iemn
Institut d'Electronique, de Microélectronique et de Nanotechnologie
UMR CNRS 8520

LEOPAR: Low-Energy On-chip Pre-processing for Activity Recognition

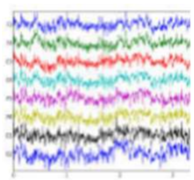
Benoit LARRAS, Kévin HÉRISSÉ, Antoine FRAPPÉ, Bruno STEFANELLI, Andreas KAISER

14/05/2019 – Colloque du GdR BioComp

Project context

- **Massive amounts of data**
- **Always-on sensing**

EEG



Images/Videos

Audio/Speech

To process locally
OR
to transfer to main computation unit (cloud)

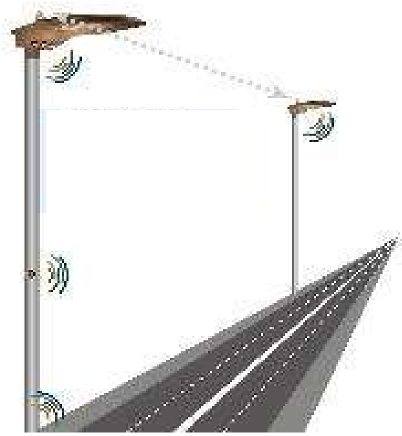
ENERGY HUNGRY

Small, cheap, no battery replacement

→ **Towards Near-Sensor Computing**

Application fields

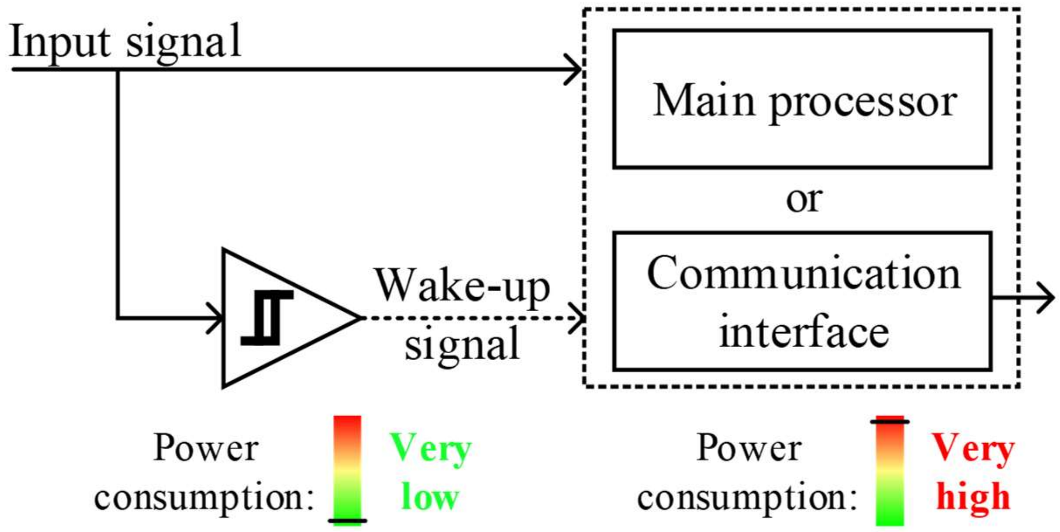
- Audio processing
    - Voice Activity Detection in noisy context
    - Vowels, words, language recognition
    - Specific feature extraction
- Human-body signal classifications
    - ECG, EEG, etc…
- Vibration and movement recognition
- Image processing
    - Motion-triggered cameras
    - Face detection / Owner-activated devices
- Automotive



iemn
Institut d'Electronique, de Microélectronique et de Nanotechnologie
UMR CNRS 8520

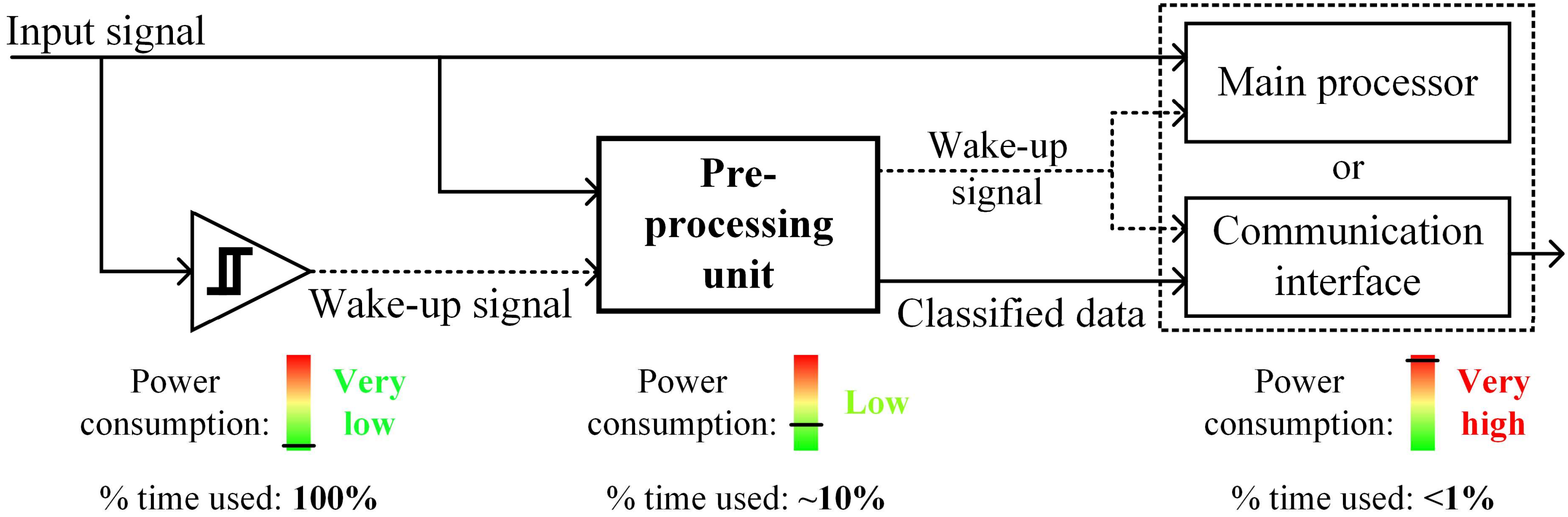
Project objectives

Standard scheme:



Non relevant data is processed if it exceeds the threshold…

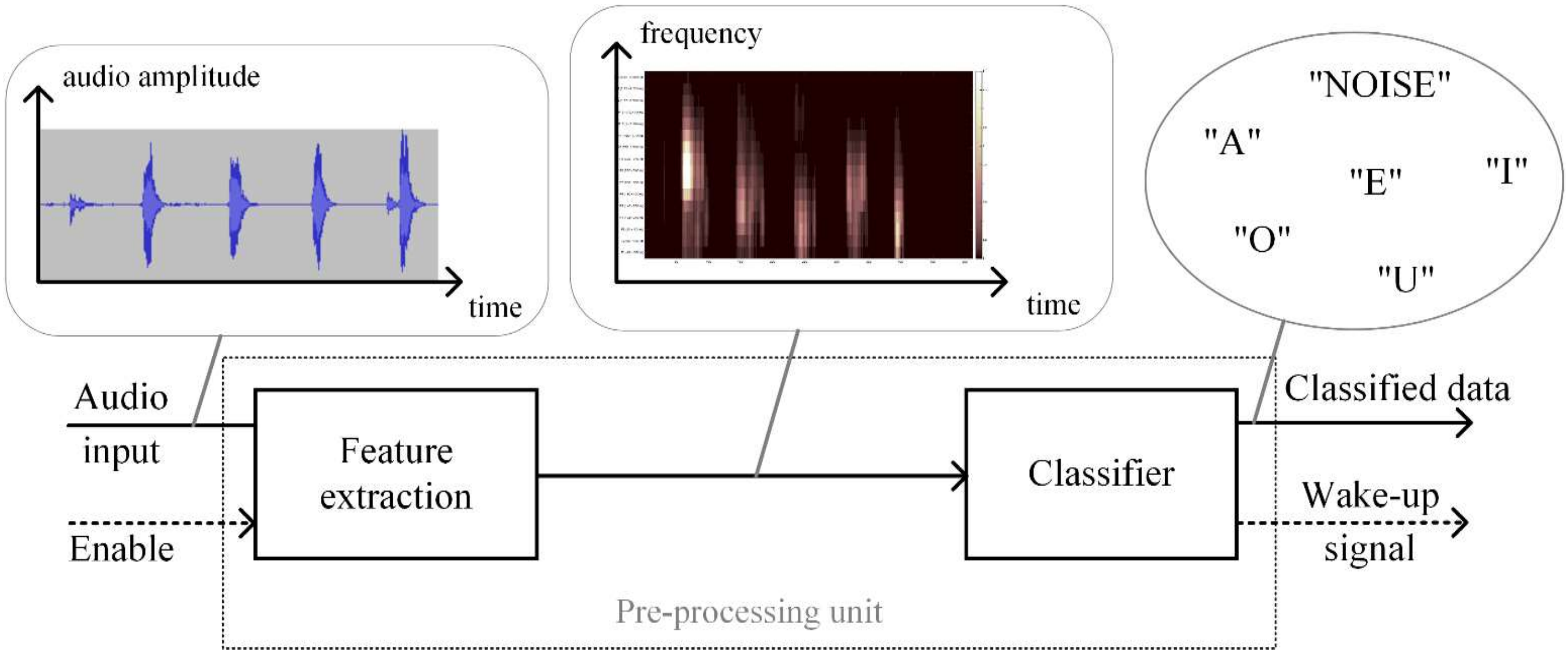
Project objectives



Near-sensor Computing: process **relevant** data **as close as the sensor** as possible

- Aggregation of a **lower amount of data**
- Need of energy-hungry processing during a **lower amount of time**

Envisionned demonstration



- Focus on **audio applications**: voice activity detection, vowels recognition, keyword detection.
- **On-chip event-driven** feature extraction
- **Small-scale neuro-inspired** classification unit

Feature extraction

Objective: **extract energy in different frequency bands**

- Analog filter bank [Badami, JSSC 2016]
  - **Low energy**
  - **Non configurable filters**
  - **High silicon area**

- Digital FFT [Price, JSSC 2017]
  - **Configurability**
  - **Audio fidelity**
  - **Latency**
  - **High complexity**
  - **High energy**

Feature extraction

- **Digital filter bank**
    - **Configurability**
    - **Low latency**
    - **Implementation capability**

Requires **preliminary always-on** A-to-D conversion and signal processing of the complete spectrum

Event-driven / Clockless ?

→ Advantages of both analog and digital implementations

Opportunity: Continuous-Time Digital Signal Processing (CTDSP)

Continuous (CT) ← **Time** → Discrete (DT)

Amplitude: Continuous (CA) ↑ / ↓ Discrete (DA)

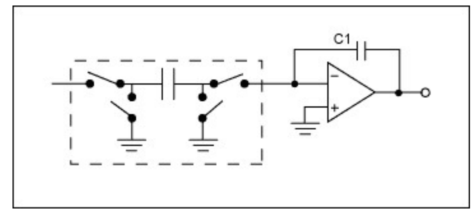
CT CA

Classical analog
Analog filters

DT CA

Sampled analog systems

Ex: Switched-capacitor filters



CT DA

Continuous-Time
Digital Signal Processing
(CT DSP)

DT DA

Digital Signal Processors
Microprocessors

iemn
Institut d'Electronique, de Microélectronique et de Nanotechnologie
UMR CNRS 8520

Opportunity: Continuous-Time Digital Signal Processing (CTDSP)

Continuous (CT) ← **Time** → Discrete (DT)

Amplitude: ↑ Continuous (CA) / ↓ Discrete (DA)

CT CA

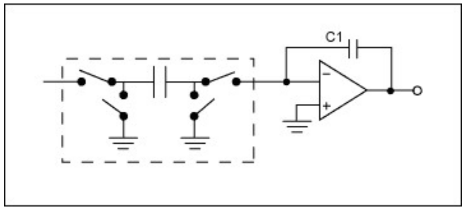
Cl

Analog filters

DT CA

Sampled analog systems

Ex: Switched-capacitor filters



CT DA

Contin   *ime*
Digital Si *al Processing*
(CT DSP)

DT DA

Digital   *essors*
Mic   *sors*

Institut d'Electronique, de Microélectronique et de Nanotechnologie
UMR CNRS 8520

Continuous-Time (CT) advantages

- **Event-driven system**
  - No clock
  - Event-driven power consumption

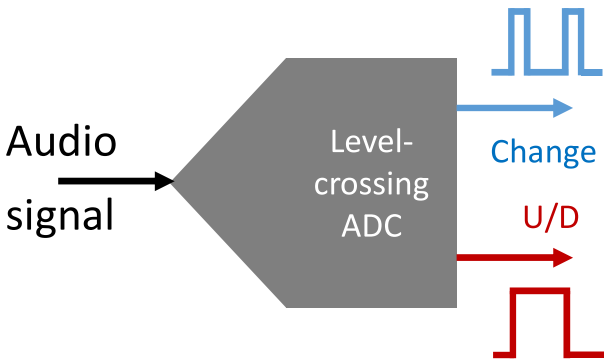
- **CMOS Digital System**
  - Configurability
  - Scalability
  - High integration level

iemn
Institut d'Electronique, de Microélectronique et de Nanotechnologie
UMR CNRS 8520

Continuous-Time (CT) advantages

- **Event-driven system**
  - No clock
  - Event-driven power consumption

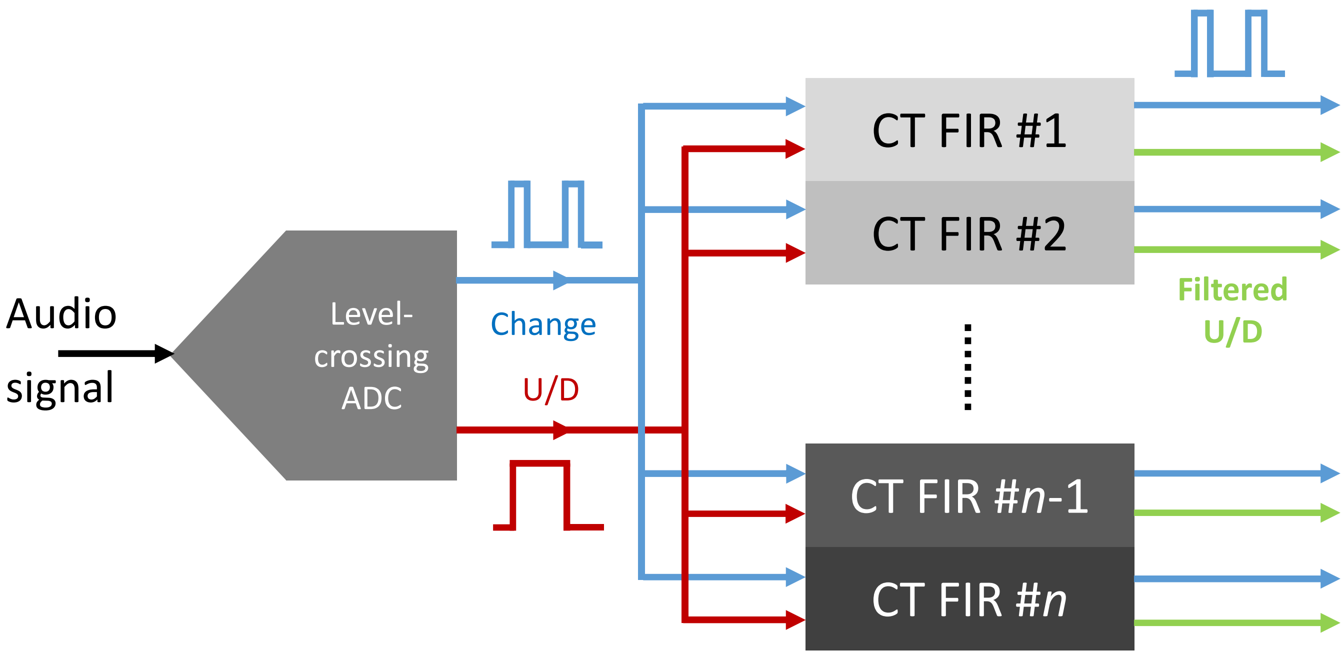
- **CMOS Digital System**
  - Configurability
  - Scalability
  - High integration level



Continuous-Time (CT) advantages

- **Event-driven system**
  - No clock
  - Event-driven power consumption

- **CMOS Digital System**
  - Configurability
  - Scalability
  - High integration level

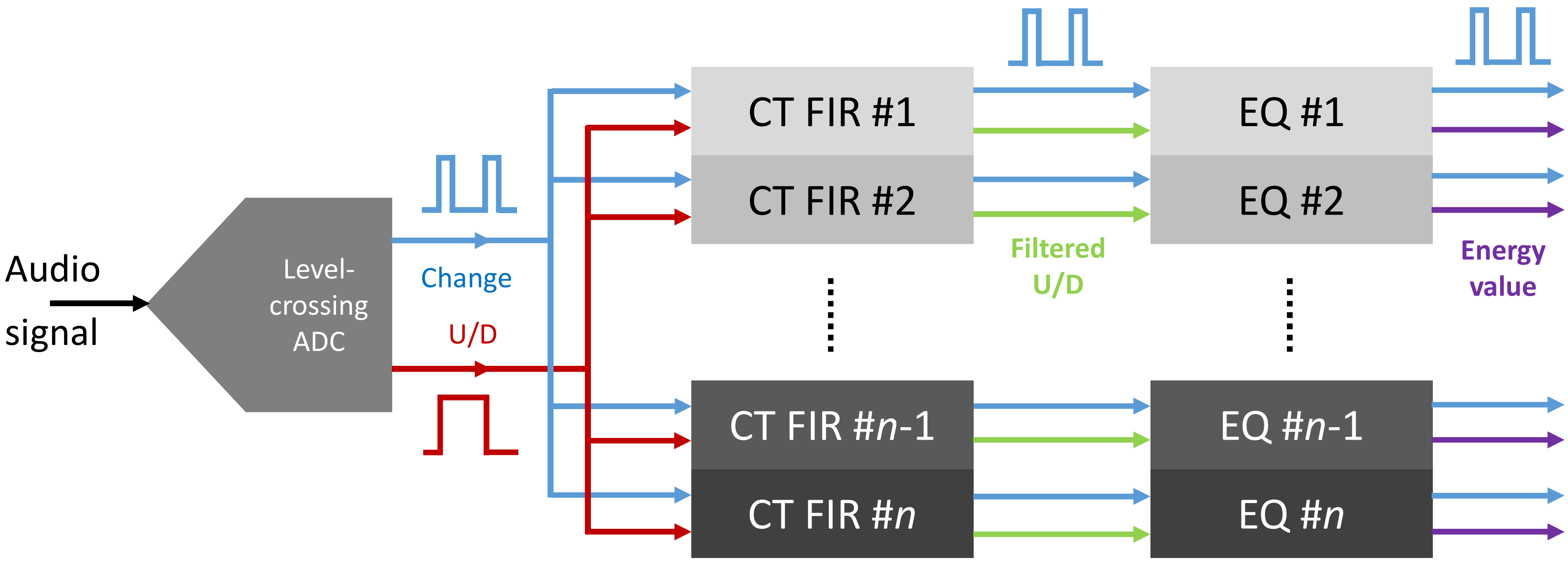


[Kurchuk, JSSC 2012]

Continuous-Time (CT) advantages

- **Event-driven system**
  - No clock
  - Event-driven power consumption

- **CMOS Digital System**
  - Configurability
  - Scalability
  - High integration level



EQ: Energy Quantifier

Classification

- Detection of a **small number of specific patterns**: voice activity, vowels, specific sounds, etc.

- **Limited amount of features** → limited amount of computing units (neurons)

- Embedded environment: **energy and complexity requirements**

→ Towards a **binarized**, **small-scale** classifier with **determined data storage**

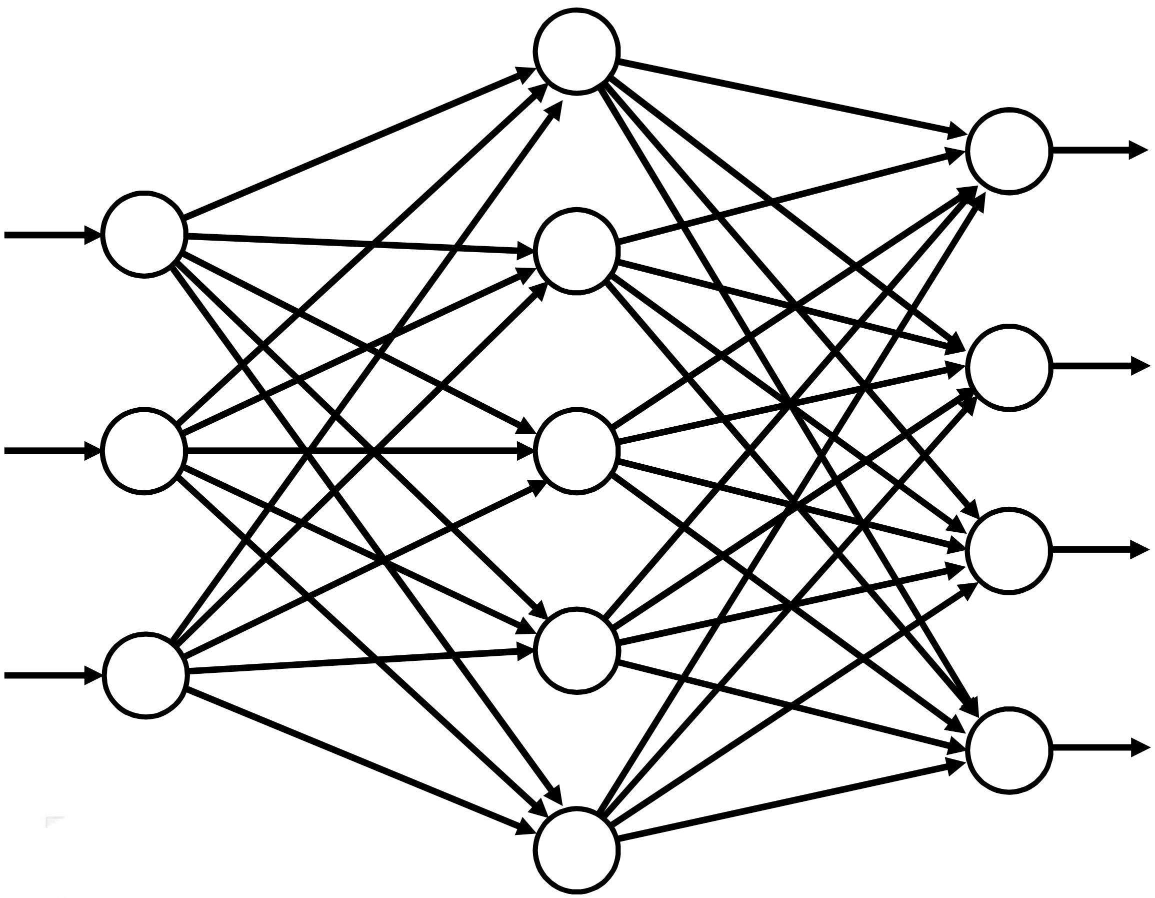
iemn
Institut d'Electronique, de Microélectronique et de Nanotechnologie
UMR CNRS 8520

Opportunity: Small-scale classifiers

- **Only necessary functions implemented**
  - Online inference only, towards binary synaptic weights
  - Activation function: e.g. local Winner-Takes-All
- **Asynchronous behavior →** Event-driven compatible
- **Short reaction time →** Real-time compatible
- **Envisioned classifier models:**
  - LSTM
  - Spiking neural networks
  - Clique-based networks

Neural networks models

Several organizations for the neurons:



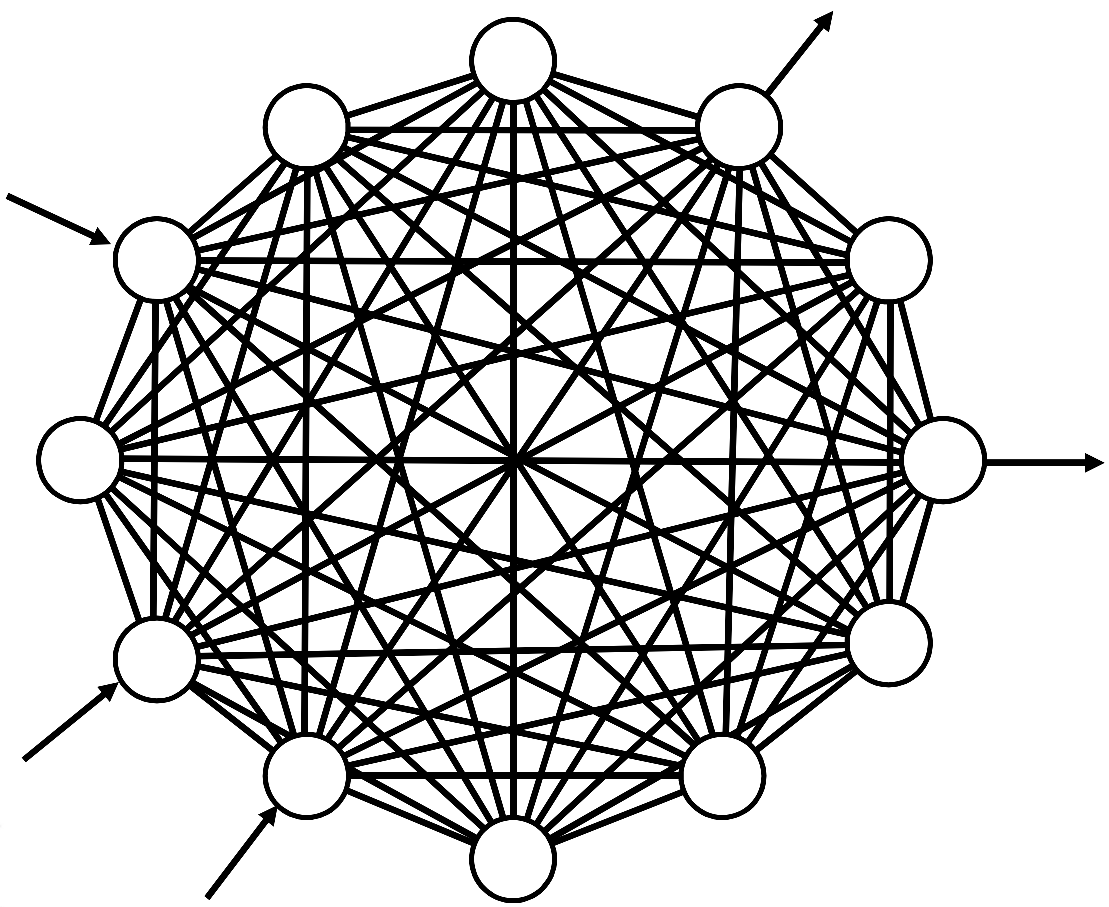
Feedforward neural networks

- Full connectivity from a layer to the next one
- Unidirectional links

iemn
Institut d'Electronique, de Microélectronique et de Nanotechnologie
UMR CNRS 8520

Neural networks models

Several organizations for the neurons:



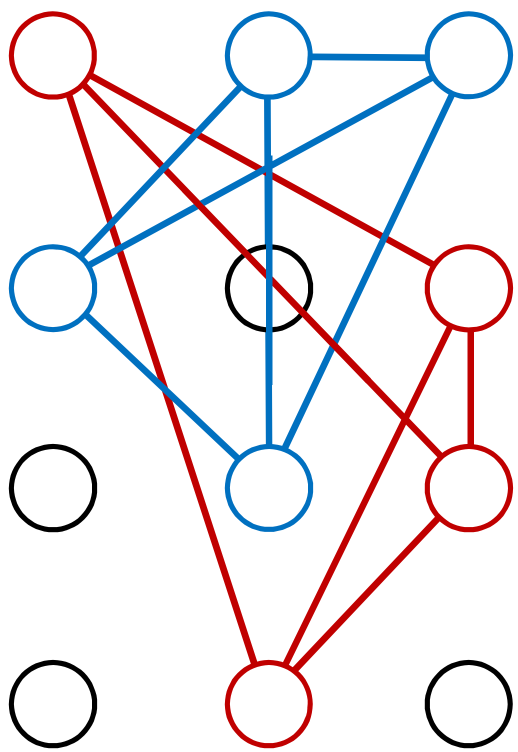
Recurrent neural networks (Hopfield)

- Full connectivity between the neurons
- Bidirectional links

iemn
Institut d'Electronique, de Microélectronique et de Nanotechnologie
UMR CNRS 8520

Neural networks models

Several organizations for the neurons:

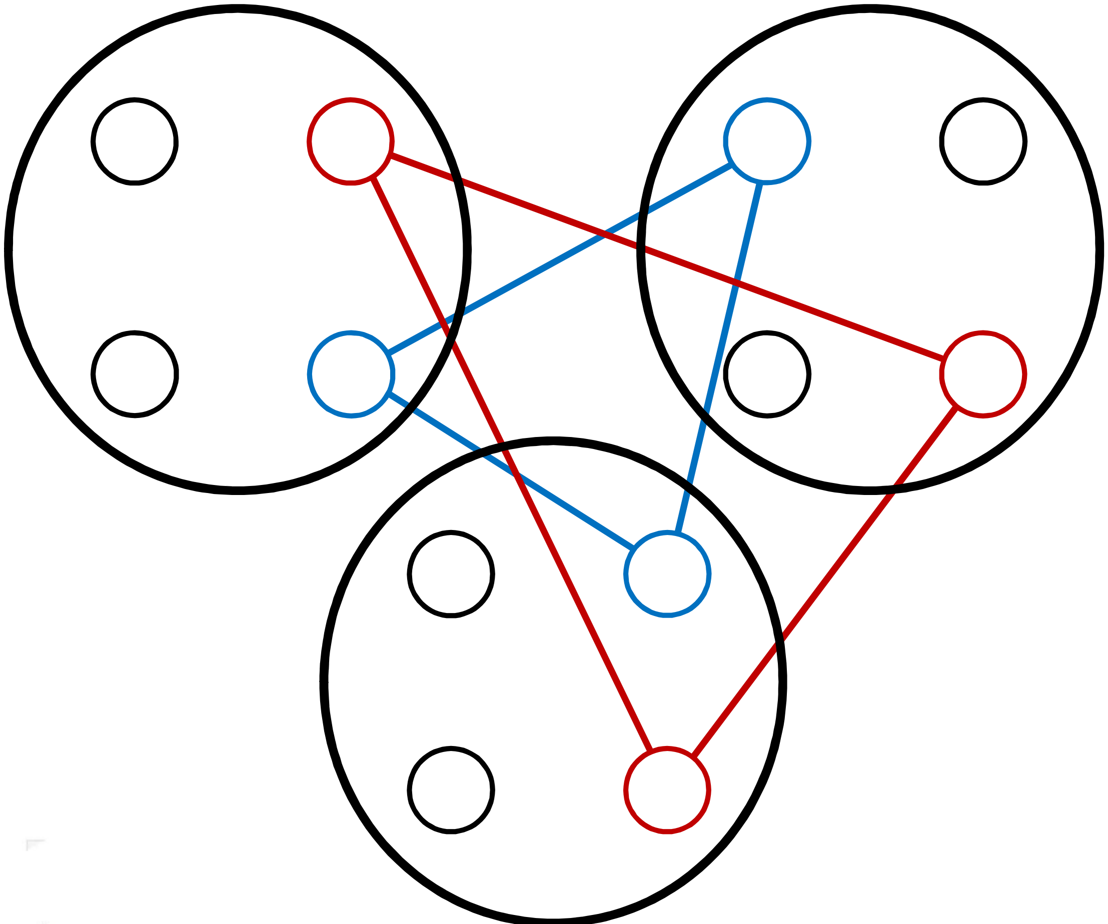


Clique-based neural networks

- Connections between neurons only through cliques
- Bidirectional links

Neural networks models

Several organizations for the neurons:



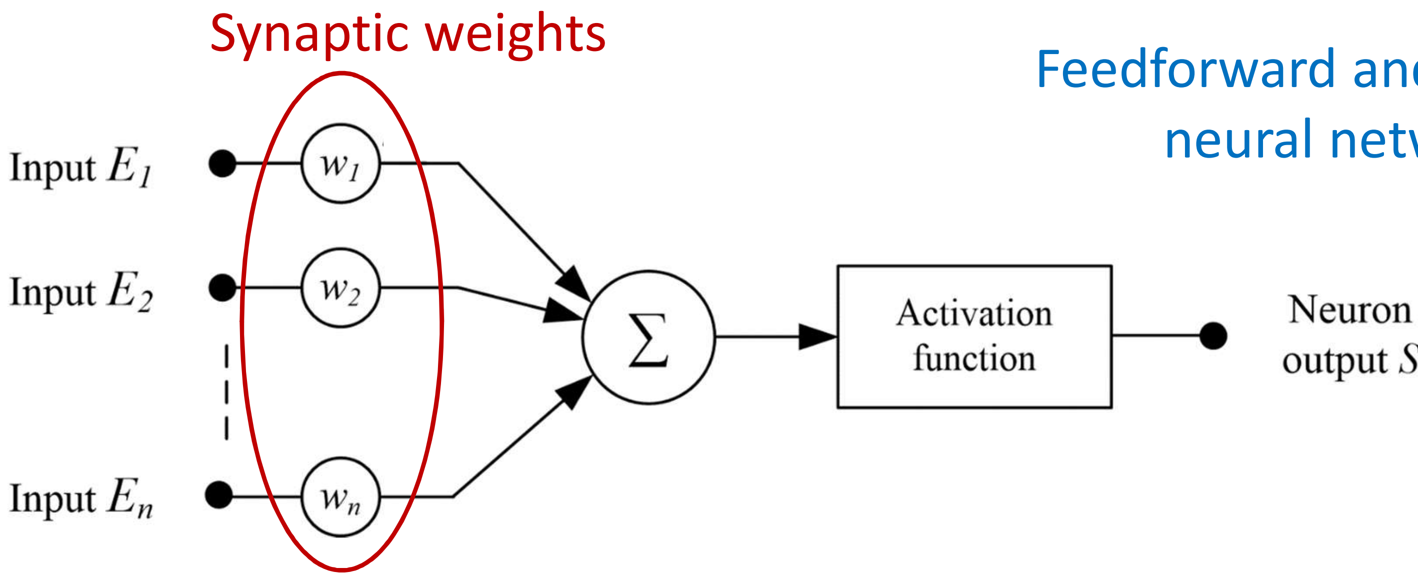
Clustered clique-based networks

- Division in clusters
- Connections between neurons from different clusters

[Gripon and Berrou, TNNLS 2011]

Inside a neuron

Structure of a neuron:



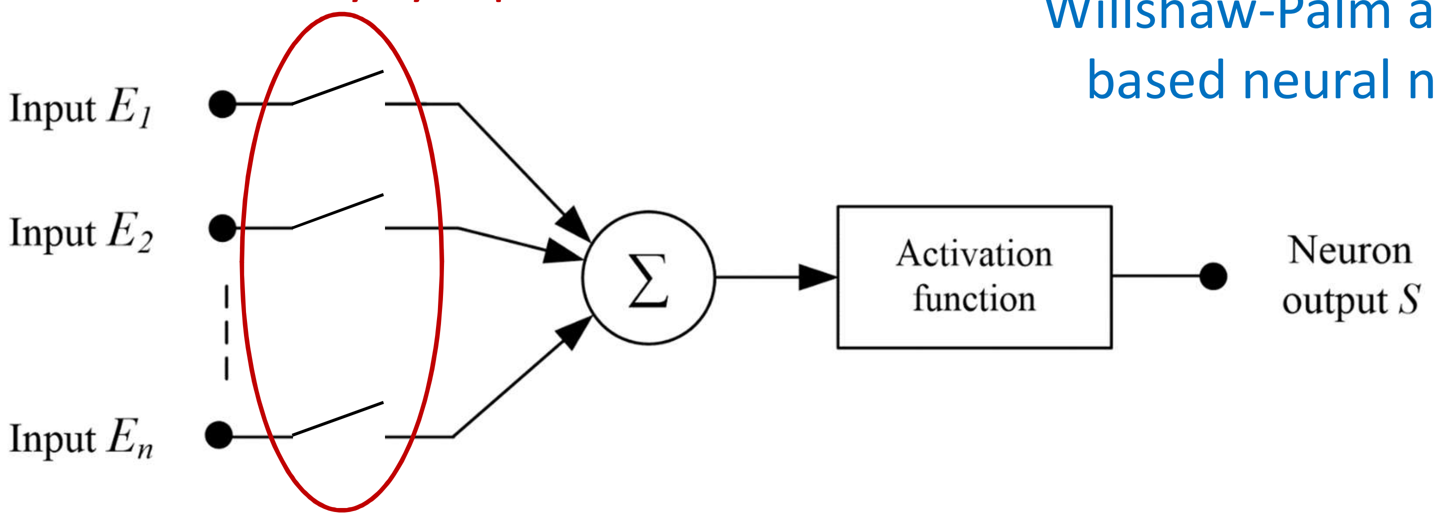
Feedforward and recurrent neural networks

Inside a neuron

Structure of a neuron:

Binary synapses

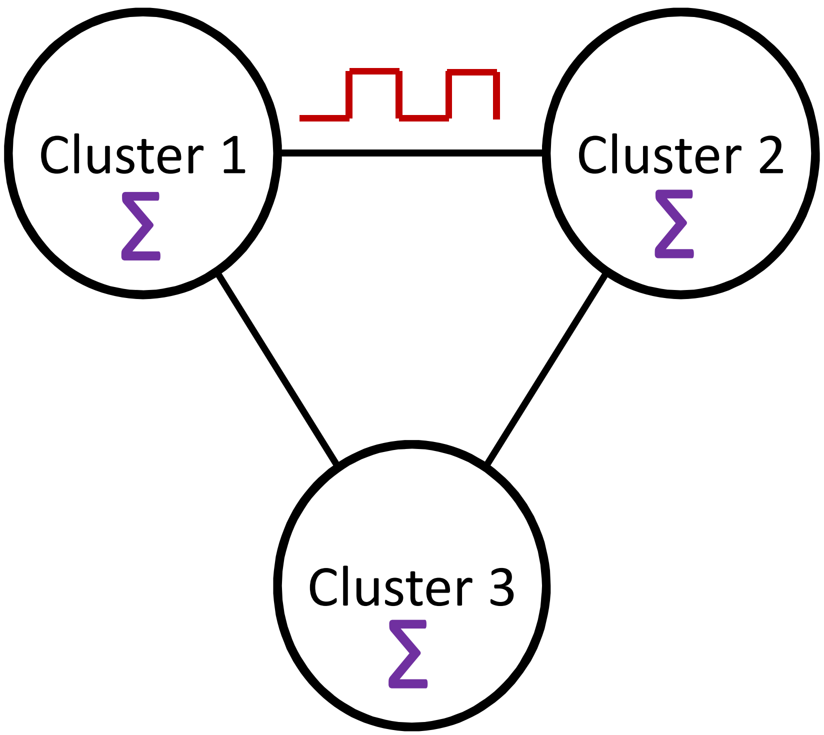
Willshaw-Palm and clique-based neural networks



Less complex activation function: WTA rule

→ comparison + activation

Implementation choices



- **Binary information** exchanged by the neurons

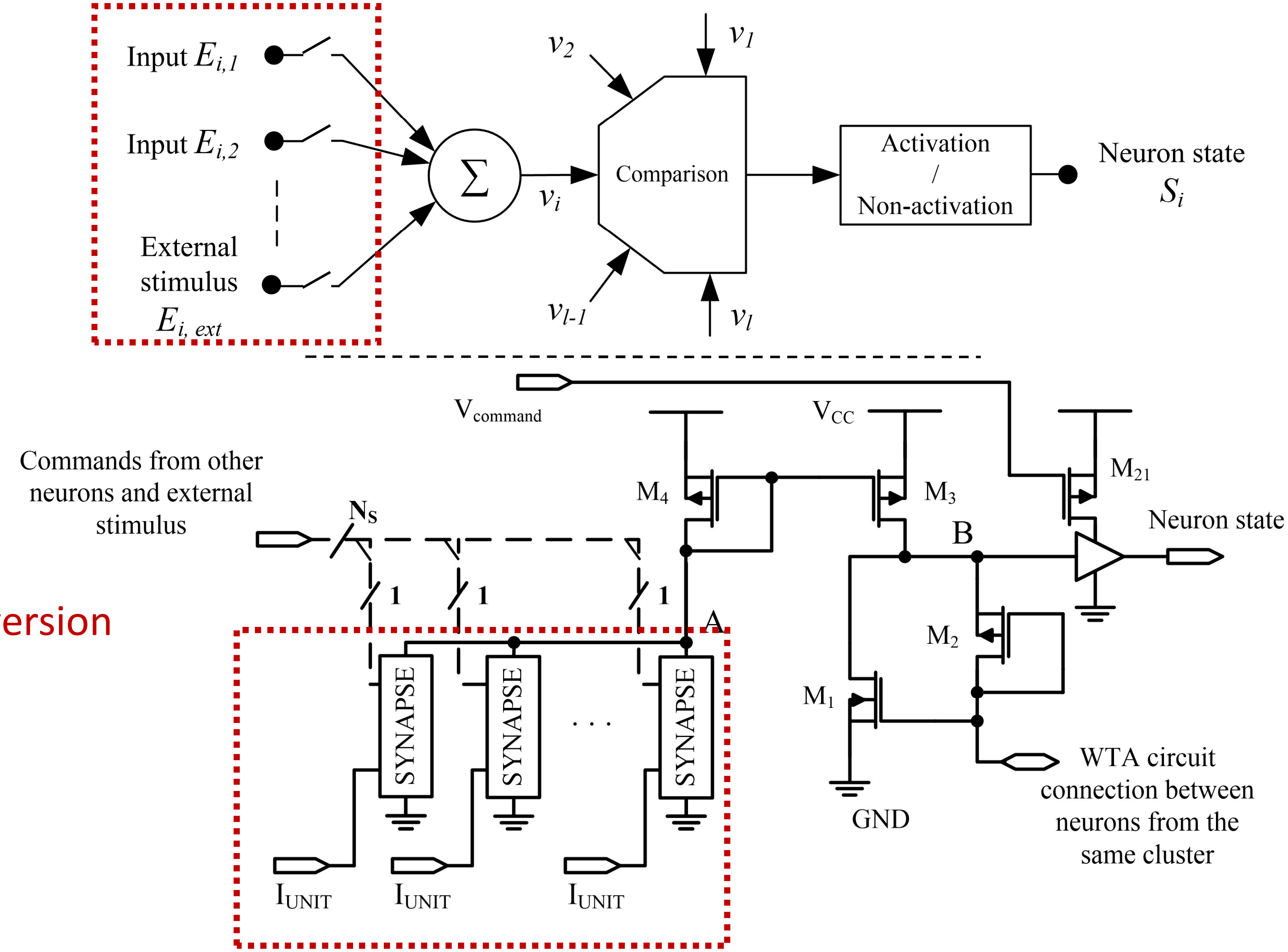
  → **Communication: digital signals**

- **Simple analog circuits** adapted to the functions in a neuron

  → **Computations: analog signals**

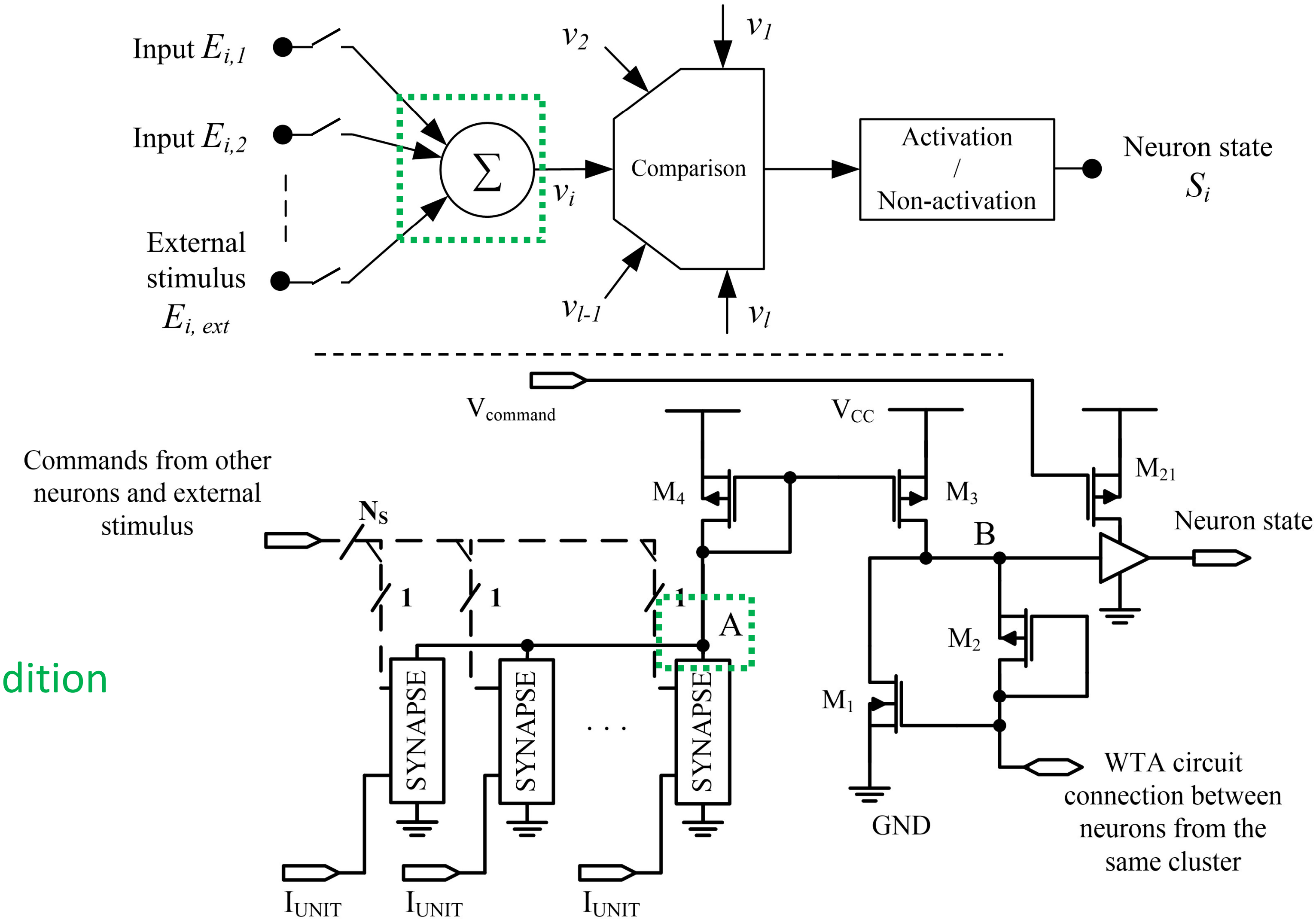
  → **Mixed-signal asynchronous implementation**

What about circuit ?



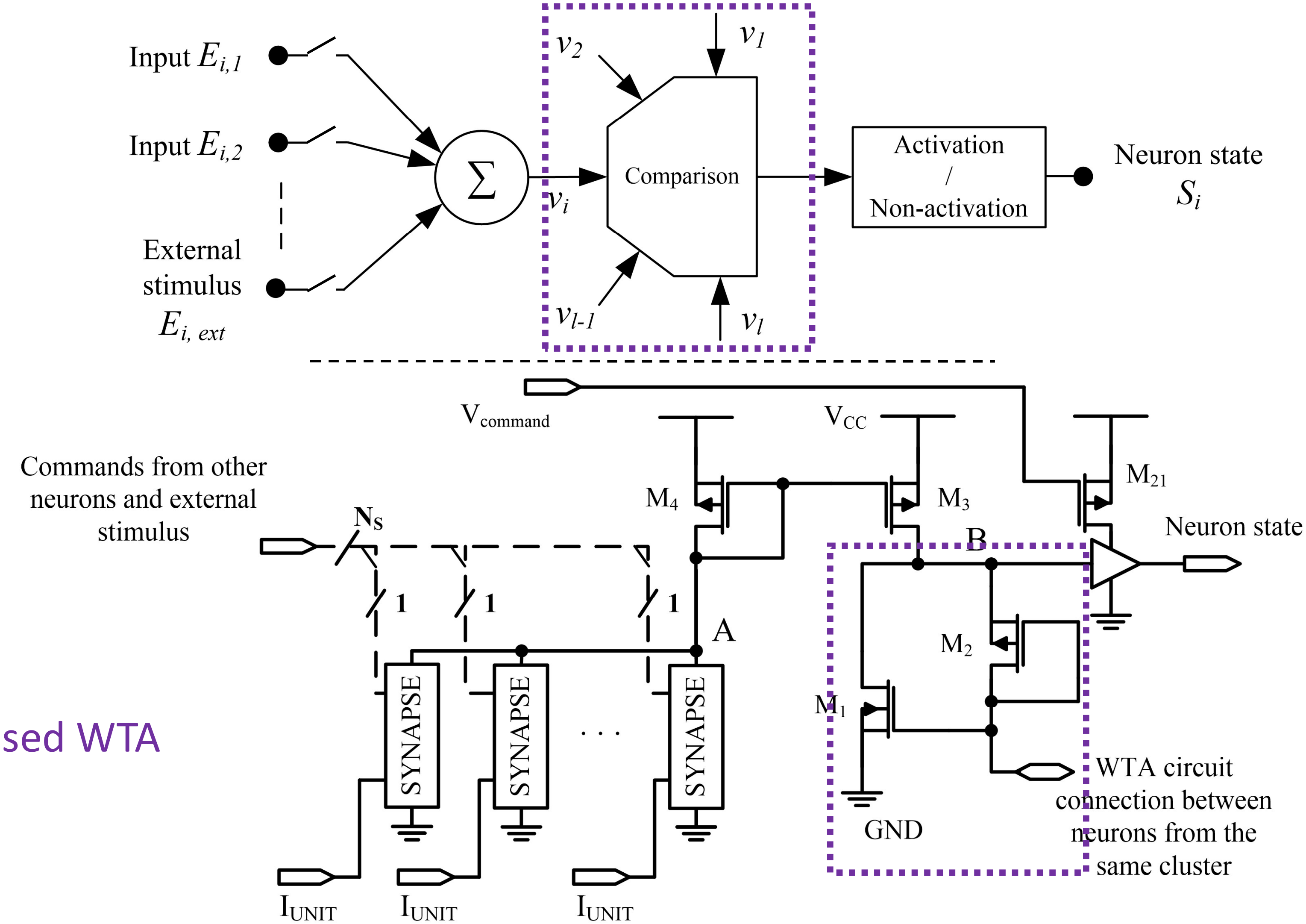
V-to-I conversion

What about circuit ?



Current addition

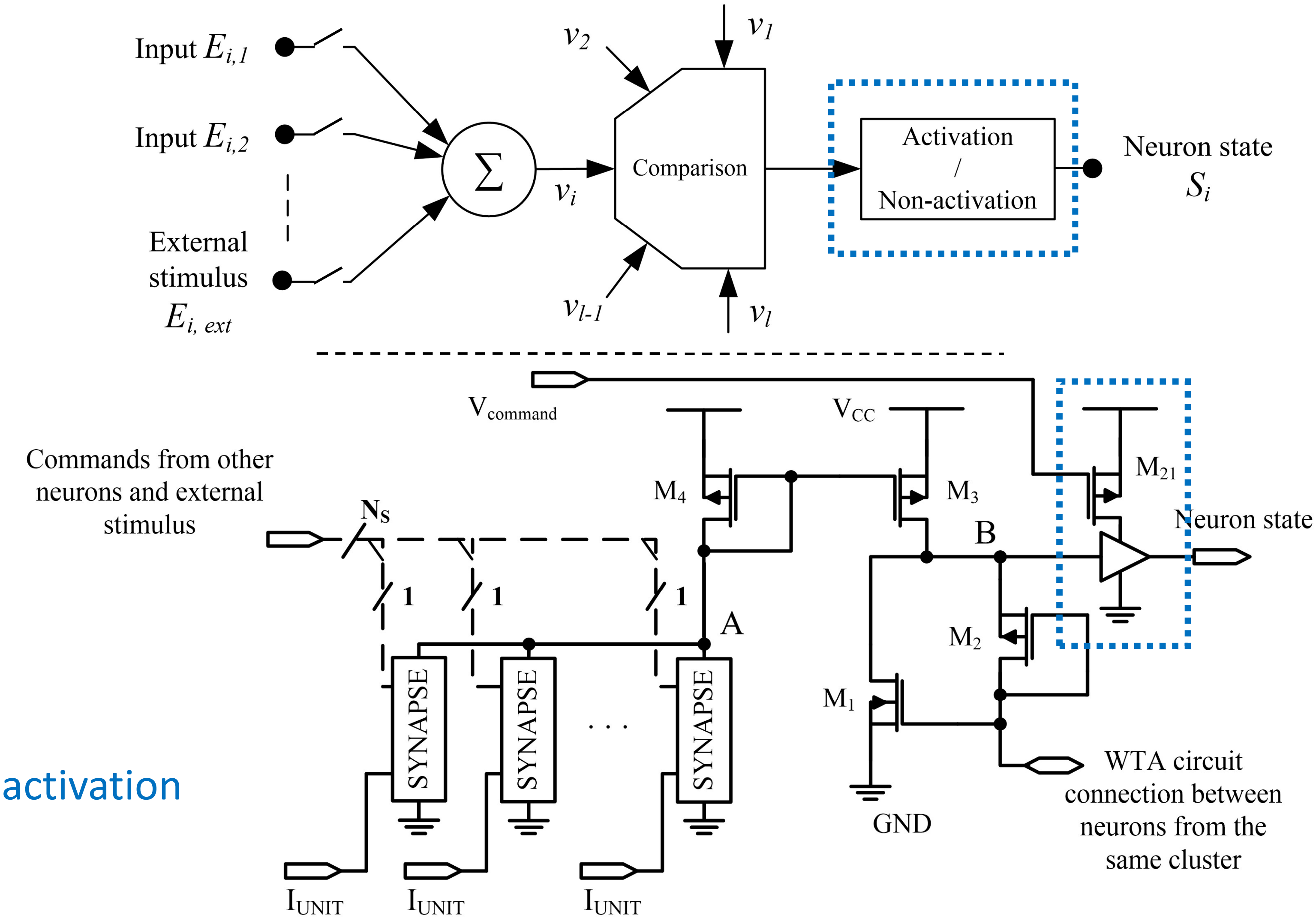
What about circuit ?



Current-based WTA

iemn
Institut d'Electronique, de Microélectronique et de Nanotechnologie
UMR CNRS 8520

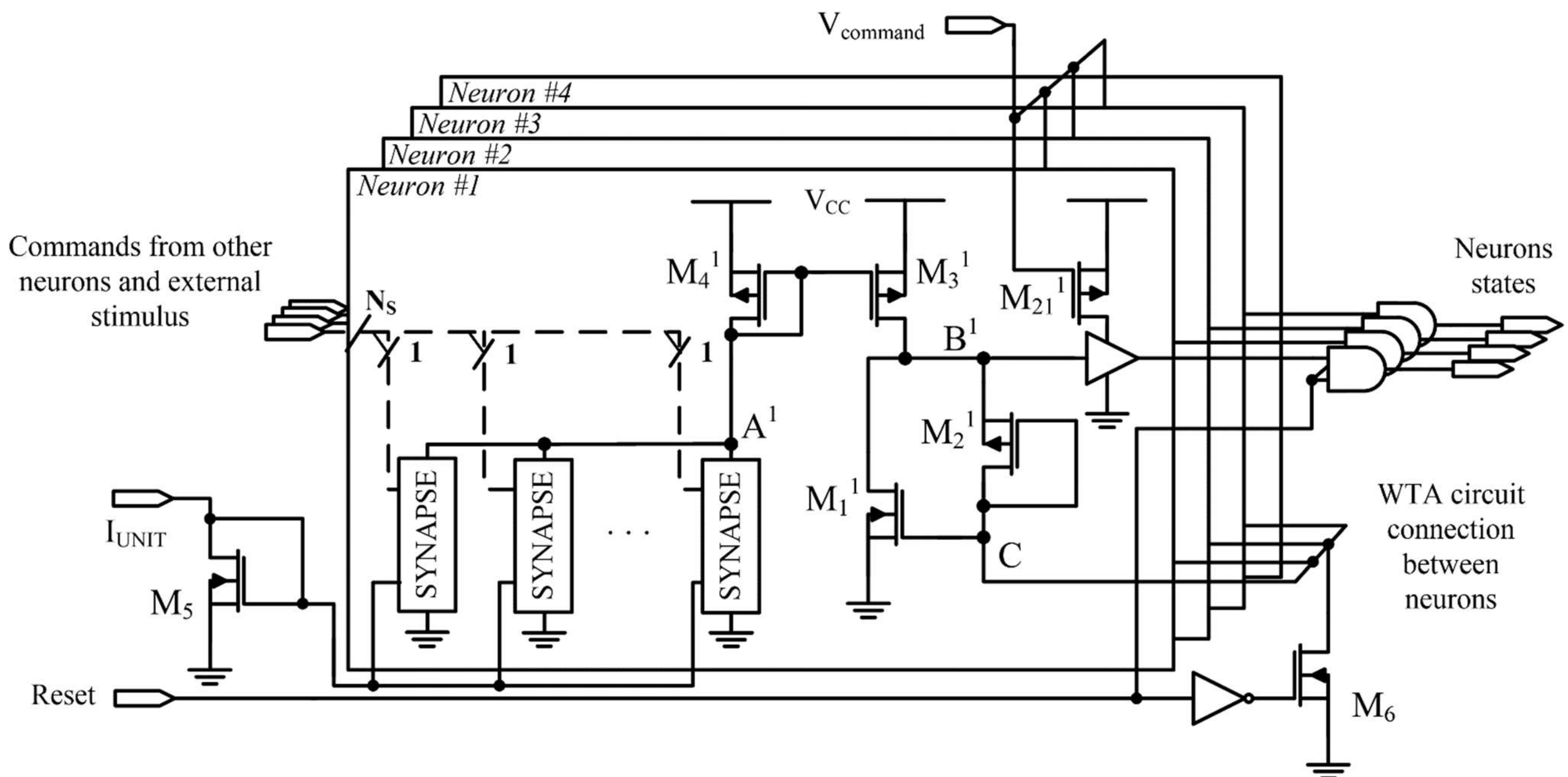
What about circuit ?



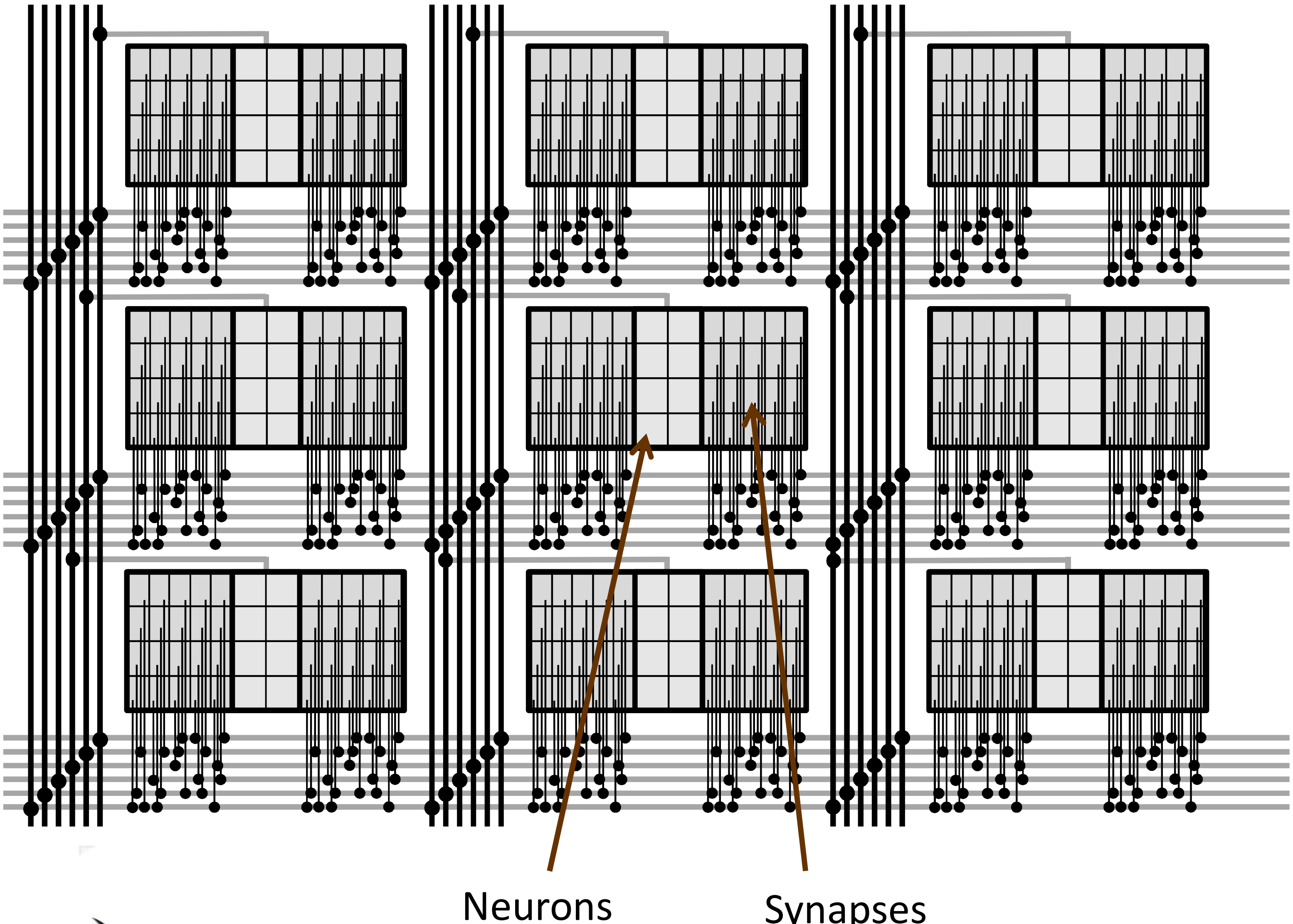
Threshold activation

What about circuit ?

Schematic of a cluster of 4 neurons:



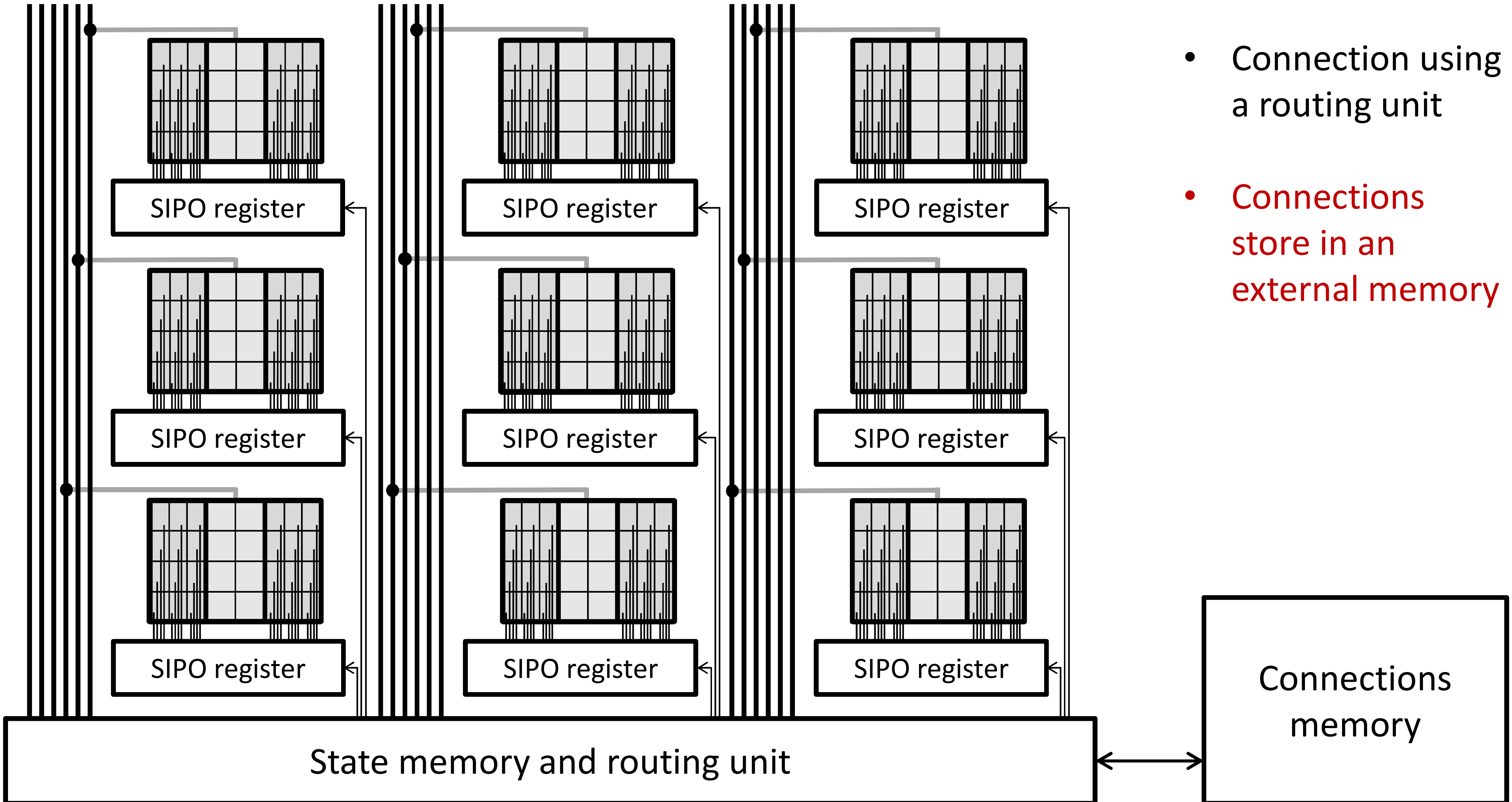
Network topologies



- Cluster matrix
- Hardwired connections between neurons
- Fastest response
- No flexibility

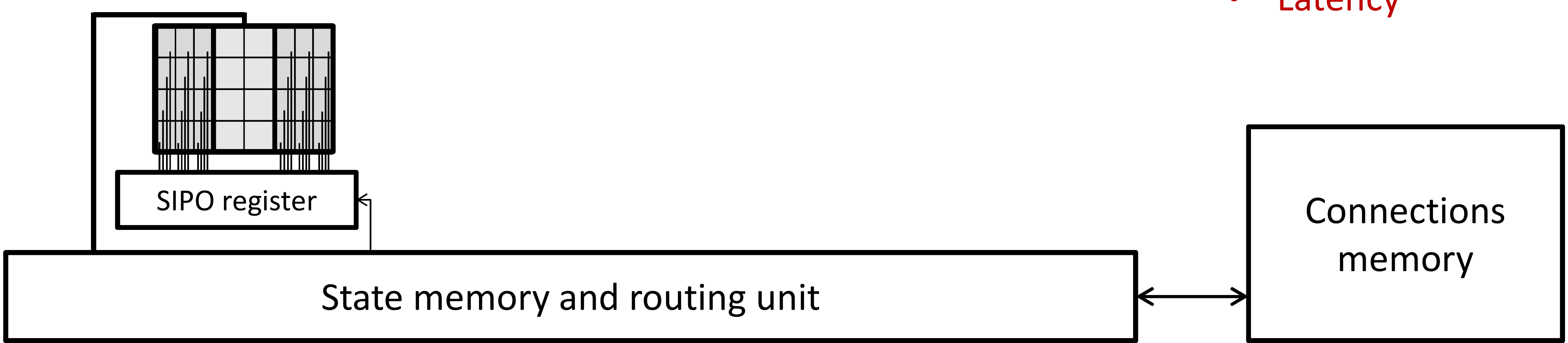
iemn
Institut d'Electronique, de Microélectronique et de Nanotechnologie
UMR CNRS 8520

Network topologies



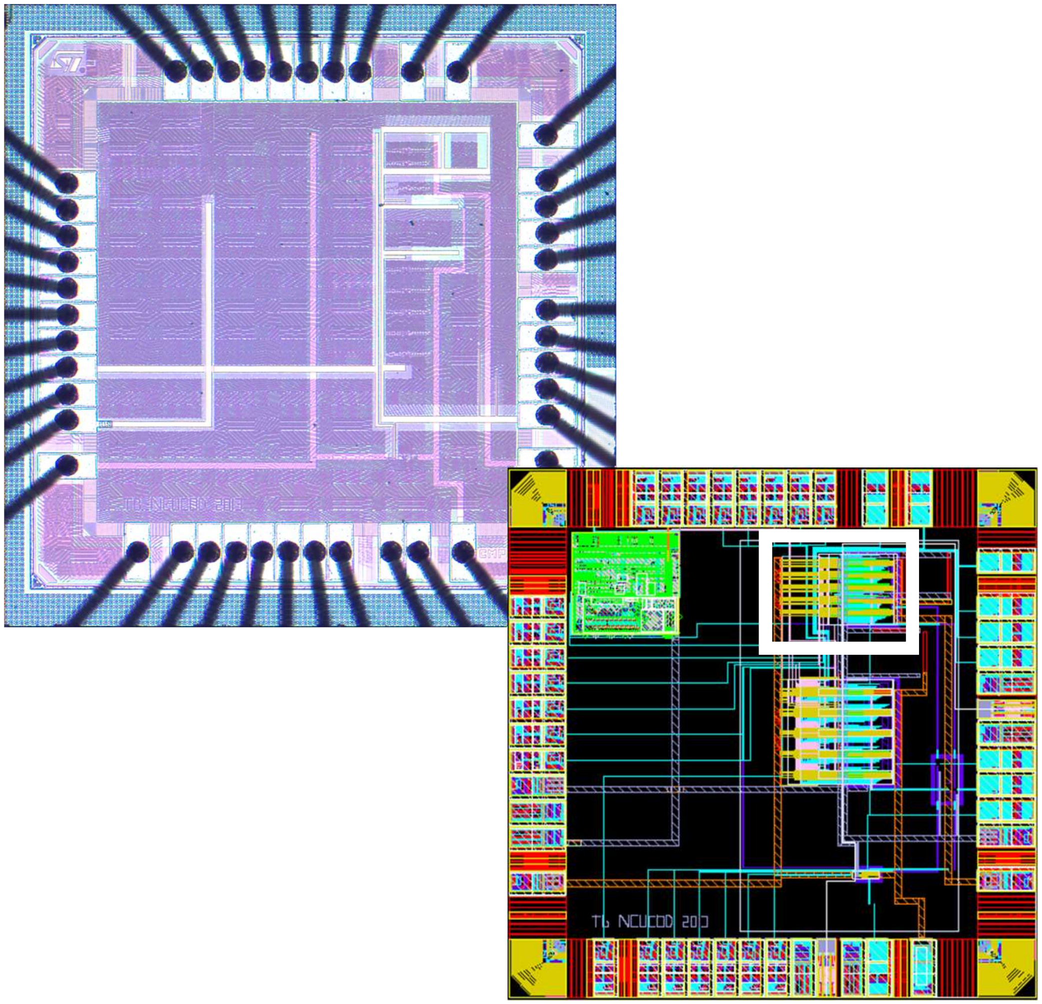
- Connection using a routing unit
- Connections store in an external memory

Network topologies



- Iteration of the process on one cluster
- Flexibility: topology changes with the number of iterations
- Latency

Hardware realizations (1/2)

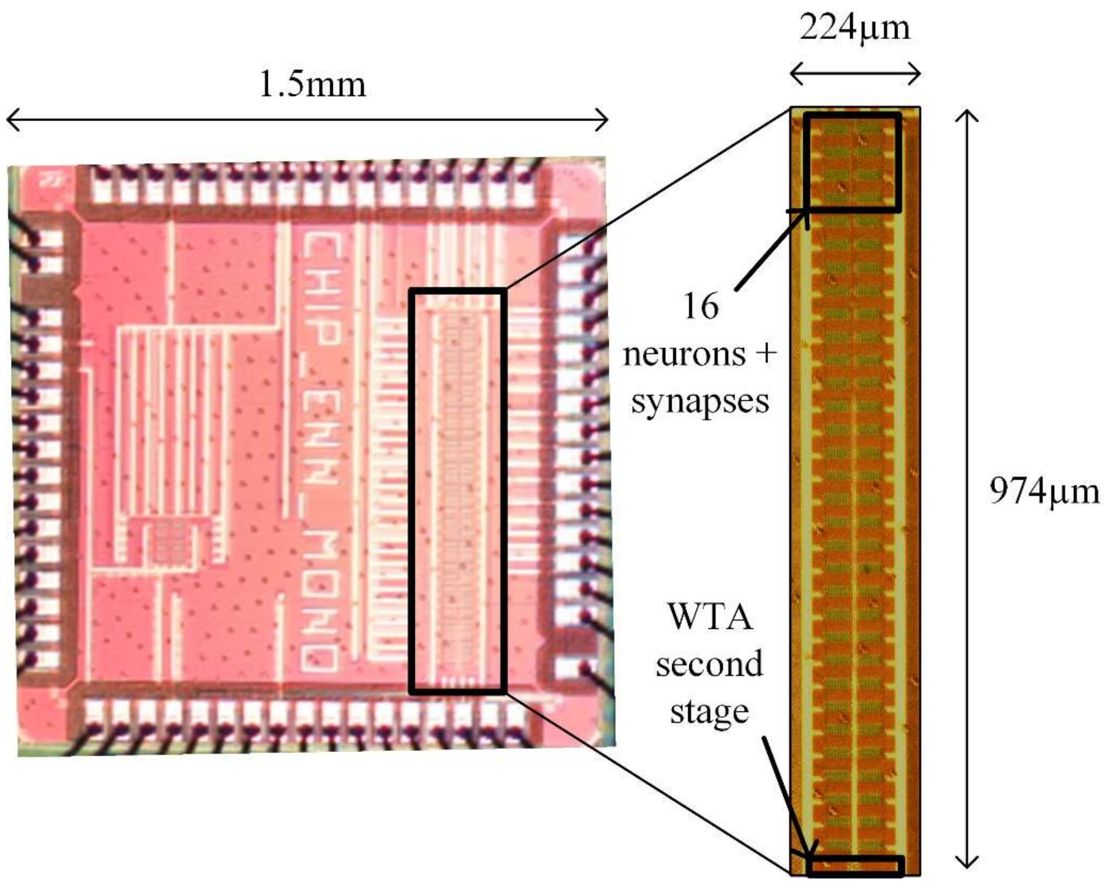


[Larras, TCAS-I 2016]

- 5 cluster of 6 neurons => 30 neurons
- Hardware connections => asynchronous
- Control signals generated by an FPGA

Technology node	**65-nm CMOS**
Silicon area occupation	**0,019 mm²**
Supply voltage	**1 V**
Synaptic current	**300 nA**
Static current	**5,4 µA**
Network response time	**58 ns**
Energy consumption per synaptic event per neuron	**48 fJ**

Hardware realizations (2/2)



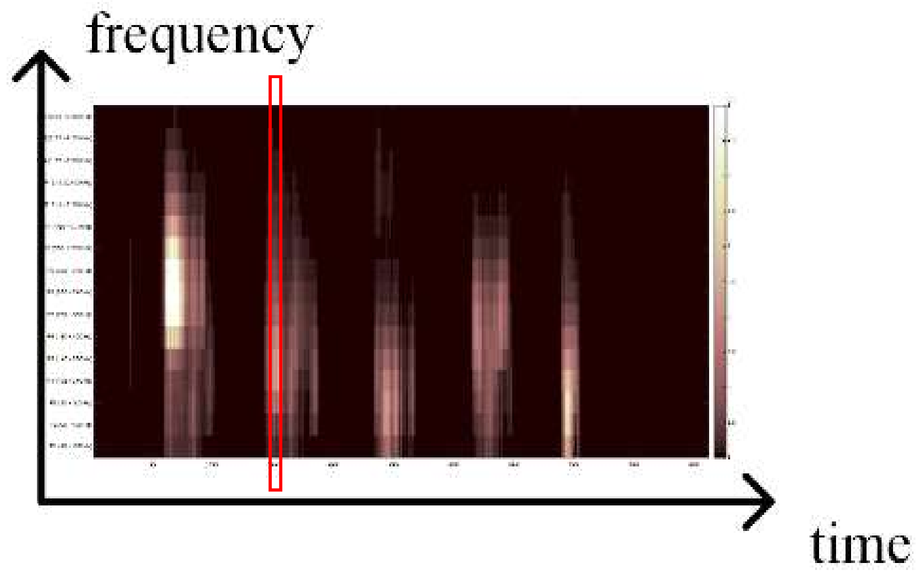
- One cluster of 128 neurons
- Time multiplexing
- Maximum of 3968 emulated neurons
- Driven by an FPGA

Technology node	65-nm CMOS
Silicon area occupation	0,21 mm²
Supply voltage	1 V
Synaptic current	300 nA
Static current	23,4 µA
Cluster response time	83 ns
Energy consumption per synaptic event per neuron	115 fJ

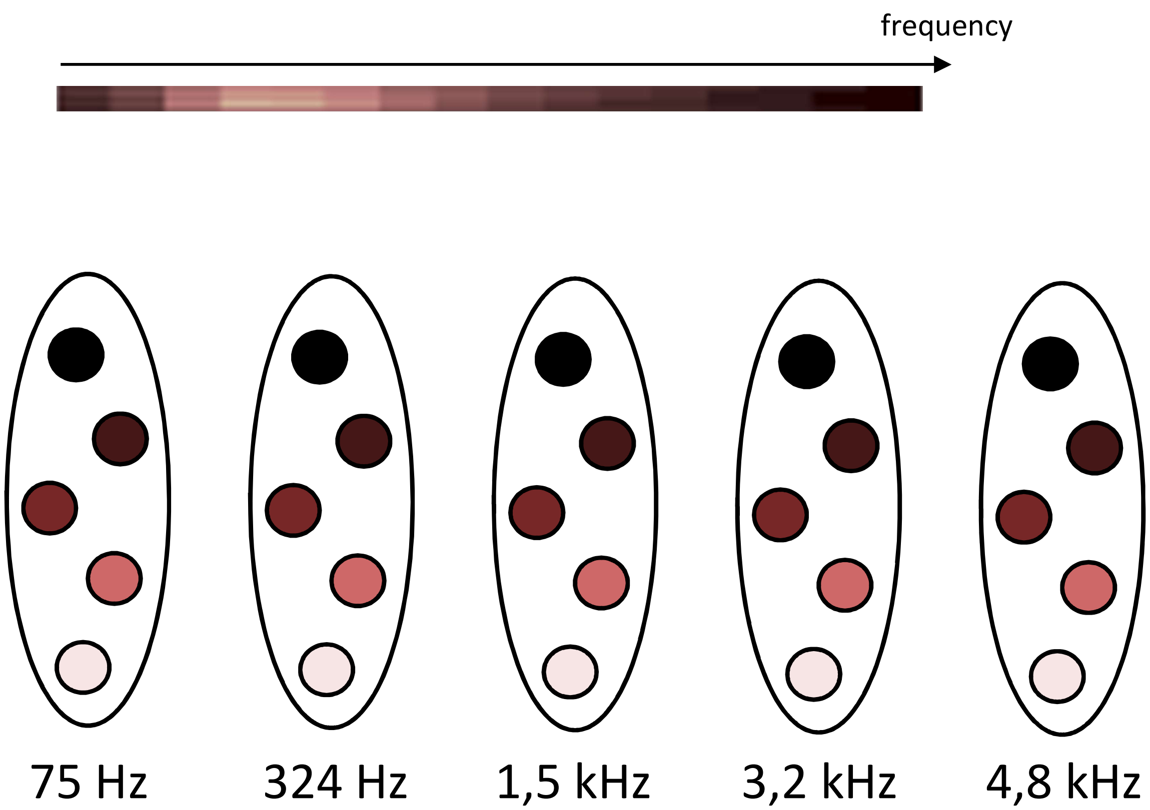
[Larras, TCAS-I 2019]

iemn Institut d'Electronique, de Microélectronique et de Nanotechnologie UMR CNRS 8520

Envisionned scheme

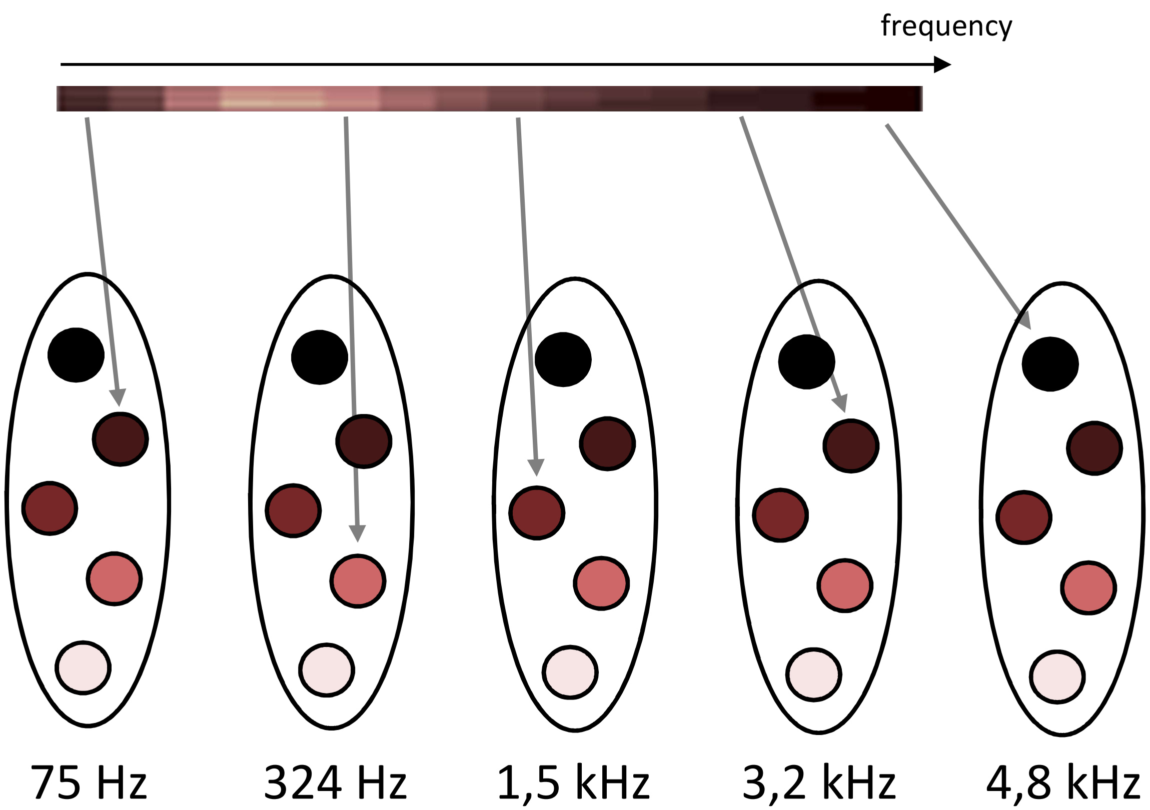


Envisionned scheme



- One feature = one cluster

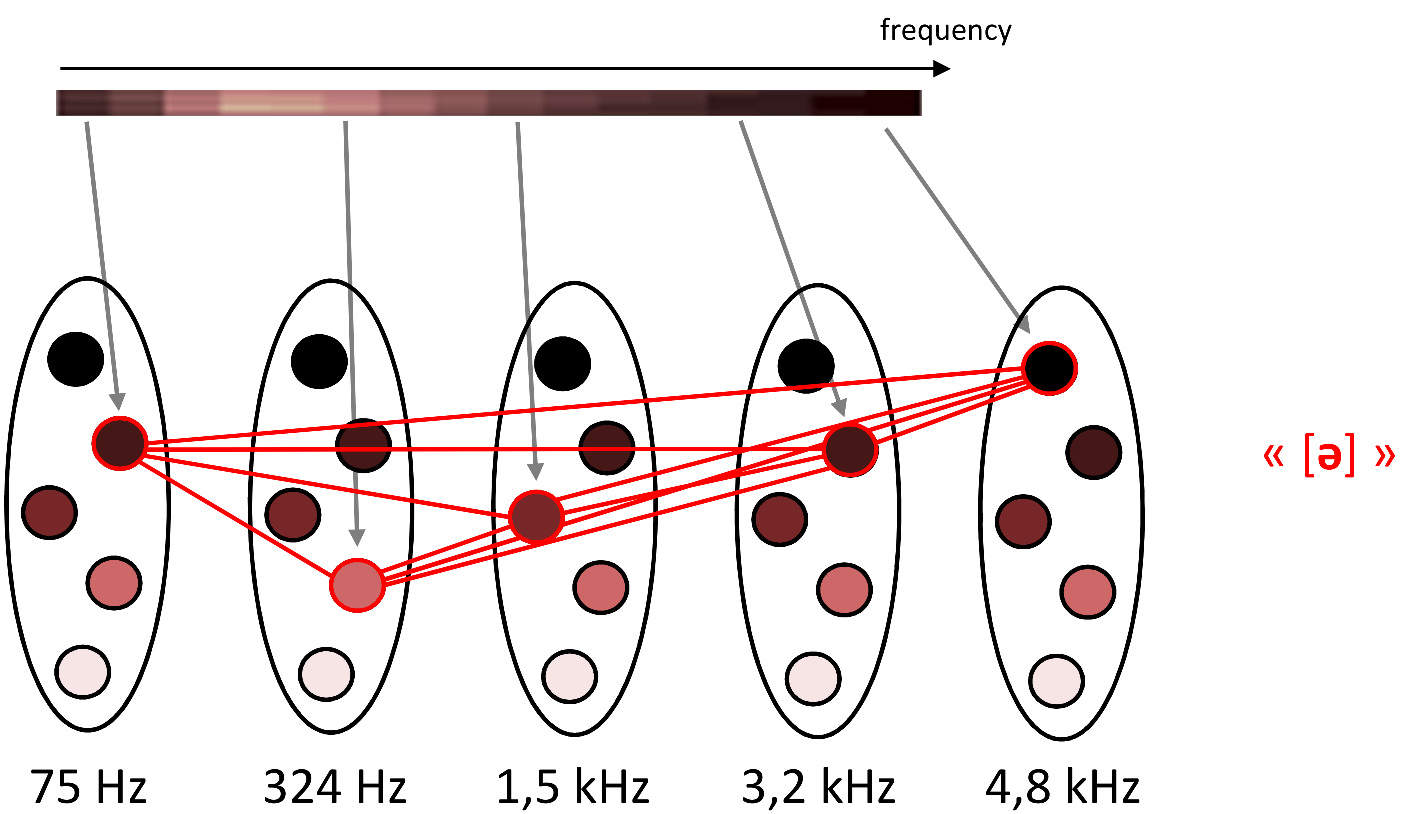
Envisionned scheme



- One feature = one cluster
- One neuron per quantization level

iemn
Institut d'Electronique, de Microélectronique et de Nanotechnologie
UMR CNRS 8520

Envisionned scheme



- One feature = one cluster
- One neuron per quantization level
- **Instantaneous detection of speech formants (cliques)**

iemn
Institut d'Electronique, de Microélectronique et de Nanotechnologie
UMR CNRS 8520

Further opportunities

- **Asynchronous formant extraction**
  - Applications: voice activity detection, phonemes detection
- **Data reduction**
  - From 2-D data to 1-D data
  - Use with LSTM stage to extract keywords
- **Circuit integrability ?**
- **Compatibility with real time ?**

iemn
Institut d'Electronique, de Microélectronique et de Nanotechnologie
UMR CNRS 8520

Challenges

- **Feature extraction unit**
  - Event-driven processing with no clocks is difficult to handle and design (concepts, tools)
  - Timing is critical…
- **Classification unit**
  - Generic topology vs. diversity of applications
  - Bridging the gap from theory to efficient hardware
- **Latency and energy consumption!**
- **Integration in advanced CMOS technology**

Conclusion

- ANR LEOPAR project targeting a breakthrough in the audio processing domain, in terms of energy efficiency
- Circuit implementation leveraging analog and digital domains
- Targeted hardware demonstration: hardware prototype and integrated circuit in 28-nm FDSOI CMOS

iemn
Institut d'Electronique, de Microélectronique et de Nanotechnologie
UMR CNRS 8520

Thank you !

Any questions ? Feel free to ask or send an e-mail to
benoit.larras@yncrea.fr

iemn
Institut d'Electronique, de Microélectronique et de Nanotechnologie
UMR CNRS 8520