Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
INFN Istituto Nazionale di Fisica Nucleare

Implementazione di Reti Neurali su FPGA con Memristori per l'Ottimizzazione dell'Elaborazione Dati in ATLAS

Martino Errico, Davide Fiacco , Stefano Giagu, Giuliano Gustavino, Valerio Ippolito, Graziella Russo

davide.fiacco@cern.ch

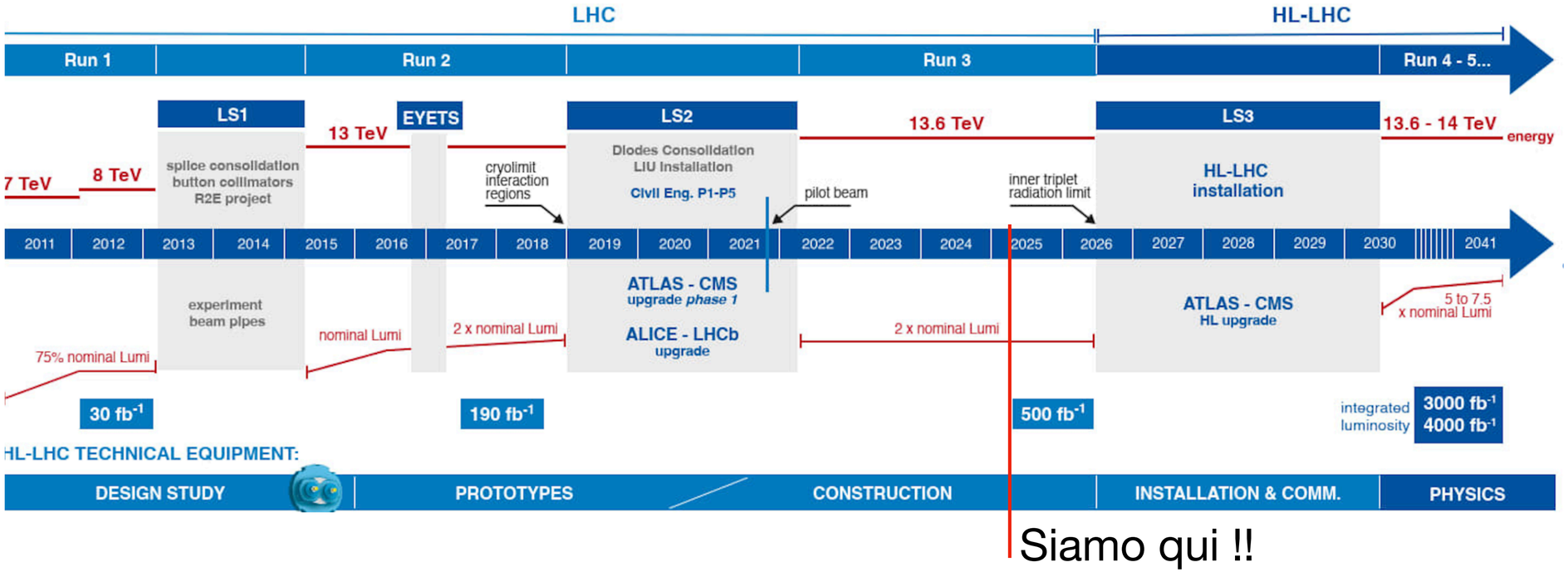
25/03/2025

ATLAS EXPERIMENT

SAPIENZA UNIVERSITÀ DI ROMA

Sfide legate all'Upgrade di Alta Luminosità di LHC

Dal 2026 inizierà un periodo di ~4 anni per l'upgrade di LHC per la fase di Alta Luminosità (HL) in cui la luminosità istantanea verrà portata fino al valore di $7.5 \cdot 10^{34} cm^{-2} s^{-1}$

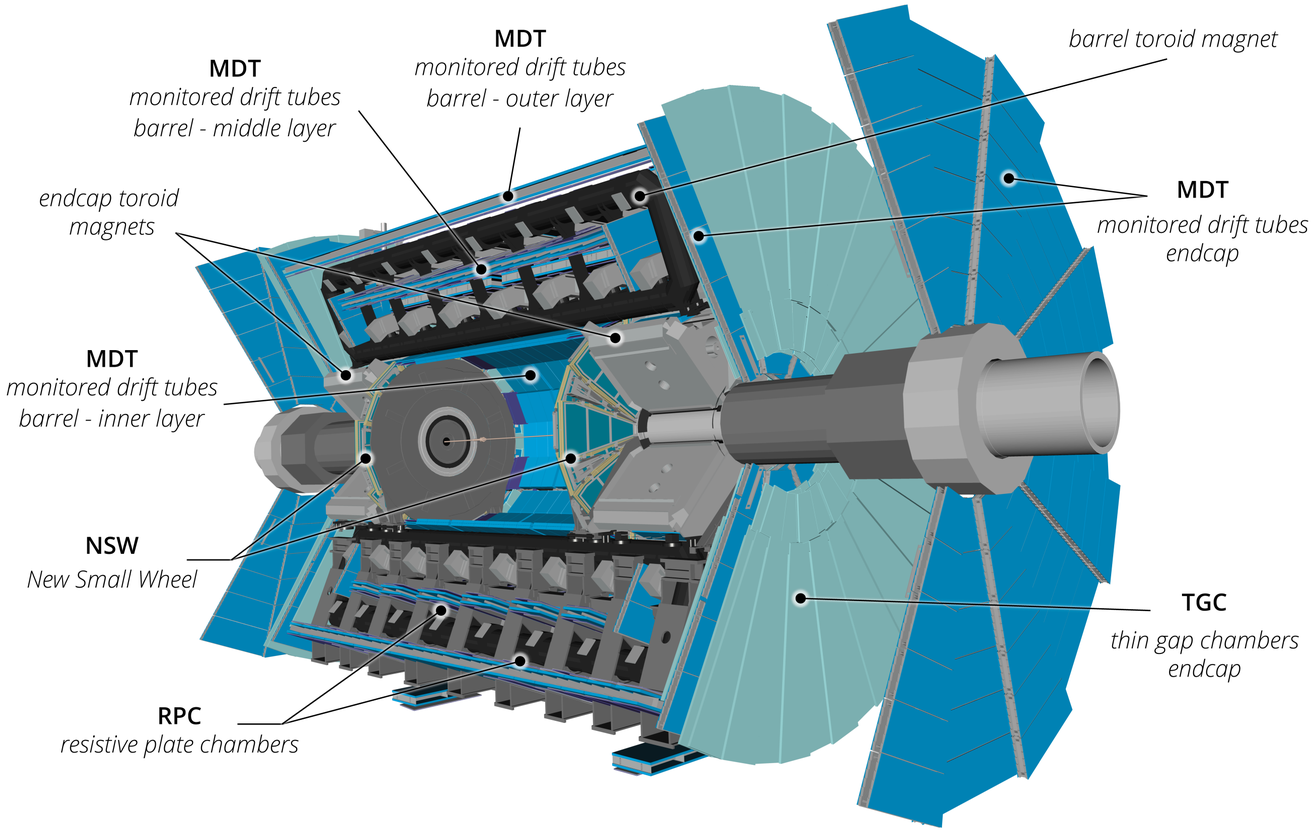


Siamo qui !!

Il detector di ATLAS dovrà essere in grado di selezionare eventi di fisica interessante da un pile-up di circa 200 eventi, il rate del Trigger L1 passerà da $100kHz$ a $1MHz$

Spettrometro a Muoni: Run3

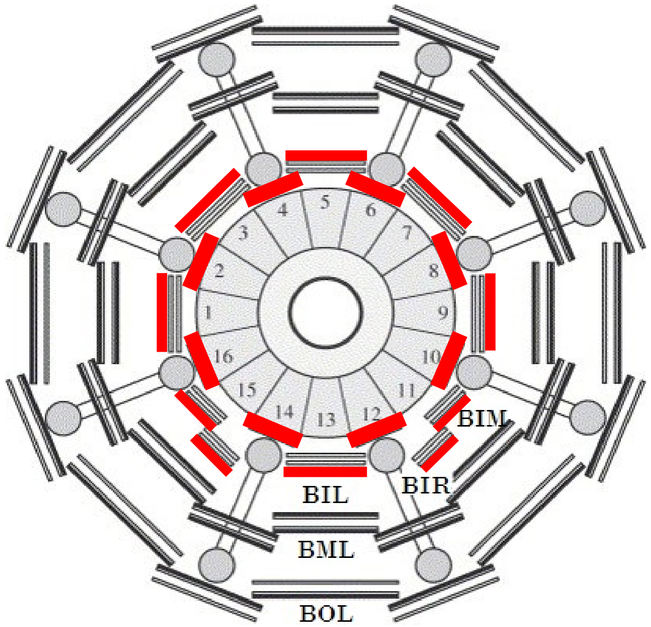
Lo spettrometro a muoni è situato all'esterno dei calorimetri di ATLAS, ed è composto da



- MDT (Monitored Drift Tubes):per il tracciamento preciso (40 μm risoluzione spaziale)
- RPC (Resistive Plate Chambers): misura rapida dei tempi di volo (10 ns risoluzione temporale)
- Micromegas, sTGC e TGC (Thin Gap Chambers): per la regione ortogonale alla traiettoria di fascio (coordinata z)

Spettrometro a Muoni: Upgrade per HL

Upgrade Hardware: Implementazione di nuovi RPC nella strato più interno dello Spettrometro



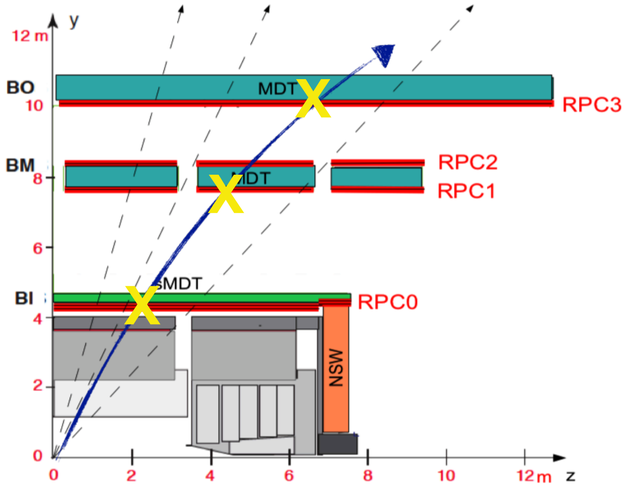
BI = Barrel Inner

BI = Barrel Outer

Trigger L0 : 3 segnali in coincidenza dagli RPC, o due dalle stazioni BI e BO (accettanza da ~70% a 96% [link])

Upgrade Firmware: Algoritmi ultrarapidi di ricostruzione di pT

Obbiettivo: ricostruzione in tempi < 400 ns



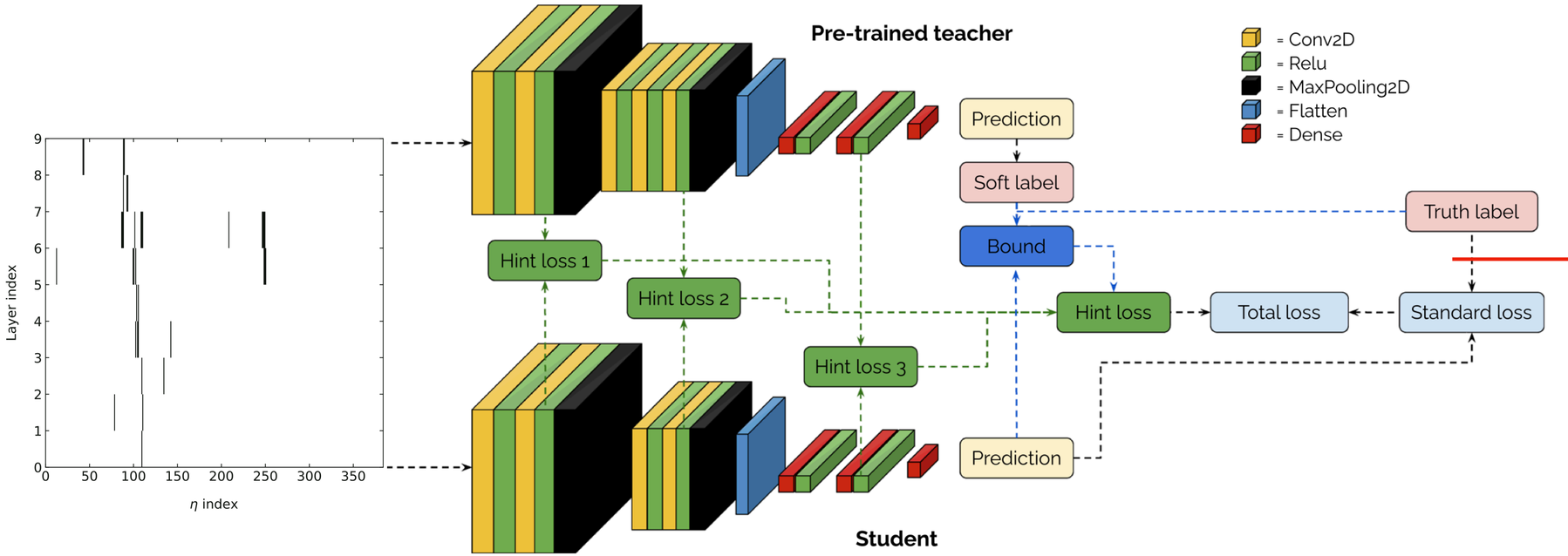
Algoritmi Tradizionali sono veloci ma poco flessibili

Algoritmi di Intelligenza Artificiale sotto studio

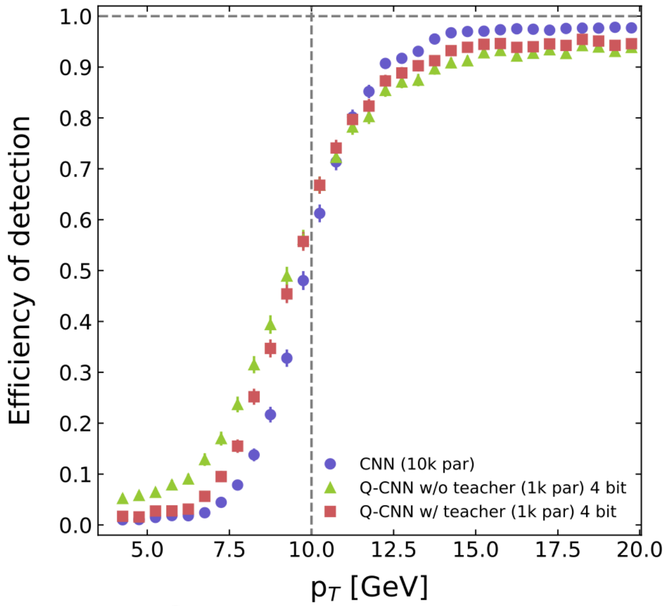
Implementazione su FPGA: CNN

Algoritmi di IA in sviluppo: 1) rete neurale Convoluzionale (CNN), 2) Rete basata su grafi (GNN)

Utilizzate tecniche di Knowledge Distillation e Quantizzazione dei parametri della rete per ridurne la dimensione e la latenza, recuperando parte delle prestazioni



[link]



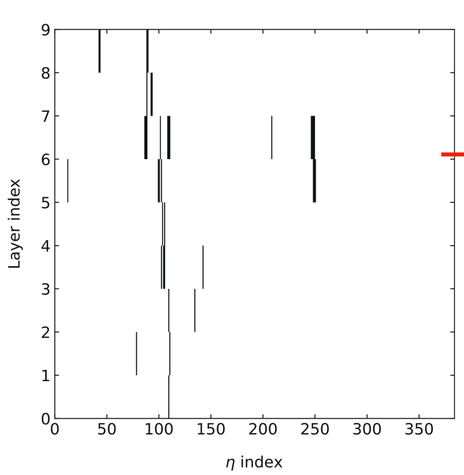
La versione per singolo muone è già stata implementata su FPGA

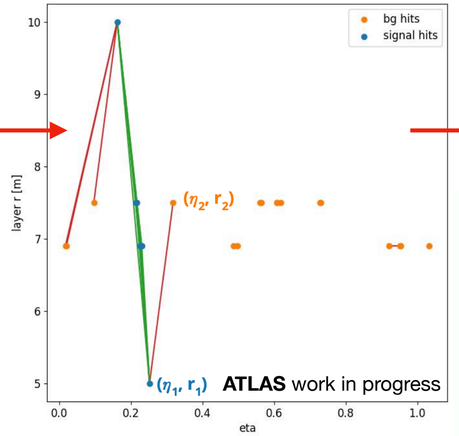
Maggiori informazioni nel poster di Martino [link] e nell'articolo [link]

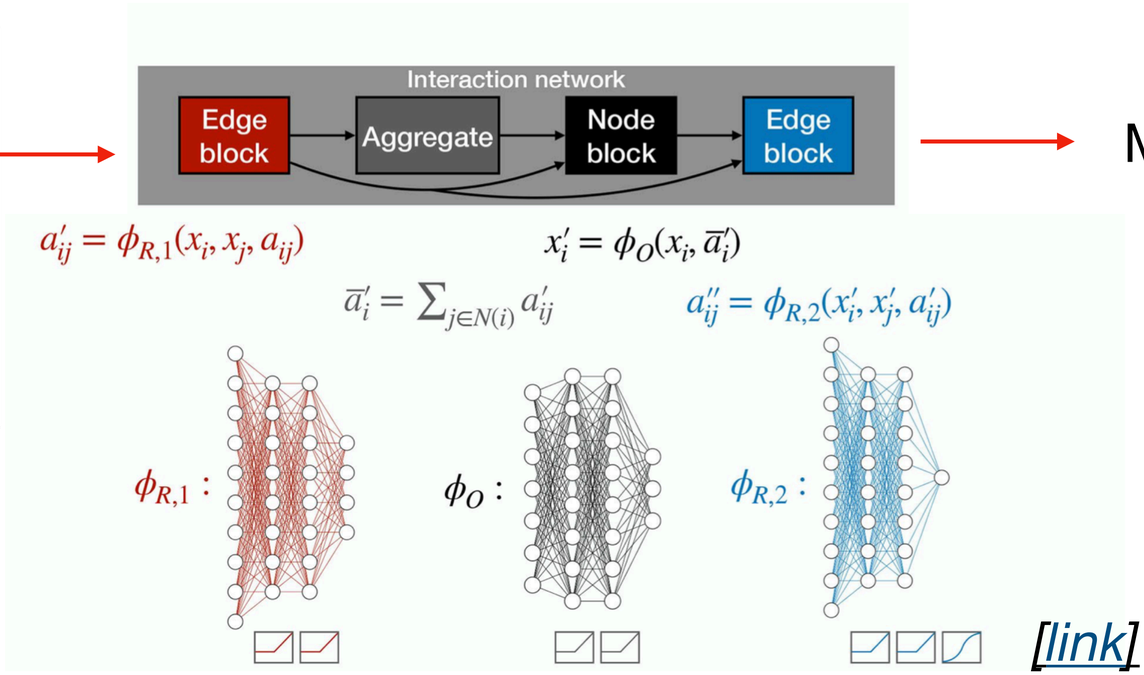
Implementazione su FPGA: GNN

Algoritmi di IA in sviluppo: 1) rete neurale Convoluzionale (CNN), 2) Rete basata su grafi (GNN)

La natura sparsa dell'input lo rende più adatto ad una rappresentazione a grafi che a griglia, usata invece come input per la CNN







Muon p_T

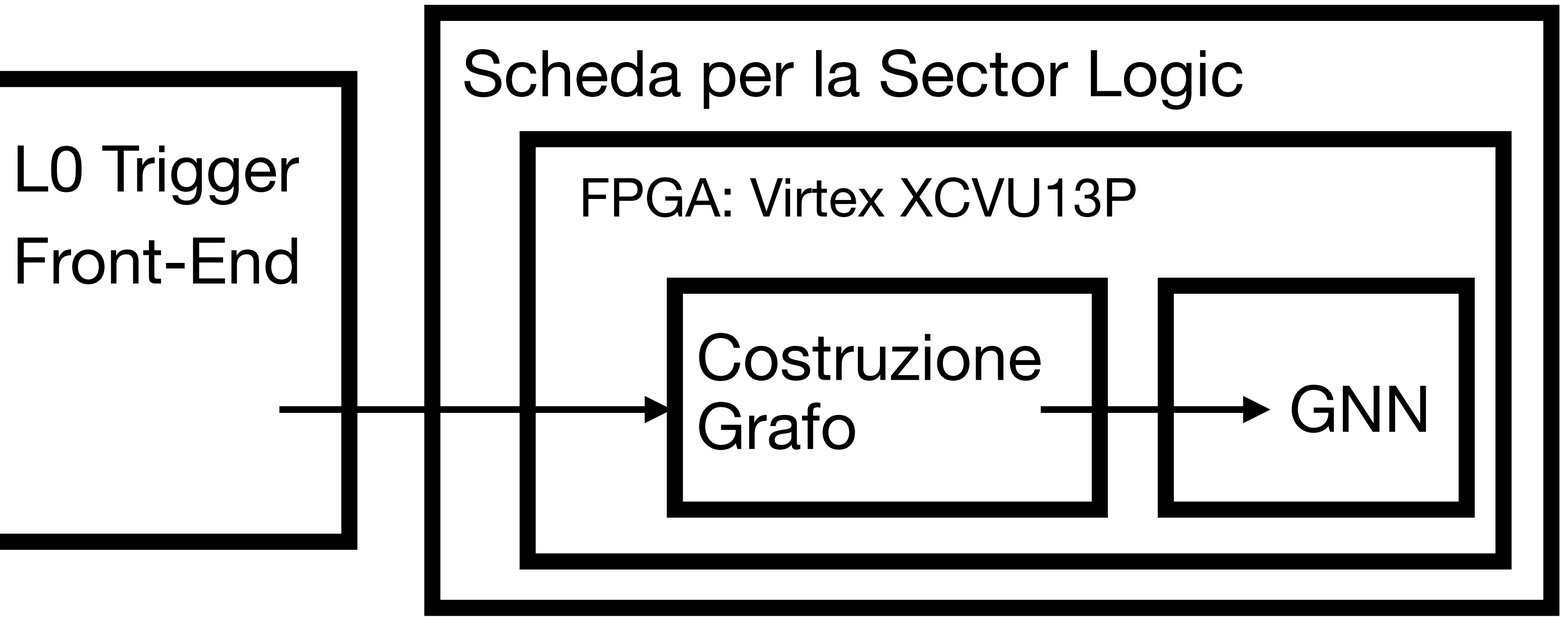
[link]

Maggiori informazioni nel poster di Martino [link]

Implementazione su FPGA: GNN

Sarà scelto l'algoritmo con la miglior performance

L'algoritmo sintetizzato dovrà rispettare una frequenza di clock di 320MHz



Strategia di implementazione:

- codice VHDL per la costruzione del grafo a partire dall'input;

- hls4ml per la rete *[link].*

Riguardo la costruzione del grafo, probabilmente si riadatterà la strategia usata in Belle II *[link]*

L’upgrade di ATLAS è solo un esempio di come le reti neurali stiano diventando centrali nei meccanismi di trigger degli esperimenti di fisica delle particelle

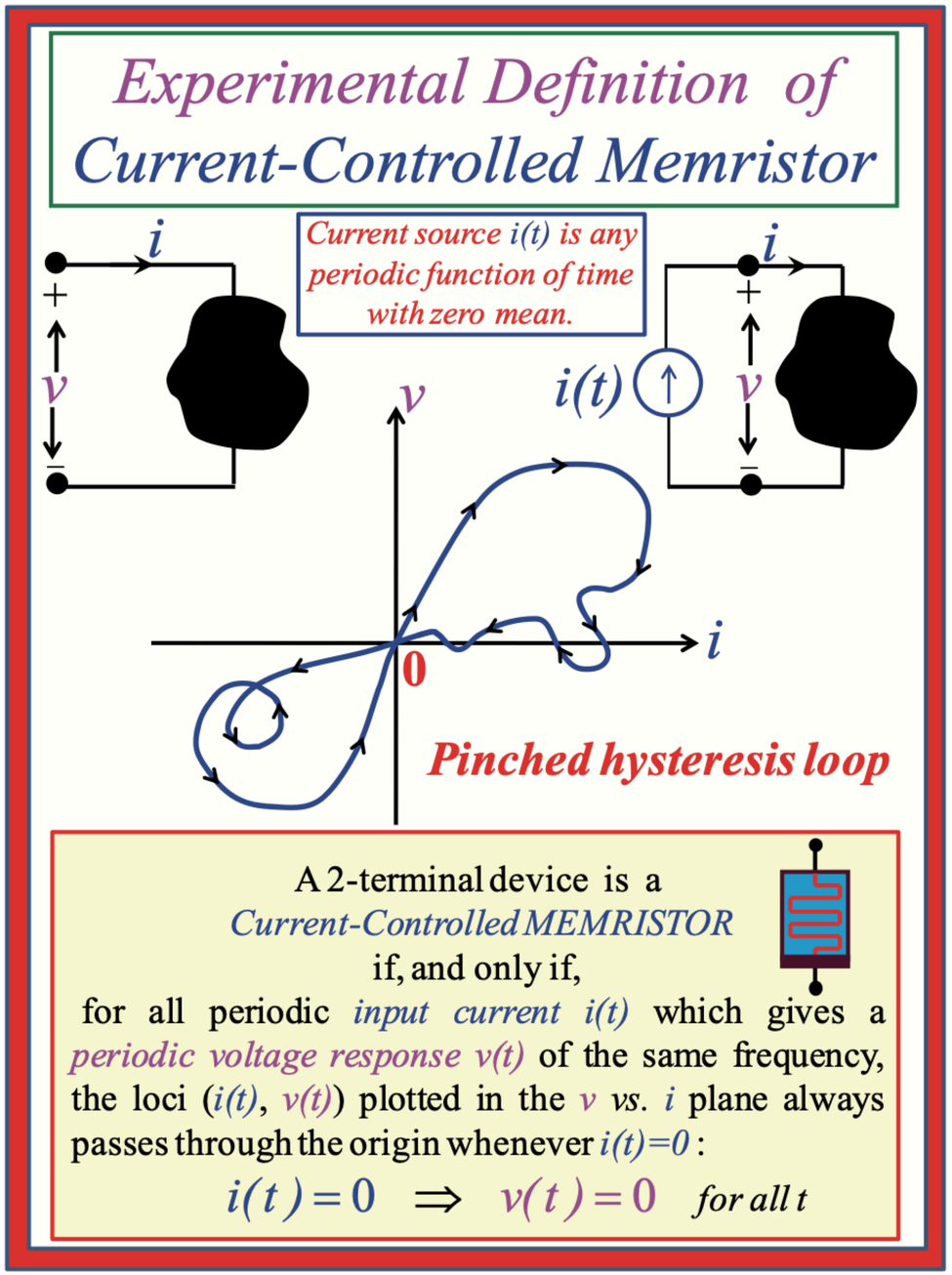
Lo sviluppo di tecnologie neuromorfe permetterà l’implementazione di reti neurali in modo da ottimizzare le performance ed i consumi energetici

Progetto di Ricerca & Sviluppo: PRIN BEEM

“ Scopo del progetto è uno studio di fattibilità di architetture neuromorfe basate sull’utilizzo dei memristori per il sistema di trigger di ATLAS ”

[link]

Cos'è un Memristore



Il Memristore (Memory Resistor) possiede una risposta non lineare che dipende dalla carica accumulata

$$V(t) = M(Q(t)) \cdot I(t)$$

Esistono numerosi modi di costruire un memristore, non è ancora stata trovata la tecnologia in grado di definire uno standard

- Materiali Ferromagnetici
- Chalcogenide Glasses
- Materiali Organici

Perché è interessante: Moltiplicazione Matriciale

Presa una matrice ‘crossbar’ di memristori, e mandato in input un vettore di voltaggi, la moltiplicazione avviene tramite legge di Ohm:

$$V \cdot \frac{1}{R} = I$$

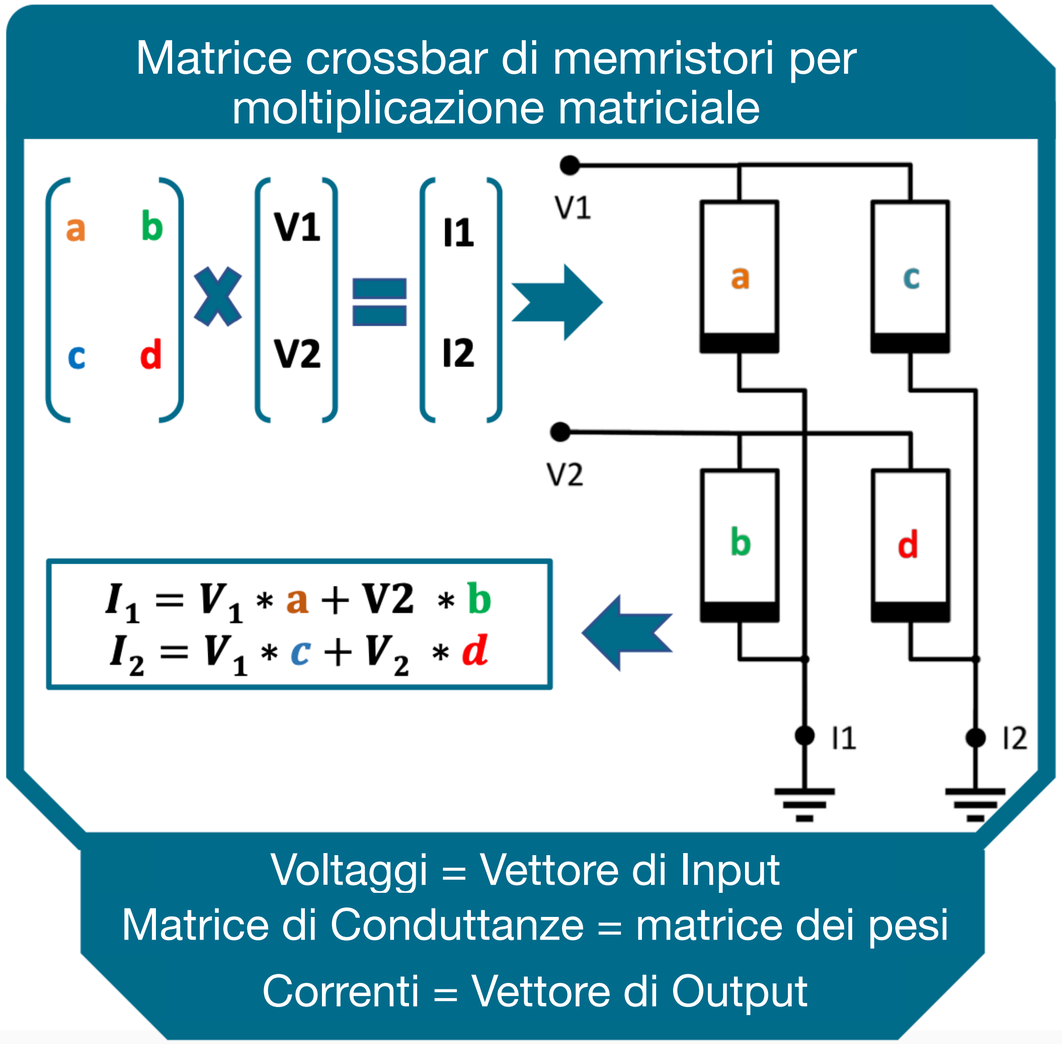
E la somma tramite legge di Kirchoff

$$I_f = I_1 + I_2$$

Il vettore delle correnti in uscita è il risultato !!

Vantaggi:
- Maggiore Efficienza energetica (~ 1 ordine di grandezza per moltiplicazione matriciale tra memristori ed FPGA *[link]*);
- Minore tempo di comunicazione (architetture non-Von Neumann);

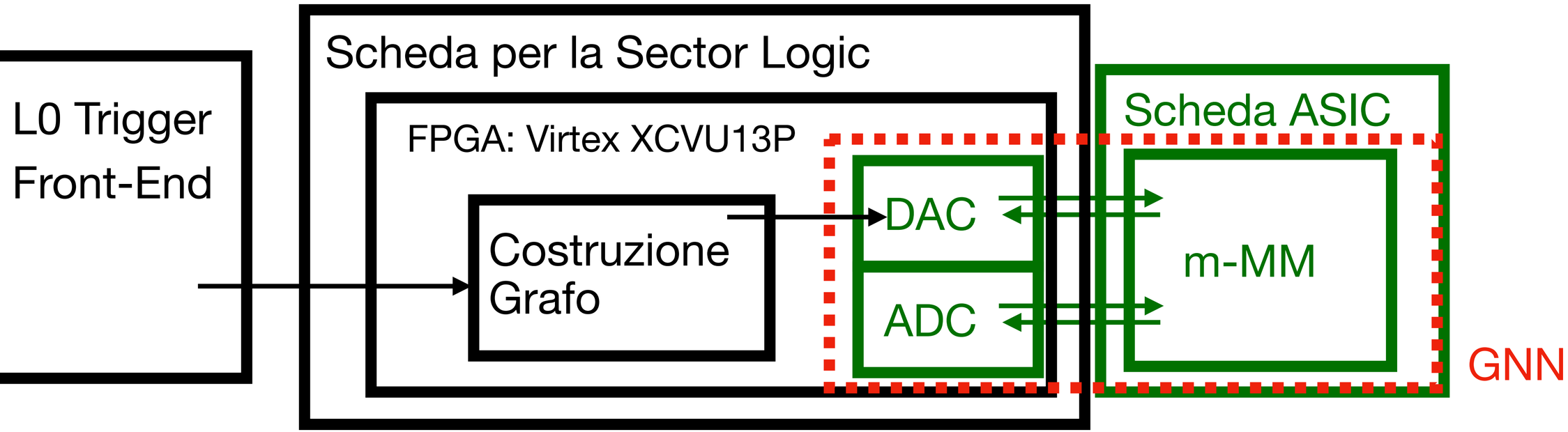
Svantaggi:
- Maggiore errore analogico;
- Assenza di uno standard.



Esempio di applicazione: Memristor Hardware-Friendly Reinforcement Learning *[link]*

Utilizzo in ATLAS e in fisica delle particelle

La Matrice di memristori può svolgere le operazioni di moltiplicazione matriciale delle reti neurali del trigger, velocizzandole



Tuttavia è necessario eseguire le operazioni di conversione Digitale Analogico in tempi ultrarapidi per evitare di annullare il guadagno ottenuto

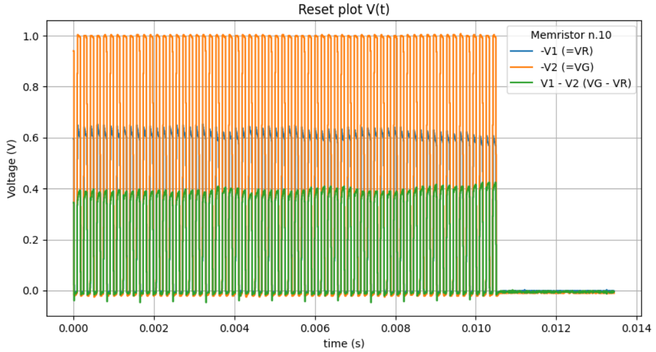
Riguardo le funzioni di attivazione della rete sono sotto studio diversi approcci, firmware o hardware

Programmazione di un memristore

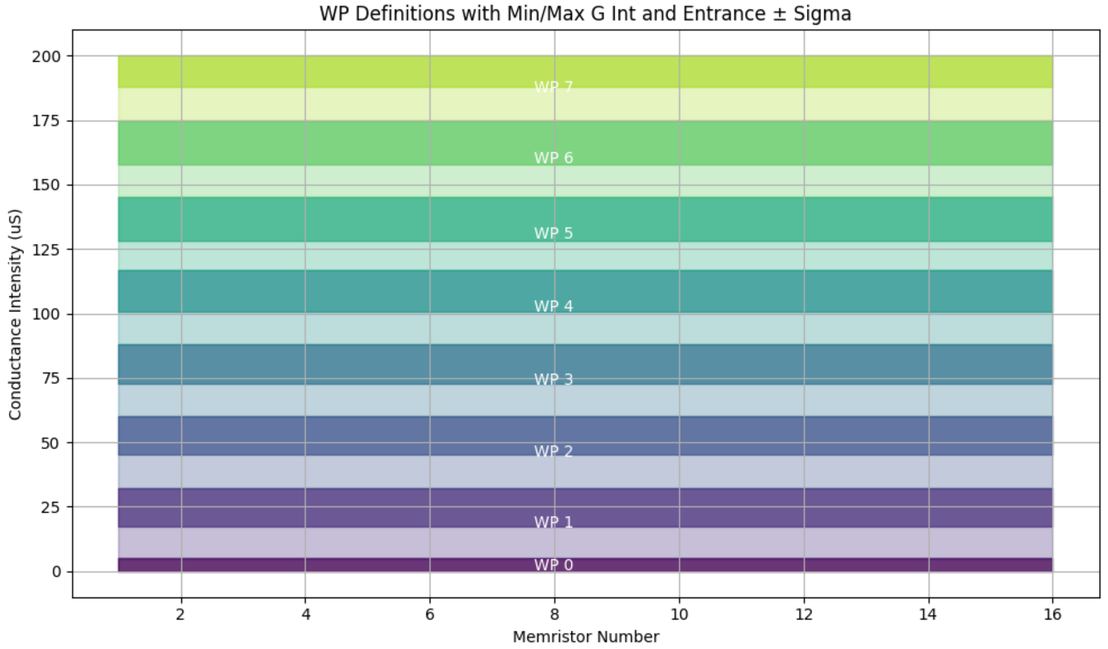
Aspetto chiave del progetto è la capacità di controllare la conduttanza dei memristori

Programmazione: Mandando impulsi di 1V ne aumentiamo la conduttanza (e viceversa)

Lettura: Con segnali nell'intervallo [-0.1, 0.1] V stimiamo la conduttanza senza modificare il memristore



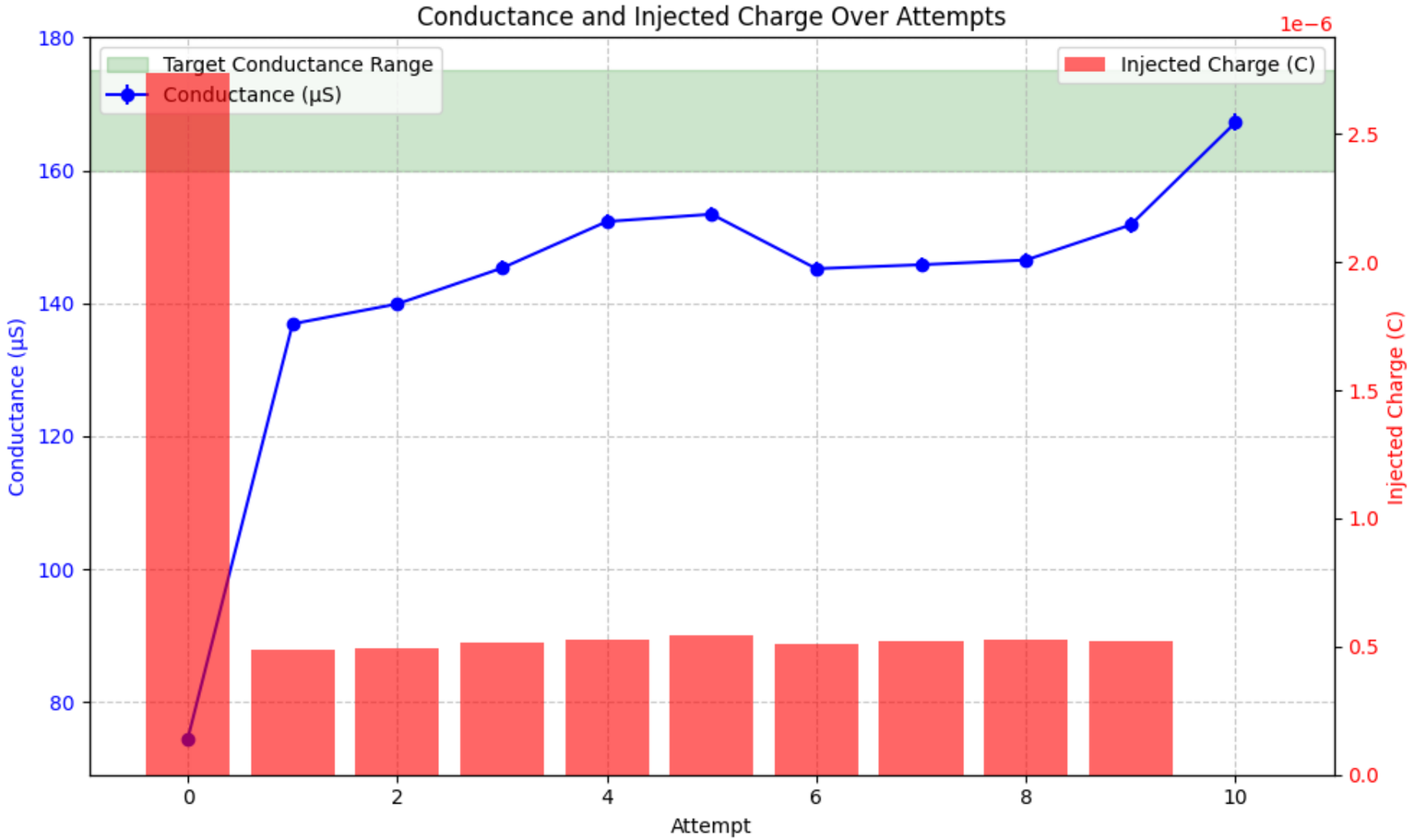
Separiamo l'intervallo utile di conduttanza in sottolivelli di lavoro



Esempio di programmazione

Algoritmo di convergenza: alternando fasi di lettura (punti) e programmazione (barre rosse) facciamo convergere il memristore al punto di lavoro

$$Q = I_{Mem} \cdot T_{1pulse} \cdot N_{pulses} \cdot 0.5$$



*Ottimizzazione del procedimento di programmazione ancora in corso…

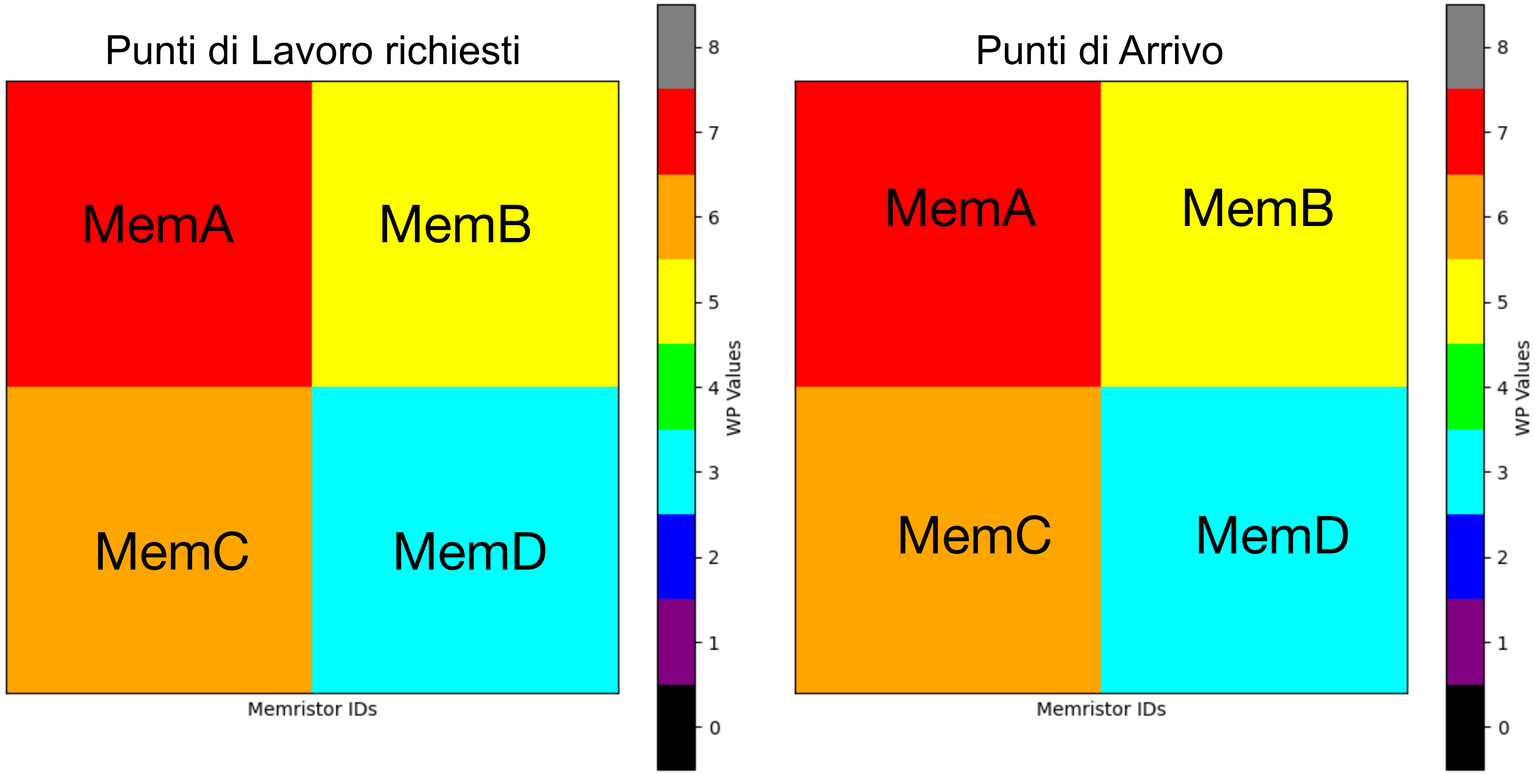
Test dei punti di lavoro

Abbiamo scelto randomicamente per ogni memristore un punto di lavoro (PL), eseguito l'algoritmo di convergenza per un massimo di 40 passi, e confrontato il punto d'arrivo (PA)

Legenda colori della tabella:

- Verde (PA == PL);
- Giallo (PA = PL +- 1);
- Rosso (PA >= PL +- 2)

Memristore	P. Lavoro	P. Arrivo
A	7	7
B	5	5
C	6	6
D	3	3



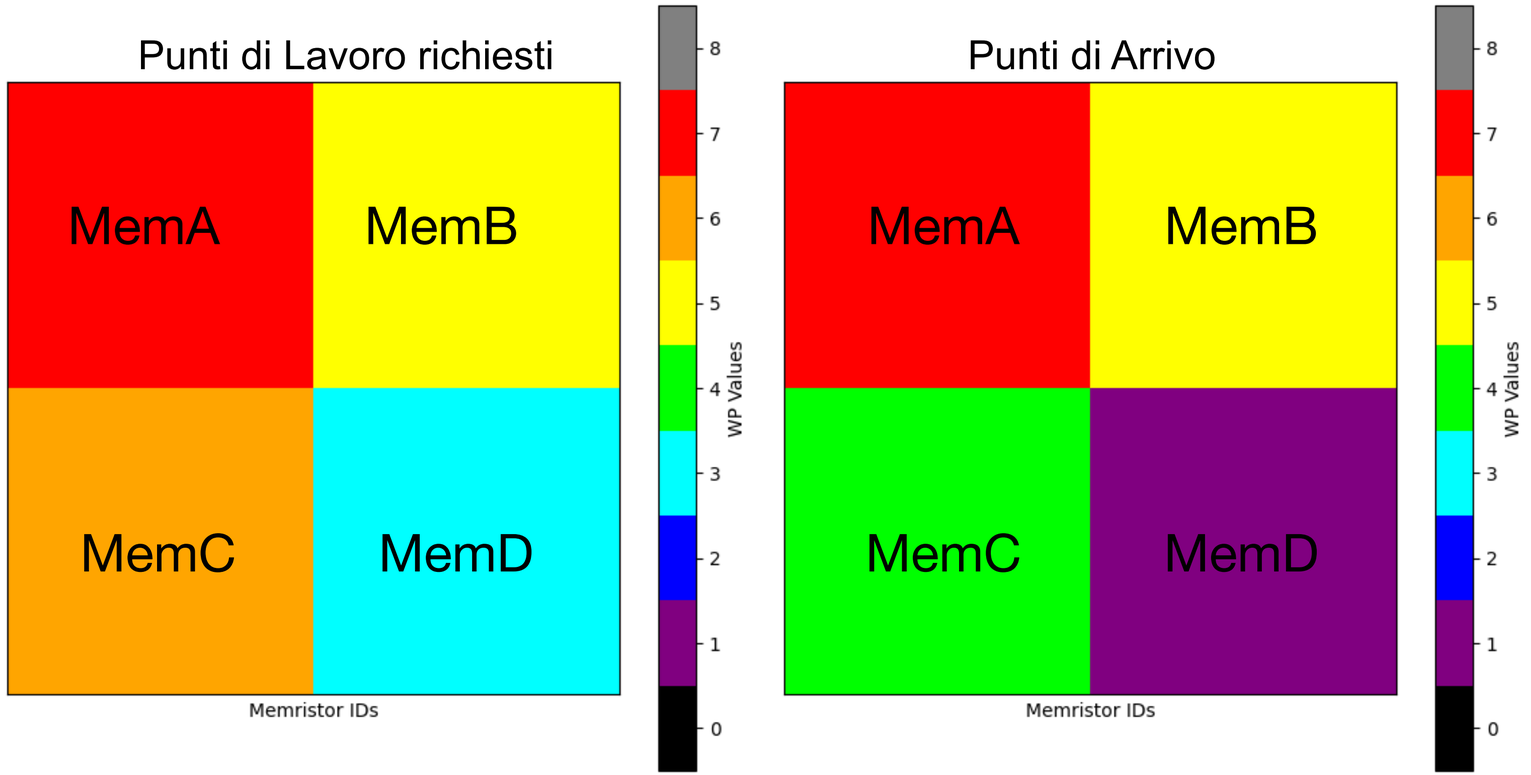
Test dei punti di lavoro: 20 minuti dopo

…dopo 20 minuti, abbiamo ristiamo il PA e riconfrontato

Legenda colori della tabella:

- Verde (PA == PL);
- Giallo (PA = PL +- 1);
- Rosso (PA >= PL +- 2)

Memristore	P. Lavoro	P. Arrivo (20 min dopo)
A	7	7
B	5	5
C	6	4
D	3	1



Conclusioni

Per l'upgrade di HL di ATLAS si stanno testando Reti Neurali Convoluzionali [link] e basate su Grafi [Poster Martino link], cosi come anche la loro implementazione su FPGA, con l'obbiettivo di inferenza ultrarapida O(100ns);

Prossimi Passi:

- Sviluppo GNN;
- Implementazione e studio di tecniche di graph building su FPGA;

È iniziato anche la sperimentazione di architetture neuromorfe basate su memristori come futuri possibili upgrade del DAQ di esperimenti di fisica delle particelle

Prossimi Passi:

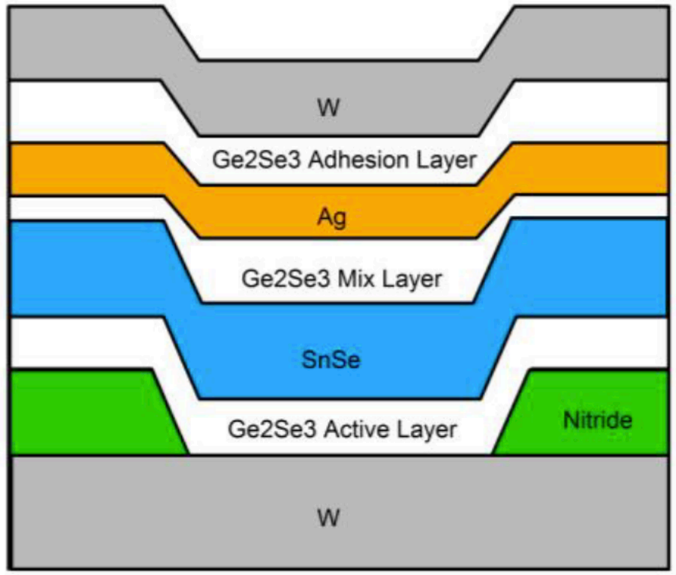
- Ottimizzazione algoritmi di Convergenza;
- Sviluppo e Design della board per la m-MM, e la connessione con l'FPGA;

Grazie dell’attenzione!

Come funzionano: l'esempio della Known

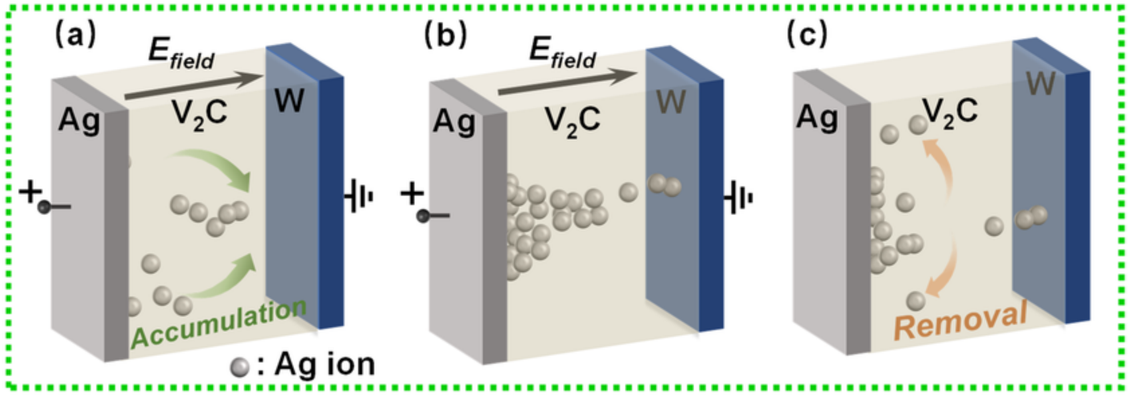
In Laboratorio abbiamo testato una matrice di memristori prodotta dalla compagnia Knowm, sono basati sulla tecnologia Chalcogenide

Rappresentazione qualitativa del memristore Known



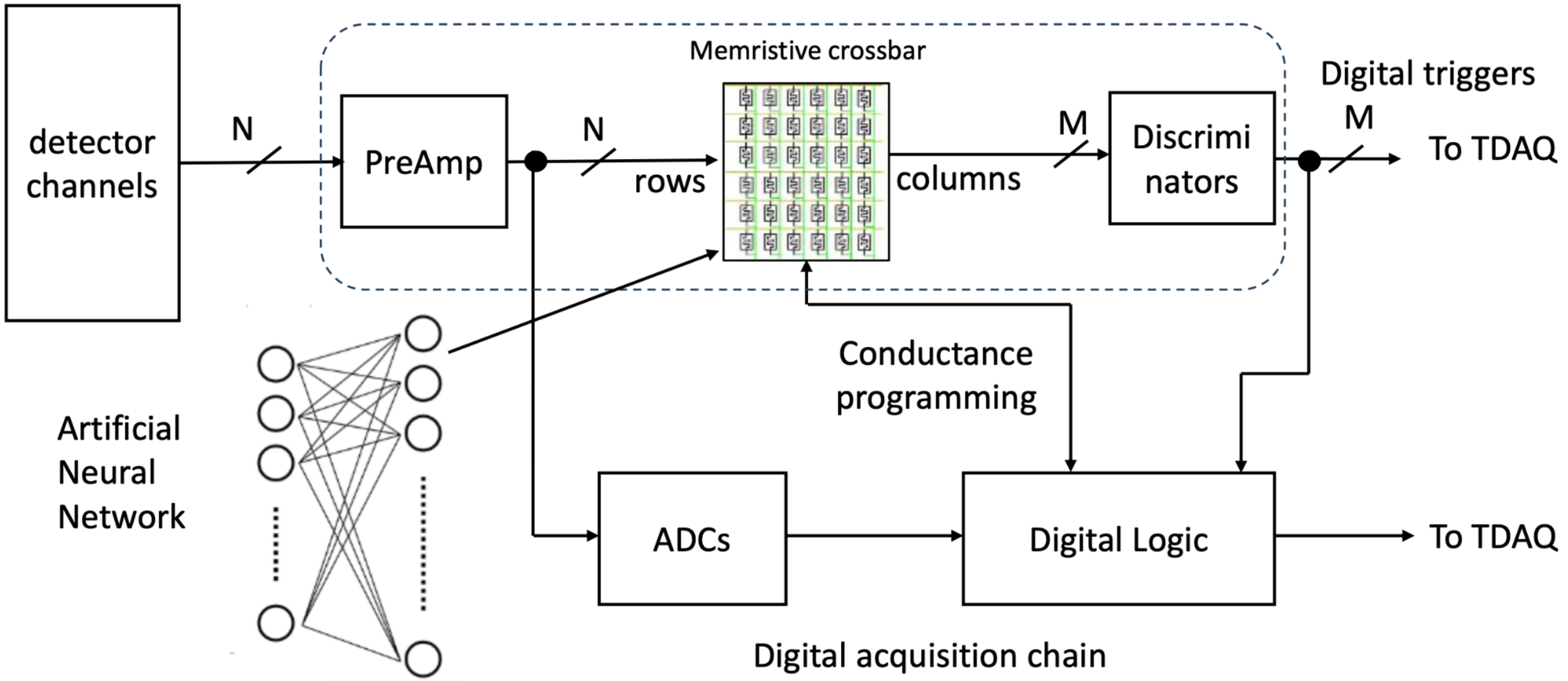
Possiamo controllarne la resistività grazie all'accumulo di Ag^+ nello strato attivo:

- mandando tensioni positive (> soglia 0.25V) aumenta l'accumulo di Ag^+ (==> minore resistività);
- mandando tensioni negative (< soglia -0.2V) diminuisce l'accumulo di Ag^+ (==> maggiore resistività);



Integrazione di un Memristore in un sistema di DAQ

Possiamo compiere l'inferenza della rete neurale del trigger direttamente tramite nela matrice memristiva

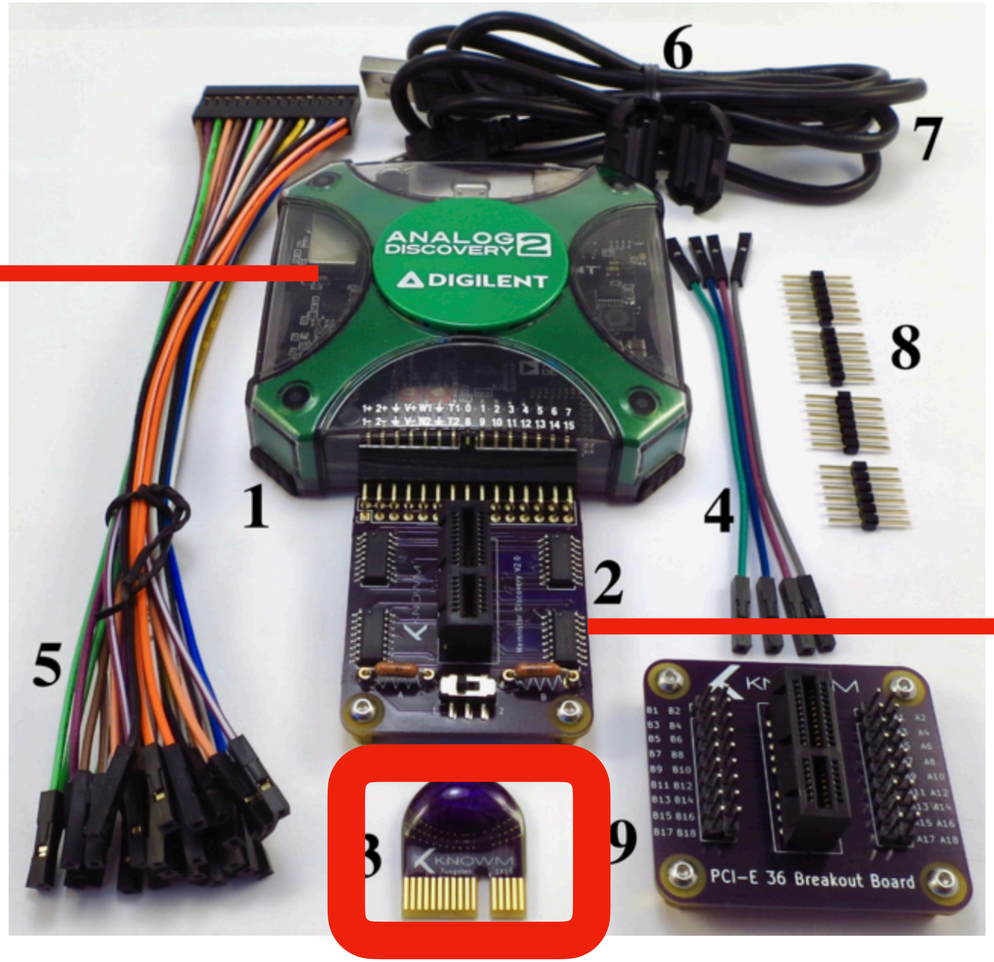


Our Setup

Analog Discovery 2 (AD2) (link)

It allows to send signals to the board and read its behaviour.

By a python package it's possible to control the input/output of the AD2 to customise the experiments (link)



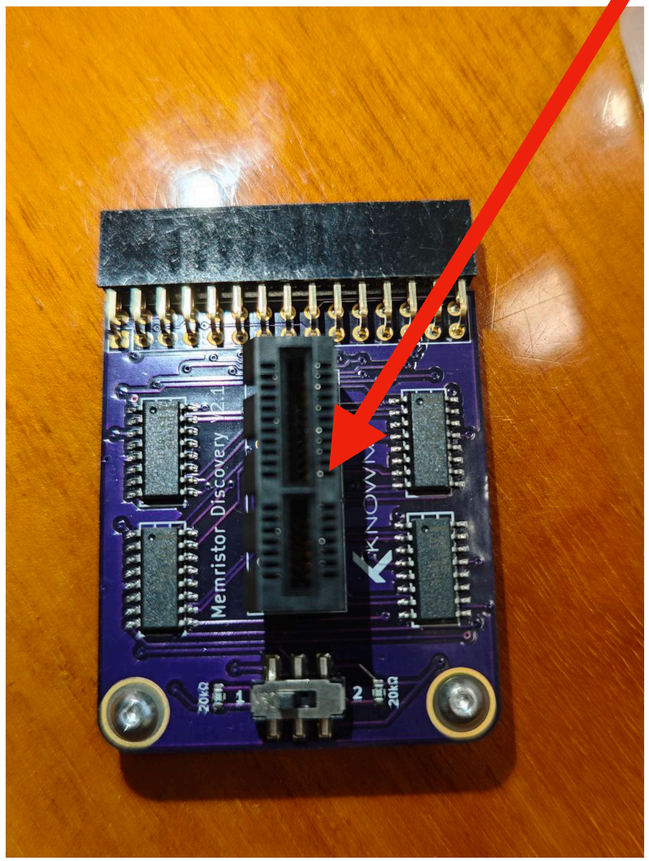
memristor chip KnowM 16 W-SDC

The board

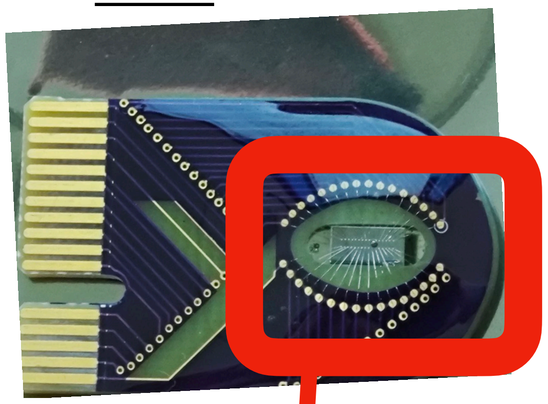
It just contains the switches and the connection to the memristor chip.

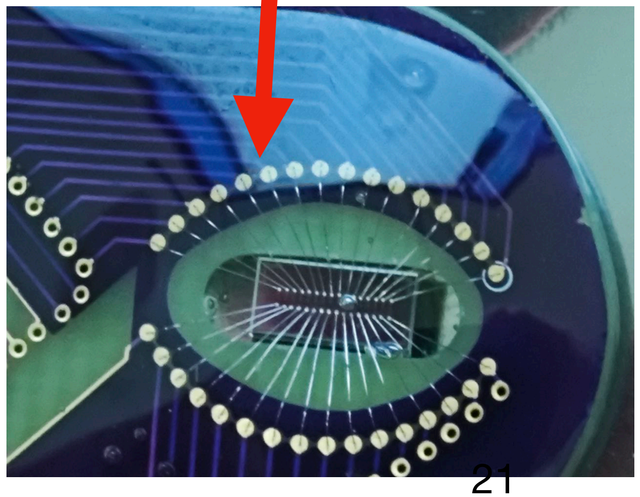
Our Memristor

Our board (link) :

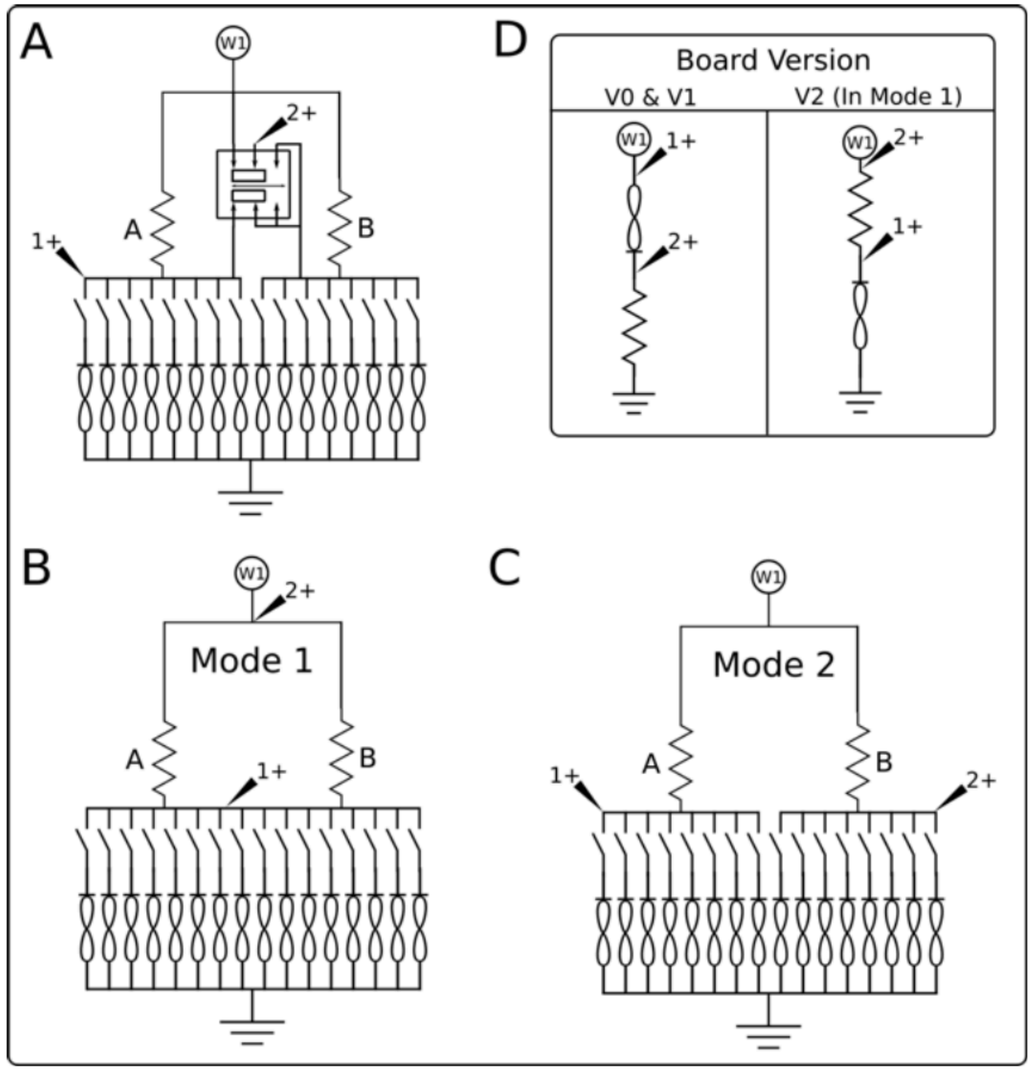


Link



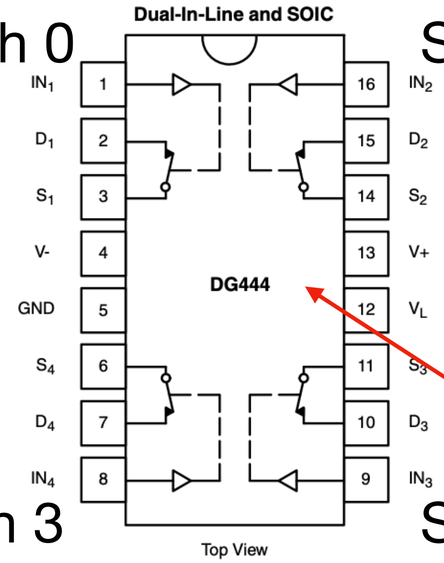


The board circuit:



Switch scheme on the board (mode1)

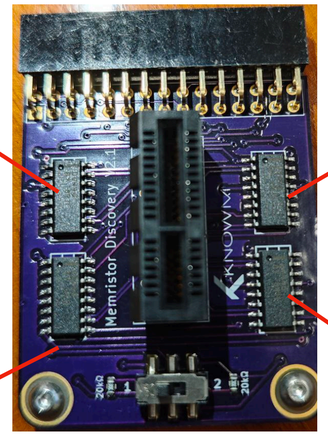
Switch 0



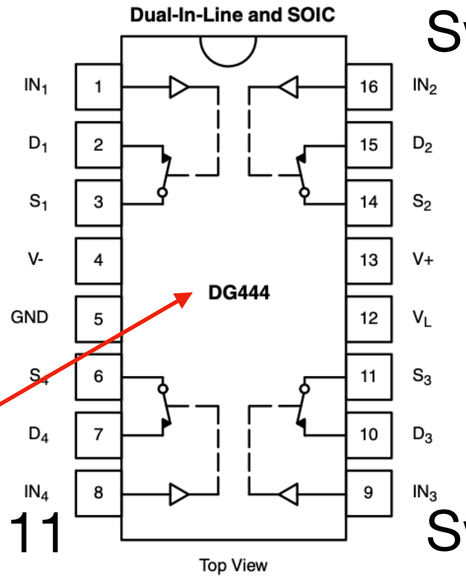
Switch 1

Switch 3

Switch 2



Switch 8

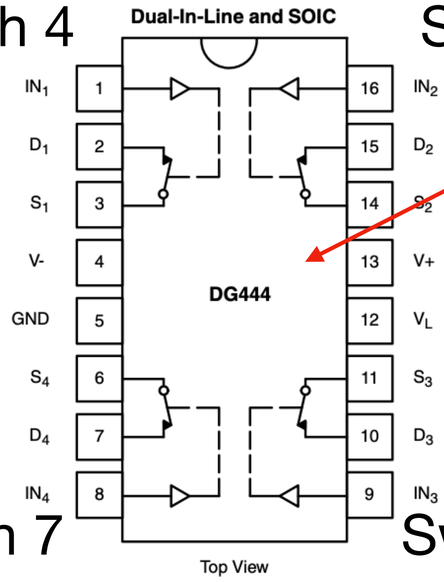


Switch 9

Switch 11

Switch 10

Switch 4

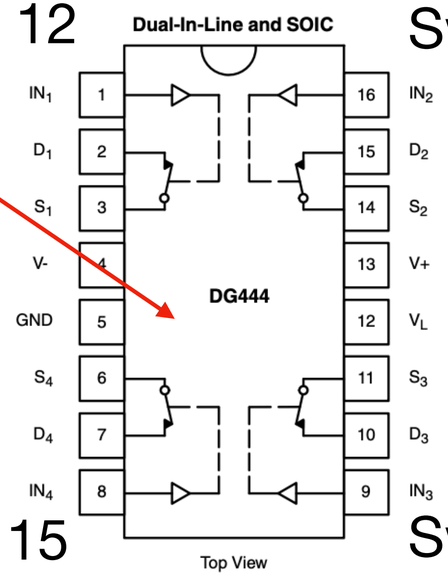


Switch 5

Switch 7

Switch 6

Switch 12



Switch 13

Switch 15

Switch 14

We control each switch from the AD2 pin:

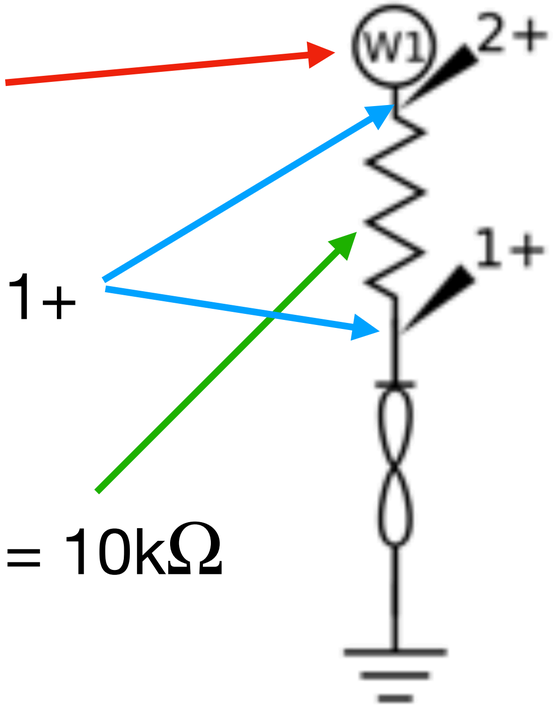
- pin 0 - switch 0
- pin 1 - switch 1
- pin 2 - switch 2
- pin 3 - switch 3
- pin 4 - switch 4
- pin 5 - switch 5
- pin 6 - switch 6
- pin 7 - switch 7
- pin 8 - switch 8
- pin 9 - switch 9
- pin 10 - switch 10
- pin 11 - switch 11
- pin 12 - switch 12
- pin 13 - switch 13
- pin 14 - switch 14
- pin 15 - switch 15

How study our memristor

We control the signal of W1 channel

AD2 has two reading channels: 2+ and 1+

We know the resistance on the board R = 10kΩ
(in Mode 1)

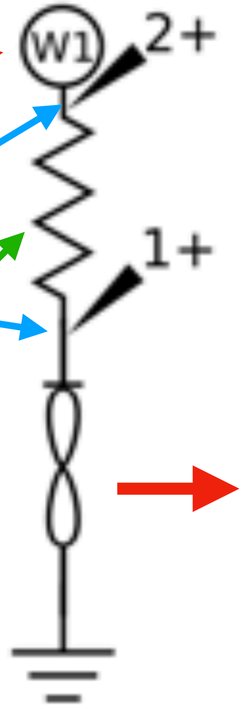


How study our memristor

We control the signal of W1 channel

AD2 has two reading channels: 2+ and 1+

We know the resistance on the board R = 10kΩ
(in Mode 1)



Simbol used by Knowm.

To decrease the resistance (Ag+ in che SDC), the lower potential has to be on the bar, that tells where is the active layer.

But because the memristor is directly connected to the ground:

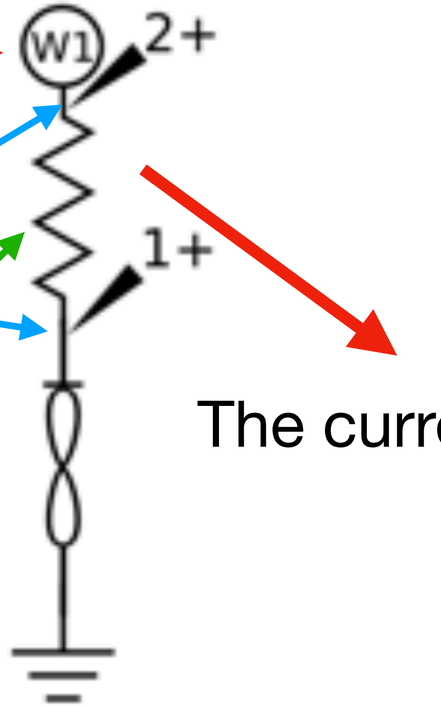
The potential of the memristor is $V_M = -V_1$

How study our memristor

We control the signal of W1 channel

AD2 has two reading channels: 2+ and 1+

We know the resistance on the board R = 10kΩ
(in Mode 1)



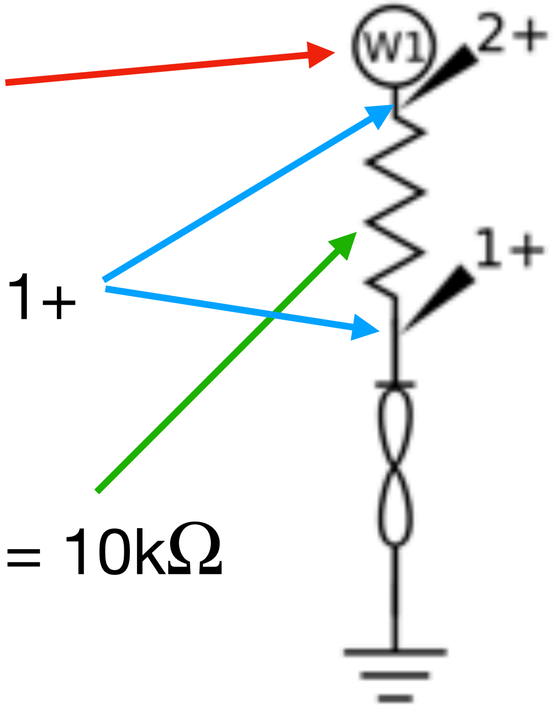
The current is $I = \frac{V_2 - V_1}{R}$

How study our memristor

We control the signal of W1 channel

AD2 has two reading channels: 2+ and 1+

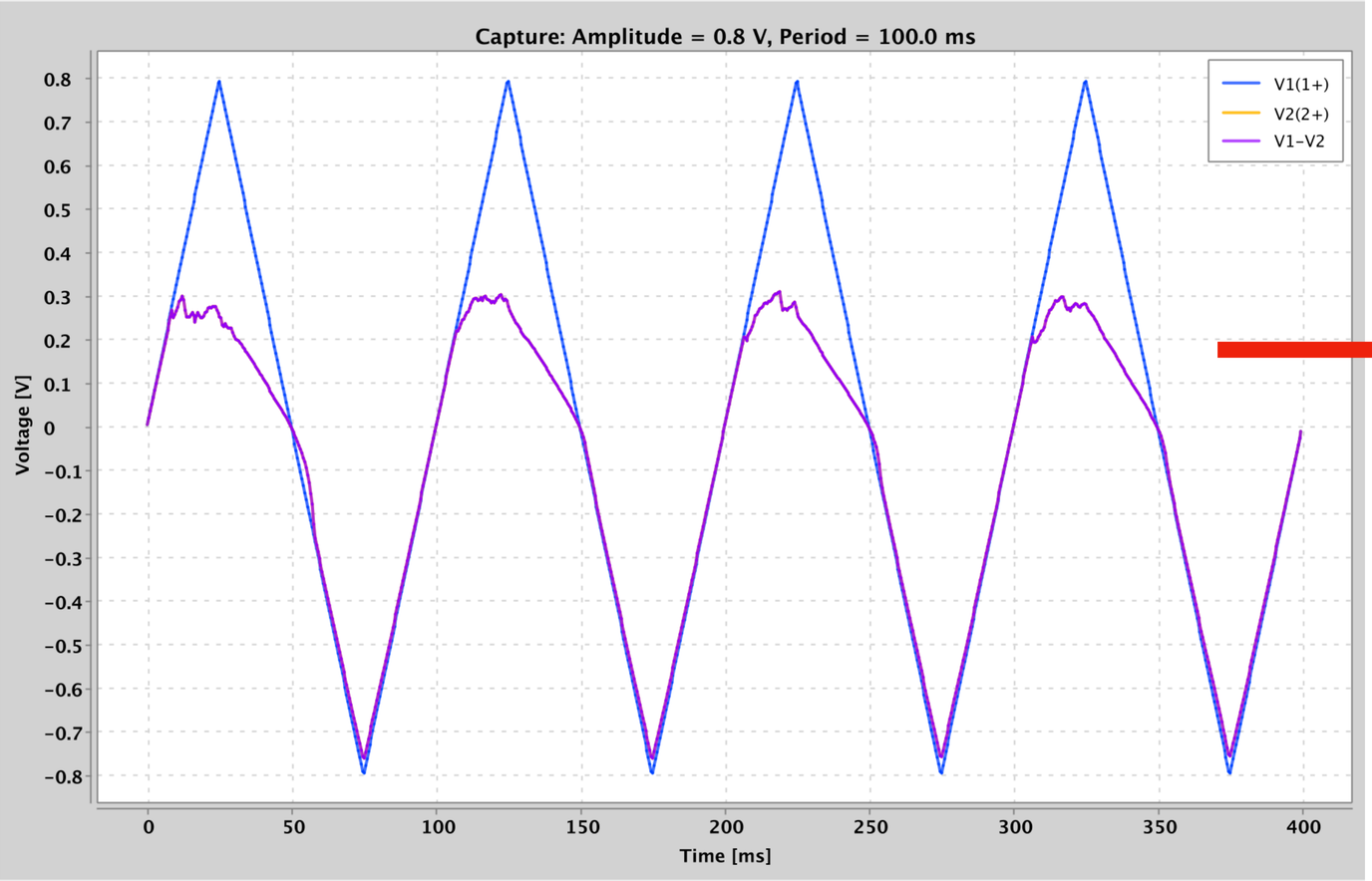
We know the resistance on the board R = 10kΩ
(in Mode 1)



In this way we can relate V_M and I

Response plot

Open all switches except 1 (i.e. we access memristor number 1), and then we send in input a triangular signal.

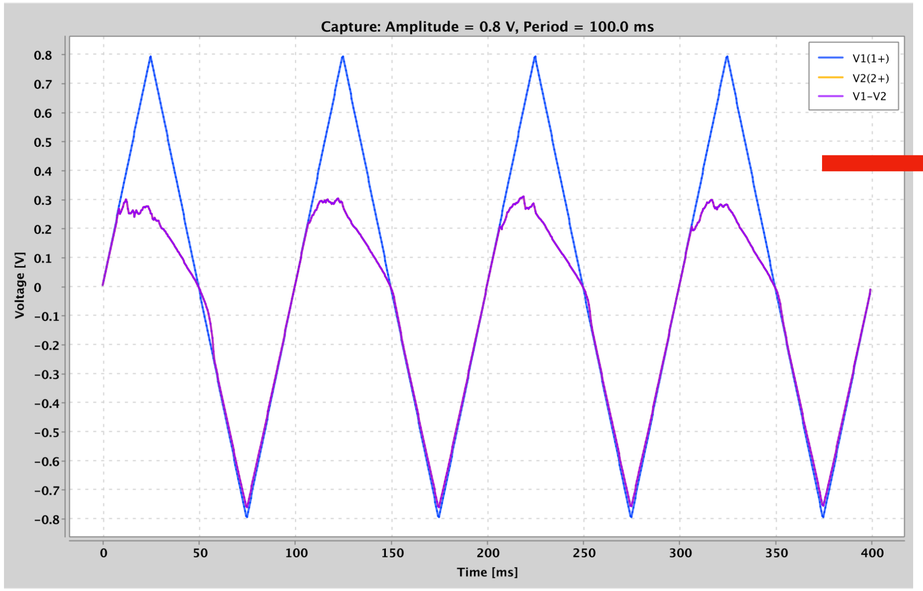


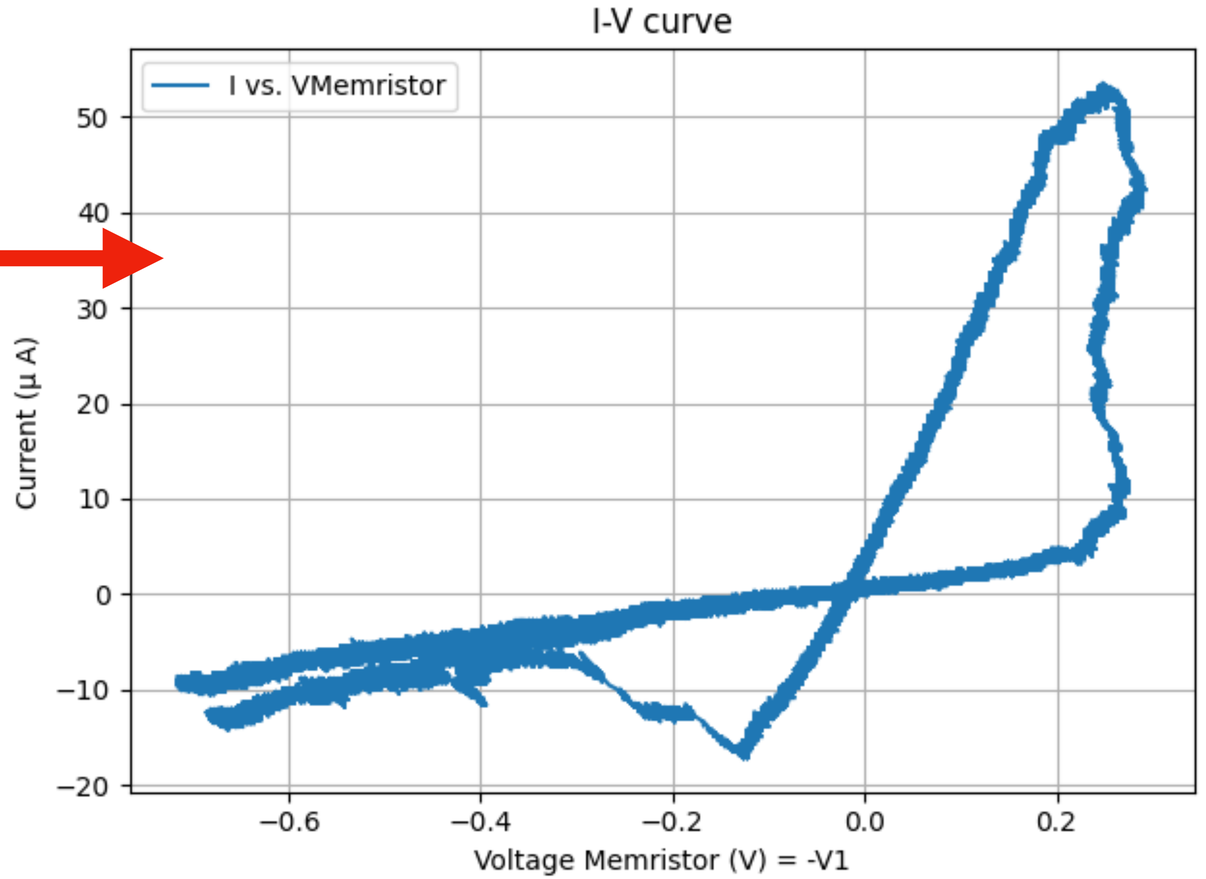
The blue line is input signal,
The purple line is the potential on the memristor

The amplitude is enough to see the change of state of the memristor.

I-V plot

Open all switches except 1 (i.e. we access memristor number 1), and then we send in input a triangular signal.



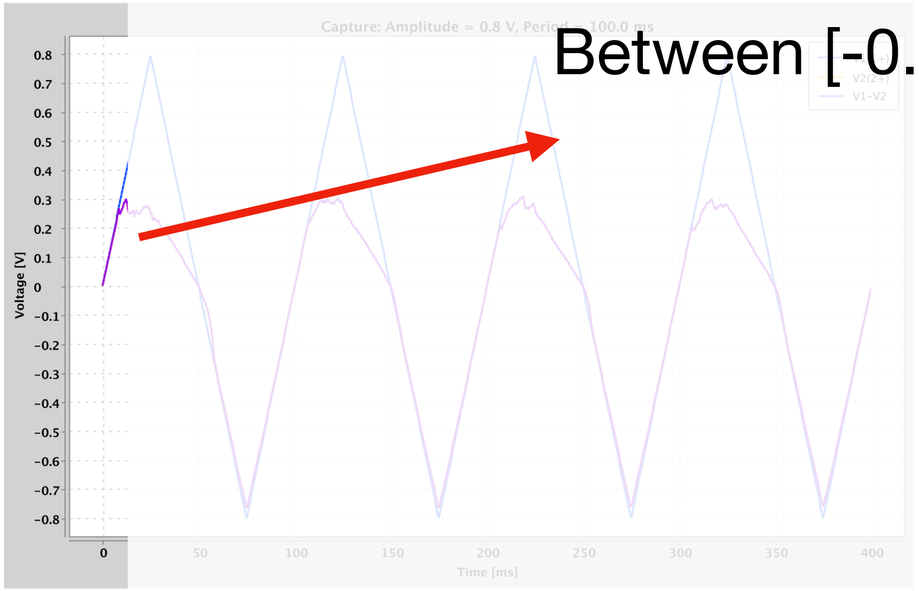


The orange line is the I-V plot of the memristor only,

The blue line is the potential of the memristor+Resistance (10kOhm)

I-V plot

Open all switches except 1 (i.e. we access memristor number 1), and then we send in input a triangular signal.



Between [-0.1, 0.25]V it behaves ~ as a standard resistance

The orange line is the I-V plot of the memristor only,

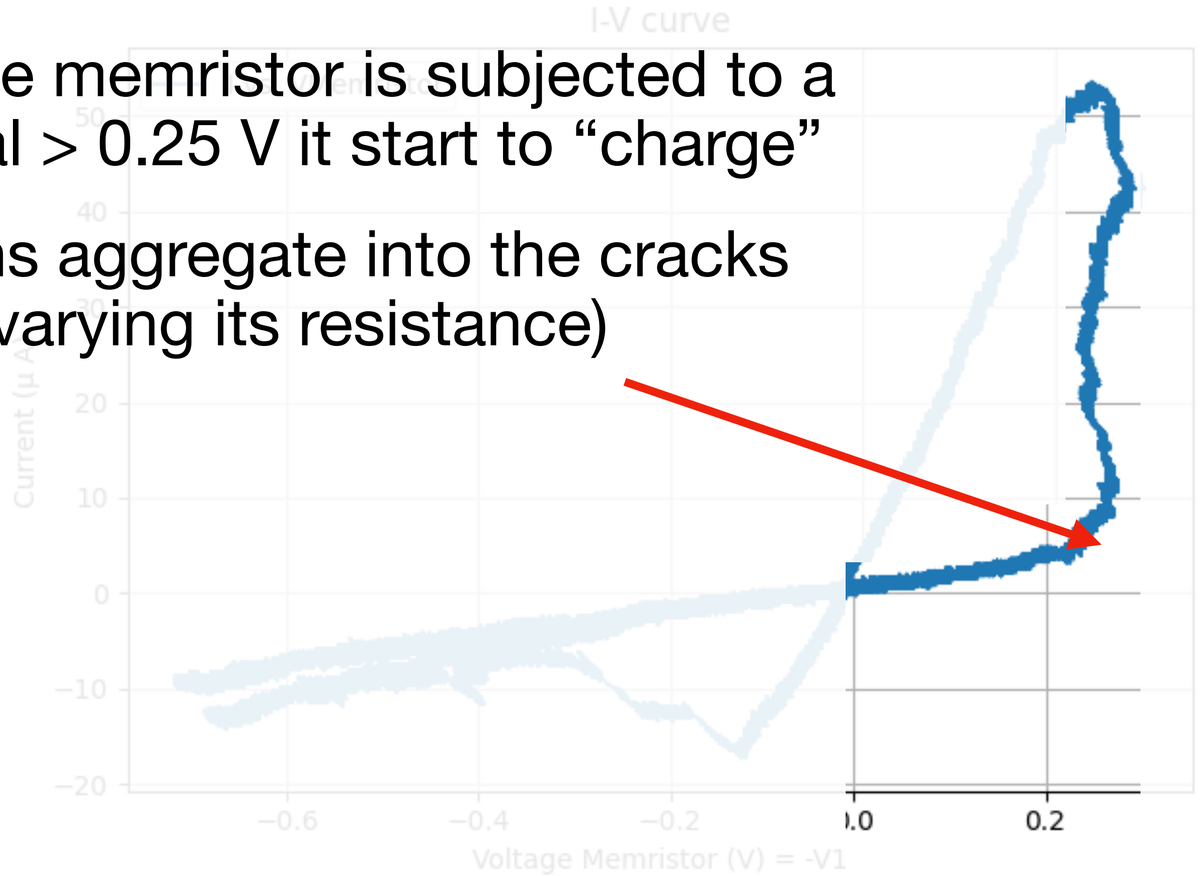
The blue line is the potential of the memristor+Resistance (10kOhm)

I-V plot

Open all switches except 1 (i.e. we access memristor number 1), and then we send in input a triangular signal.

When the memristor is subjected to a potential > 0.25 V it start to “charge”

(Ag ions aggregate into the cracks varying its resistance)



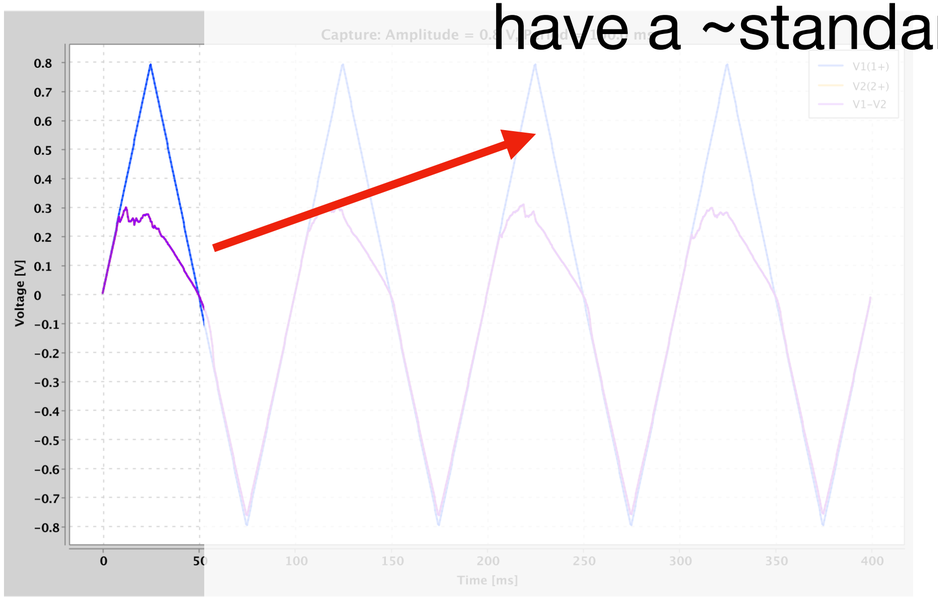
The orange line is the I-V plot of the memristor only,

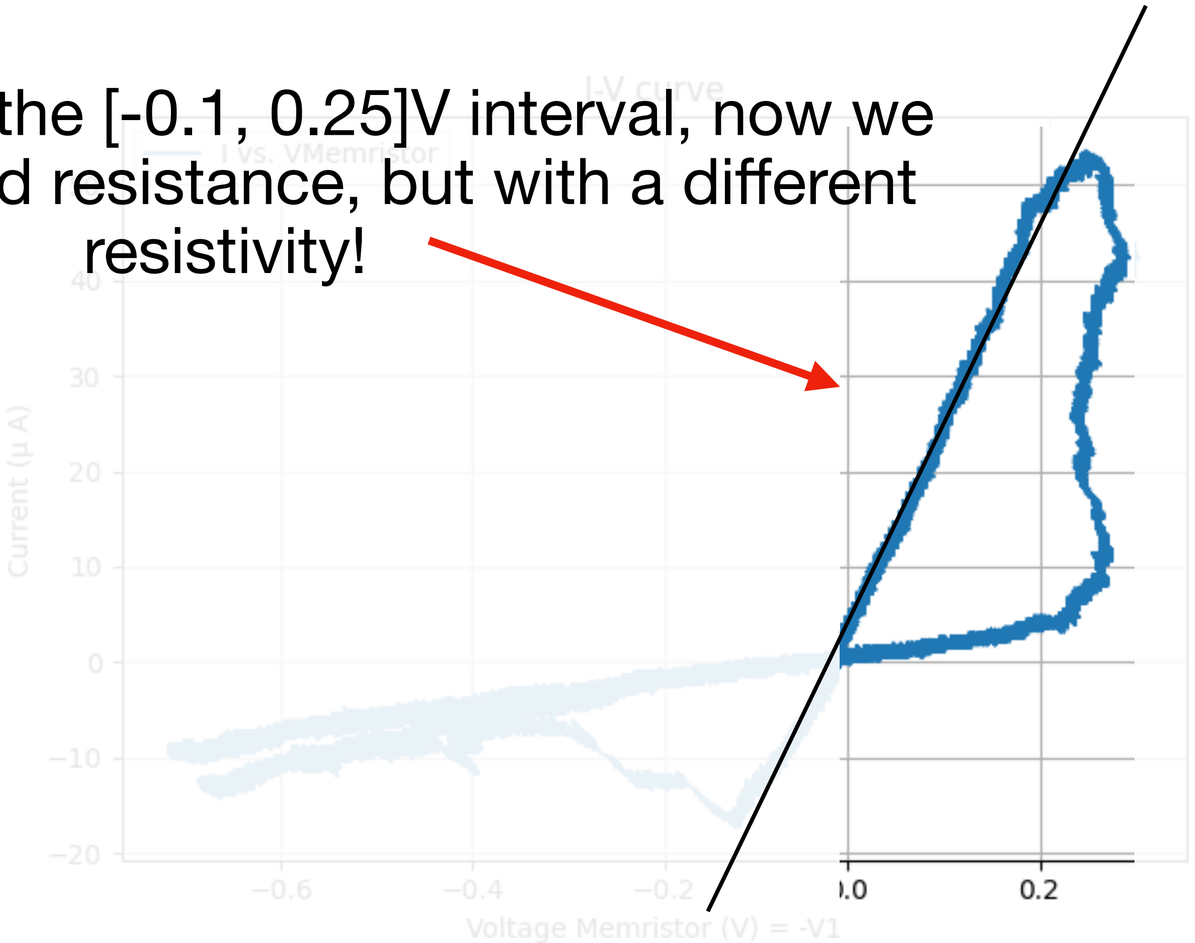
The blue line is the potential of the memristor+Resistance (10kOhm)

I-V plot

Open all switches except 1 (i.e. we access memristor number 1), and then we send in input a triangular signal.

Coming back to the [-0.1, 0.25]V interval, now we have a ~standard resistance, but with a different resistivity!



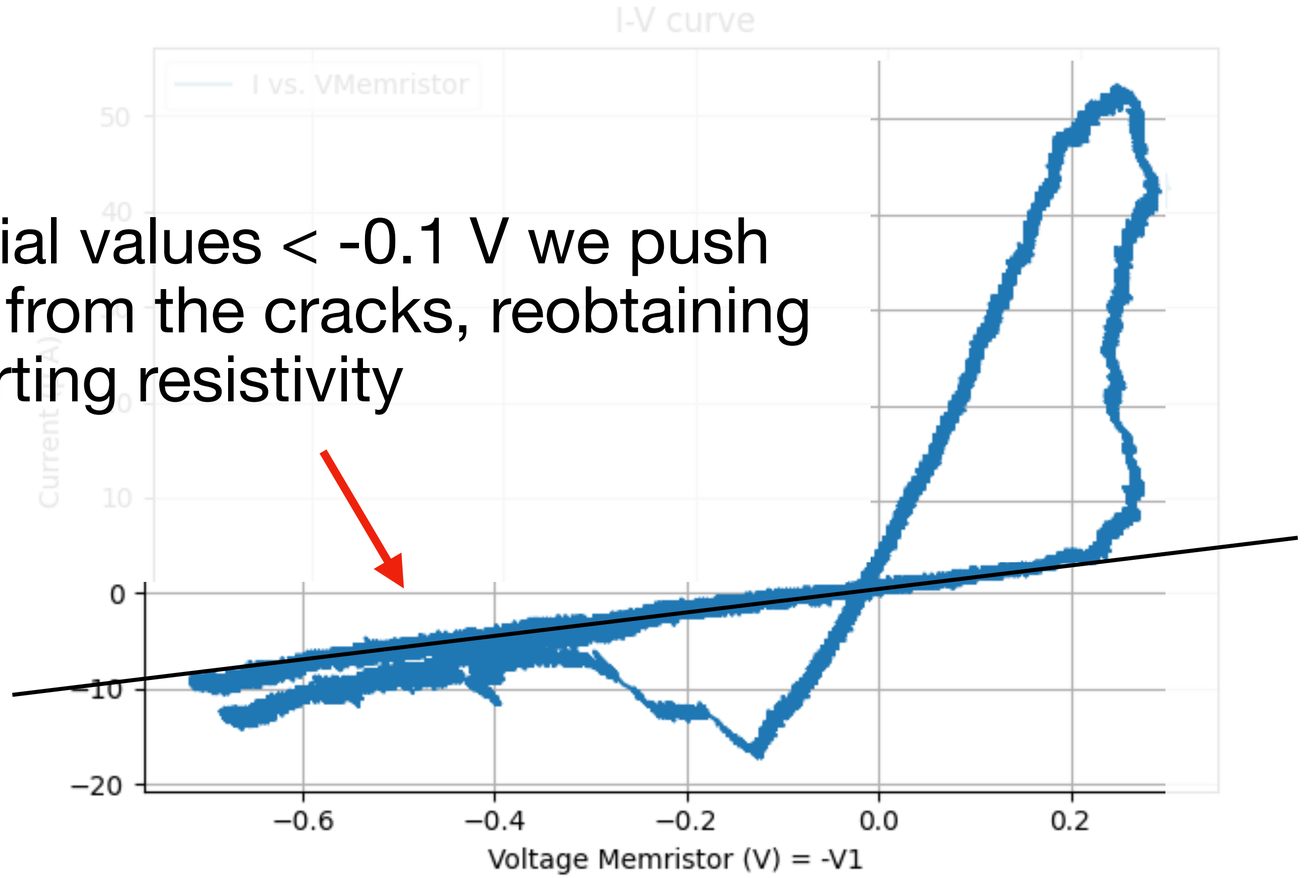


The orange line is the I-V plot of the memristor only,
The blue line is the potential of the memristor+Resistance (10kOhm)

I-V plot

Open all switches except 1 (i.e. we access memristor number 1), and then we send in input a triangular signal.

If we use potential values < -0.1 V we push away the Ag ions from the cracks, reobtaining starting resistivity

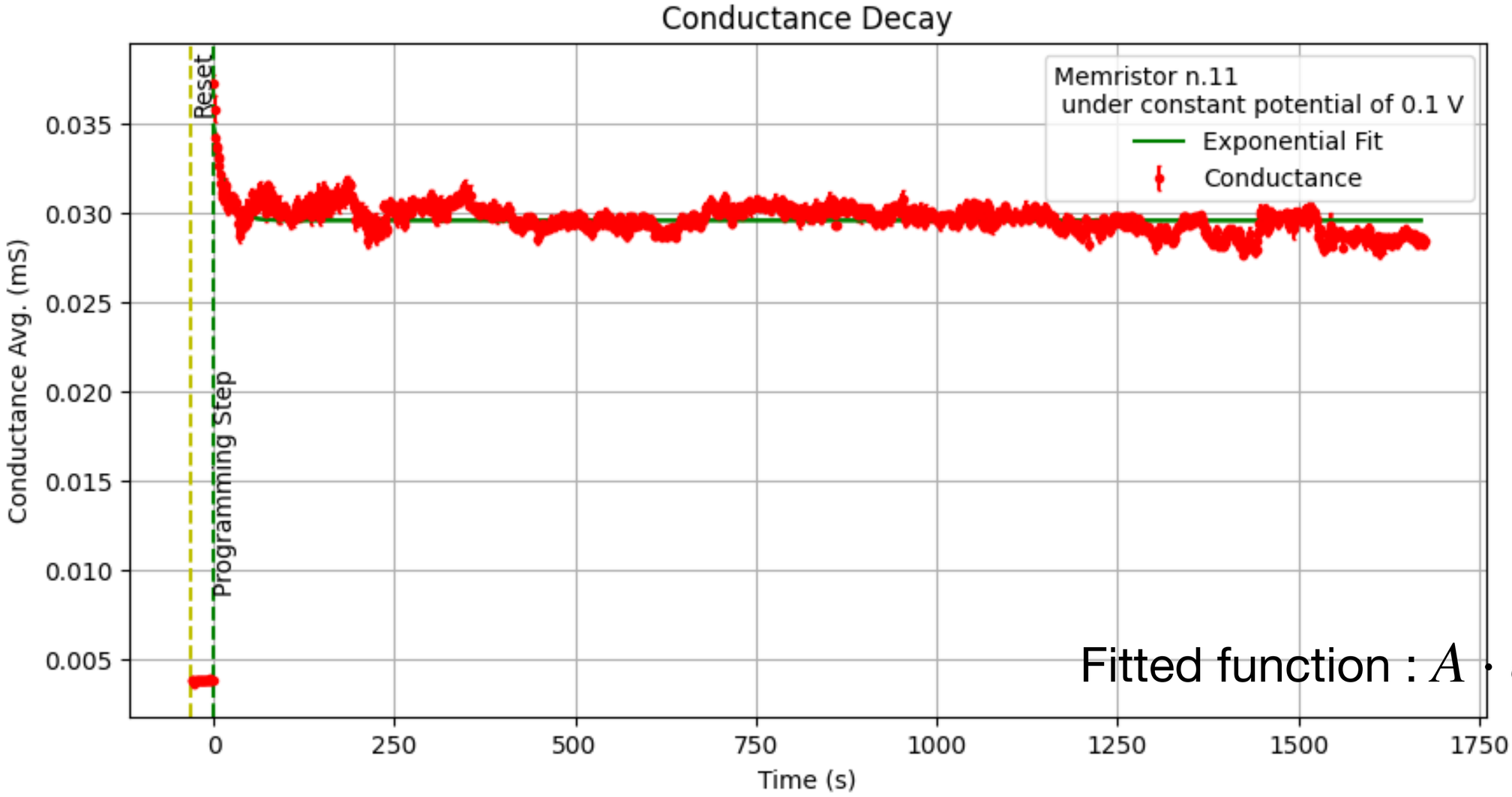


The orange line is the I-V plot of the memristor only,

The blue line is the potential of the memristor+Resistance (10kOhm)

Cundactance Decay

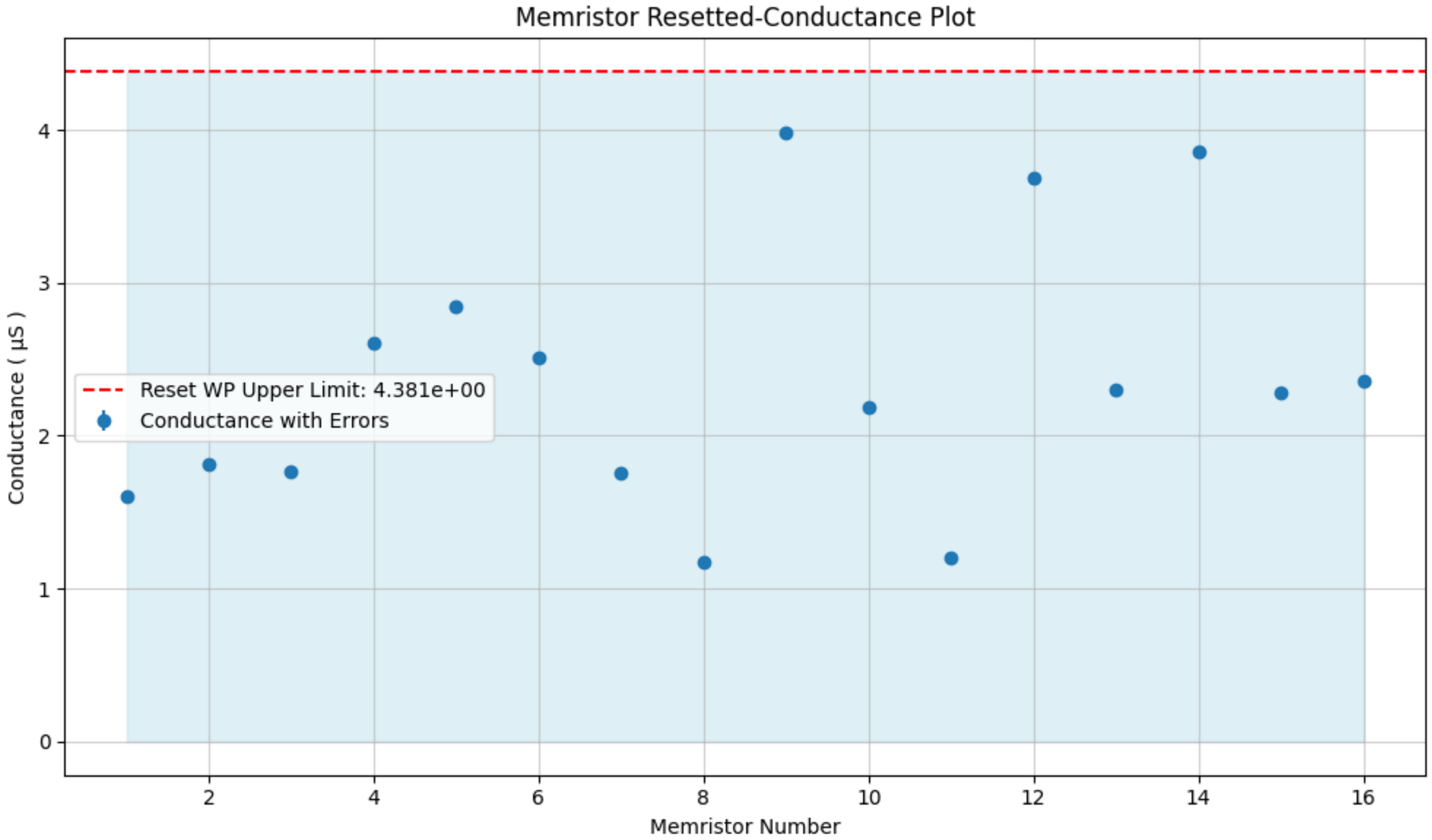
Let’s look at one of good cases (not even the best!)



Fitted function : $A \cdot e^{-\frac{t}{\tau}} + C$

Reset state Definition (WP0)

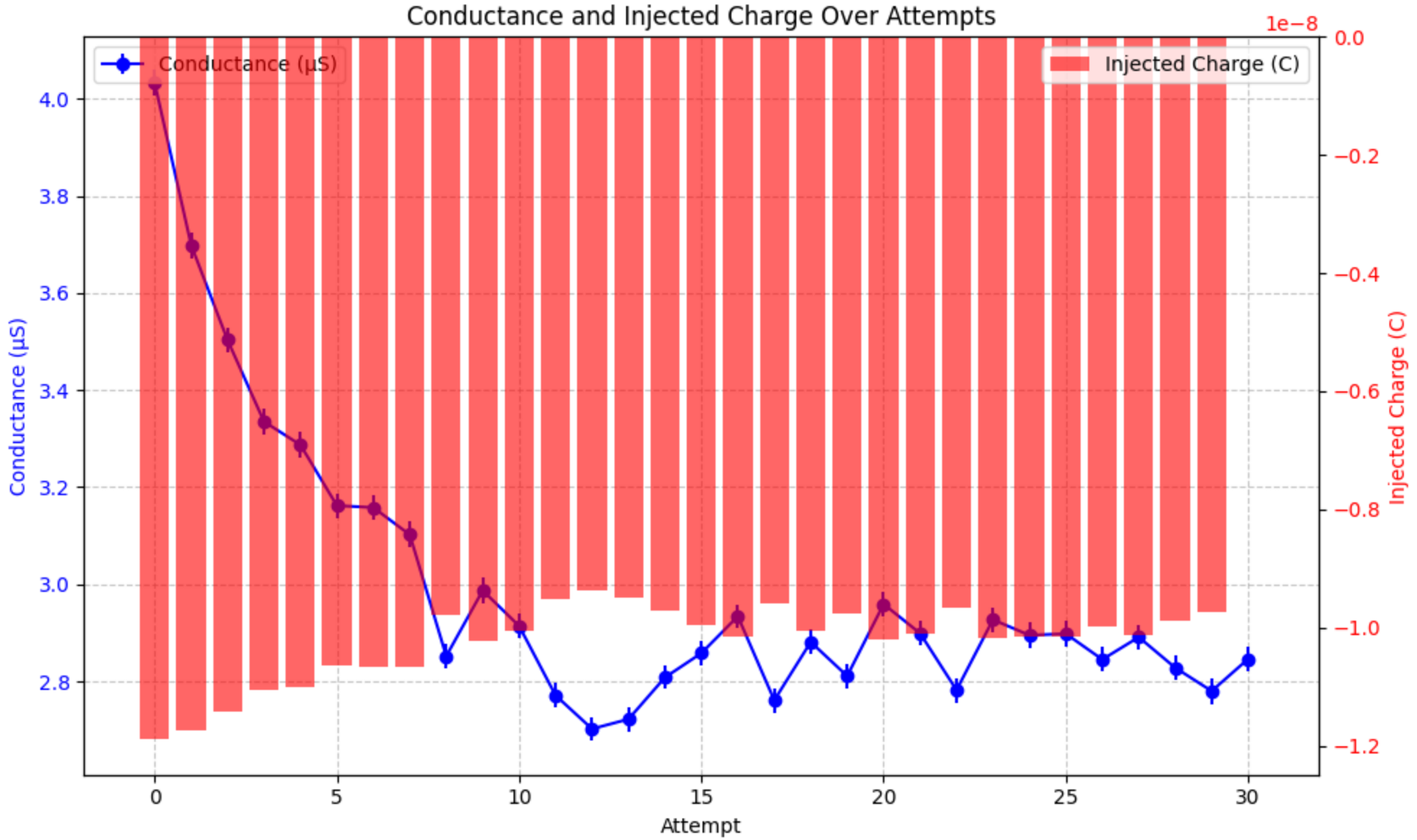
I tried 30 iterations to reach the lower conductance value possible for each memristor



Story of a memristor (mem5)

We keep the loop of injected charge and memristor conductance

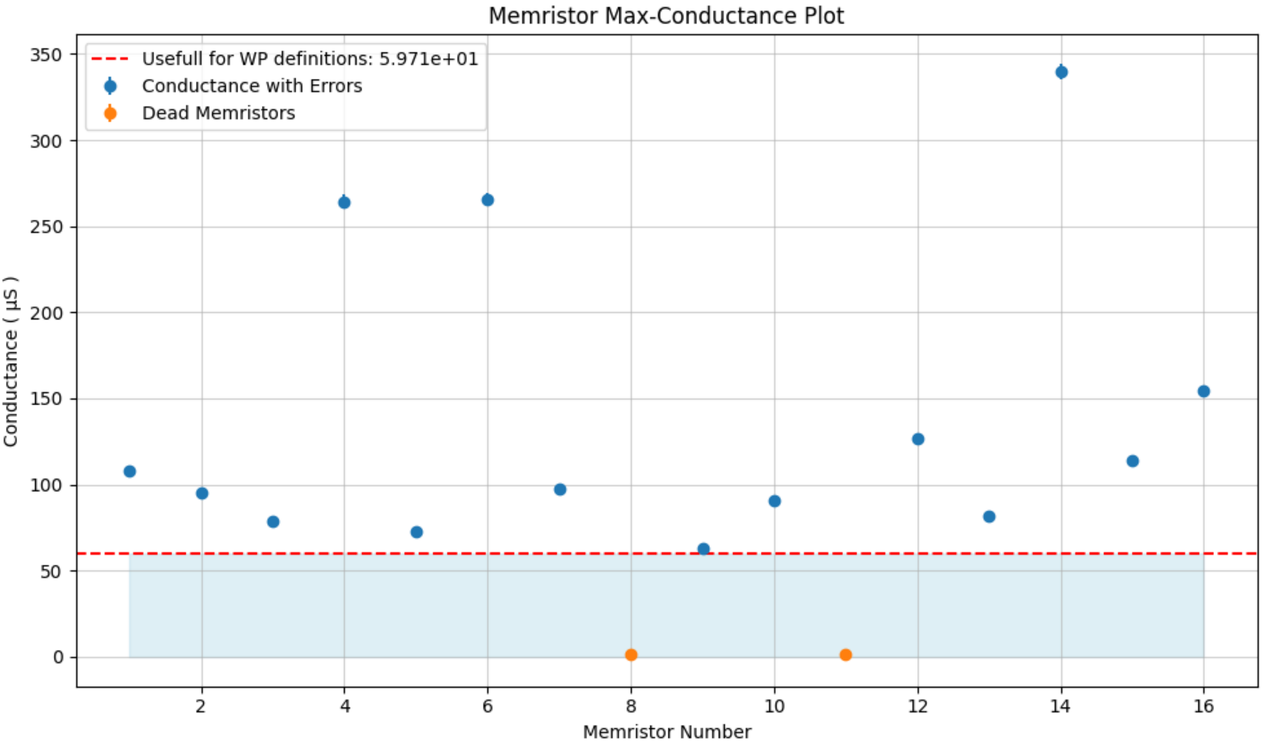
$$Q = V_{Mem}^{Max} \cdot G_{Mem} \cdot T_{1pulse} \cdot N_{pulses}$$



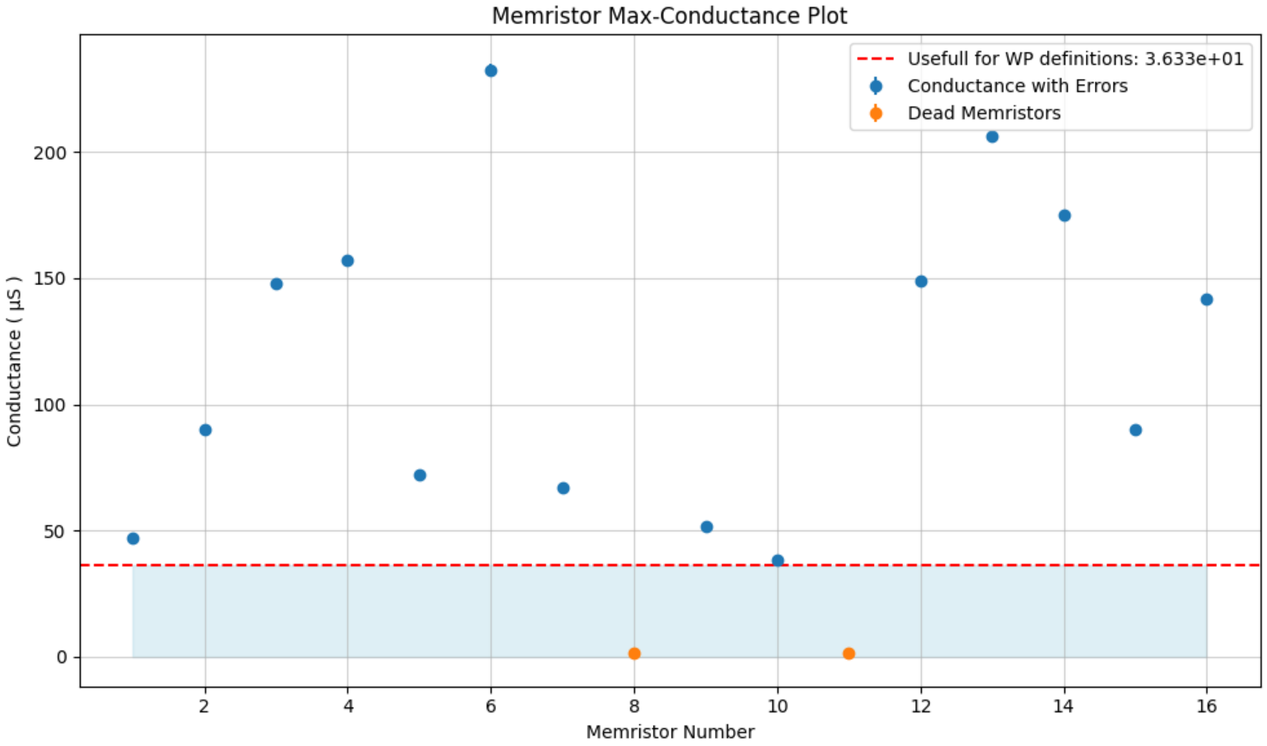
Reset state Definition (WP0)

I tried 30 iterations to reach the lower conductance value possible for each memristor

December 2024 test



January 2025 test



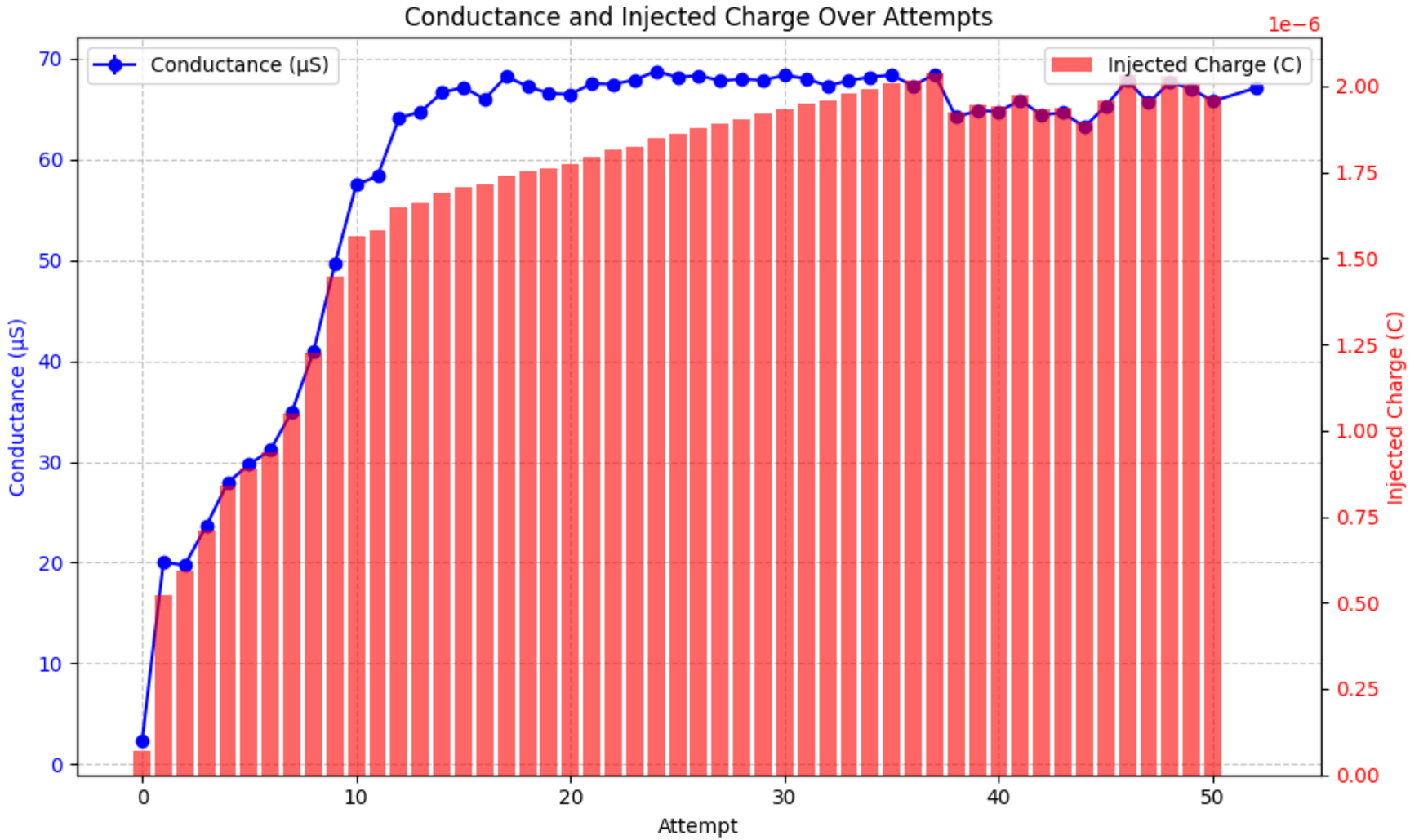
The maximum conductance values seems quite different

Because last time it was ~60uS, I tried to use same value and check the pattern-test results

Story of a memristor (mem7)

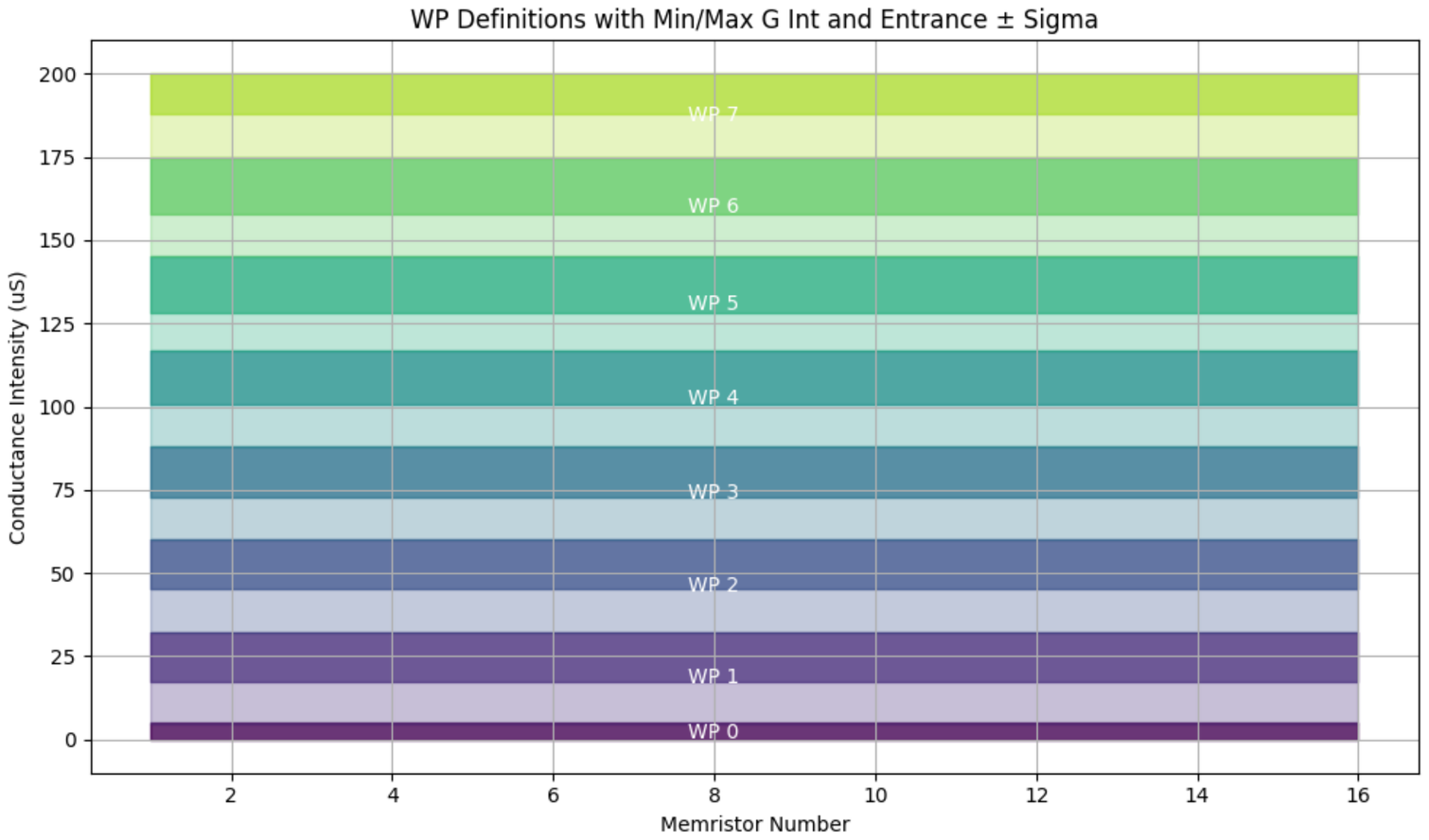
We keep the loop of injected charge and memristor conductance

$$Q = V_{Mem}^{Max} \cdot G_{Mem} \cdot T_{1pulse} \cdot N_{pulses}$$



Attempt of WP definitions

As a good first goal we can imagine we want study a 3-bit quantised NN; so we need 8 WP for each mem



WP	Min G Int (uS)	Max G Int (uS)	Acc Interval (uS)
0	0.0	4.5	0 - 4
1	4.5	32	18 - 32
2	32	60	46 - 60
3	60	88	74 - 88
4	88	117	103 - 117
5	117	145	130 - 145
6	145	175	160 - 175
7	175	200	187 - 200