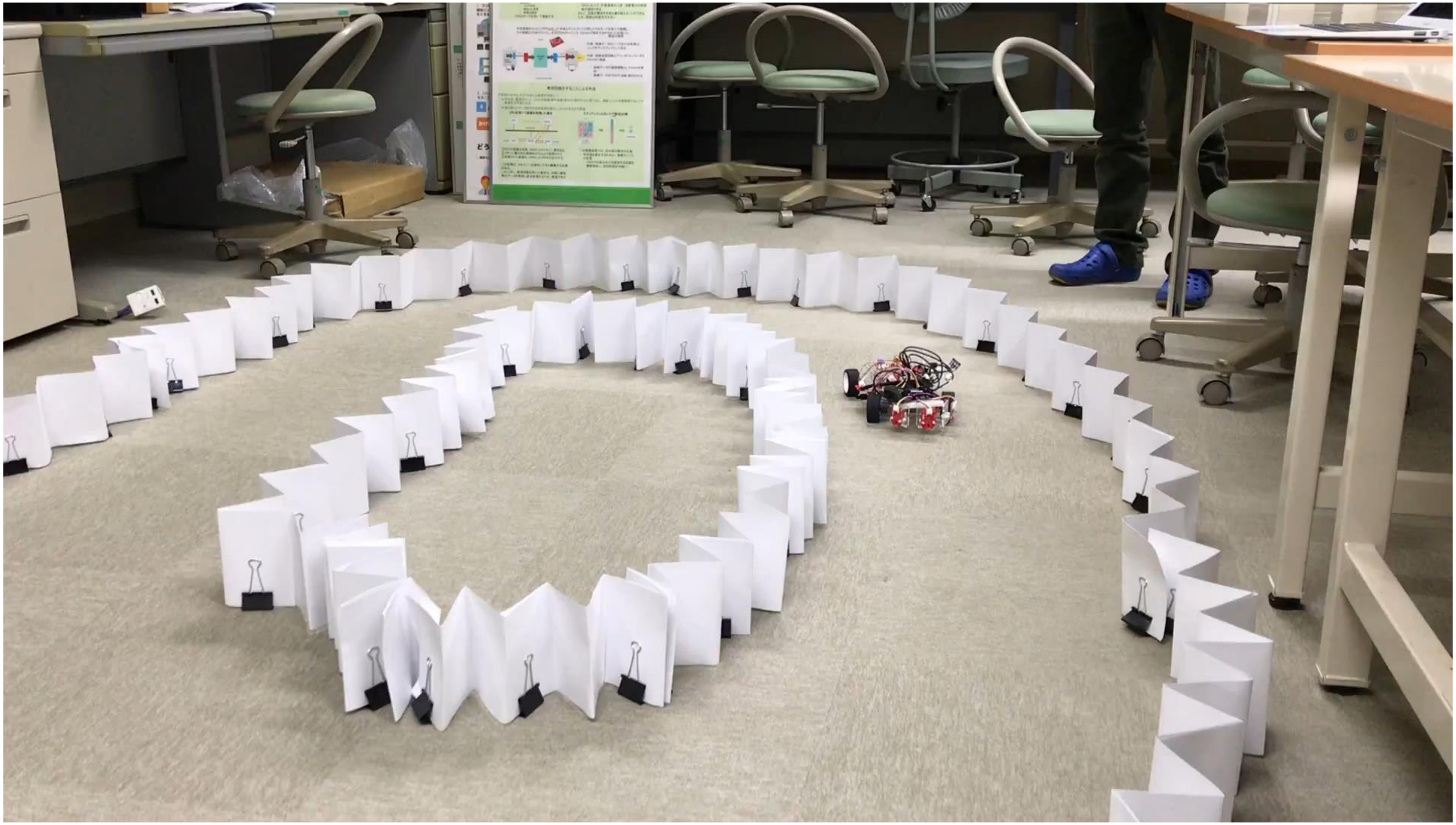


Table 1 : ANN Performance Evaluation

ALUs	Registers	Pins	Fmax
10,989 (33%)	5,814 (18%)	432 (89%)	76.02 MHz
Memory	DSP Block	Power Consumption	
4,956 (1%)	54 (77%)	286.84 mW	

- Yuji Murakami, " Design of a Neural Network Architecture for Traffic Light Detection Towards Autonomous Driving Vehicles," Master's Thesis, Graduate School of Computer Science and Engineering, The University of Aizu, 3/2019
- Yuji Murakami, Yuichi Okuyama, Abderazek Ben Abdallah, "SRAM Based Neural Network System for Traffic-Light Recognition in Autonomous Vehicles", Information Processing Society Tohoku Branch Conference, Feb. 10, 2018

Demo 1



Application III
Neuro-inspired System for Wild Animals Monitoring

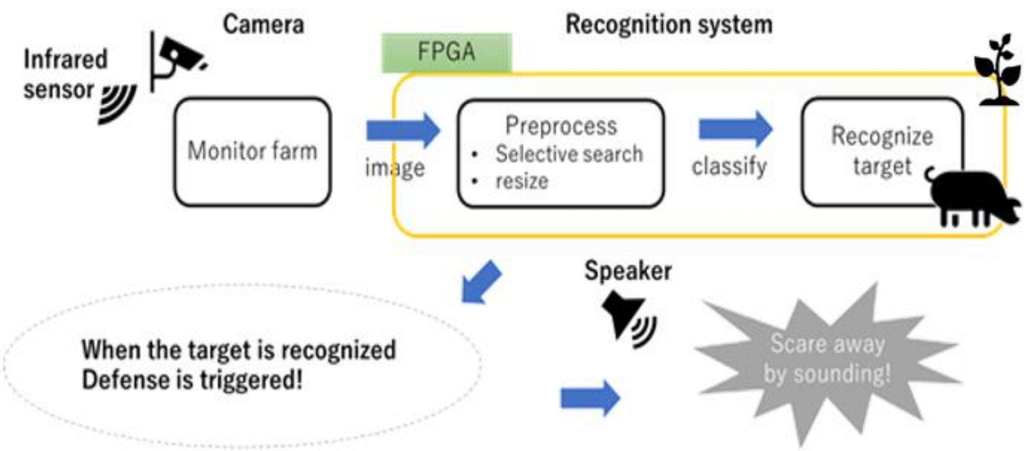
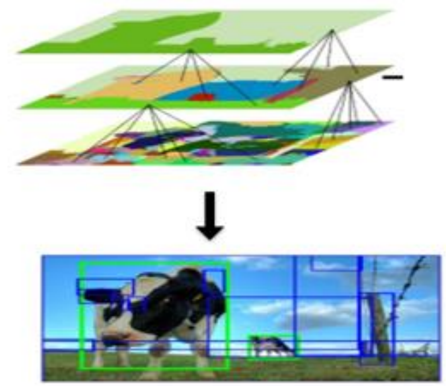


Fig 4. System overview: OASIS FMS-1



出典:「Rich feature hierarchies for accurate object detection and semantic segmentation」

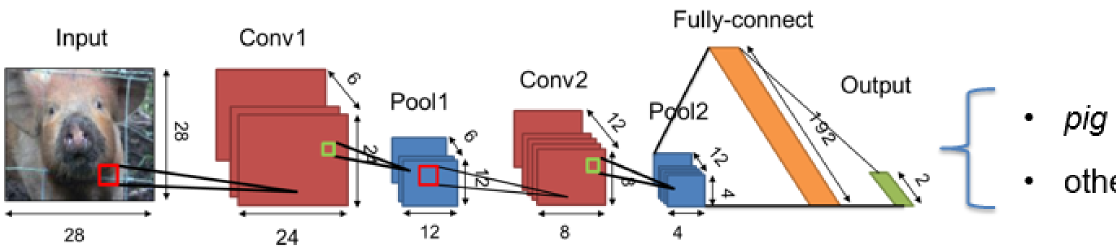


Fig 3. CNN example

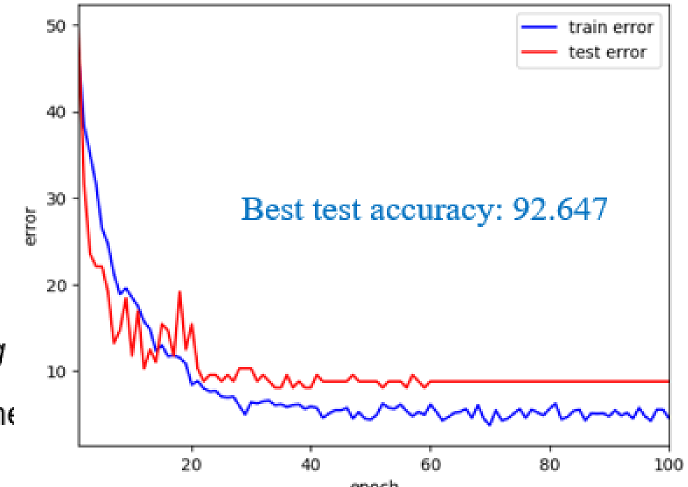
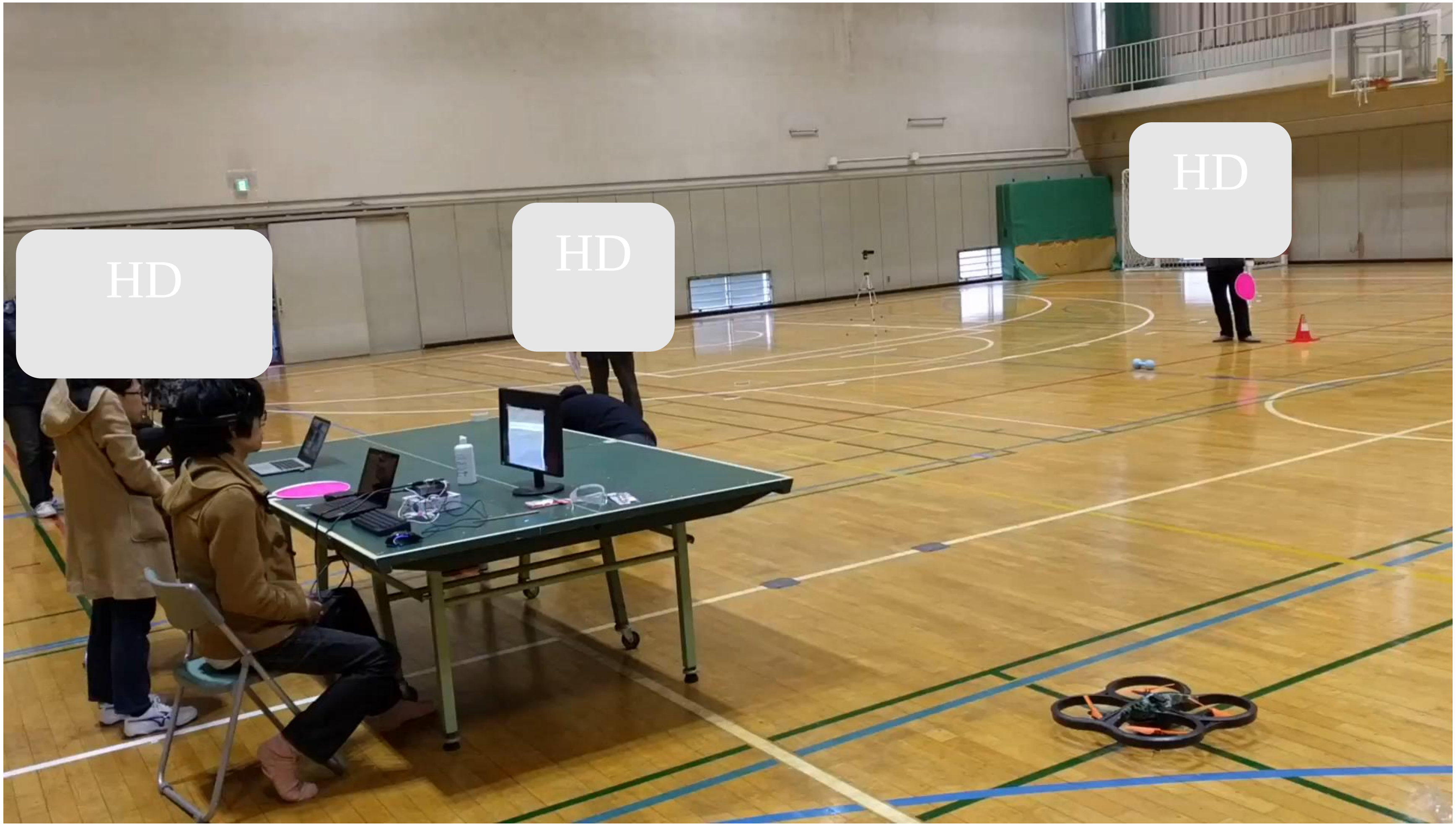


Fig 6. Learning result

Ryunosuke Murakami, Yuichi Okuyama, Abderazek Ben Abdallah, "Animal Recognition and Identification with Deep Convolutional Neural Networks for Farm Monitoring", Information Processing Society Tohoku Branch Conference, Feb. 10, 2018

Demo 2



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

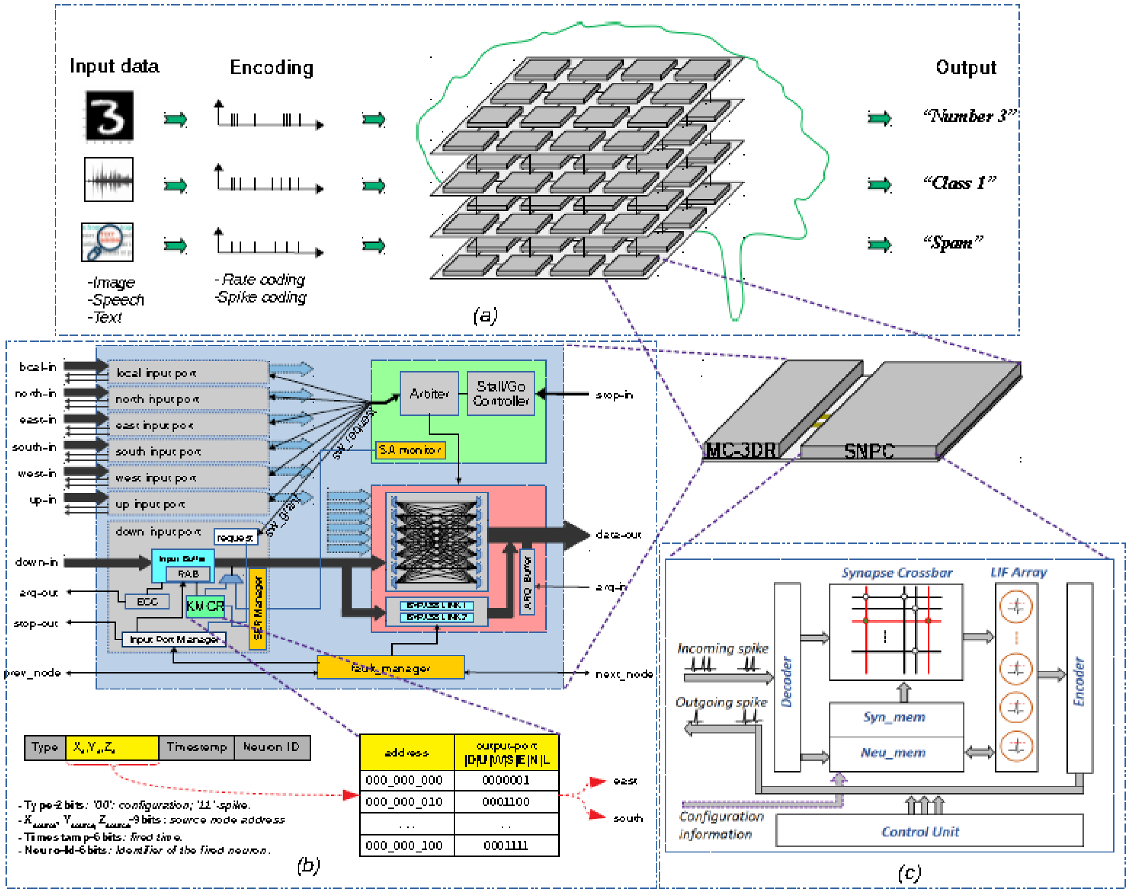
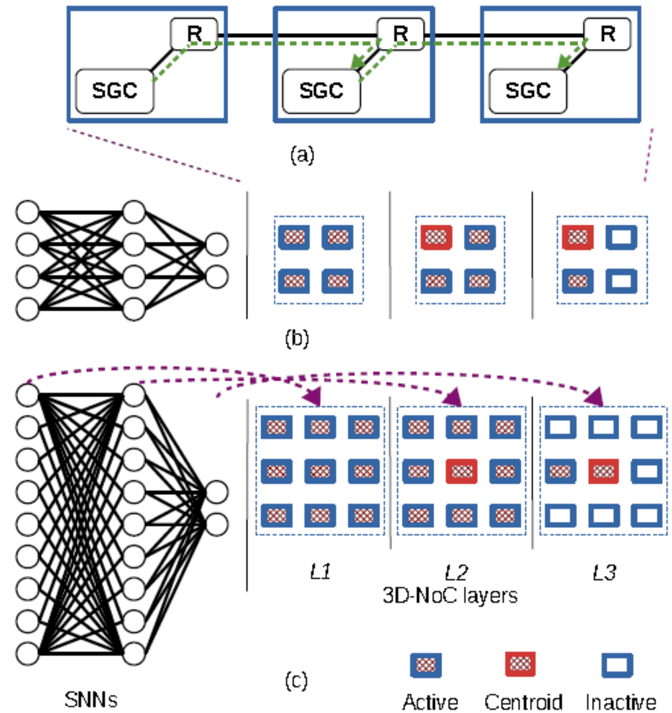
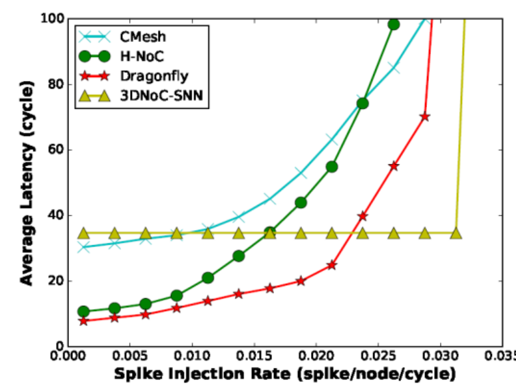
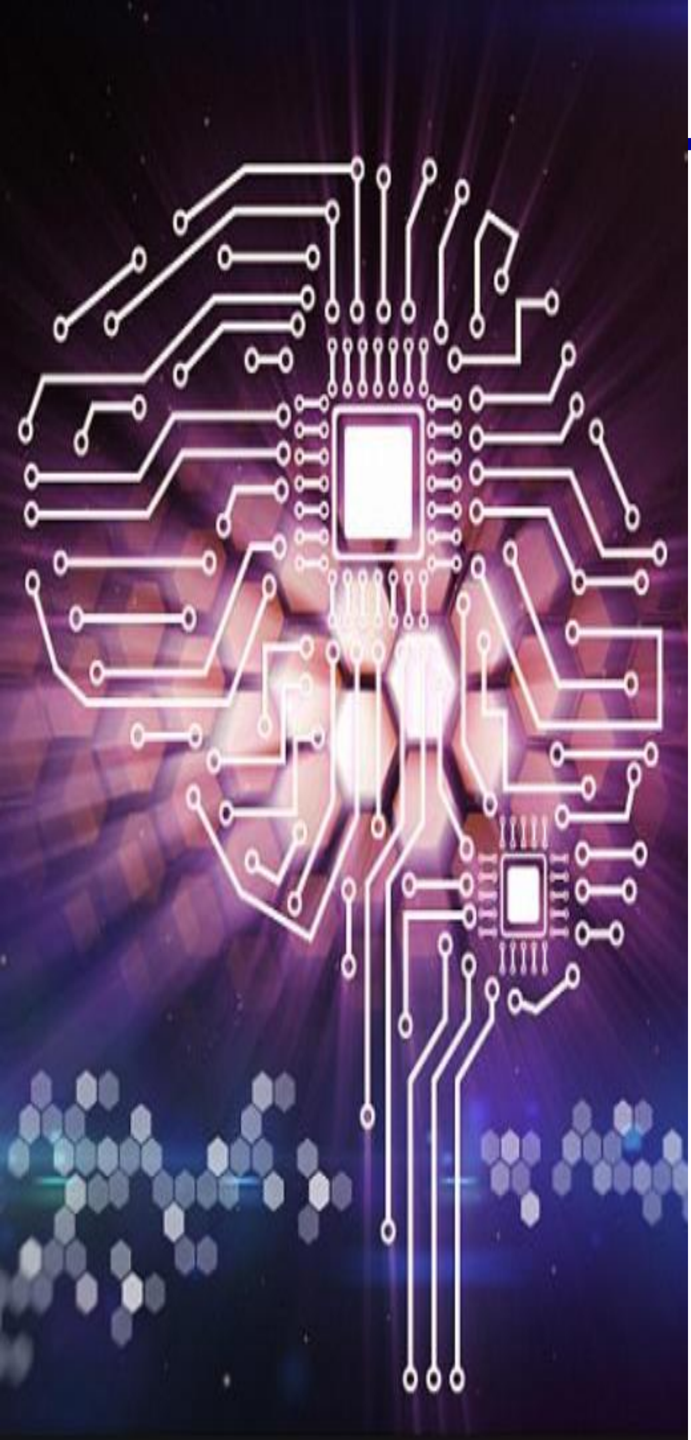


Fig. 5: System architecture: (a) 3DNoC-SNN organization, (b) Multicast router architecture (MC-3DR), (c) Spiking neuron processing core (SNPC).





Average latency evaluation and comparison over various SIRs.



Agenda

- **Fundamental Trends**
- **AI – The Emerging Industrial Revolution**
- **AI at the Edge**
- **ASL Neuromorphic Chips**
- **Conclusions** 54

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)

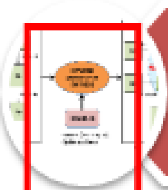
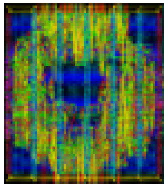
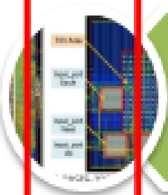
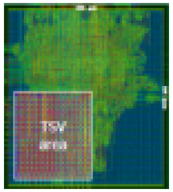
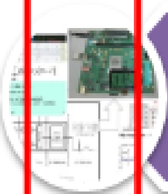
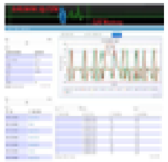
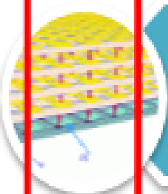
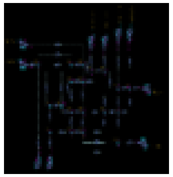
❖ Reduce size of operands for storage/compute

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

❖ Reduce number of operations for storage/compute

- ✓ Network Pruning; Compact Network Architectures

ASL SoCs , AI-Chips

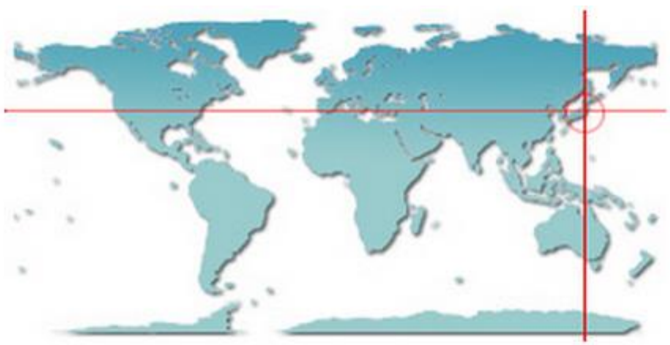
Year		Project	Publications	
2006		OASIS-1 – Scalable Packet-Switched Network-on-Chip	JASSST06, MCSOC12, JPDC14, SUP14	
2013		OASIS-2 - Fault-Tolerant Network-on-Chip	MCSOC14, JPDC14, SUP16	
2014		BANSMOM - Bio-Chip for Elderly Monitoring	ES2016,ACHRAF-MS1, KIMEZAWA-MS	
2015		PHENIC- High-bandwidth Photonic NoC	SUP16, MCSOC15, CANDAR16,	
2018		MASH - Neuromorphic AI Chips	BigComp18, BigCom19, SC19	

Thank you.

Ben Abdallah Abderazek

***A**daptive **S**ystems **L**aboratory*

benab@u-aizu.ac.jp



to Advance Knowledge for Humanity