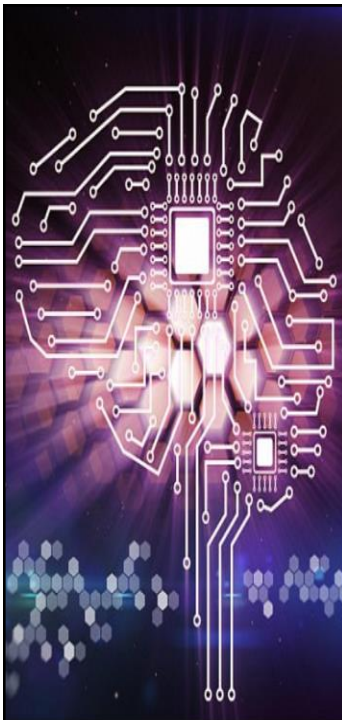


# AI-Chips: Artificial Intelligence Chips for Intelligent Systems



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

1



## Agenda

- **Fundamental Trends**
- **AI – The Emerging Industrial Revolution**
- **Intelligent AI-Chips**
- **ASL AI-Chips and Systems**
- **Conclusions**

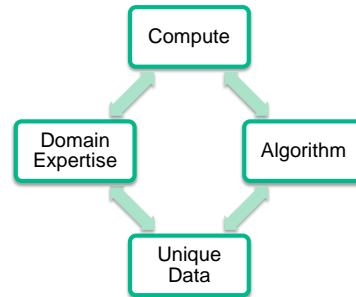
2

## Why do we need Intelligent Systems (IS)?

Experts do not scale → we need intelligent workflows.

### What is Intelligence?

The ability of a system to **compute**, **reason**, **perceive** relationships and analogies, **learn** from experience, **store** and **retrieve** information from memory, solve problems, classify, generalize, and **adapt** to new environments.



The value chain of intelligent systems

3

## AI-Chips are ... everywhere

### Self-driving Car



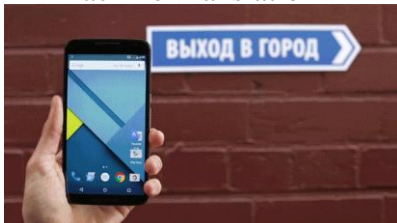
Bottom Image source: [edition.cnn.com](http://edition.cnn.com)

### Smart Robots



Image source: [roboticsbusinessreview.com](http://roboticsbusinessreview.com)

### Machine Translation



Bottom Image source: [missqt.com](http://missqt.com)

### Gaming

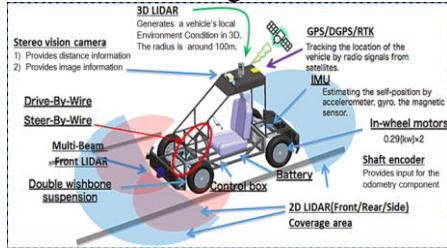


Bottom Image Source: [newatlas.com](http://newatlas.com)

4

# AI-Chips are ... everywhere

## Self-driving Car



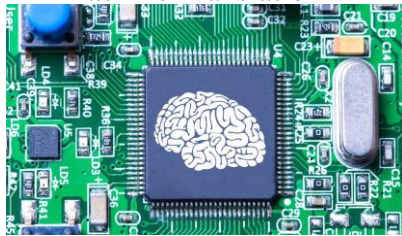
Bottom Image source: [edison.chm.com](http://edison.chm.com)

## Smart Robots



Image source: [roboticsbusinessreview.com](http://roboticsbusinessreview.com)

## Machine Translation



Bottom Image source: [missqt.com](http://missqt.com)

## Gaming



Bottom Image Source: [newatlas.com](http://newatlas.com)

5

# AI-Chips are ... everywhere

## Brain implant allows paralysed monkey to walk

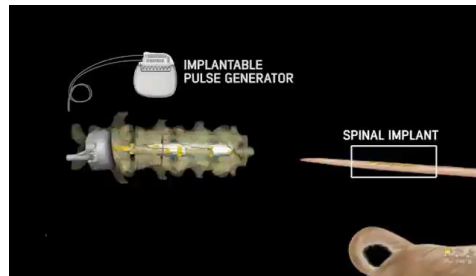
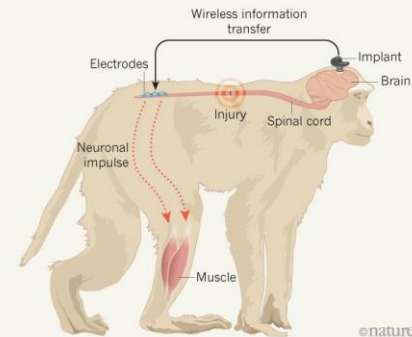
*There really is a kind of intelligence inside the spinal cord. We are not just talking about reflexes that automatically activate muscles. In the spinal cord there are networks of neurons able to take their own decisions*

*-Grégoire Courtine-*

*Neuroscientist, Federal Institute of Technology, Lausanne*

### PARALYSED PRIMATES WALK

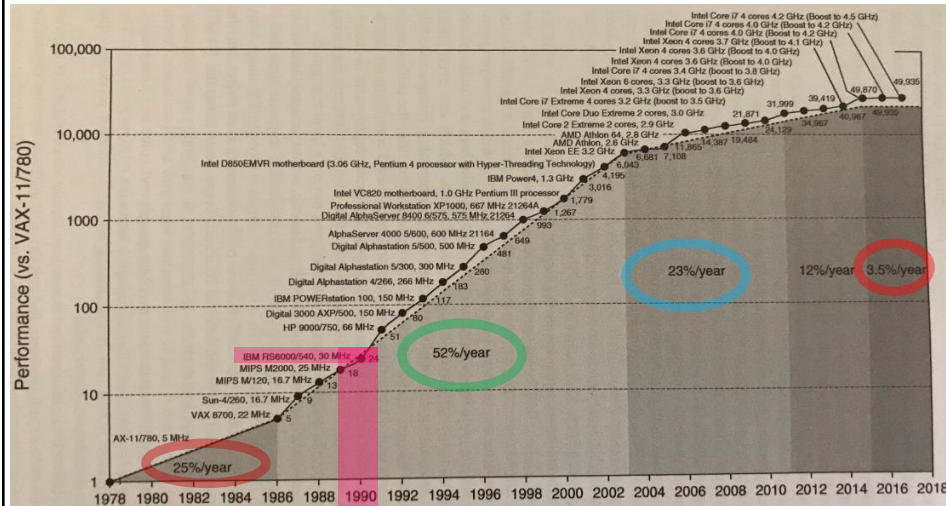
A wireless implant bypasses spinal-cord injuries in monkeys, enabling them to move their legs.



Nature volume539, pages284–288 (10 November 2015)

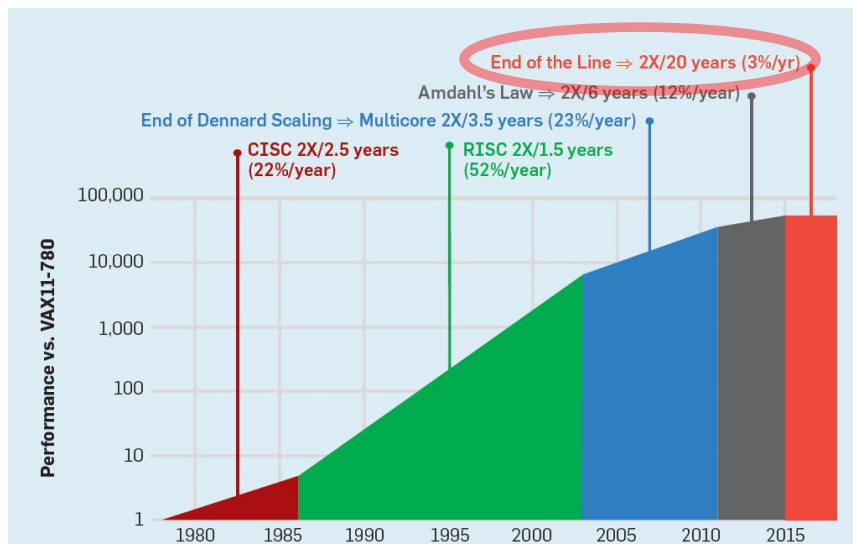
6

## Moore's law is no longer providing more Compute



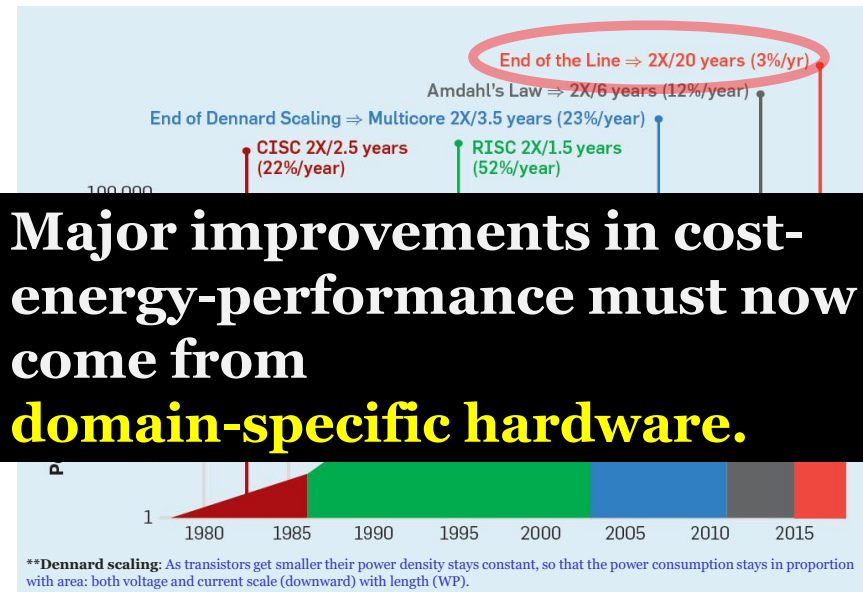
Source: Norman P. Jouppi, Cliff Young, Nishant Patil, David Patterson, Communications of the ACM, September 2018, Vol. 61 No. 9, Pages 50-59

## Moore's law is no longer providing more compute



\*\*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).

## Moore's law is no longer providing more compute



9

## DNN Compute Requirements is Steadily Growing

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
<b># of CONV Layers</b>	<b>2</b>	<b>5</b>	<b>16</b>	<b>21 (depth)</b>	<b>49</b>
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
<b># of FC layers</b>	<b>2</b>	<b>3</b>	<b>3</b>	<b>1</b>	<b>1</b>
# of Weights	58k	58.6M	124M	1M	2M
# of MACs	58k	58.6M	124M	1M	2M
<b>Total Weights</b>	<b>60k</b>	<b>61M</b>	<b>138M</b>	<b>7M</b>	<b>25.5M</b>
<b>Total MACs</b>	<b>341k</b>	<b>724M</b>	<b>15.5G</b>	<b>1.43G</b>	<b>3.9G</b>

Source: Joel Emer, ISCA Tutorial, 2017

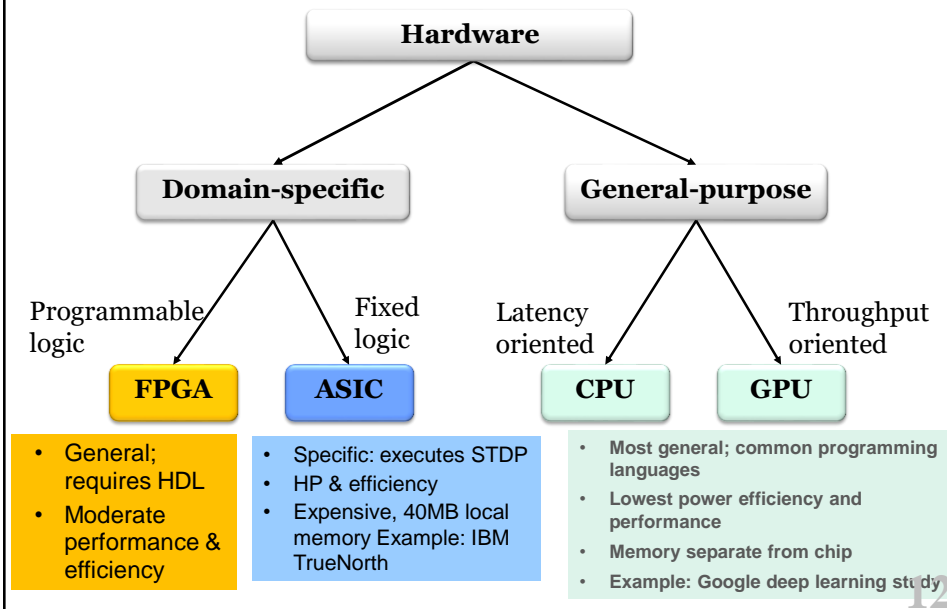
10

## What does it mean ?

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

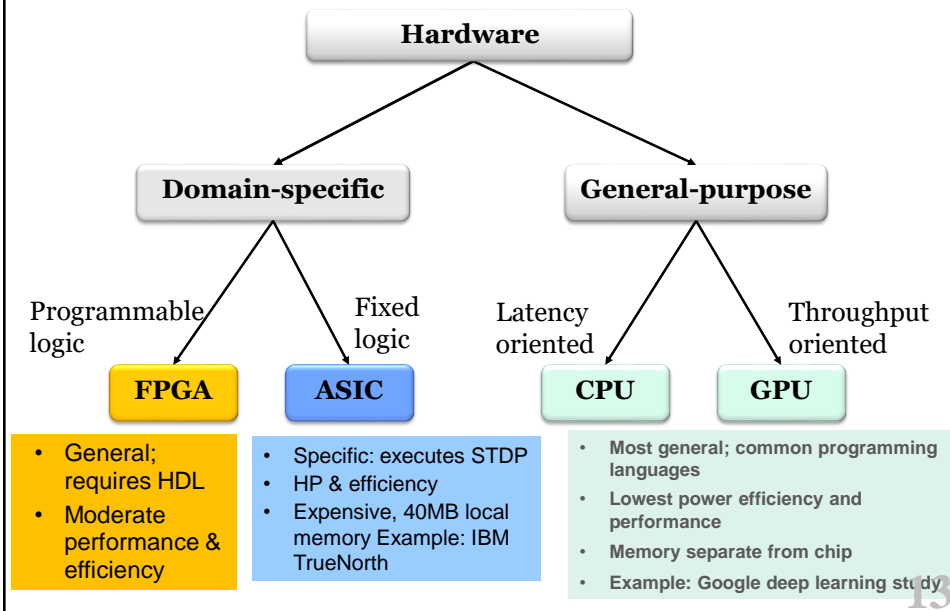
11

## Current State of the Art in Neural Algorithms HW Computing

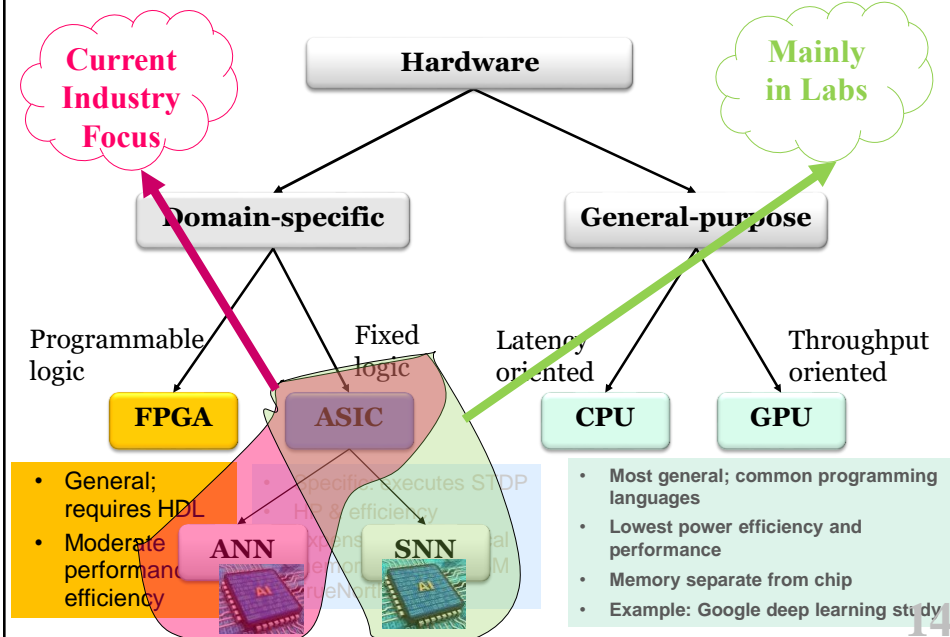


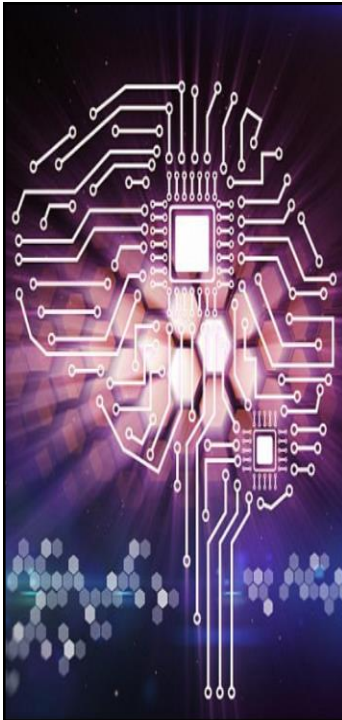
12

## Current State of the Art in Neural Algorithms HW Computing



## Current State of the Art in Neural Algorithms HW Computing





# Agenda

- Fundamental Trends
- **AI – The Emerging Industrial Revolution**
- Intelligent AI-Chips
- ASL AI-Chips and Systems
- Conclusions

15

## Four Main Factors in Promoting AI/AI HW



Image: kduggets.com

AI algorithms are being applied to nearly everything we do.



Image: sas.com

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

Strong Gov. & Industry Engagements



Image: kduggets.com

Growth of computational power



Image: spectrum.ieee.org

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

16

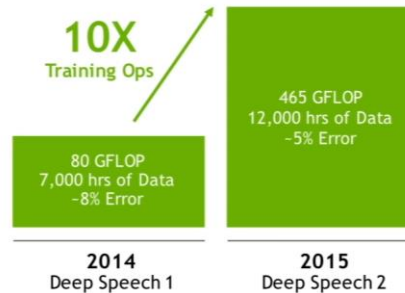
## Hardware & Data Enable DNNs

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

### IMAGE RECOGNITION



### SPEECH RECOGNITION



Microsoft

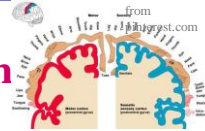
Baidu 百度

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

17

## AI HW is inspired by Nature – Biological neuron

AI Chips and systems are inspired by biology → parallel computation



18

## AI HW is inspired by Nature – Biological neuron

**AI Chips and systems are inspired by biology → parallel computation.**

- ❖ # of neurons:  $\sim 10^{11}$
- ❖ # of synapses:  $\sim 10^{15}$
- ❖ Power consumption:  $\sim 20$  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  **$\sim 500$  s** on state-of-the-art supercomputer

Latest digital DL processors:

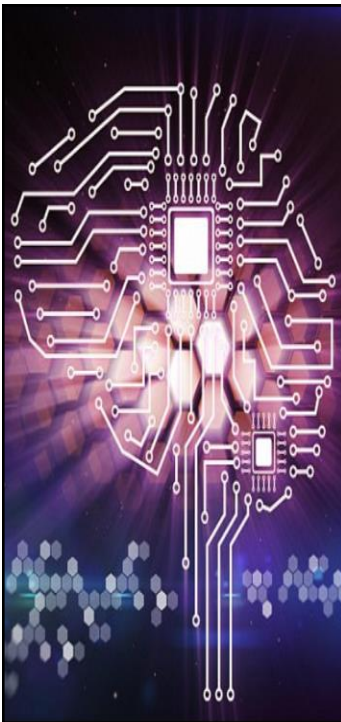
$\sim 10$  TOPS/W

Synapse op. in brain: 0.1~1 fJ/op

1,000~10,000 TOPS/W

=1~10 POPS/W

19



## Agenda

- Fundamental Trends
- AI – The Emerging Industrial Revolution
- **Intelligent AI-Chips**
- ASL AI-Chips and Systems
- Conclusions

20

## Different approaches to AI Chips

Poor/Simple

Good/Complex

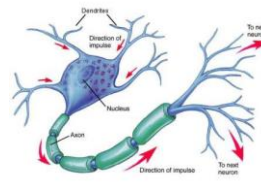


**Neuron**    Digital, Analog, LIF.    . . .    Izhikevich model    Huxley-Hodgkin model    . . .

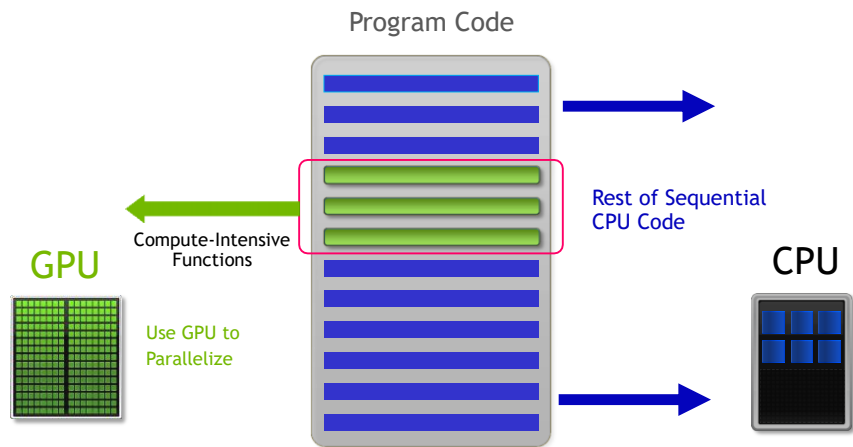
**Synapse**    MAC (weighted . . . sum)    Spiking STDP    . . .    Many nonlinear properties    . . .

Generally Used in DL algorithms

**Frequency**    10~100 Hz (brain)

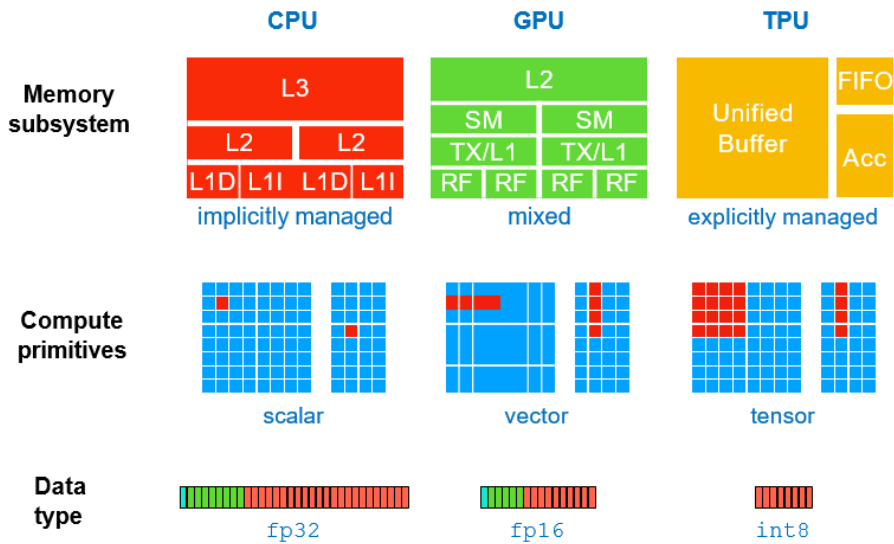


## Current AI Chip = Accelerator/Co-processor



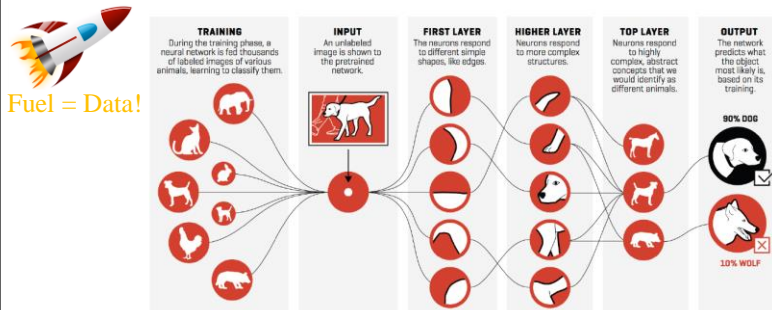
Acceleration with GPU

## Accelerator Characteristics



[Ref 3] 23

## ... Deep Learning is considered as a sophisticated "rocket" of Machine Learning!!

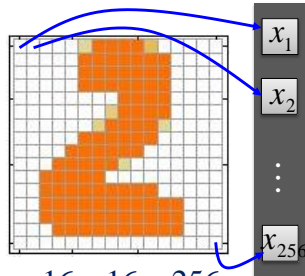


1. "Deep Learning" means using a neural network with several layers of nodes between input & output
2. the series of layers between input & output do feature identification and processing in a series of stages, just as our brains seem to.

24

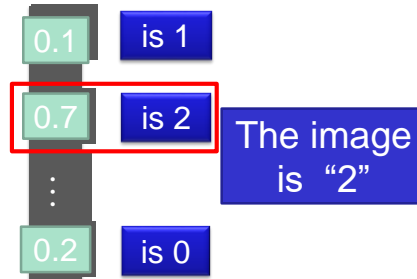
## Example1: Handwriting Digit Recognition on FPGA

Input



16 x 16 = 256  
Ink → 1  
No ink → 0

Output

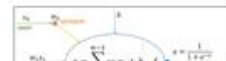
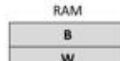
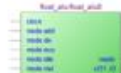


Each dimension represents the confidence of a digit.

25

## Example2: Character Recognition on FPGA

Character Recognition with BP training



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	



'O' letter



26

## The are two AI Chip Models: ANN and SNN

- The output of ANN Chip depends only on the current stimuli, the output of SNN depends on previous stimuli also
- The SNN/Neuromorphic Chip operates on biology-inspired principles to improve performance and increase energy efficiency

**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)

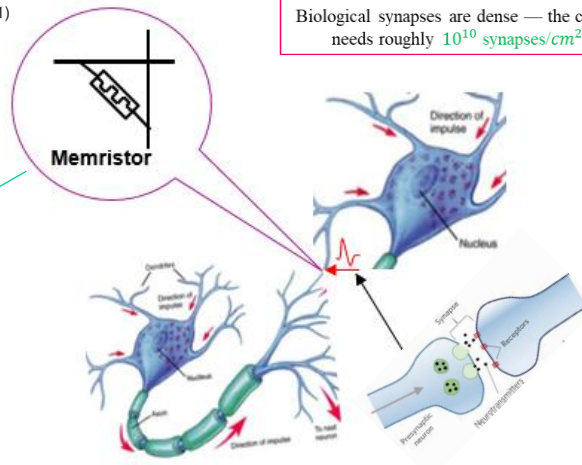
27

## Memristor for Synapse Design

(Chua, 1971)

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

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

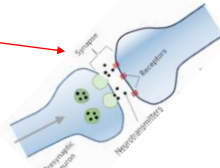


❖ Voltage **pulses** can be applied to a **memristor** to change its **resistance**, just as **spikes** can be applied to a **synapse** to change its **weight**.

28

## How biological neurons learn?

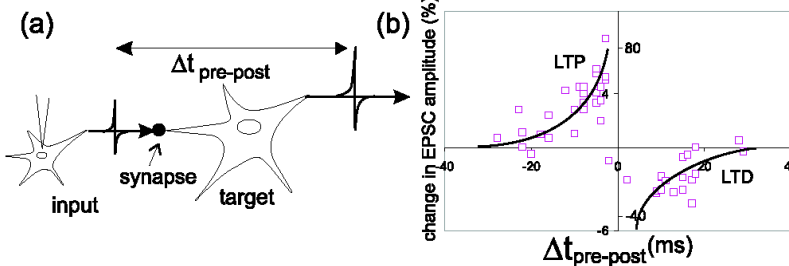
Brain is a large network of neurons connected and communicating via **synapses**



29

## 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.



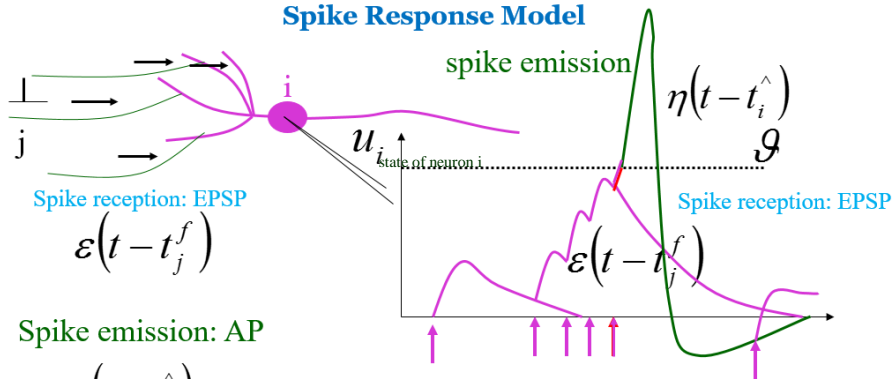
Spike-timing-dependent plasticity (STDP)

- 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

30

# Spiking Neuron Model

## Spike Response Model



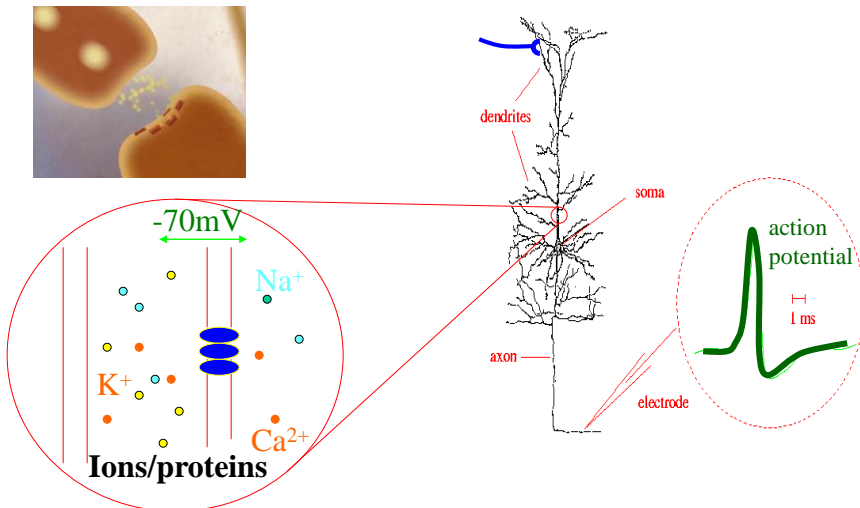
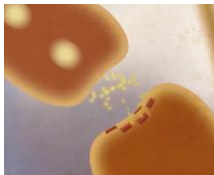
$$u_i(t) = \eta(t - \hat{t}_i) + \sum_j \sum_f w_{ij} \varepsilon(t - t_j^f)$$

$$u_i(t) = \mathcal{G} \Rightarrow \text{Firing: } \hat{t}_i = t$$

[Ref. 18]

31

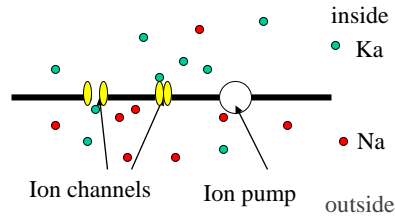
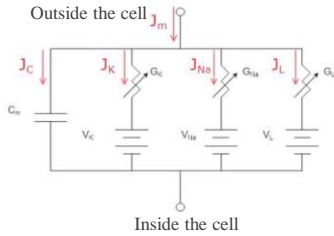
# Spiking Neuron Model- Molecular Basis



[Ref. 18]

32

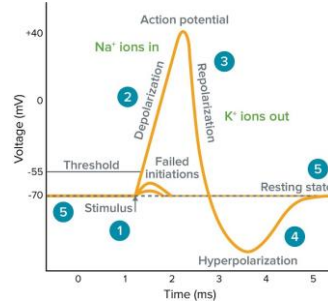
# Hodgkin-Huxley Model



$$J_c = C_m \frac{\partial V_m}{\partial t} \quad J_{Na^+} = G_{Na^+} (V_m - V_{Na^+})$$

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

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

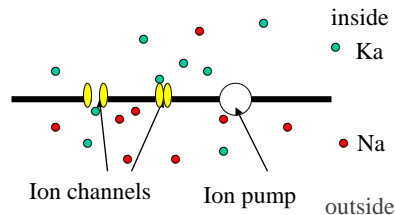
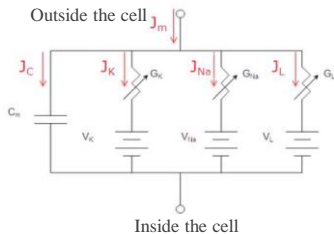


$$J_m = C_m \frac{\partial V_m}{\partial t} + G_{K^+} (V_m - V_{K^+}) + G_{Na^+} (V_m - V_{Na^+}) + G_L (V_m - V_L)$$

[Ref. 18]

33

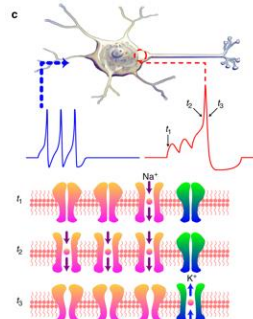
# Hodgkin-Huxley Model



$$J_c = C_m \frac{\partial V_m}{\partial t} \quad J_{Na^+} = G_{Na^+} (V_m - V_{Na^+})$$

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

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



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

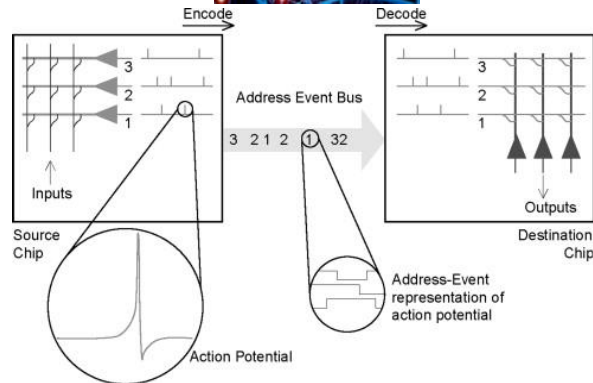
[Ref. 18]

34

## Wiring via AER (Address Event Representation)



(Courtesy: iStock/Henrik5000)



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

34

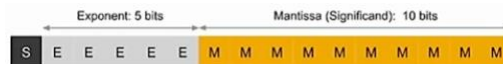
## NN Training Works with Low-precision FP

### 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

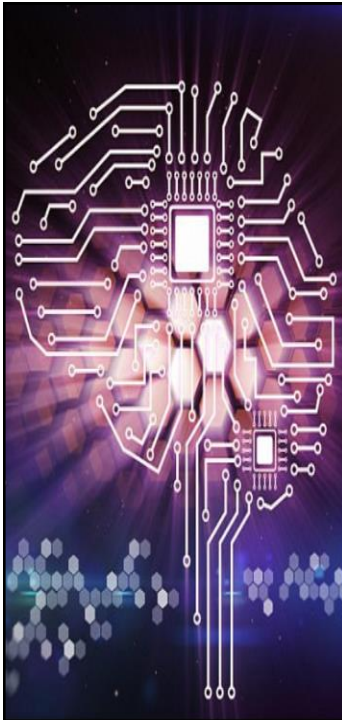
### bfloat16: Brain Floating Point Format



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.

35

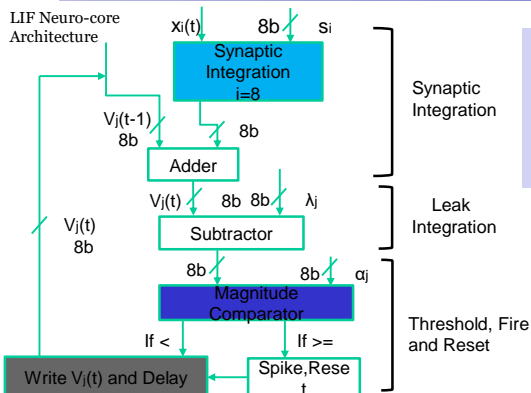


# Agenda

- Fundamental Trends
- AI – The Emerging Industrial Revolution
- Intelligent AI-Chips
- **ASL AI-Chips and Systems**
- Conclusions

36

## LIF Neuro-core for NASH System



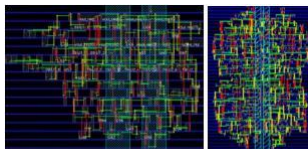
- $X_i(t)$  – Spike input to the synapse
- $S_i$  – synaptic weight
- $V_j(t)$  – Membrane potential
- $\alpha_j$  – Neuron threshold
- $\lambda_j$  – Leak value

Table 1: Area Evaluation

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

Table 1: Power Evaluation

Item	NC-1N	NC-4N
Combinational Area	186.998 $\mu$ m	562.856001 $\mu$ m
Non-Comb Area	47.88002 $\mu$ m	213.864000 $\mu$ m
Total Cell Area	234.878002 $\mu$ m	776.720001 $\mu$ m



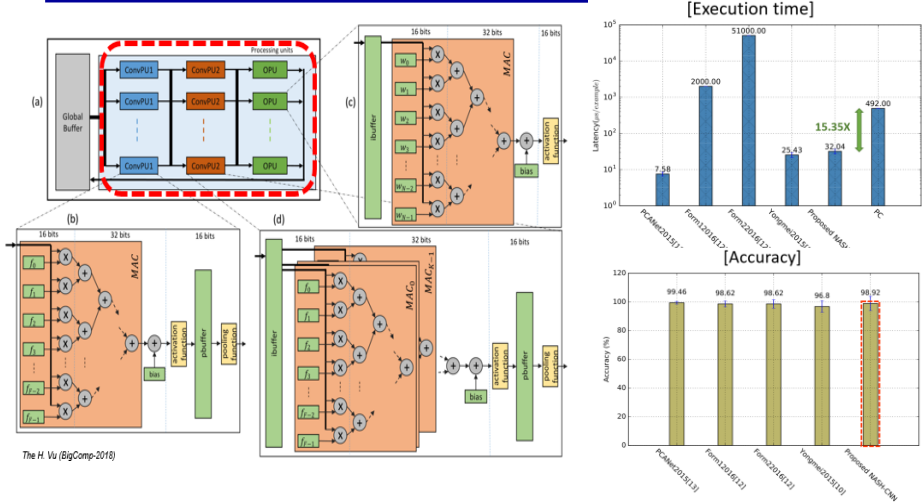
Placement of LIF-1N (Left) and LIF-4N (right)

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

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

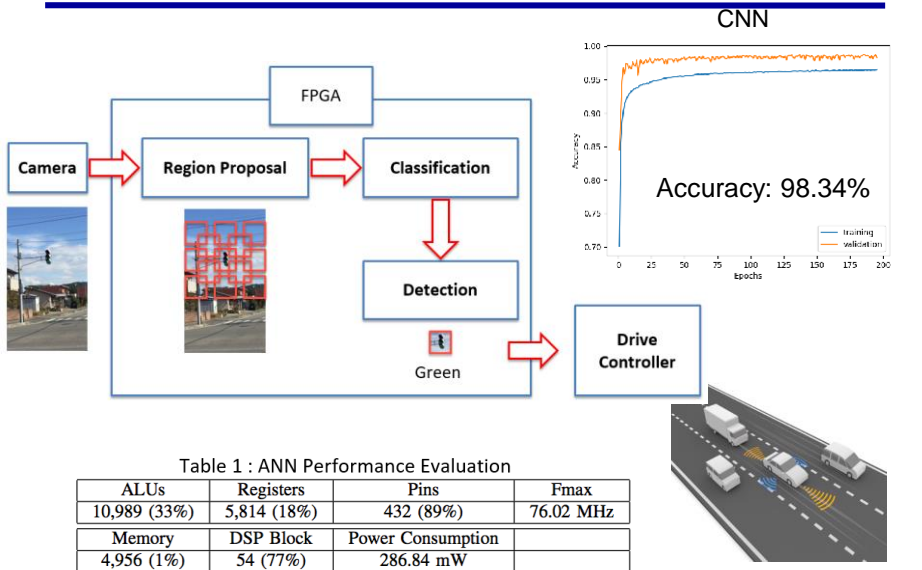
37

## Application I Neuro-inspired Hardware System for Image Recognition



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



- 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



40

## Application III Neuro-inspired System for Wild Animals Monitoring

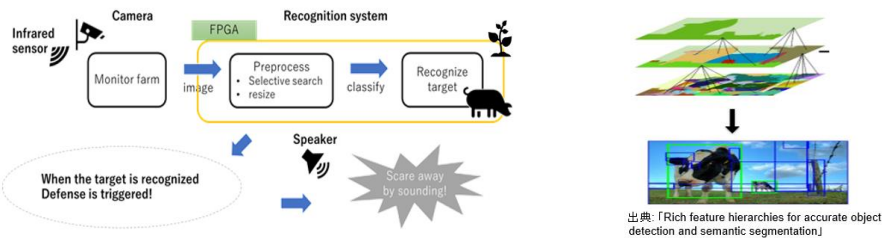


Fig 4. System overview: OASIS FMS-1

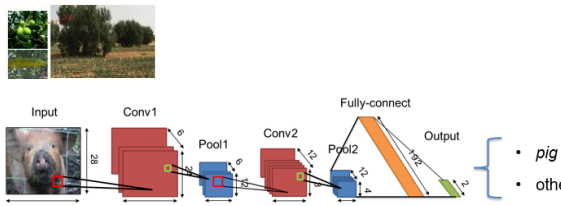
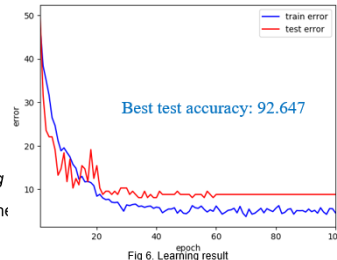


Fig 3. CNN example



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

41

# Application IV Brain-inspired Drone Control with BCI

Flight navigation logic (in C++/Python)

Emotiv community SDK

Parrot ARDrone SDK

Other libraries

Facial command  
Mental command  
Raw EEG  
Gyro

Drone control  
Video inputs  
Sensor inputs

Spiking neurons  
Neural networks  
SLAM (mapping)

**Brain to Brain drone system**

**Numerical computation with SNNs**

**Demo**

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

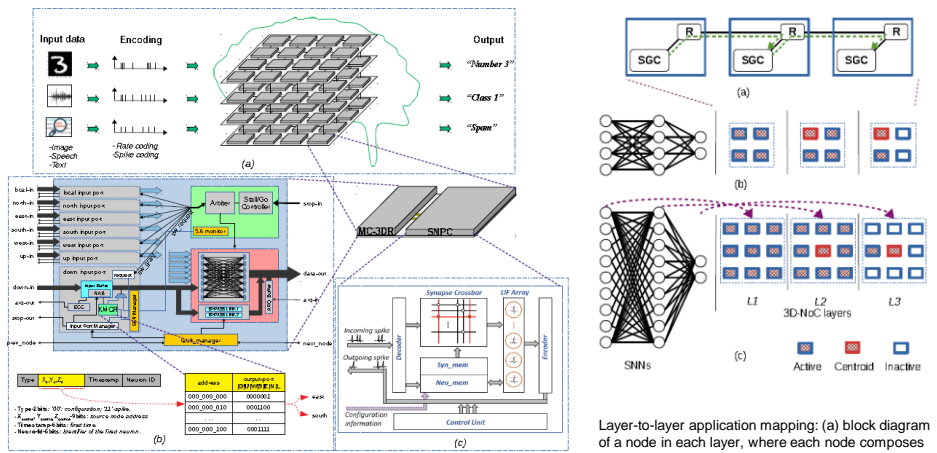
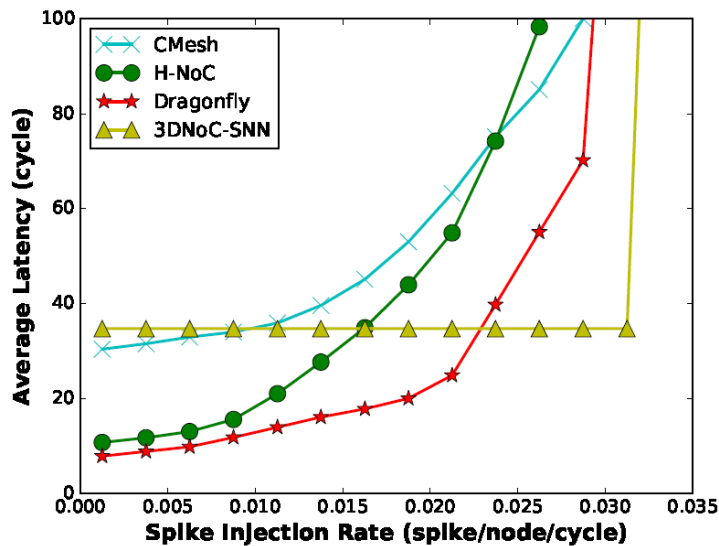


Fig. 5: System architecture: (a) 3DNoC-SNN organization, (b) Multicast router architecture (MC-3DR), (c) Spiking neuron processing core (SNPC).  
 Layer-to-layer application mapping: (a) block diagram of a node in each layer, where each node composes of a spike generator/counter (SGC) and a router, (b) for the Inverted pendulum application: neurons in SNN (left-side) are mapped onto the proposed system in a layer-to-layer manner, (c) mapping method for Wisconsin data-set

The H. Vu, Abderazek Ben Abdallah, "Comprehensive Analytic Performance Assessment and Low-latency Algorithm for Spike Traffic Routing in 3D-NoC of Spiking Neurons (3DNoC-SNN)", *ACM Journal on Emerging Technologies in Computing (JETC)*, (to appear)

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



Average latency evaluation and comparison over various SIRs.

44

## 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

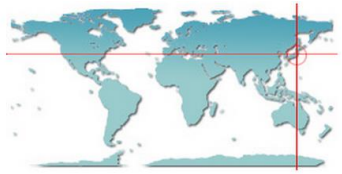
46

# Thank you!

**Ben Abdallah Abderazek**

*Adaptive Systems Laboratory*

[benab@u-aizu.ac.jp](mailto:benab@u-aizu.ac.jp)



to Advance Knowledge for Humanity