SiO_X-Based ReRAM Memory Window Optimization

Giuseppe PICCOLBONI

R&D Engineer

Outline

- Weebit-nano Presentation
- Device Description
- Initial Resistance Tuning
- Memory Window Optimization
- SiO_X-based ReRAM Neuromorphic Application
- Conclusions

Outline

- **Weebit-nano Presentation**
- **Device Description**
- **Initial Resistance Tuning**
- **Memory Window Optimization**
- **SiO_X-based ReRAM Neuromorphic Application**
- **Conclusions**

Weebit - Overview

CEO

Coby Hanoch

40 years of experience in the semiconductor industry

CEO of PacketLight

CHAIRMAN

David Perlmutter

Ex-Intel EVP
IEEE Fellow

Brought to Market: Centrino™ mobile technology

EXECUTIVE DIRECTOR

Dr. Yoav Nissan-Cohen

CEO of Tower semiconductors

Co-founder of Saifun Semiconductor

CTO

Amir Regev

45nm NOR Flash Technology Development at Micron

Was part of Intel's' Automotive division

DIRECTOR

Atiq Raza

Chairman and CEO of NextGen Inc

President, COO of AMD

Weebit - Market Opportunities

Analog

Replacing EEPROMS

Key Applications:
All types of sensors, PMIC, LED drivers, Audio

Uses:
Trimming
Data storage
Code Storage

Capacities:
64bits – 1/2Mb

IoT

Replacing NOR Flash

Key Applications:
Wearables, security, smart cities

Uses:
Data storage
Code Storage

Capacities:
16Kb –1Mb

Artificial Intelligence

Replacing DRAM usage

Key Applications:
Facial & object recognition

Uses:
Inference
Learning tasks

Capacities:
Mb-Gb

Potential game-changer in a wide range of applications

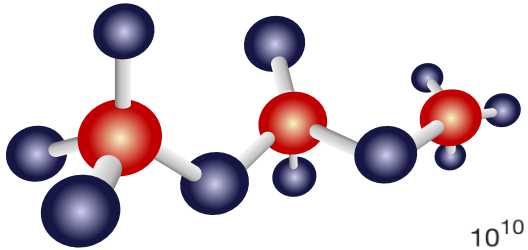
Why SiO_x

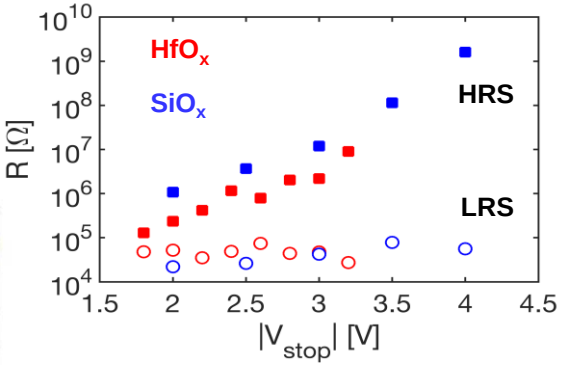
<u>Physical Characteristics</u>

- High bandgap material - large resistive window
- Low leakage
- Low HRS variability
- High temperature stability

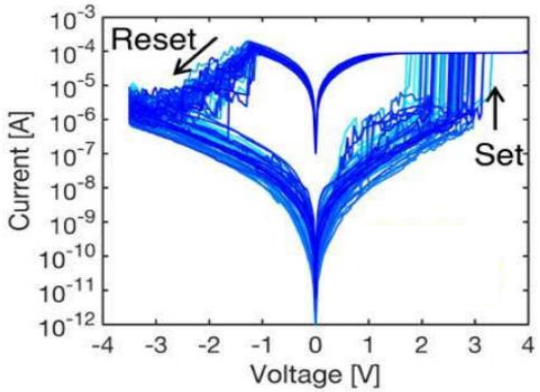
<u>Manufacturability characteristics</u>

- Full CMOS compatibility
- High manufacturability
- Any Fab / process / deposition technique
- Easily tunable
  - Thickness
  - Stoichiometry
- Cost effective





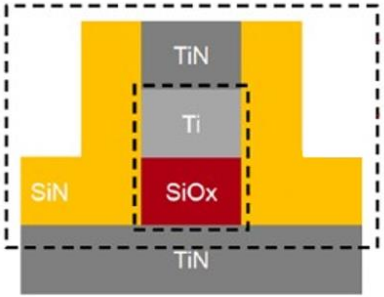
E. Ambrosi, *Faraday Discussions 2019*

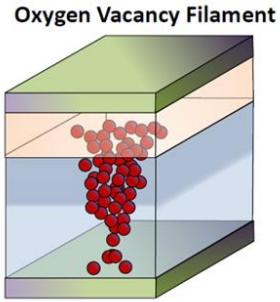


Outline

- Weebit-nano Presentation
- Device Description
- Initial Resistance Tuning
- Memory Window Optimization
- SiO_X-based ReRAM Neuromorphic Application
- Conclusions

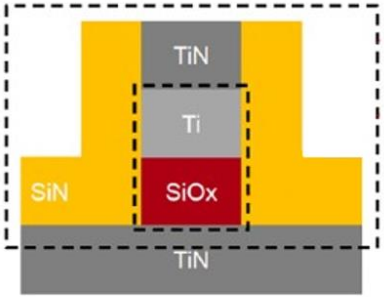
Resistive RAM (ReRAM)

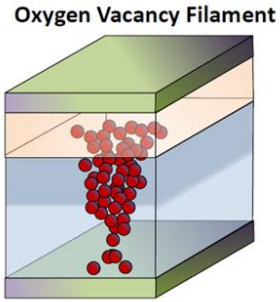


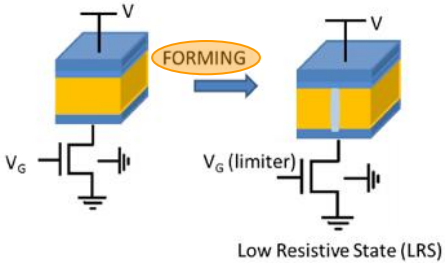


- Memory technology based on the formation/disruption of a conductive filament (CF) to encode binary information
- In the case of Oxide-based ReRAM (OxRAM) the CF is made of oxygen vacancies

Resistive RAM (ReRAM)

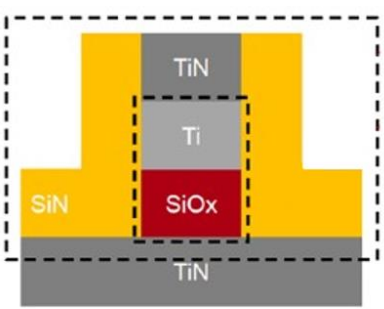


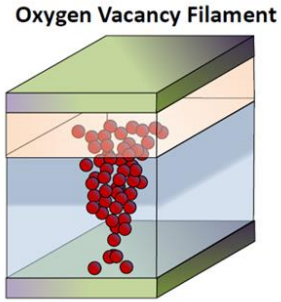


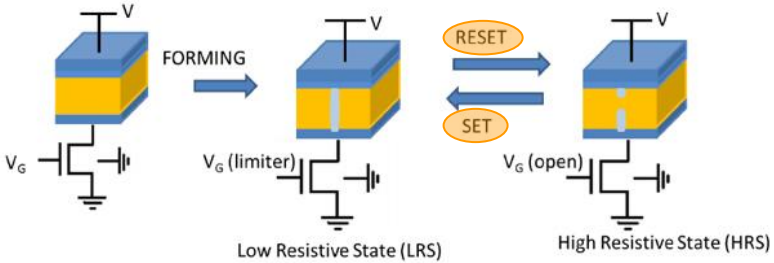


- Memory technology based on the formation/disruption of a conductive filament (CF) to encode binary information
- In the case of Oxide-based ReRAM (OxRAM) the CF is made of oxygen vacancies
- Operations:
  - This technology requires a first step called FORMING where the CF is formed in a pristine oxide.

Resistive RAM (ReRAM)

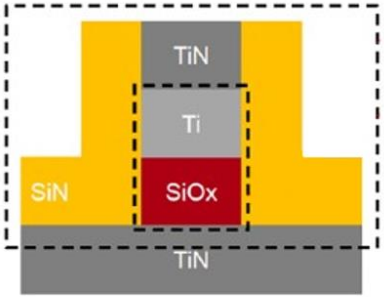


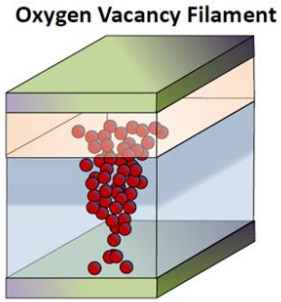


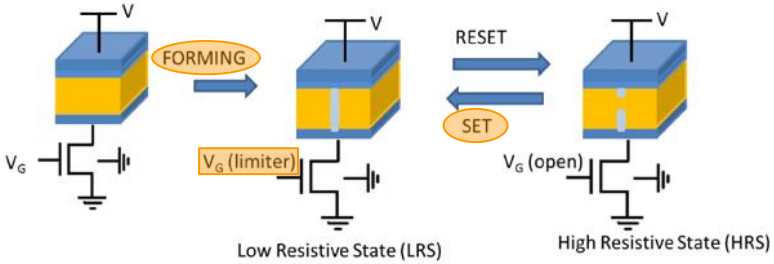


- Memory technology based on the formation/disruption of a conductive filament (CF) to encode binary information
- In the case of Oxide-based ReRAM (OxRAM) the CF is made of oxygen vacancies
- Operations:
  - This technology requires a first step called FORMING where the CF is formed in a pristine oxide.
  - Once formed the CF can be partially erased with a RESET operation leading to the High Resistance State (HRS) and re-created with the SET operation leading to the Low Resistance State (LRS)

Resistive RAM (ReRAM)







- Memory technology based on the formation/disruption of a conductive filament (CF) to encode binary information
- In the case of Oxide-based ReRAM (OxRAM) the CF is made of oxygen vacancies
- Operations:
  - This technology requires a first step called FORMING where the CF is formed in a pristine oxide.
  - Once formed the CF can be partially erased with a RESET operation leading to the High Resistance State (HRS) and re-created with the SET operation leading to the Low Resistance State (LRS)
  - FORMING and SET operations require a current limiting device such as a transistor to limit the current flowing into the device

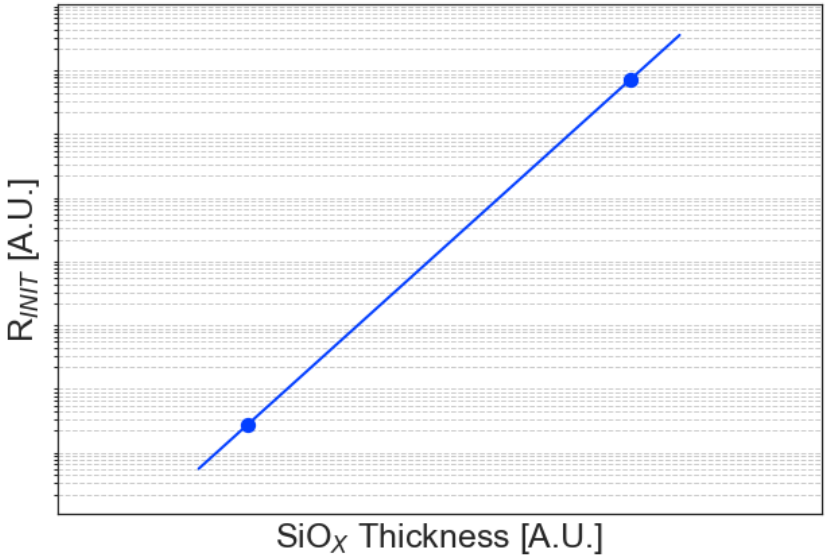
Outline

- Weebit-nano Presentation
- Device Description
- Initial Resistance Tuning
- Memory Window Optimization
- SiO_X-based ReRAM Neuromorphic Application
- Conclusions

Initial Resistance Tuning

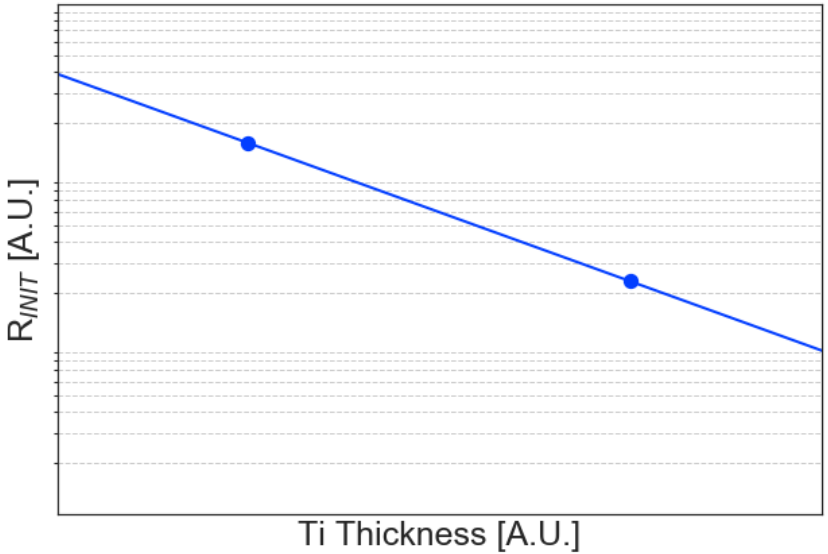
- Initial Resistance (R_{INIT}) tuning is of great importance to optimize memory behavior:
  - *If R_{INIT} too Low* → HRS and hence memory window (HRS/LRS) will be limited. In general HRS < R_{INIT}
  - *If R_{INIT} too High* → high Forming Voltage (V_F) degrades the memory cell and poses problem from a design point of view
- Knobs to tune Rinit:
  - SiO_X thickness
  - TE Thickness
  - Stoichiometry

SiO_X Thickness effect on R_{INIT}



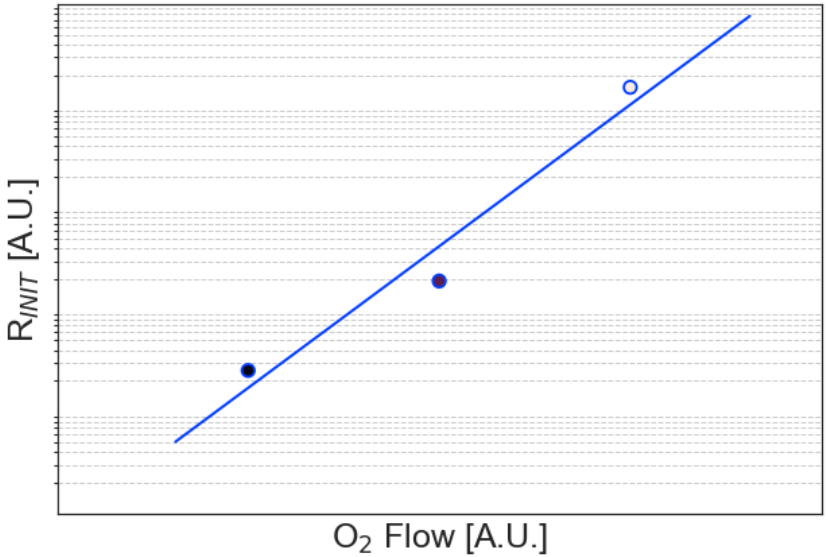
- Increasing the oxide thickness leads to a R_{INIT} increase
- The R_{INIT} has the following dependence with the Oxide thickness
  - $Log_{10}(R_{INIT}) = A*SiO_X + B$

Ti Thickness effect on R_{INIT}



- Increasing the Ti Top Electrode (TE) thickness leads to a R_{INIT} reduction thanks to the O gettering action of the Ti on the Oxide.
- The R_{INIT} has the following dependence with the Ti TE thickness
  - $Log_{10}(R_{INIT}) = -A*Ti + B$

O_2 flow effect on R_{INIT}

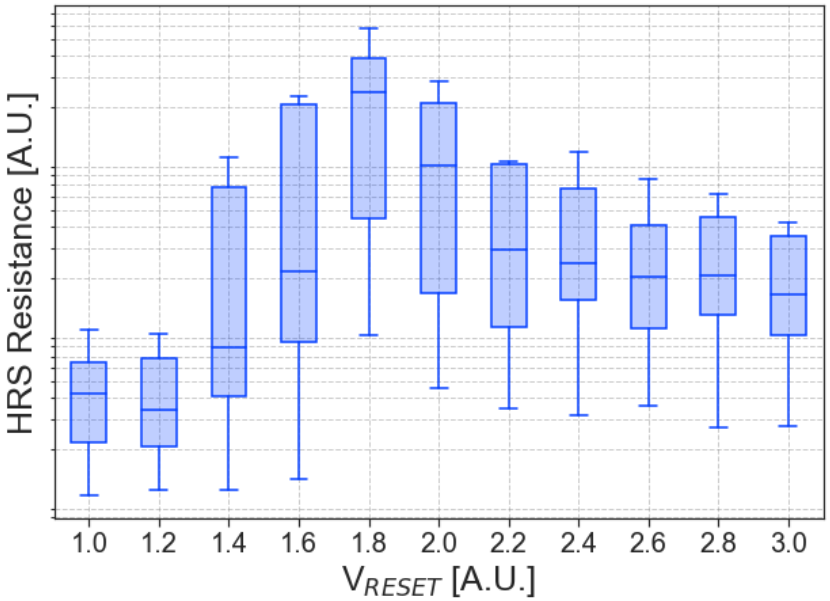


- Higher O_2 flow in the PVD chamber gives a less sub-stoichiometric SiO_X, therefore the R_{INIT} results higher.
- The R_{INIT} has the following dependence with the O_2 Flow thickness
  - $Log_{10}(R_{INIT}) = A*O_2 + B$

Outline

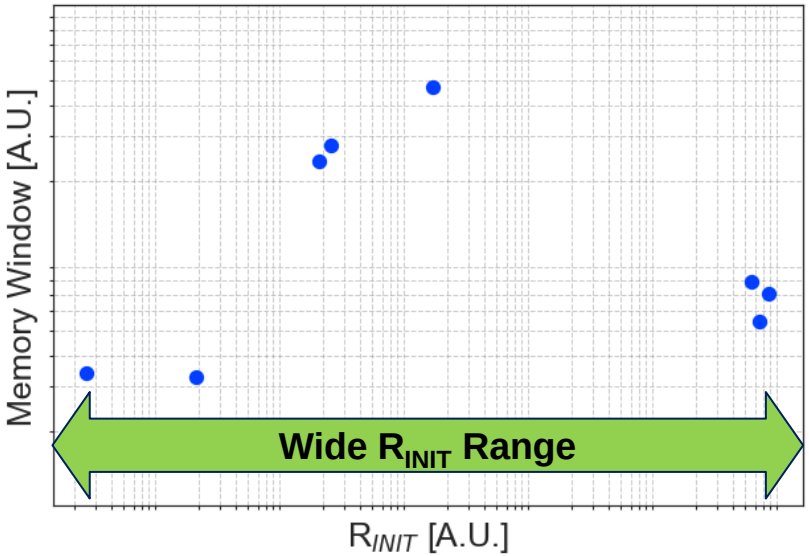
- Weebit-nano Presentation
- Device Description
- Initial Resistance Tuning
- Memory Window Optimization
- SiO_X-based ReRAM Neuromorphic Application
- Conclusions

Memory Window optimization



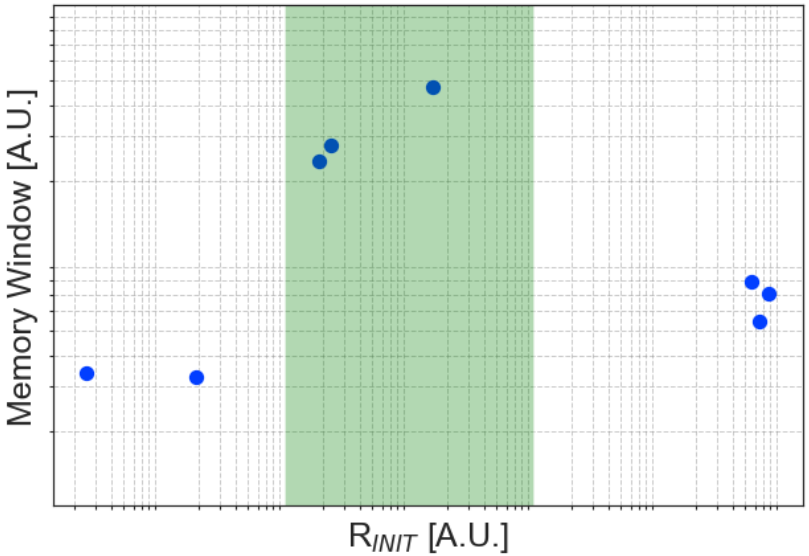
- Memory Window = HRS/LRS
- For each split studied we obtain the best memory window by varying the RESET voltage (V_{RESET}) applied to the cell being the LRS governed solely by the compliance current (I_C) used during SET operation.
- In the left figure an example of HRS Resistance as a function of the V_{RESET}. There is an optimum value beyond which the reset operation becomes less effective.

Memory Window optimization



- By combining variations of the 3 parameters above-mentioned we are able to easily tune R_{INIT}

Memory Window optimization



- By combining variations of the 3 parameters above-mentioned we are able to easily tune R_{INIT}
- Measurements show how there is an optimal R_{INIT} range that maximizes the memory window margin (defined as HRS/LRS)

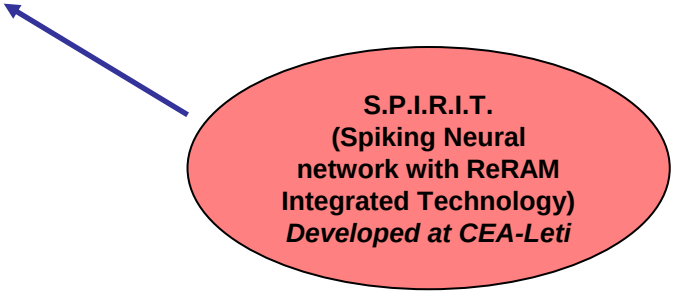
Outline

- Weebit-nano Presentation
- Device Description
- Initial Resistance Tuning
- Memory Window Optimization
- SiO_X-based ReRAM Neuromorphic Application
- Conclusions

SPIRIT

Motivations:

Data-centric workloads exhibited by Deep Neural Network (DNN) applications require circuit architectures that minimize data transfer.



SPIRIT

Motivations:

Data-centric workloads exhibited by Deep Neural Network (DNN) applications require circuit architectures that minimize data transfer.

ReRAM as Possible Candidate:

- Fast access
- High density
- Ease of integration in BEOL
- Good endurance and window
- Limited resistance drift

S.P.I.R.I.T.
(Spiking Neural network with ReRAM Integrated Technology)
Developed at CEA-Leti

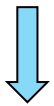
SPIRIT

Application:

- Hand-written digit classification
- MINST Database

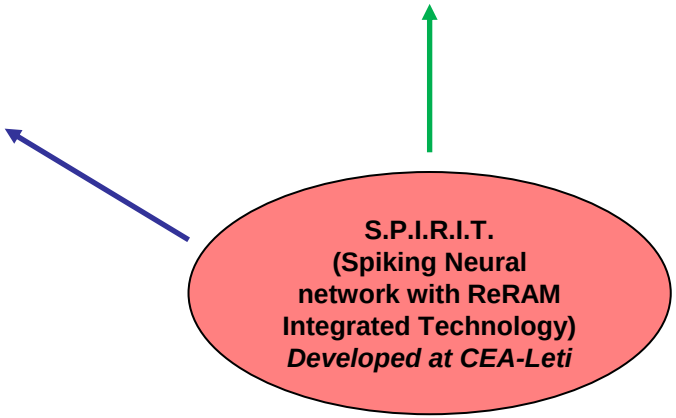
Motivations:

Data-centric workloads exhibited by Deep Neural Network (DNN) applications require circuit architectures that minimize data transfer.



ReRAM as Possible Candidate:

- Fast access
- High density
- Ease of integration in BEOL
- Good endurance and window
- Limited resistance drift



SPIRIT

Application:

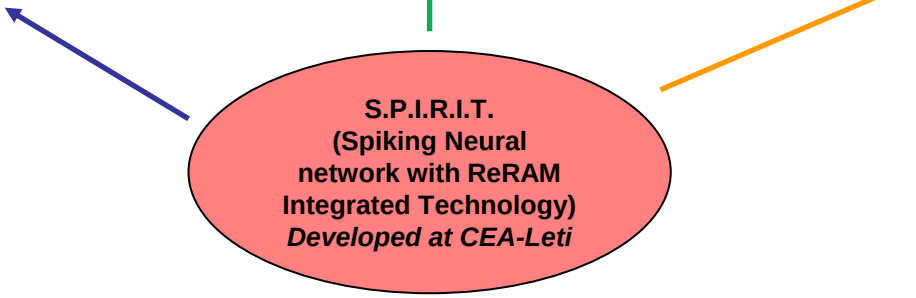
- Hand-written digit classification
- MINST Database

Motivations:

Data-centric workloads exhibited by Deep Neural Network (DNN) applications require circuit architectures that minimize data transfer.

Architecture:

- Spiking Neural Network (SNN)
- Fully connected topology
- 1-neuron/class
- 1 synapse ← 8 ReRAM in //



ReRAM as Possible Candidate:

- Fast access
- High density
- Ease of integration in BEOL
- Good endurance and window
- Limited resistance drift

SPIRIT

Motivations:

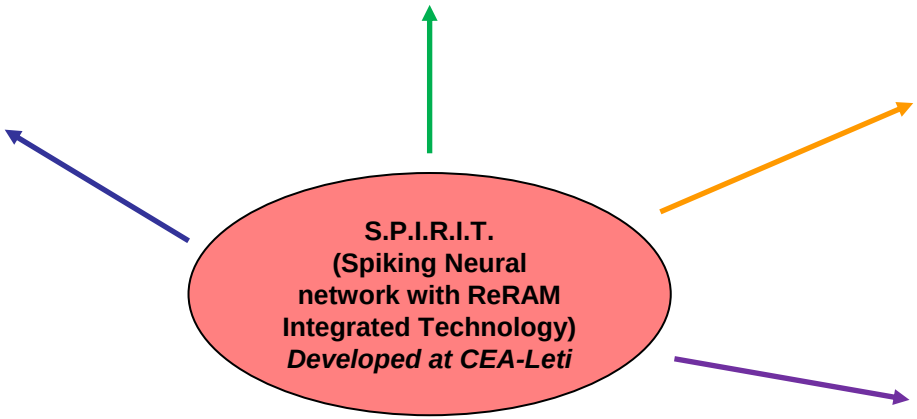
Data-centric workloads exhibited by Deep Neural Network (DNN) applications require circuit architectures that minimize data transfer.

ReRAM as Possible Candidate:

- Fast access
- High density
- Ease of integration in BEOL
- Good endurance and window
- Limited resistance drift

Application:

- Hand-written digit classification
- MINST Database



Architecture:

- Spiking Neural Network (SNN)
- Fully connected topology
- 1-neuron/class
- 1 synapse ← 8 ReRAM in //

Use of ReRAM:

- 1T-1R Structure
- Only 2 states are encoded (HRS/LRS)
- Analogue weights are obtained with more cells in parallel

SPIRIT

Motivations:

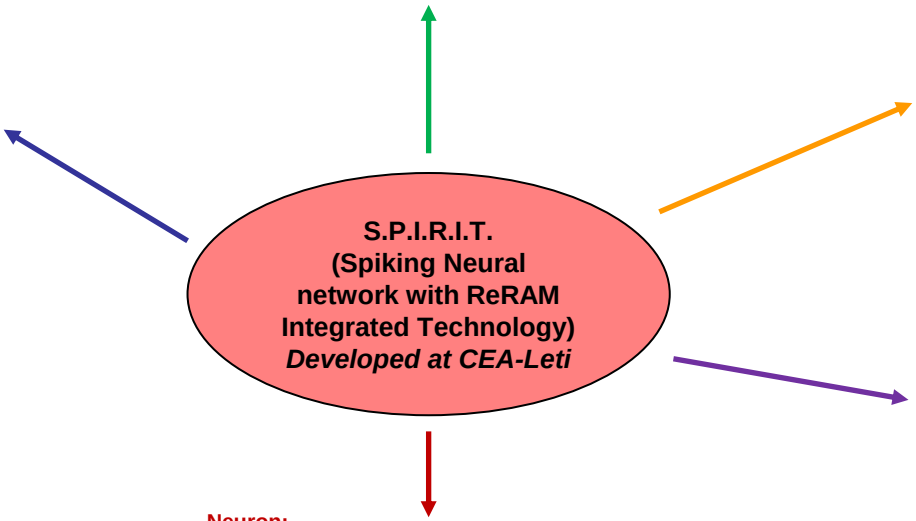
Data-centric workloads exhibited by Deep Neural Network (DNN) applications require circuit architectures that minimize data transfer.

ReRAM as Possible Candidate:

- Fast access
- High density
- Ease of integration in BEOL
- Good endurance and window
- Limited resistance drift

Application:

- Hand-written digit classification
- MINST Database



Architecture:

- Spiking Neural Network (SNN)
- Fully connected topology
- 1-neuron/class
- 1 synapse ← 8 ReRAM in //

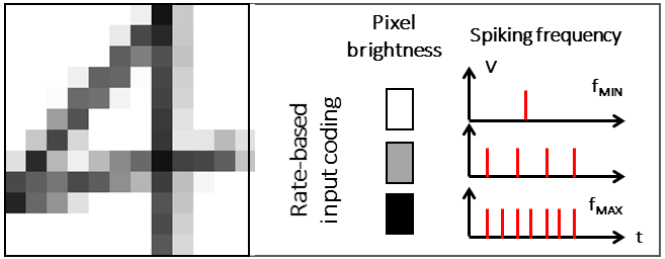
Use of ReRAM:

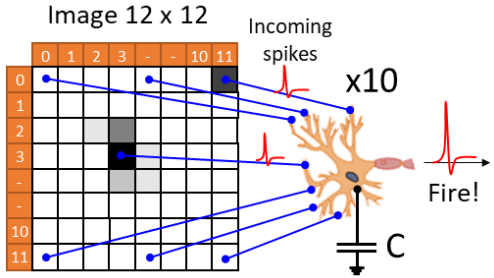
- 1T-1R Structure
- Only 2 states are encoded (HRS/LRS)
- Analogue weights are obtained with more cells in parallel

Neuron:

- Integrate and Fire Analog Neuron

Workflow





User input

Image conversion

- Image cropped and downscaled to 12 x 12 pixels
- Greyscale converted in spike frequency

Recognition

- Spikes are sent to neurons (fully connected topology)
- Neurons integrate weighted spikes
- Neurons fire if membrane potential threshold is reached

Outline

- Weebit-nano Presentation
- Device Description
- Initial Resistance Tuning
- Memory Window Optimization
- SiO_X-based ReRAM Neuromorphic Application
- Conclusions

Conclusions

- Experimental results show how just by varying 3 parameters namely, O_2 Flow, SiO_X and Ti thickness we can tune the Initial Resistance

- Initial Resistance is a key parameter for the memory behavior after the forming operation and the ability to tune it allow us to easily optimize the memory window

- Weebit-nano arrays were successfully used in a SNN architecture for handwritten digit classification

Conclusions

- Experimental results show how just by varying 3 parameters namely, O_2 Flow, SiO_X and Ti thickness we can tune the Initial Resistance
- Initial Resistance is a key parameter for the memory behavior after the forming operation and the ability to tune it allow us to easily optimize the memory window
- Weebit-nano arrays were successfully used in a SNN architecture for handwritten digit classification

Conclusions

- Experimental results show how just by varying 3 parameters namely, O_2 Flow, SiO_X and Ti thickness we can tune the Initial Resistance
- Initial Resistance is a key parameter for the memory behavior after the forming operation and the ability to tune it allow us to easily optimize the memory window
- Weebit-nano arrays were successfully used in a SNN architecture for handwritten digit classification

Conclusions

- SPIRIT Demonstrator with Weebit-nano technology
- Flash Memory Summit, August 2019

- Weebit-nano arrays were successfully used in a SNN architecture for handwritten digit classification

THANK YOU