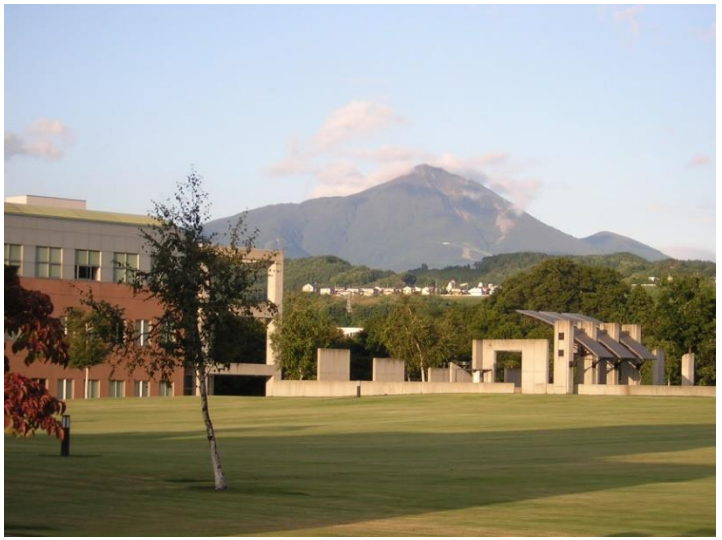
# Neuro-Inspired Adaptive Manycore SoCs and Applications

**Abderazek Ben Abdallah**
University of Aizu
Division of Computer Engineering
Adaptive Systems Laboratory
benab@u-aizu.ac.jp

# Aizu University

**Selected as a Top Global University by Ministry of Education, Culture, Sports, Science and Technology (MEXT)**

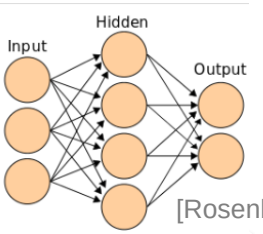




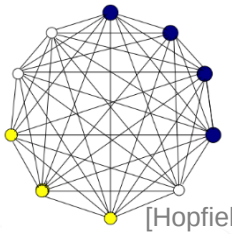
# Computing Trends: Opportunities & Challenges

- Huge **progress in IC technology**
- Emergence of **IoT, Embedded/Ubiquitous/Pervasive** applications.
- **Large bandwidth and** low power requirements.
- Data become more knowledge intensive (unstructured).
- Performance is limited by the **Communication** Network rather than the Computation Logics.

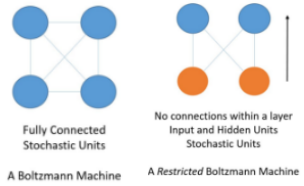
# Neural Networks



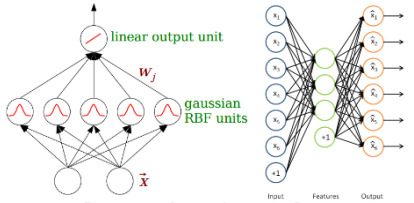

[Rosenblatt,1958]



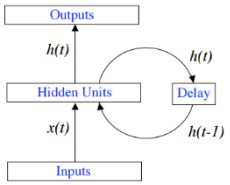
[Hopfield,1982]




Hinton,1986]




[Broomhead,1988]




[Elman,1990]

**1954**
- F-Forward (FFNN)

**1982**
- Hopfield (HN)

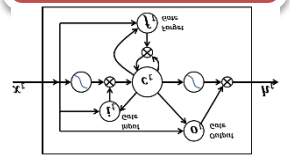
**1986**
- **Boltzmann, Restricted Boltzmann (RBM)**

**1988**
- Radial Bas. Func, Autoencoder (AE)

**1990**
- Recurrent Neural (RNN)

[Hochreiter,1997]

**1997**
- Long/Short Term Mem (LSTM), BiRNN



**1998**
- Conv. Neural Network (CNN)

[LeCun, 1998]

**2004**
- Echo State Network (ESN)

[Jaeger, 2004]

**2007**
- **Deep Belief Net (DBN), SAE**

[Bengio, 2007]

Ex.
I: red flower
O: red flower image

**2010**
- **Deconvolutional Network (DN)**

[Zeiler, 2010]

[Hayes, 2013]

**2013**
- **Markov Chain (MC), VAE, ELM**

**2014**
- Generartive adversarial net(GAN, GRU, NTM

[Goodfellow, 2015]

**2015**
- Deep Residual Net (DRN), DCIGN,

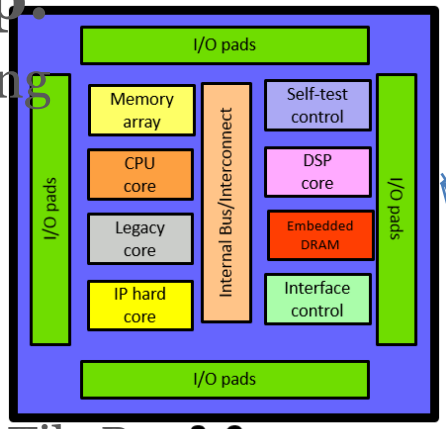
[He, 2015]

**2015**
- **Deep Convol. Inverse graphics Net (DCIGN)**

[Kulkarni, 2015]

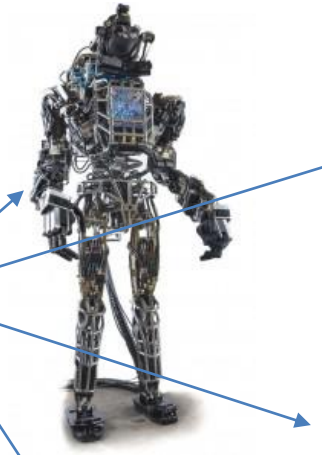
# System-on-Chip

- **Multi/Many cores integrated on a single chip:**
  - Technology scaling
  - 3D integration
  - Many examples
    - STMicro
    - picoChip
    - Tilera Tile GX, Tile Pro
    - Intel Polari , …
- **Complex apps, strict constraints**
  - Massively parallel applications
    - Big-data Analysis, Pattern Recognition, Deep Learning

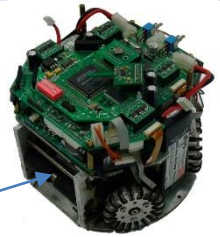


SoC

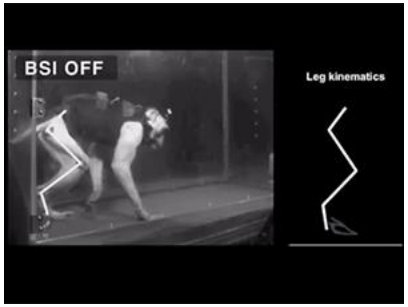
Robots Soldier

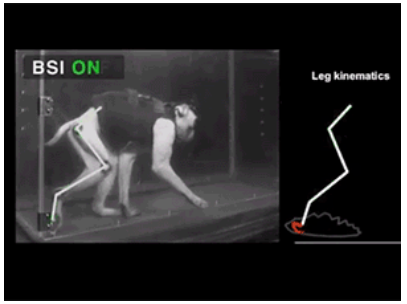


http://www.footnote1.com/robots-soldiers-cyborgs-the-future-of-warfare/atlas_4437-3/

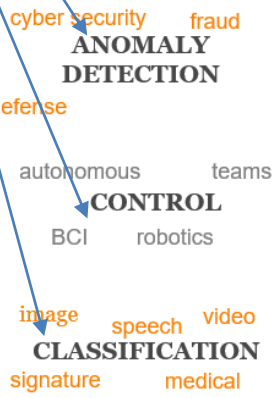


The Soccer Robot based on SoCs.
https://www.intechopen.com/



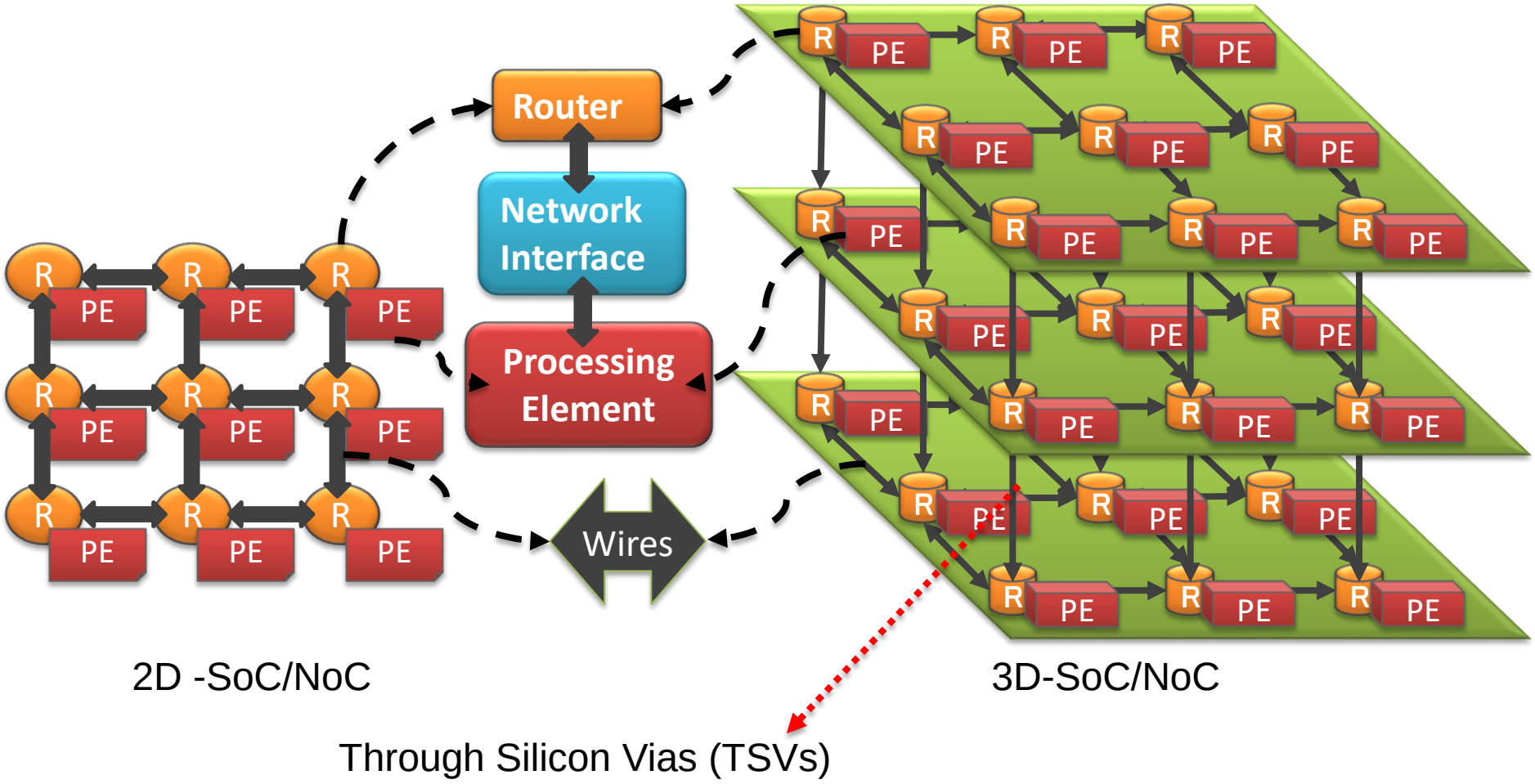

[NATURE, VOL 539, 10 NOVEMBER 2016]



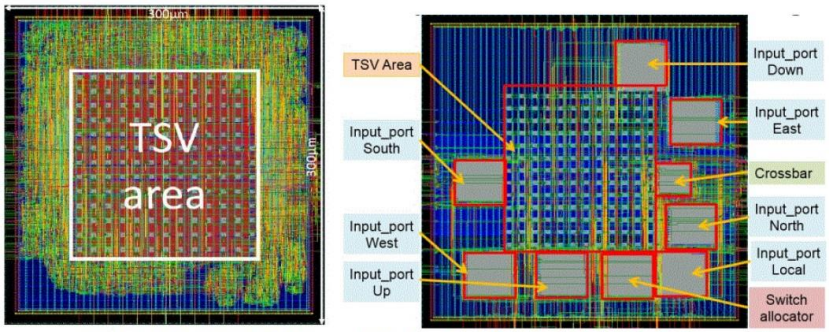

# Adaptive Systems-on-Chip (AsoC)

- **Compute** (High-performance, Real-Time)
- **Adapt** (needs to run in dynamic environments, where physical context, network topologies, and workloads are always changing)
- **Learn (No programming!)**

# ASoC - Approach I:
# Scalable Packet-Switched SoC/NoC



# ASL Approach I: Scalable Packet-Switched SoC/NoC



OASIS-3DFTRV1 (Right) and OASIS-3DRV1 (left) Test-Chip Layouts

- Technology: 45nm CMOS Process
- Chip Size: 2.205X2.220 (micron)
- Frequency: 0.91 GHZ Confirmed
- Supply voltage: 1.1V
- Power Dissipation: 222.387 uW
- Number of Pins: 557

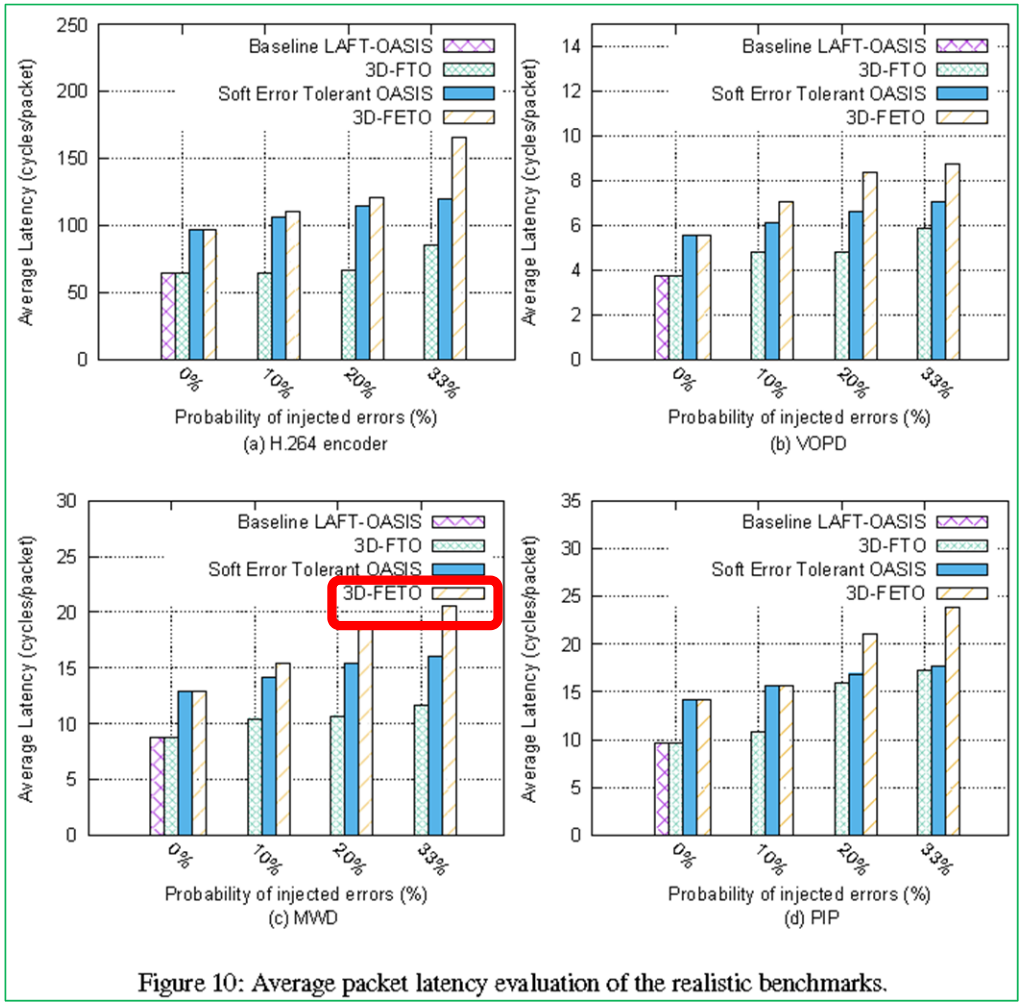
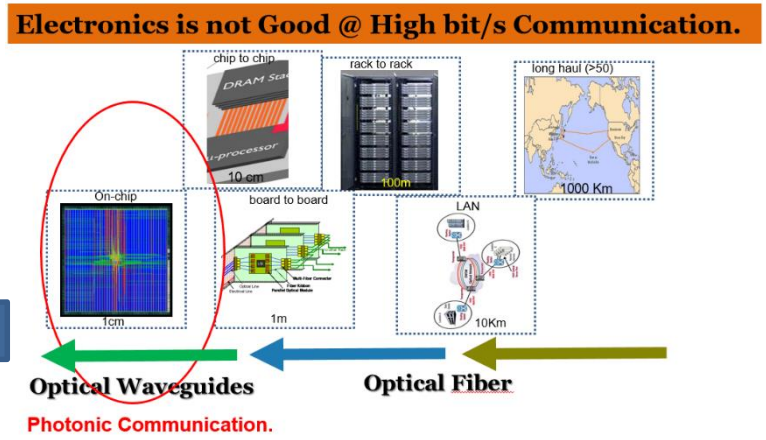
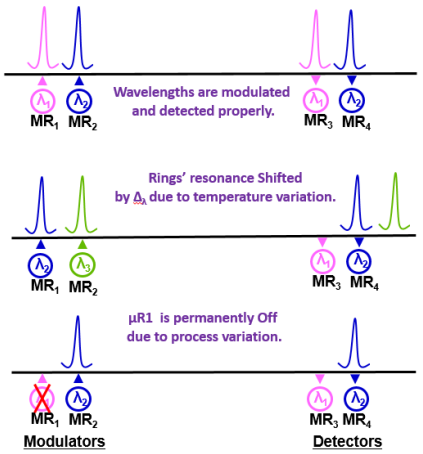


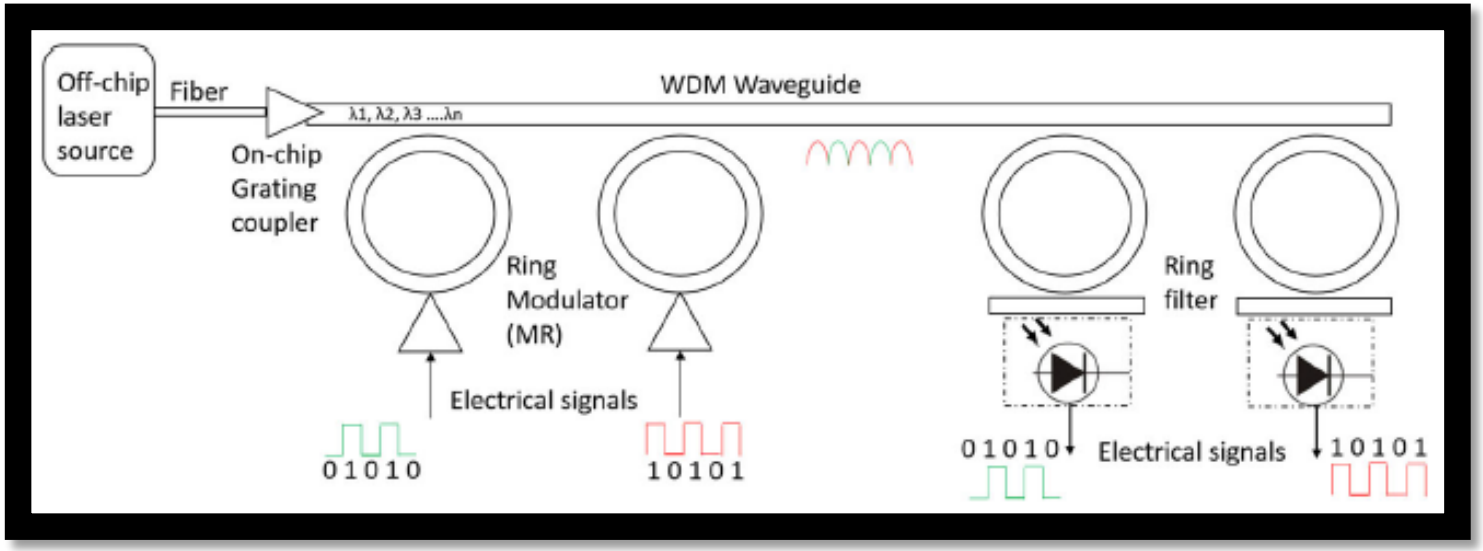
Figure 10: Average packet latency evaluation of the realistic benchmarks.

# ASoC - Approach II: Photonic Packet-Switched SoC/NoC

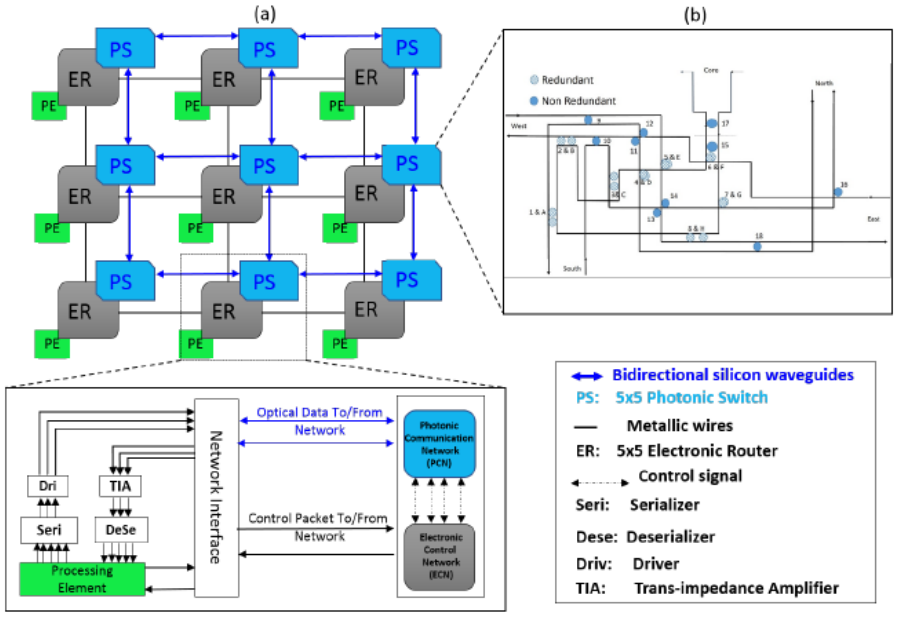



*Photonic Link affected by process and temperature variations.*



# ASoC - Approach II: Photonic Packet-Switched SoC/NoC



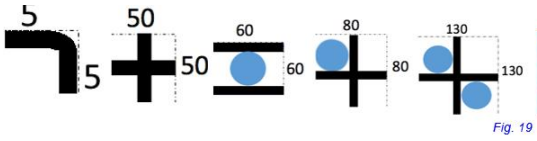




Fig. 19 Parallel Redundant MRs

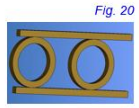


Fig. 20 Perpendicular Redundant MRs

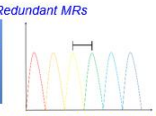
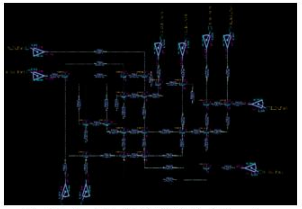


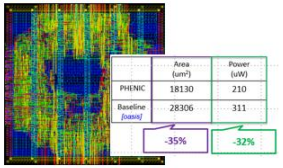
Fig. 21 Wavelength Spacing

$$LOSS_{router} = Loss_{bend} \times N_{bend} + Loss_{cross} \times N_{cross} + Loss_{MR_{OFF}} \times N_{MR_{OFF}} + Loss_{MR_{ON}} \times N_{MR_{ON}}$$

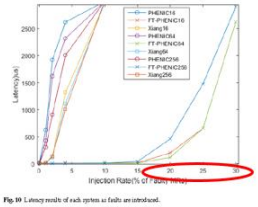
$$Delay_{router} = Delay_{bend} \times N_{bend} + Delay_{cross} \times N_{cross} + Delay_{MR_{OFF}} \times N_{MR_{OFF}} + Delay_{MR_{ON}} \times N_{MR_{ON}}$$



Model of a 5 Port FTTDOR Switch and a Wavelength Shifting Controller




PHENIC's electronic controller layout in 45 nm process



(a) 3x3 Mesh-based System, (b) 5x5 non-blocking photonic switch, (c) Unified Title.

- Micoring fault-tolerant
- Can support enormous intrinsic data bandwidths.
- Immune to the electrical interference.
- Bit-rate transparency
- Transmission latency is very small – Depends on the group velocity of light

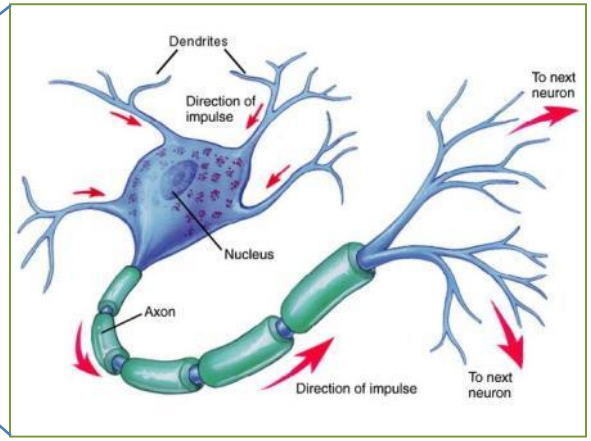

# ASoC - Approach III: Neuro-inspired Manycore SoC

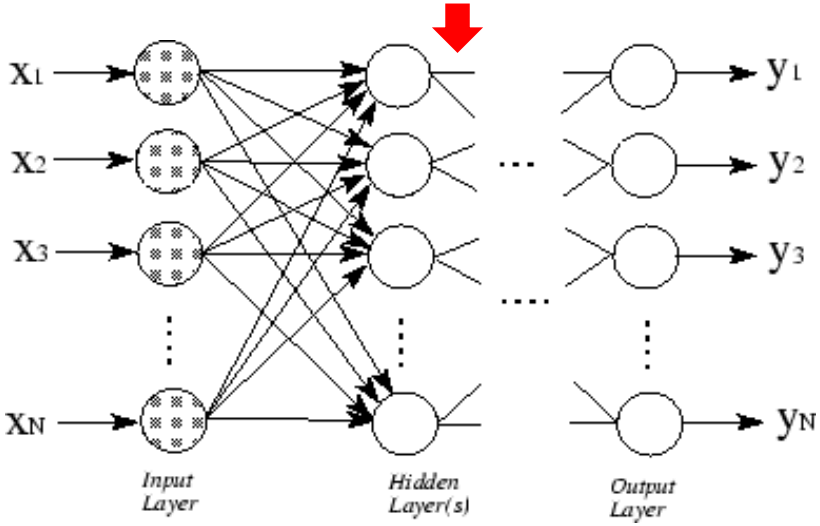
- Neurons accumulating charge and firing over periods of time.
- Different from traditional ANN models, which are static organizations of mathematic formulas.
- Adaptive, spike-based (time dependent) and its structure can change over time:
  - Neurons and synapses may be created or destroyed over time.

# The return to the neuro-inspired computing – Why Now?

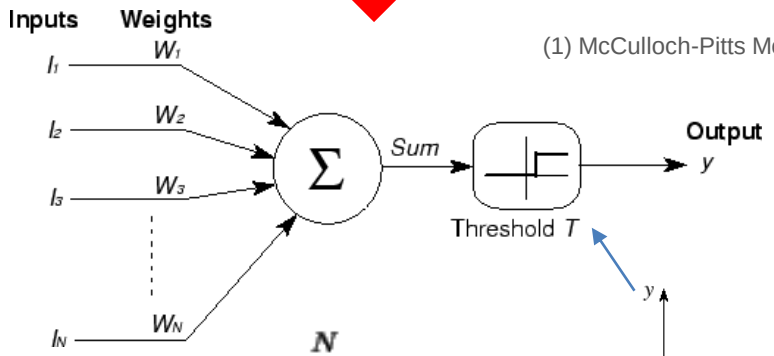
- Huge progress in IT technologies
  - > Number of PCs, cell phones doubles every 5 years
- Emergence of nano-devices
- Better understanding of neural functions
- IC power consumption is reaching its limit
- Brain is extremely energy efficient
  - > ~$10^{-16}$ j/op/s vs. $10^{-16}$ j/op/s for the best computers today [NCL J. A. Bullinaria 2015]

# Deep Neural Network for DL
## Parallel and Deep

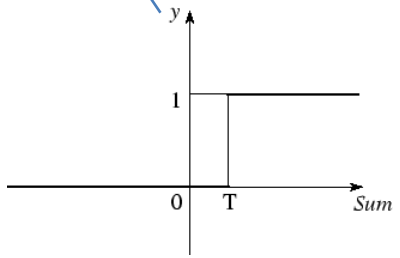




(1) McCulloch-Pitts Model, 1943

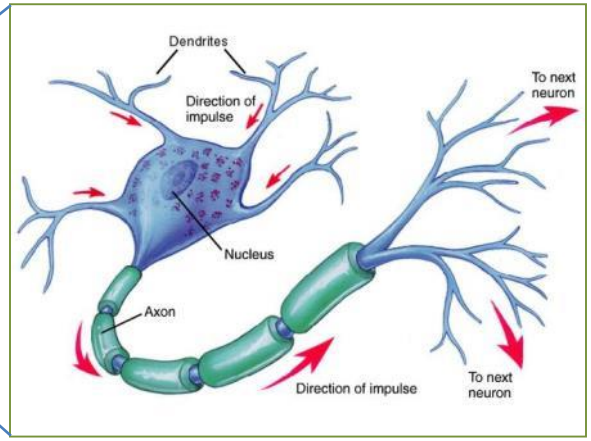


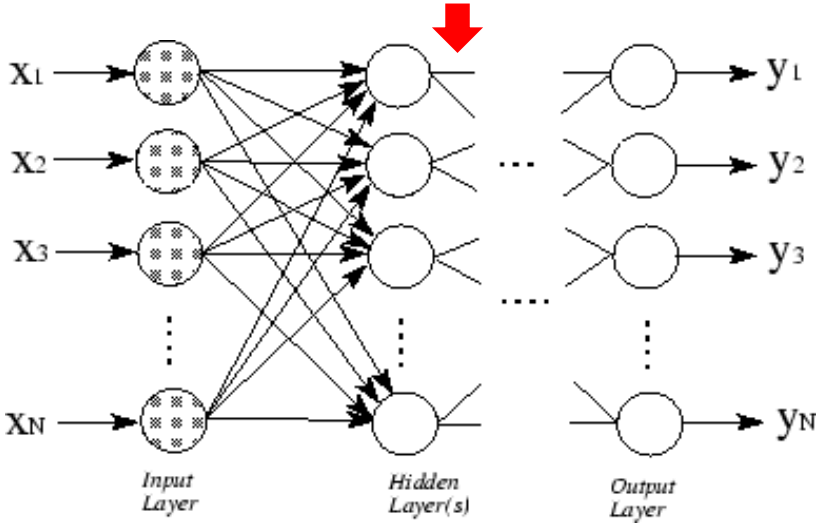

$$Sum = \sum_{i=1}^{N} I_i W_i,$$

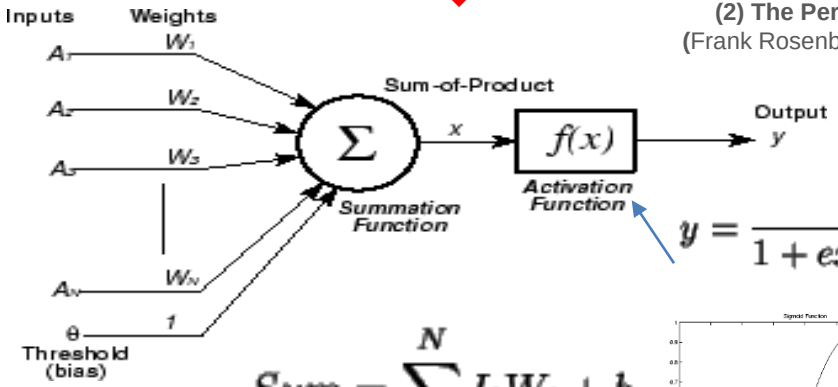
$$y = f(Sum).$$



# Deep Neural Network for DL
## Parallel and Deep

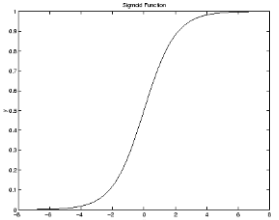




**(2) The Perceptron**
(Frank Rosenblatt, 1950s)

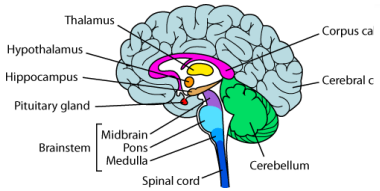
$$y = \frac{1}{1 + exp(-x)}.$$

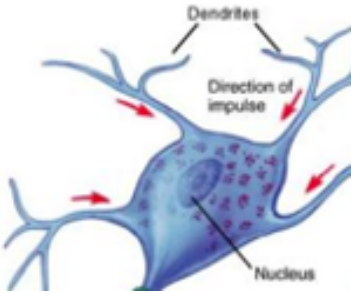
$$Sum = \sum_{i=1}^{N} I_i W_i + b,$$

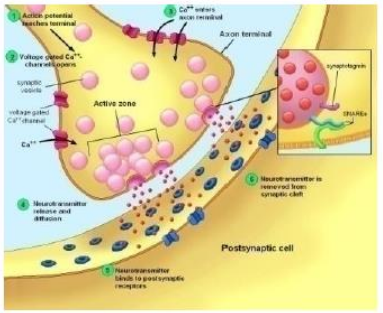


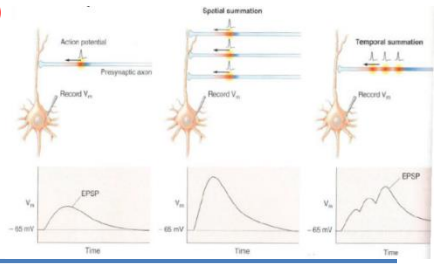
# Brain Features

- Ten billion ($10^{10}$) neurons
- Neuron switching time $>10^{-3}$secs
- On average, each neuron has several thousand connections
- Hundreds of operations per second
- Face Recognition ~0.1secs
- High degree of parallel computation, Distributed representations
- Each neuron is connected to the others through 10000 synapses
- It can learn, reorganize itself → adaptive



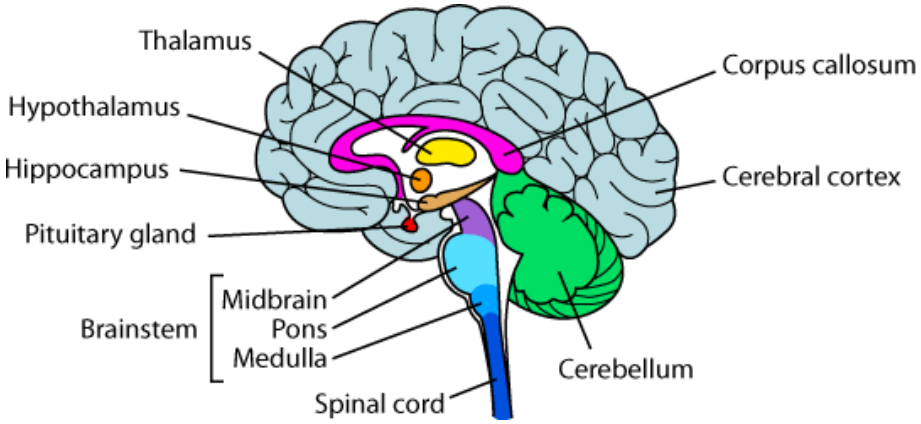



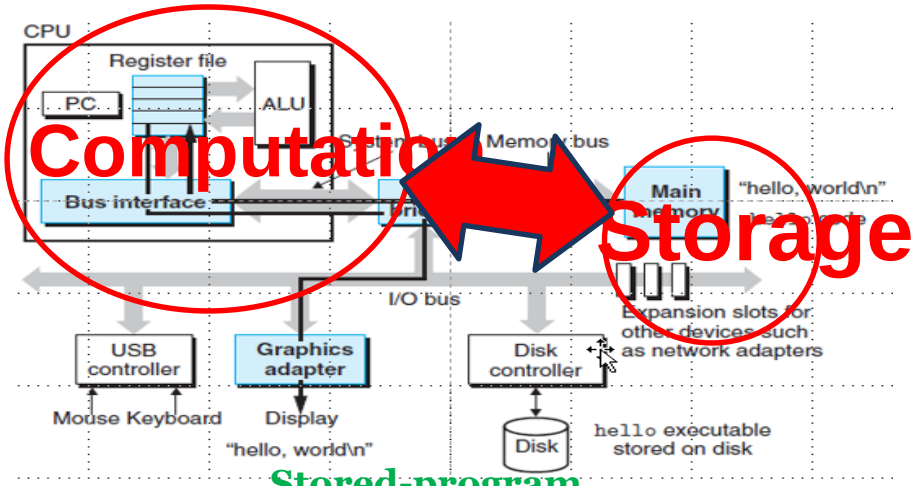



# Computer Vs. Brain

- **Computer**
  - Calculation
  - Precision
  - Logic
- **Brain**
  - Pattern Recognition
  - Noise Tolerance
  - Complexity

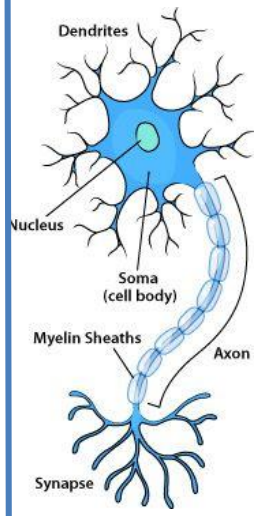


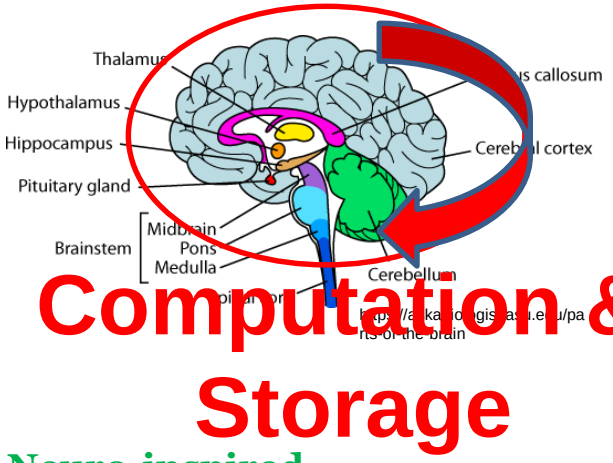
# Computer Vs. Brain




**Stored-program**

- High Power
- Storage and computation are separated
- Poor at recognition



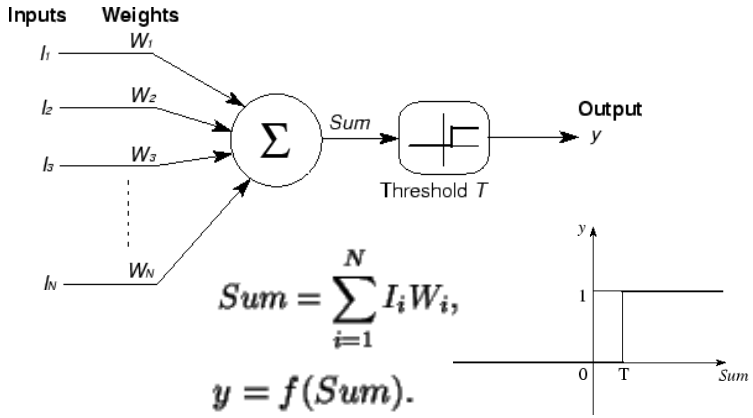




**Neuro-inspired**

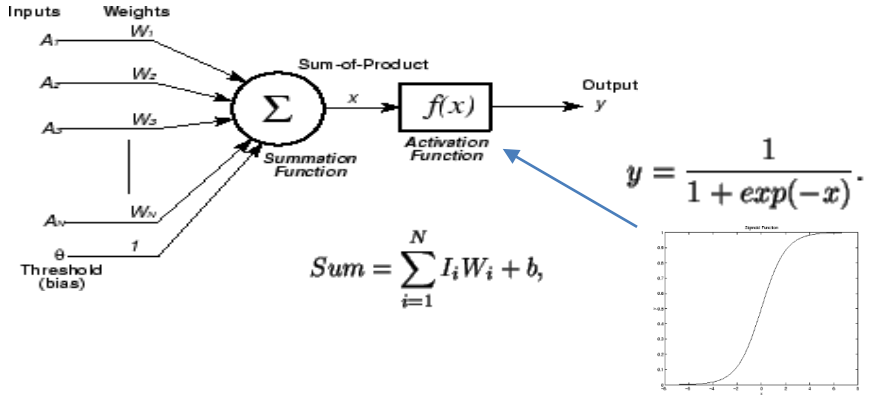
- Low Power
- Storage and computation are not separated
- Good at recognition

# Simplistic Neural Network Models

## (1) Linear threshold gate
(McCulloch-Pitts Model, 1943)




## (2) The Perceptron
(Frank Rosenblatt, 1950s)

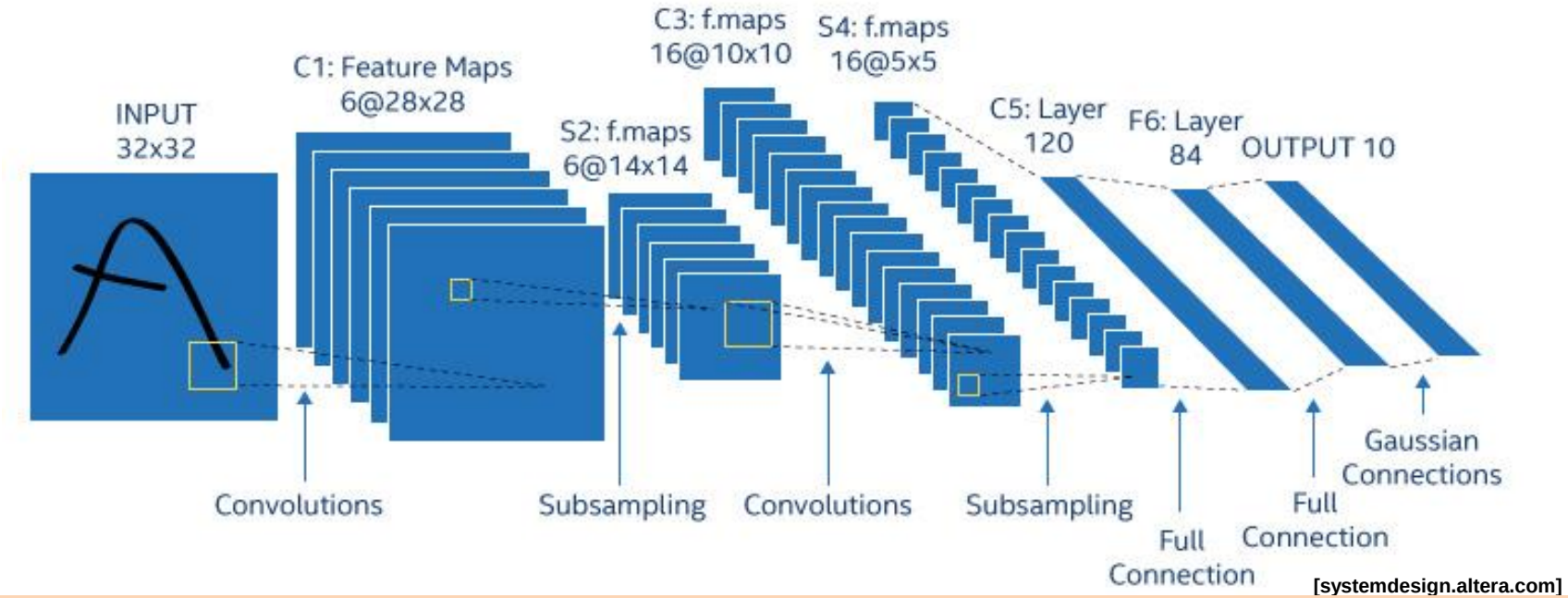



W1,W2…Wm are weights normalized in (0,1) or (-1,1), *Sum* is the weighted sum, and $T$ is a threshold constant, the function $f$ is a linear step function at threshold $T$.

**This model is so simplistic that it only generates a binary output and also the weight and threshold values are fixed.**

- Merging between McCulloch-Pitts model and **Hebbian learning** rule of adjusting weights.
- In addition to the variable weight values, the perceptron model added an extra input that represents **bias.**

# Deep Neural Network for DL
## Parallel and Deep

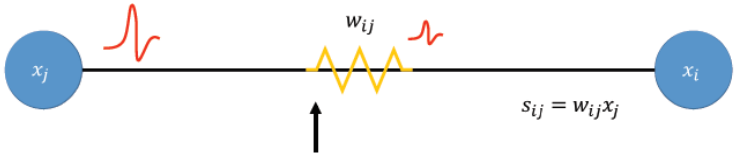


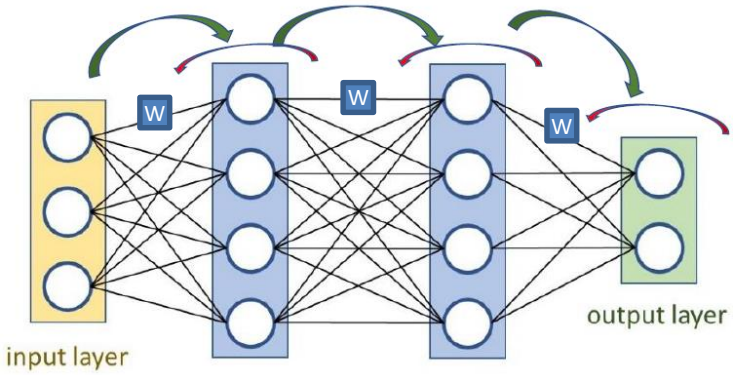

[systemdesign.altera.com]

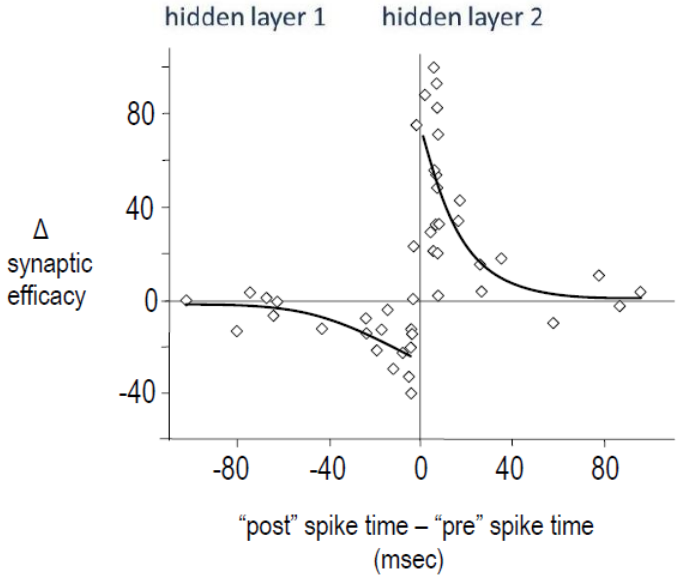
- DNN/DL systems have been highly successful in the areas of image classification and customer preference determination.
- They are not designed for applications that are time-dependent/dynamic, which is our focus.

# Spike Time Dependent Plasticity

- Large number of highly interconnected neurons with **small local memory**, communicating via **spike timing**.
- Connections (synapses) holds knowledge (**weights**)
- Need **training** on examples (supervised/un-supervised learning) to adjusts weights (learning rule)

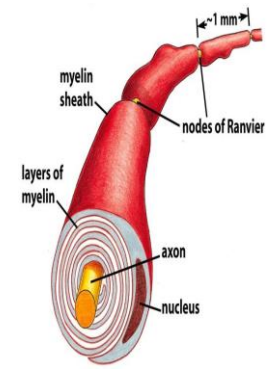


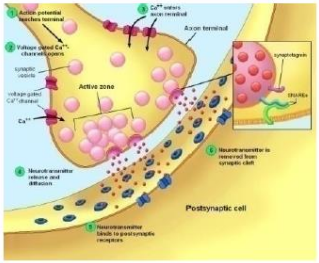



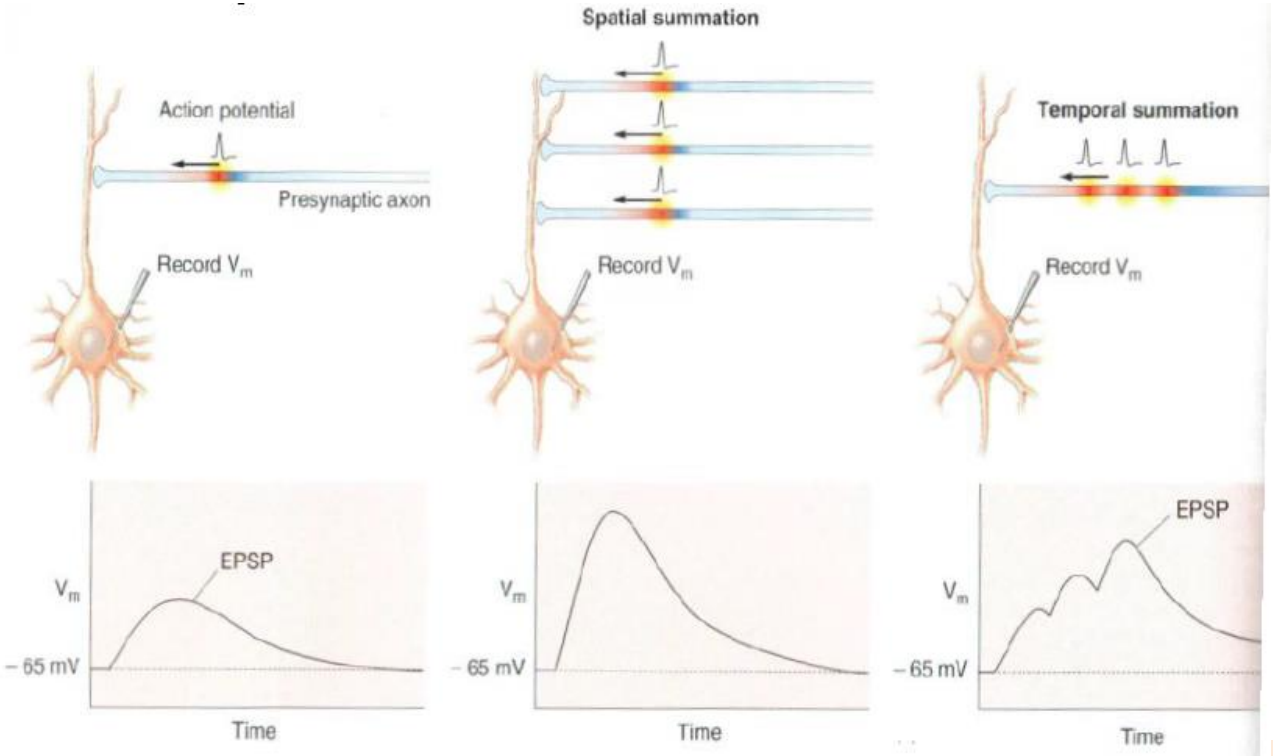




Spike Time Dependent Plasticity

# Short term and long term plasticity for adaptive architecture

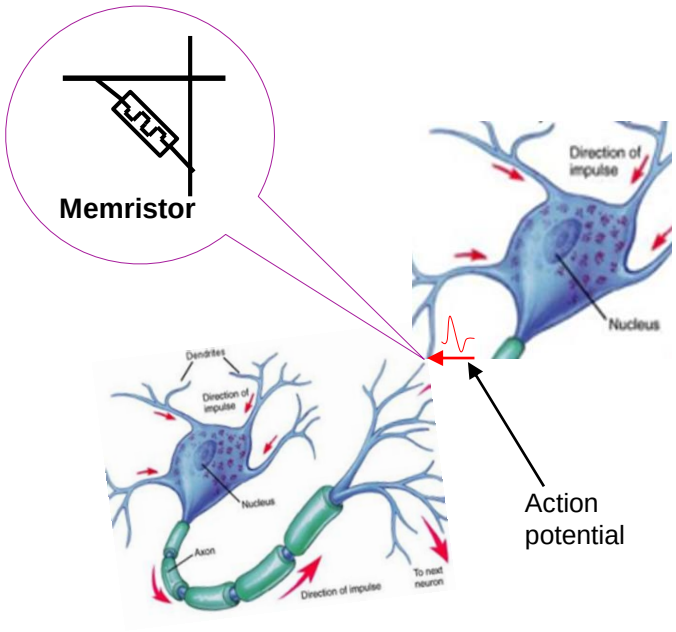


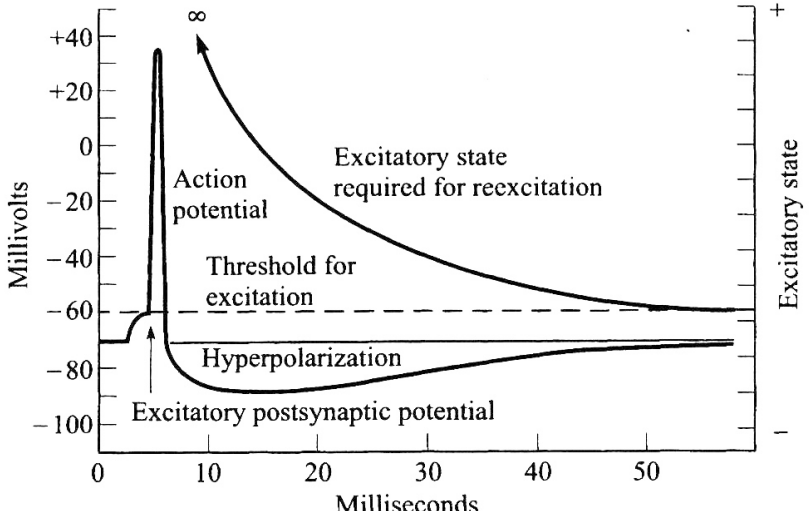



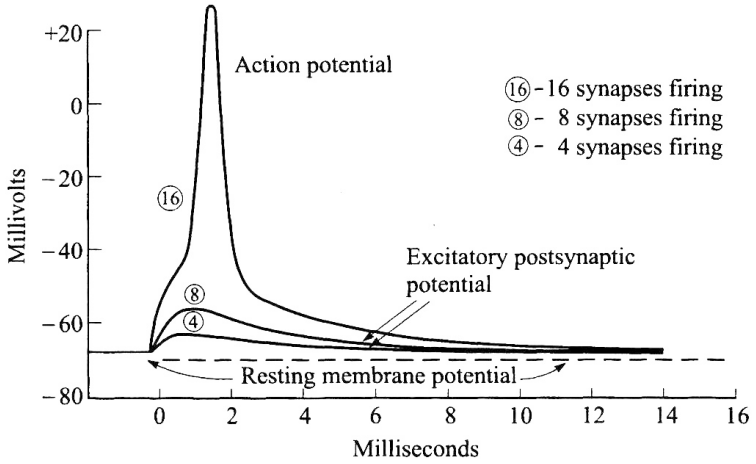




- Time multiplexing → Dynamically reconfigurable interconnect
- Short term and long term Plasticity → Adaptive architecture
- Time/space summation→ Memory/logic operations at interconnect
- Time/space synchronization → Noise/defect tolerant signal processing

# Neuron Action Potential



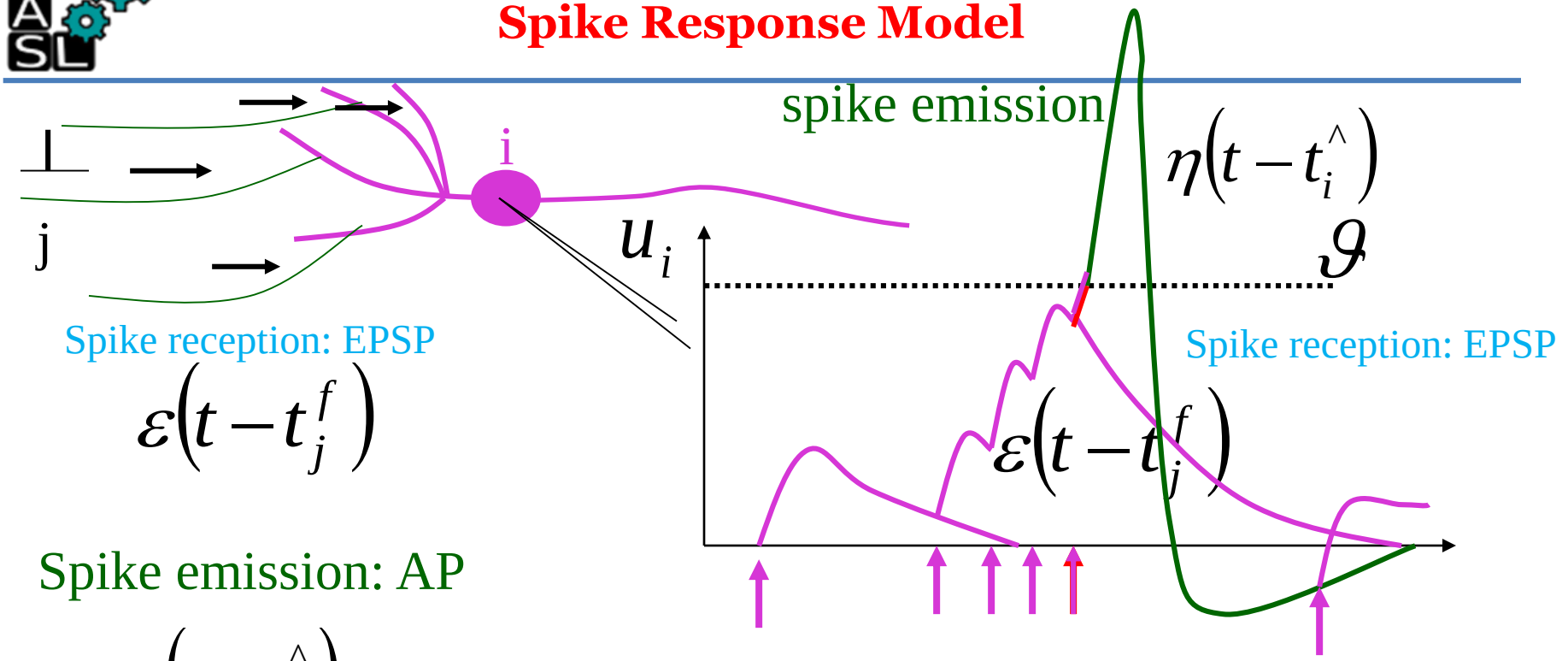




Principles of Neurocomputing for Science & Engineering, Ham & Kostanic, McGraw-Hill, 2001.

# Spiking Neural Networks (SNNs)

## Spike Response Model




Spike emission: AP

$$\eta(t-t_i^{\wedge})$$

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

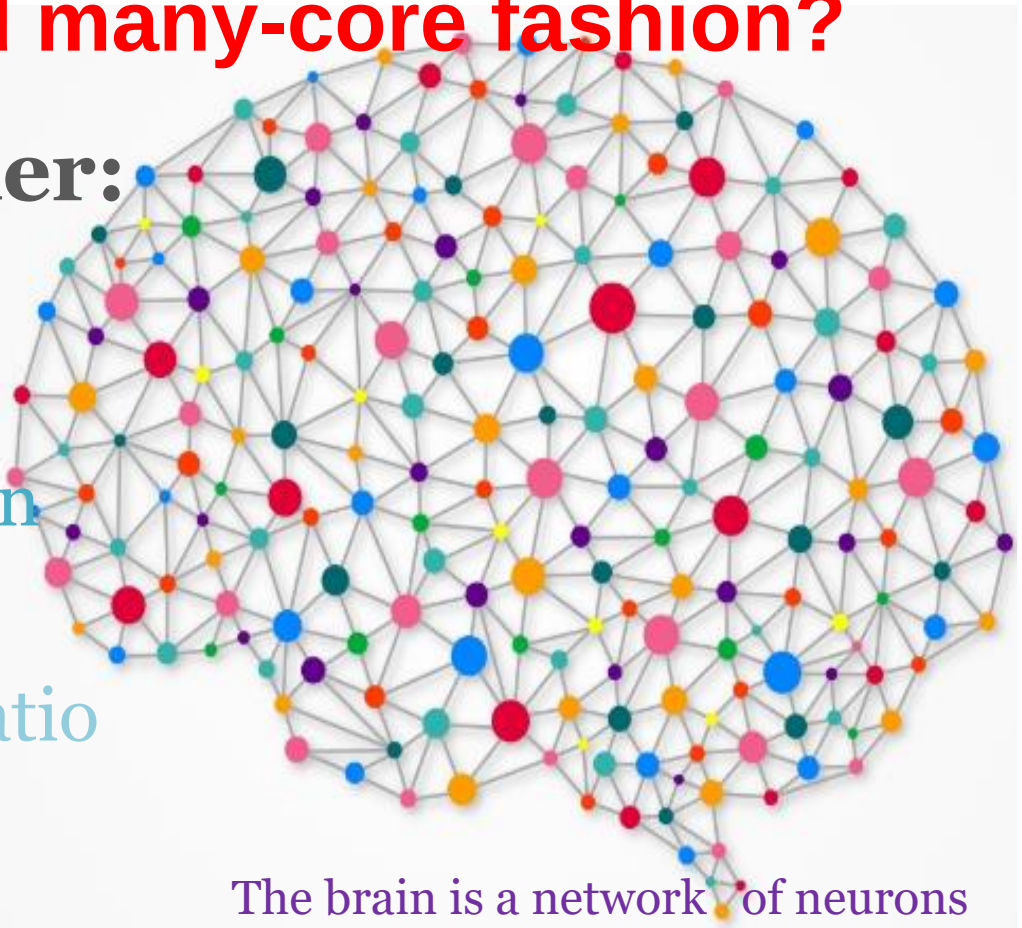
$$u_i(t) = \vartheta \Rightarrow \text{Firing:} \quad t_i^{\wedge} = t$$

# ...but great opportunities come with great challenges

**How do we interconnect a large number of SNs in a networked many-core fashion?**

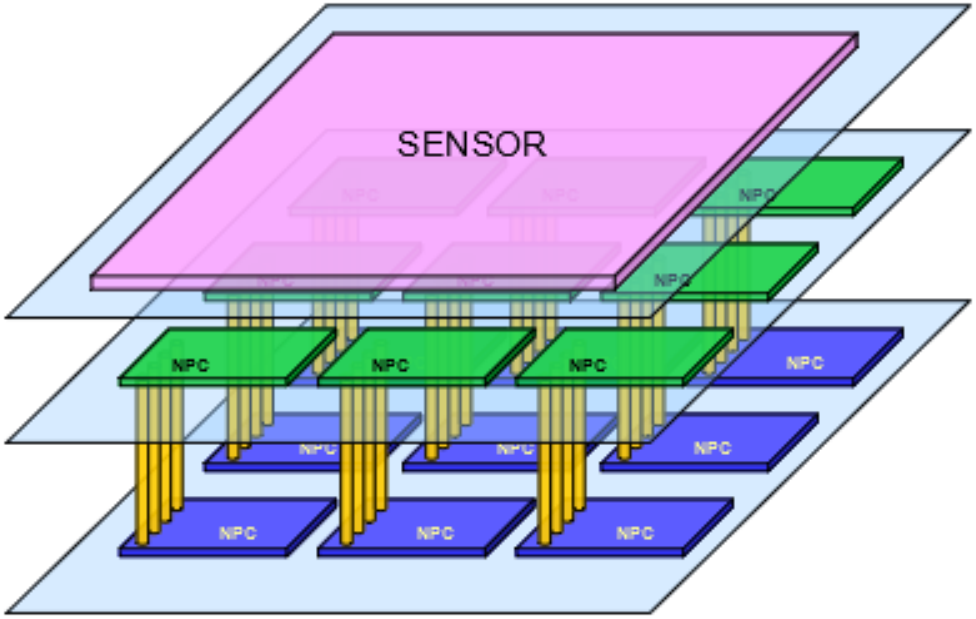
## We should consider:

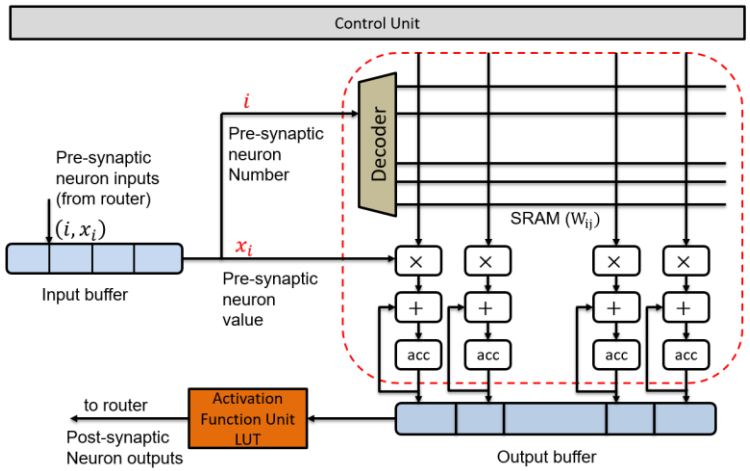
- Scalability
- Area utilisation
- Power consumption
- Throughput
- Synapse/neuron ratio



The brain is a network of neurons

# Neuro-inspired ArchitectureS in Hardware - NASH

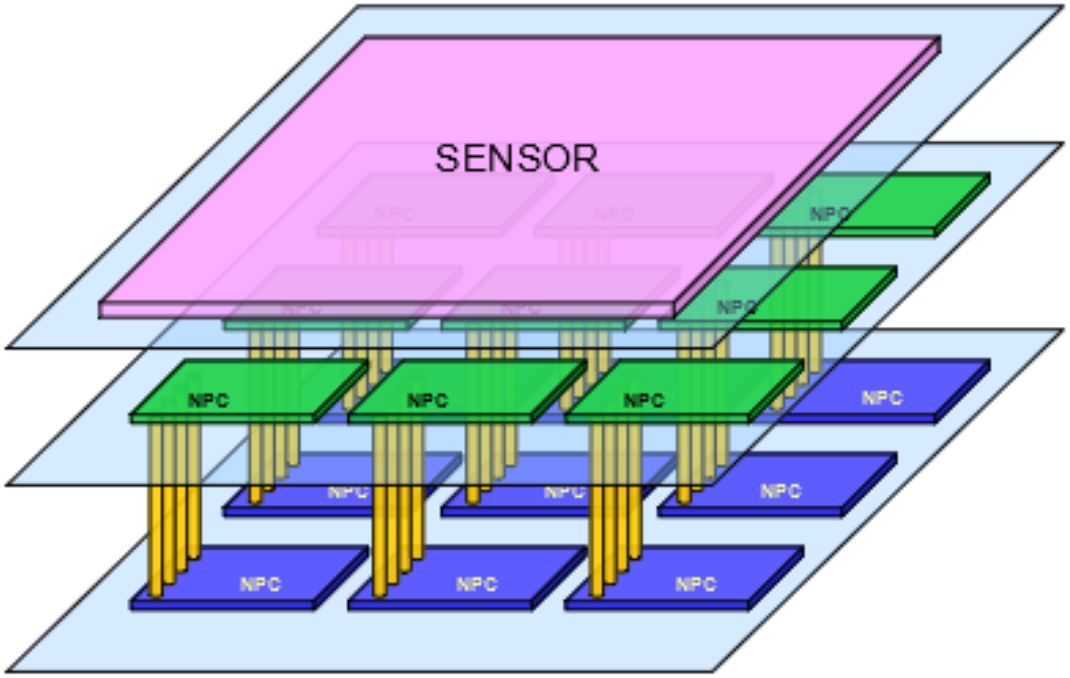


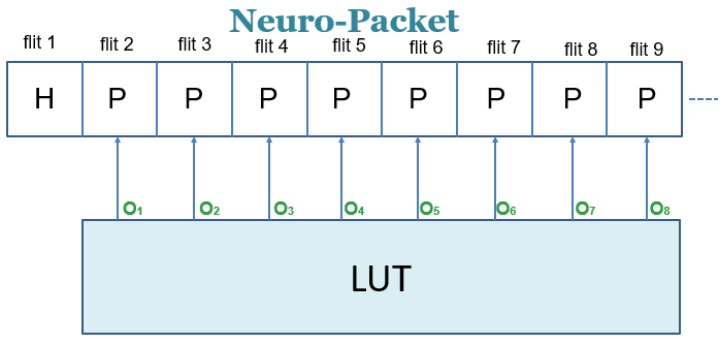

A Single NPC based on SRAM

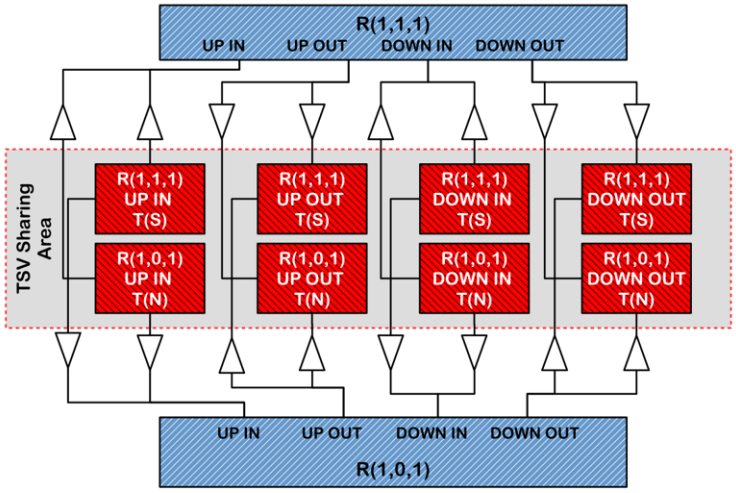
- NASH processes data coming from sensors via 3D-TSV.
- Each neuro core processes a collection of N neurons each with M synaptic weights Wij.

# Neuro-inspired ArchitectureS in Hardware - NASH



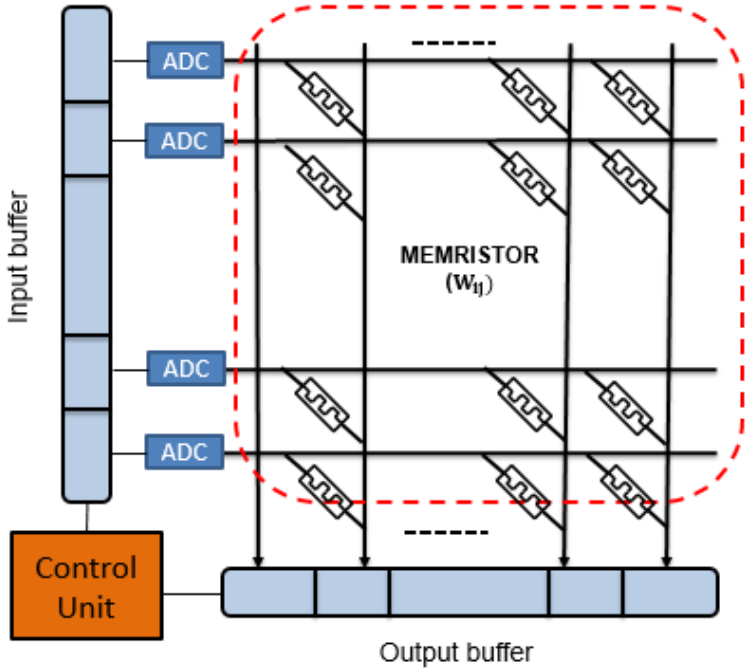

Clustered Neuro-inspired Architecture



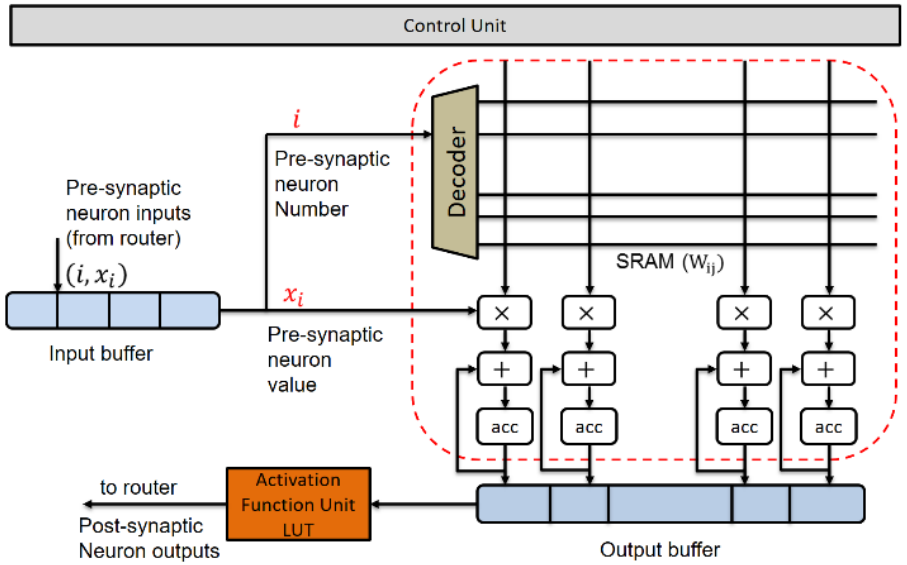




TSV-Sharing Mechanism

# Neuro Processing Core (NPC)

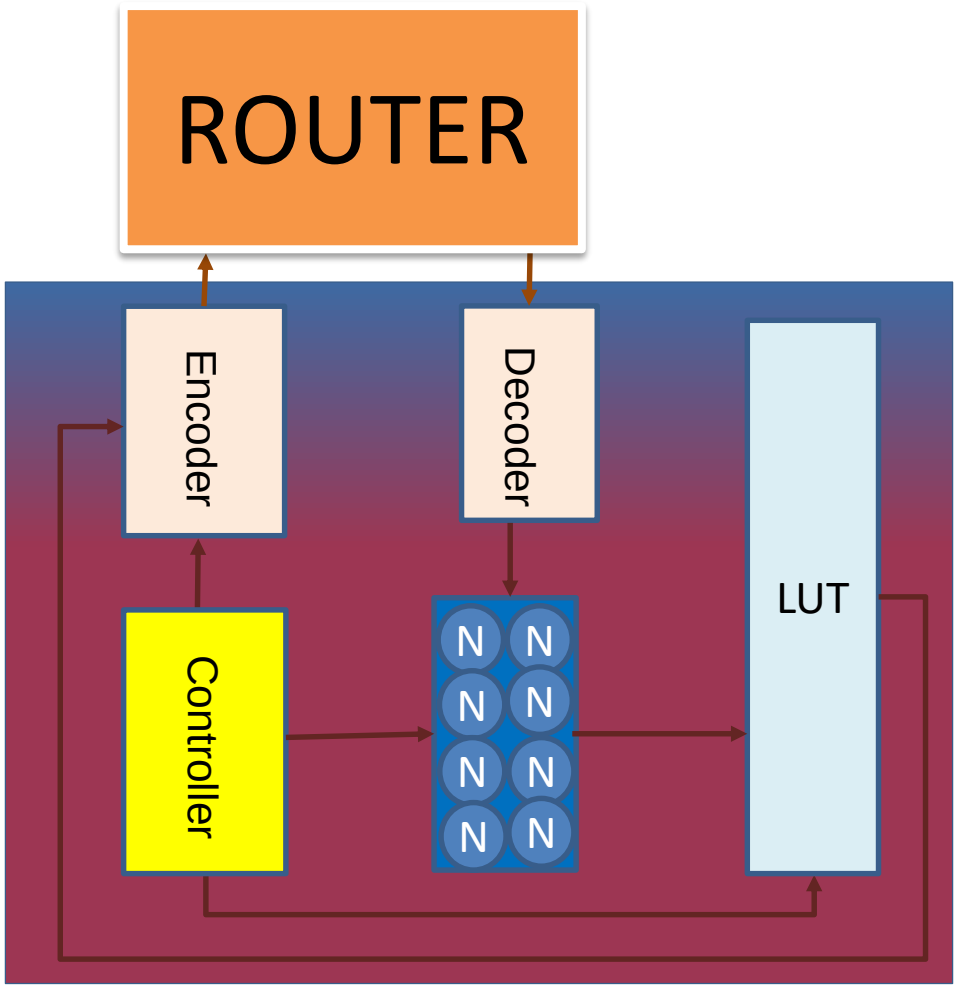



(a) MEMRISTOR based Neuro Processing Core (NPC)

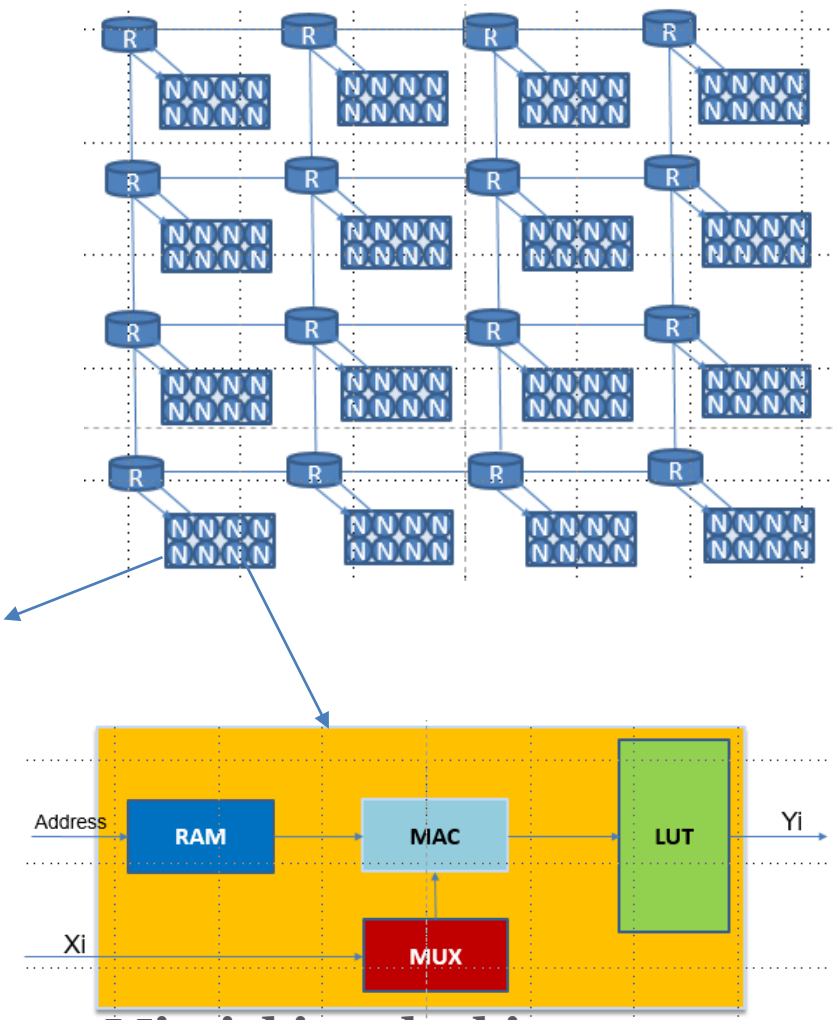



(b) SRAM based Neuro Processing Core (NPC)

# Neuro-inspired ArchitectureS in Hardware - NASH

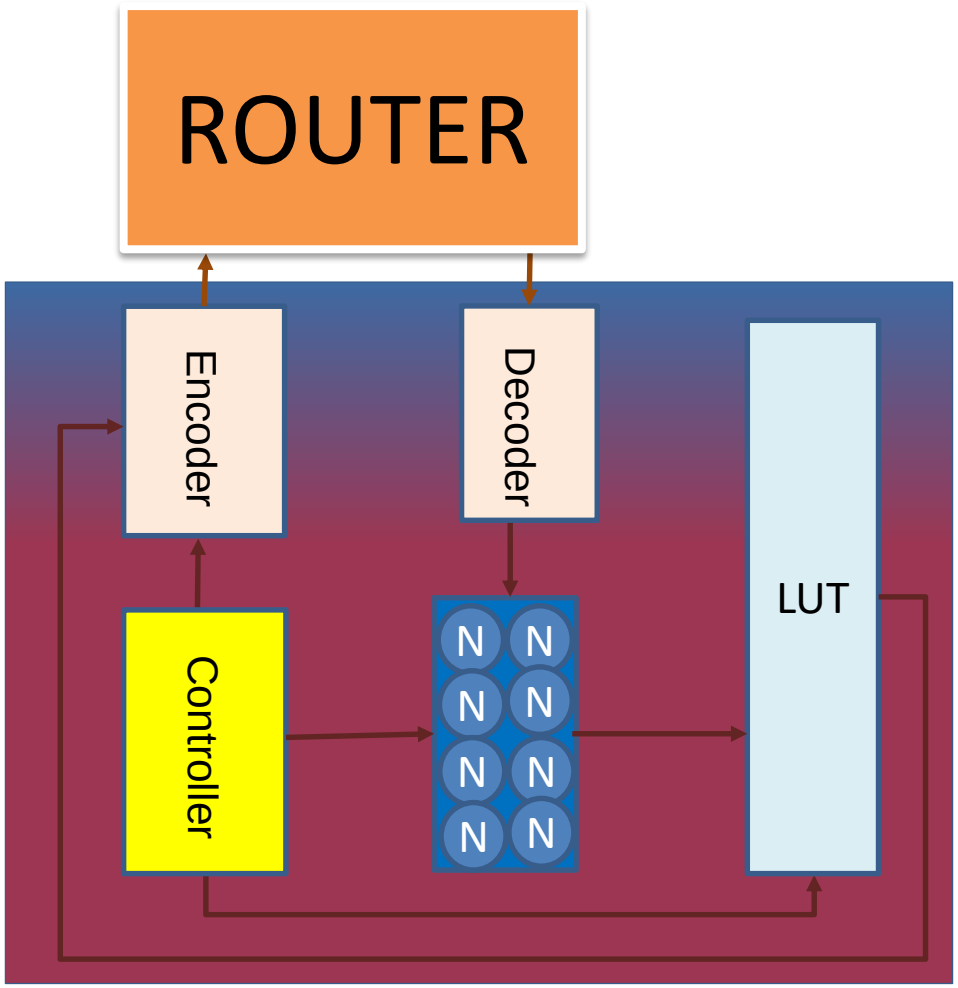



Structure of one PE

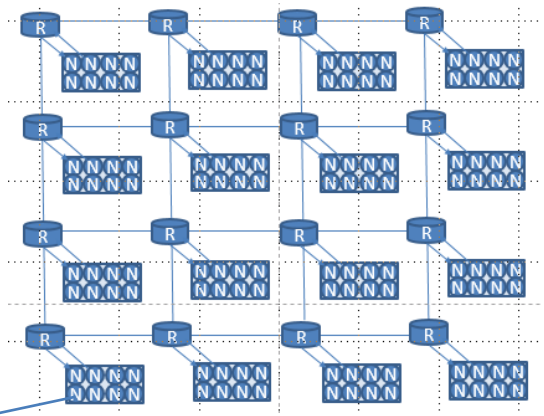


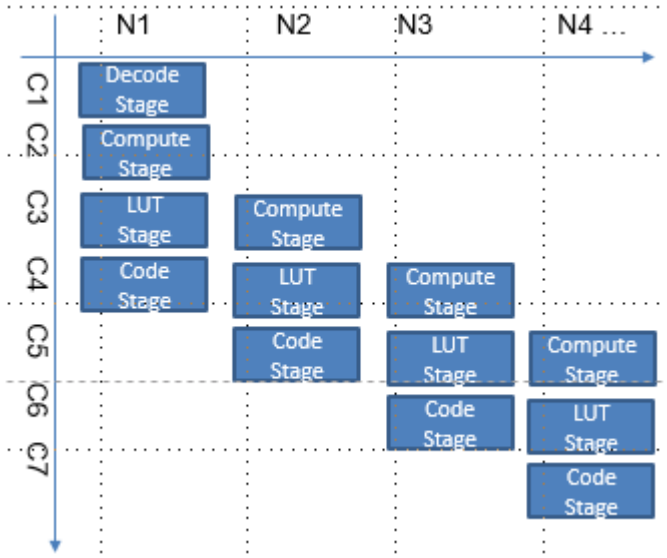

Mimicking the bio-neuron

# Neuro-inspired ArchitectureS in Hardware - NASH




Structure of one PE

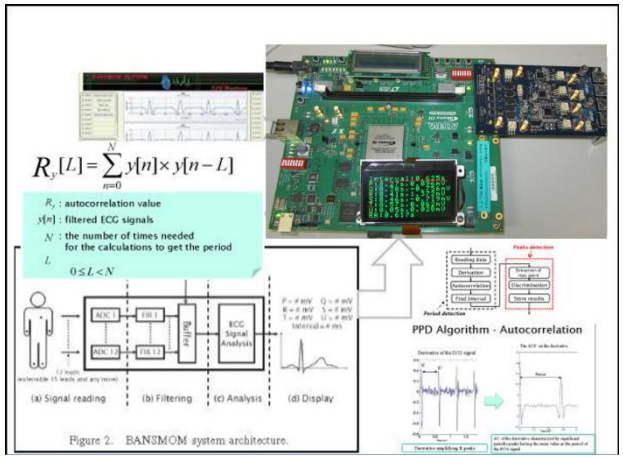


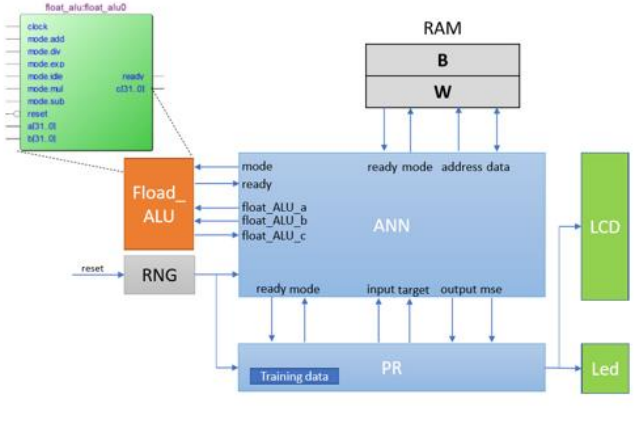

Processing Stages in one PE

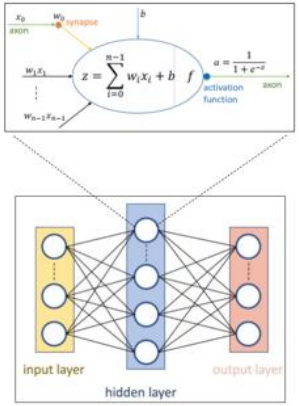
# Some ASoC Applications

Demo 1



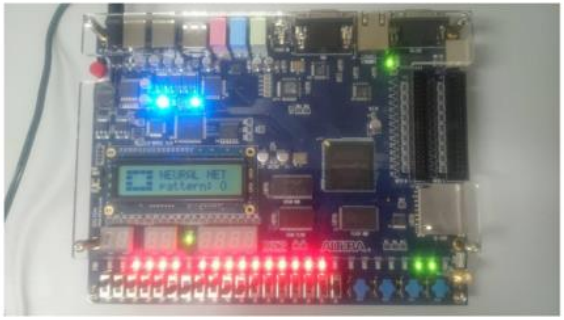

BANSMOM System (Conventional)




| | | | |
|---|---|---|---|
| 1 | 2 | 3 | 4 |
| 5 | 6 | 7 | 8 |
| 9 | 10 | 11 | 12 |
| 13 | 14 | 15 | 16 |

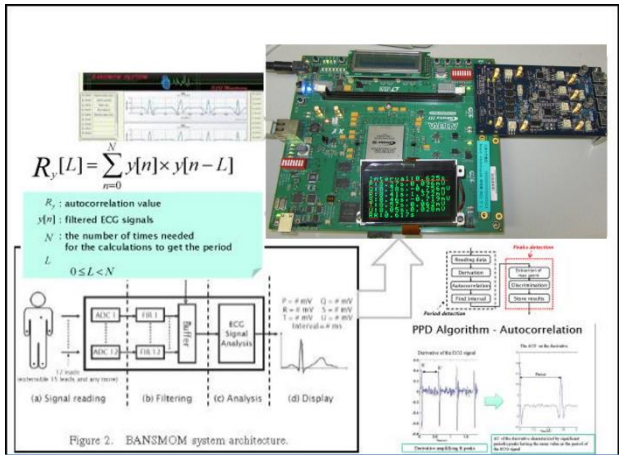
'O' letter



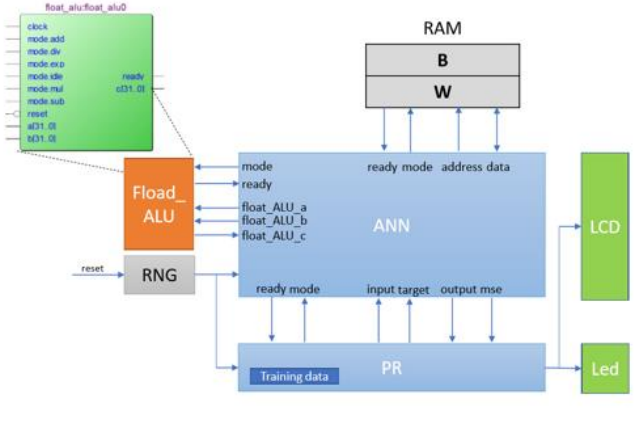
Character Recognition based on BP training

# Some ASoC Applications

Demo 2

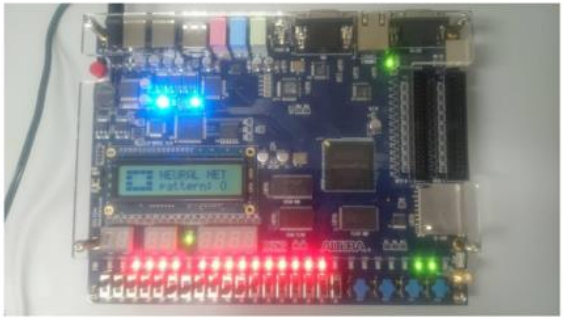



BANSMOM System (Conventional)



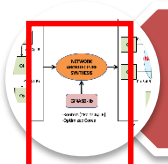
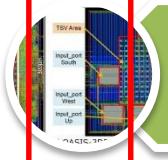
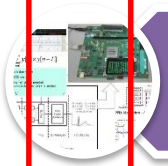
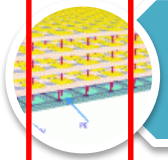

| | | | |
|---|---|---|---|
| 1 | 2 | 3 | 4 |
| 5 | 6 | 7 | 8 |
| 9 | 10 | 11 | 12 |
| 13 | 14 | 15 | 16 |

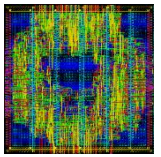
'O' letter

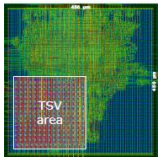


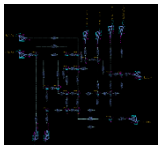
Character Recognition based on BP training

# ASL Adaptive SoCs

| | | | |
|---|---|---|---|
| 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, |
| 2016 |  | Adaptive Neuro-inspired SoC and Platform, NASH | TR-OASIS-NP-042015 |

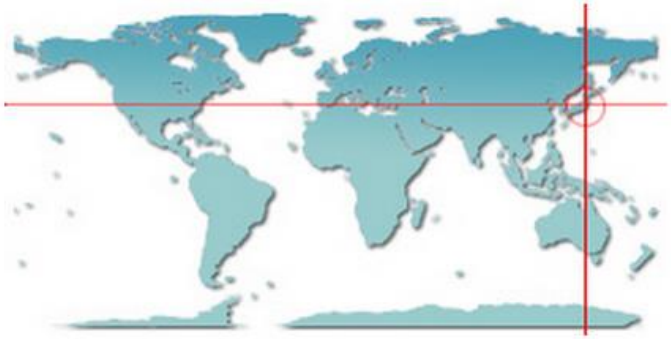


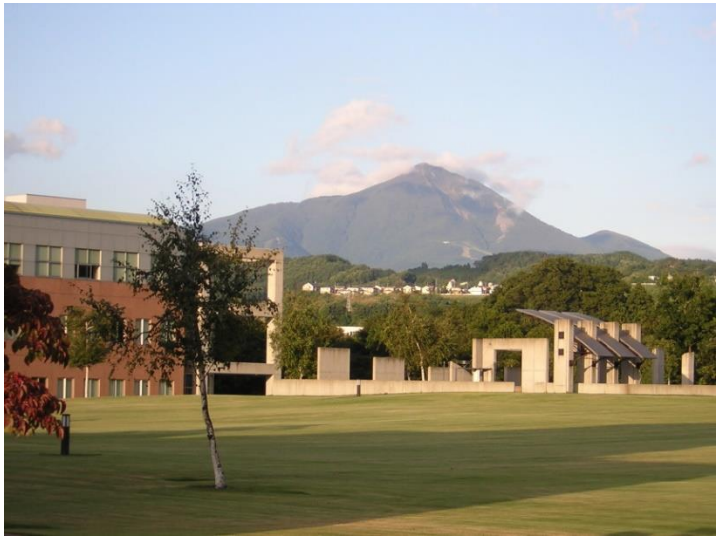


# Thank you!

Abderazek Ben Abdallah
benab@u-aizu.ac.jp



to Advance Knowledge for Humanity