RRAM-BASED NEURO-INSPIRED COMPUTING FOR UNSUPERVISED TEMPORAL PREDICTIONS

R. DEGRAEVE, D. RODOPOULOS, A. FANTINI, D. GARBIN, D. LINTEN, P. RAGHAVAN

OUTLINE & PURPOSE

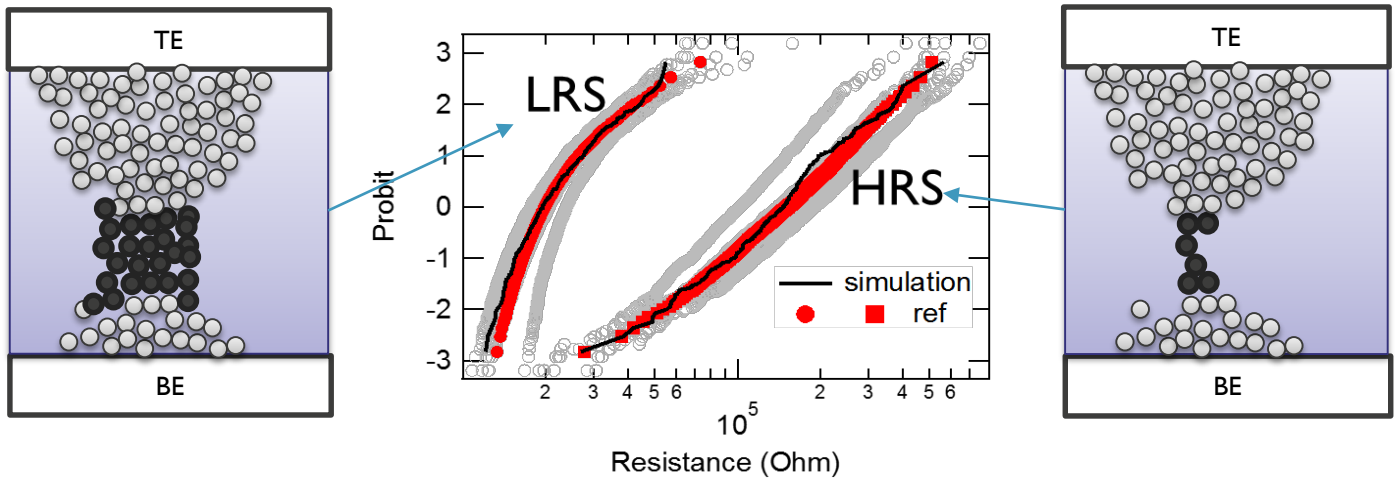
Purpose

Show a learning algorithm that exploits the physical properties of RRAM as true imitation of synaptic connections

- Resistive RAM properties
- Algorithm
- Discussion and analysis
- Conclusions

O-VACANCY RRAM = TUNABLE CONDUCTIVE FILAMENT

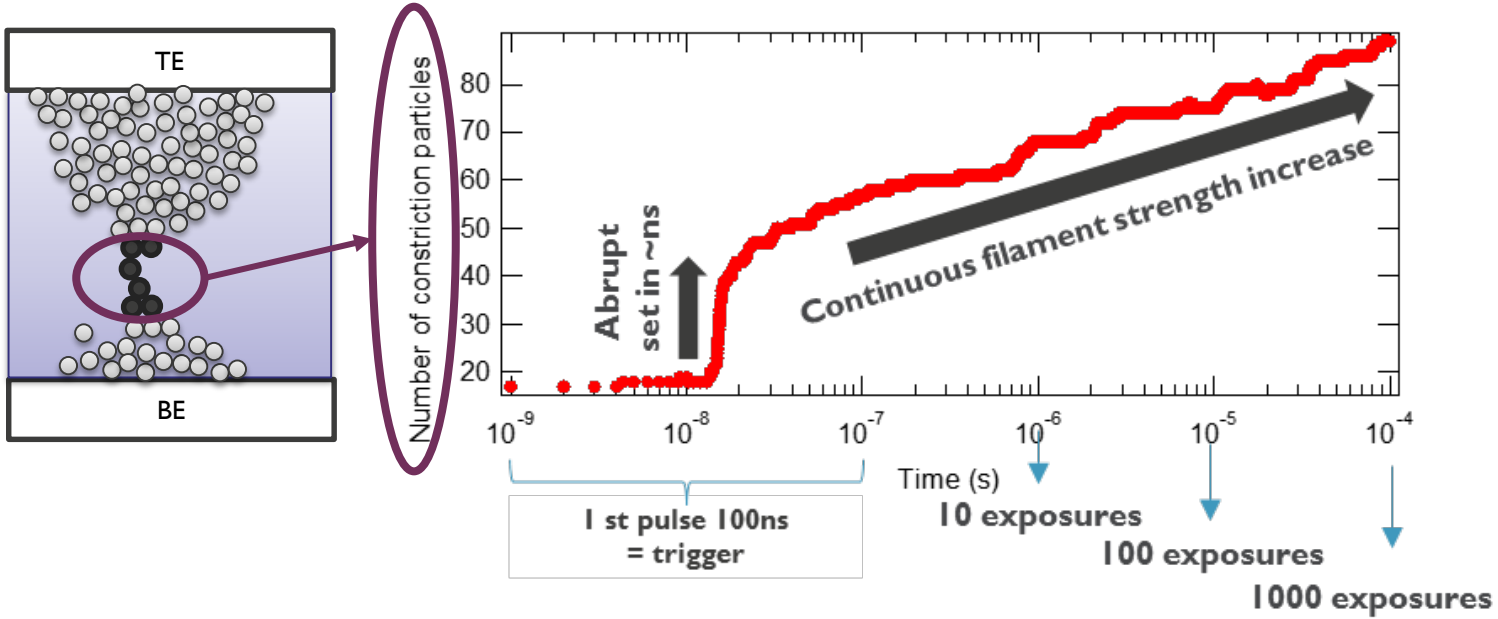
BUT STOCHASTIC BEHAVIOR MAKES IT AN UNRELIABLE STORAGE ELEMENT



- Broad distribution at both low and high resistive state
- Stochastic behavior after program, after read and as a function of time
- Extremely **unsuited as a stable 'weight' element**

TRUE RRAM PROPERTY = STABLE MEAN BEHAVIOR

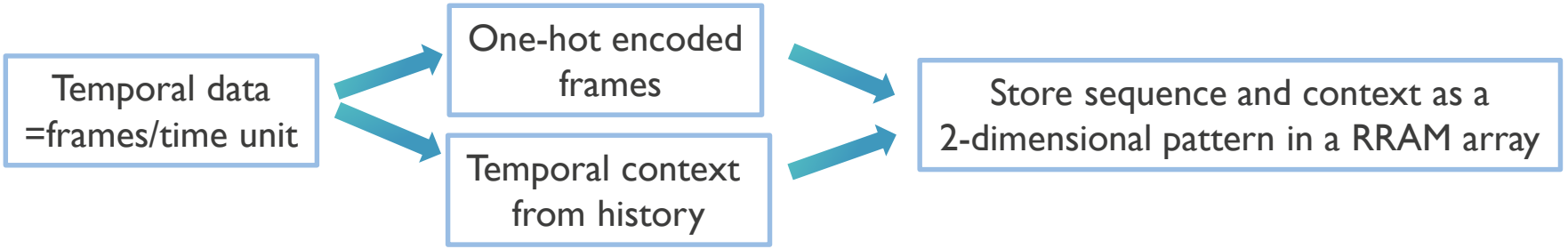
PHYSICS = STABLE NUMBER OF VACANCIES IN FILAMENT CONSTRICTION



- Switch from high to low resistance = abrupt due to positive feedback effect
- Repeated stimuli result in **stable mean filament growth**
- BUT = read-out remains wide distribution & stochastic

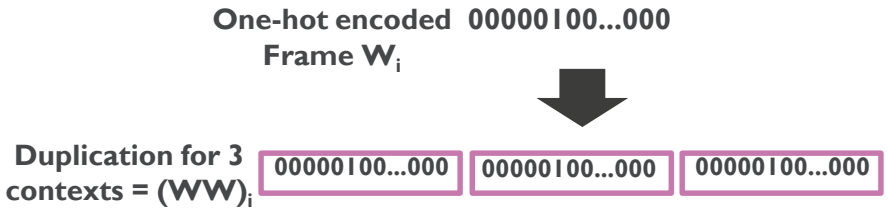
WHAT ALGORITHM TAKES BEST ADVANTAGE OF RRAM PROPERTIES?

CONTINUOUS LEARNING OF TEMPORAL DATA



Step (i) | Algorithm example

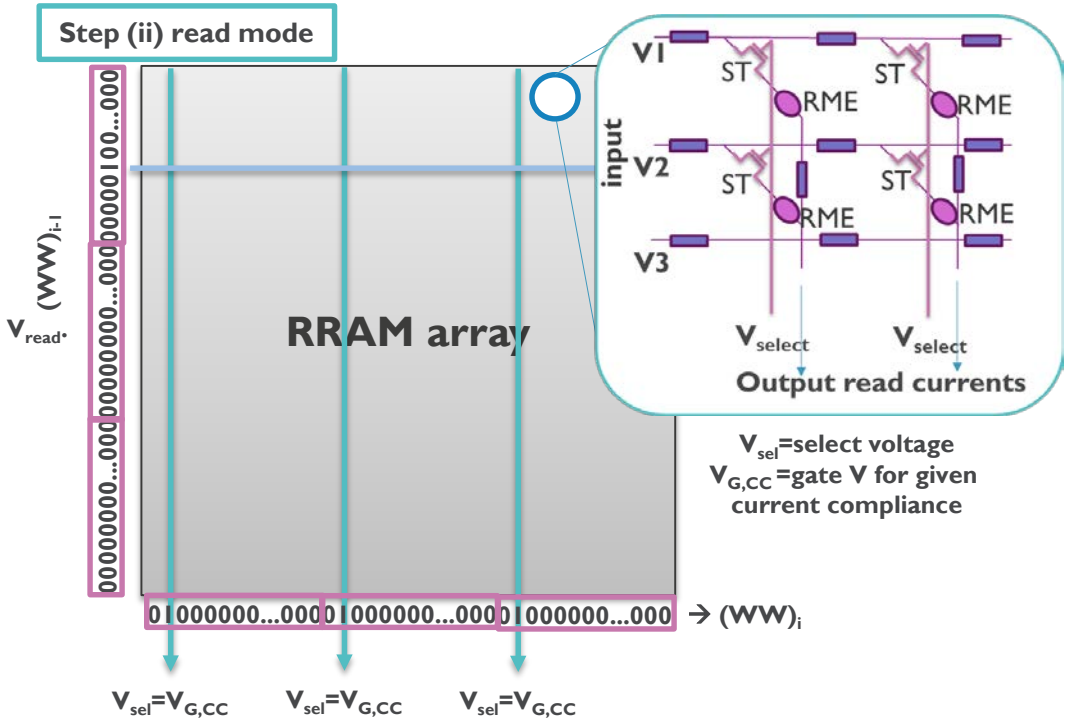
- STEP 1: open the new frame to all contexts



STEP 2: CHECK PREVIOUSLY SEEN SEQUENCE STEPS

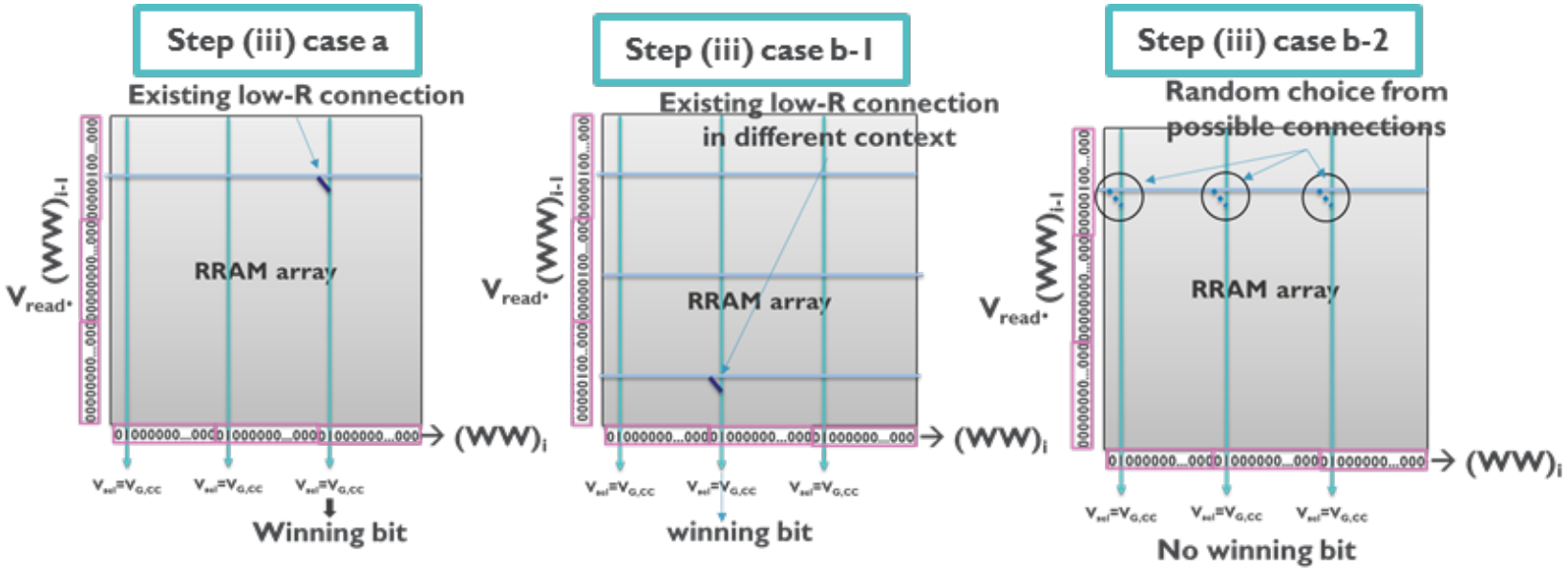
RRAM ARRAY IN READ MODE @ LOW VOLTAGE

- Read possible existing connections between new state and previous state
- Parallel reading
- Simply checking whether current > threshold



STEP 3: FIND EXISTING CONNECTIONS

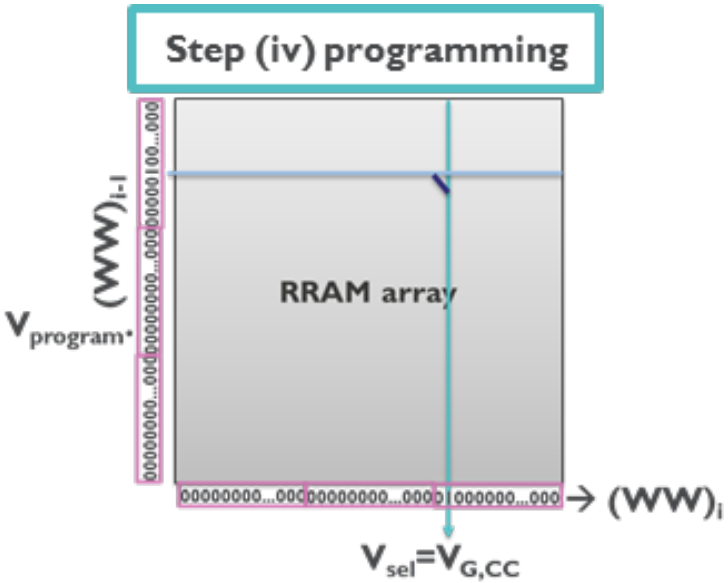
AIM= EACH CONNECTION IS PROGRAMMED ONLY ONCE



- Existing connection is detected in same context
- Existing connection is detected in a different context
- No connection exists yet

STEP 4 : PROGRAMMING SEQUENCE AS RRAM CONNECTION

PROGRAM VOLTAGE AND PULSE LENGTH ARE ALWAYS THE SAME

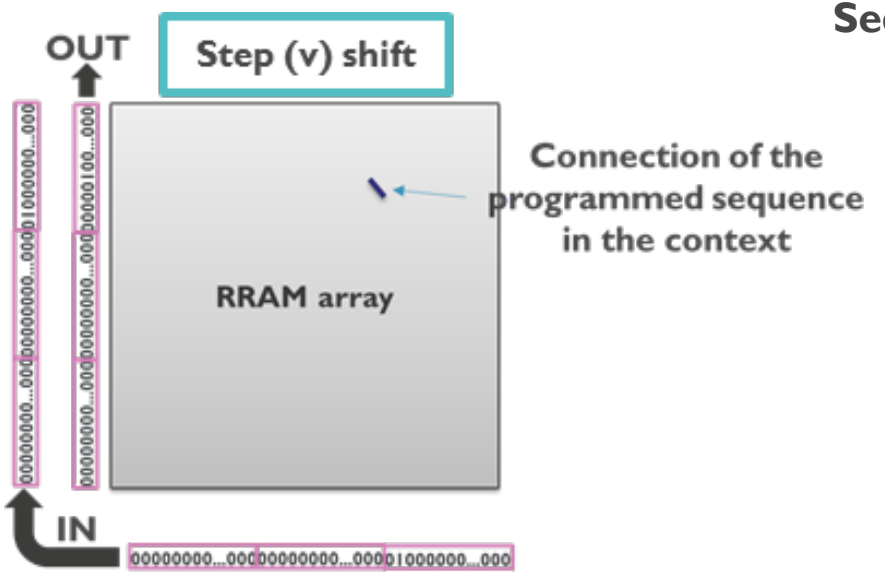


- New input is restricted to the selected column
- Programming step
  - Either makes a **new connection** between previous and new state
  - Either **confirms an existing connection** and strengthens it
- Current is limited by driving transistor

$V_{program}$ = 1.5V for t=100ns

STEP 5: SHIFT OPERATION

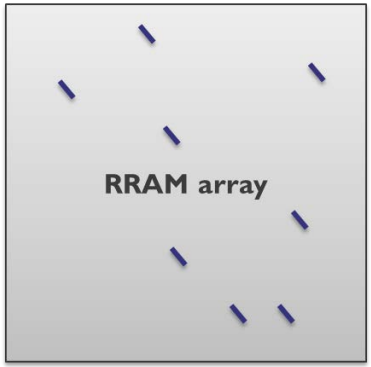
SYSTEM IS READY FOR NEXT INPUT



Sequence of abstract data= A-F-R-S-D-E-A-X-B

→

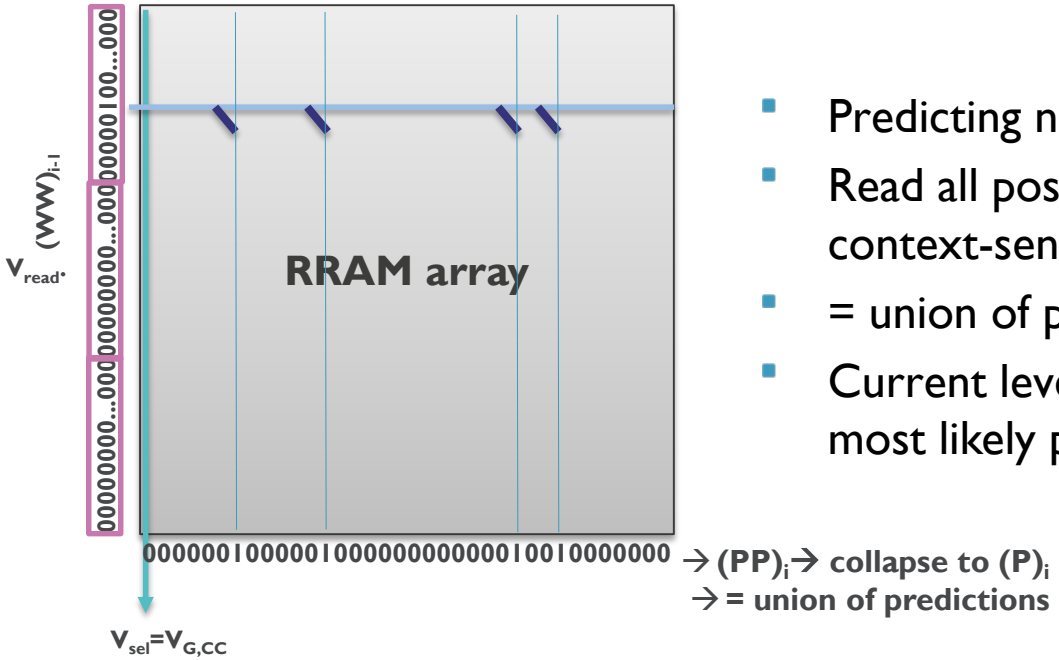
2D- pattern of connections



- Each step in the time sequence adds 1 connection or strengthens an existing one
- Time sequence is casted into the RRAM array as a 2D pattern

PREDICTING DATA =READING STEP

DATA CAN BE PREDICTED BY READING WITH ALL COLUMNS SELECTED



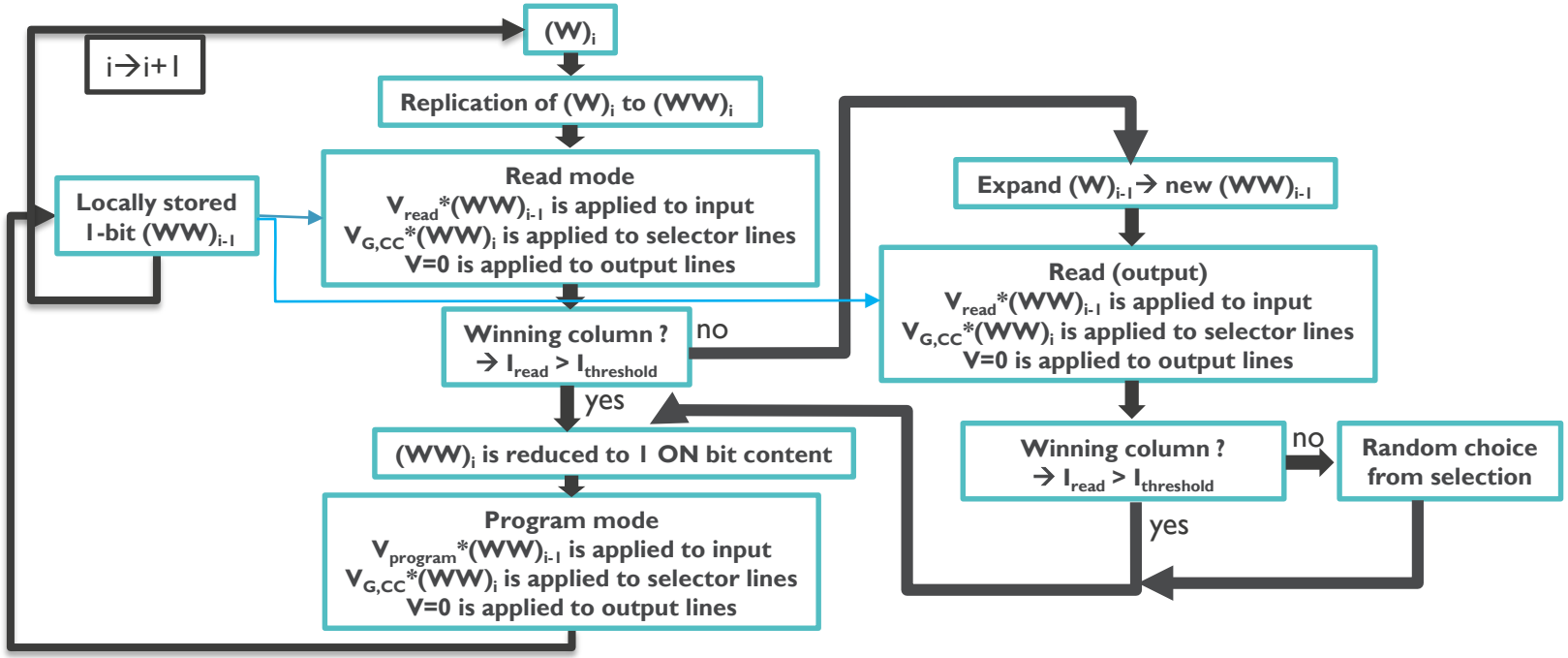
→ $(PP)_i$ → collapse to $(P)_i$
→ = union of predictions

READ output current

- Predicting next data frame
- Read all possible connections starting from the context-sensitive last known frame
- = union of predictions
- Current levels has statistical meaning identifying most likely prediction

FLOW CHART ALGORITHM

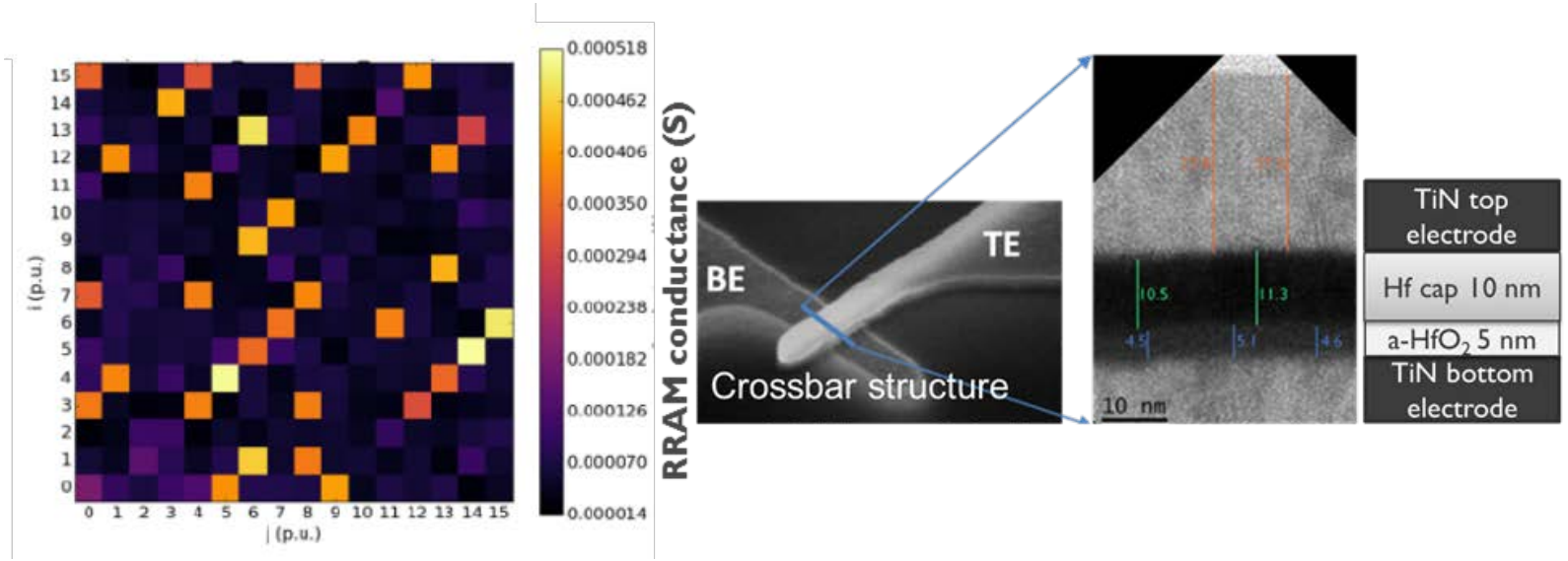
FOR CONTINUOUS UNSUPERVISED LEARNING



- See my poster stand for details and discussion

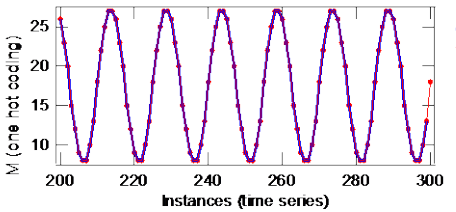
HARDWARE PROOF ON RRAM TEST CHIP

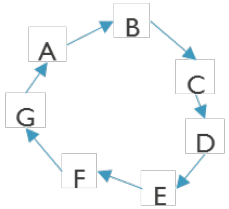
SIMPLE VERSION OF ALGORITHM DEMONSTRATES FEASIBILITY

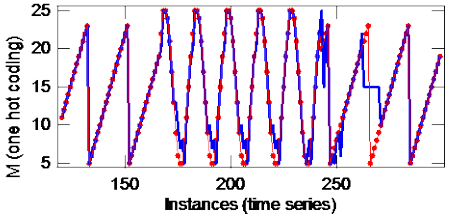


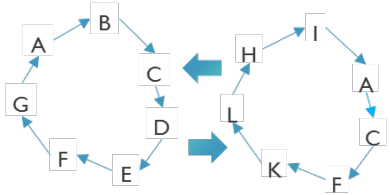
- Technology with capped 5 nm HfO or TaO layer and TiN electrode
- Crossbar test structures
- Simple example of algorithm and measured corresponding conductance

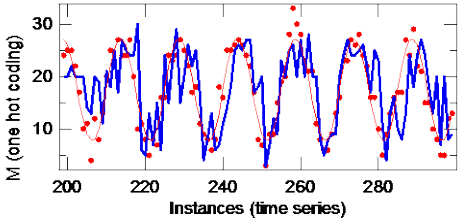
SIMULATED CONTROL EXAMPLES SHOW OPERATION

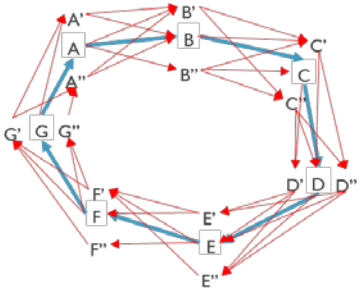








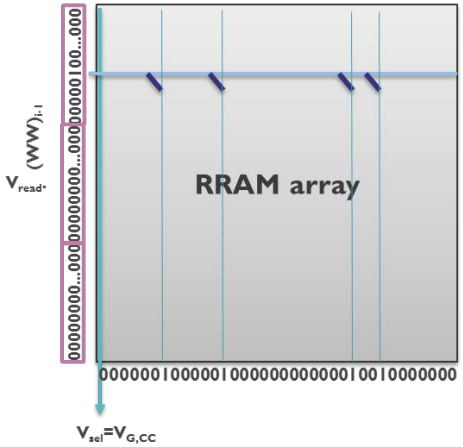




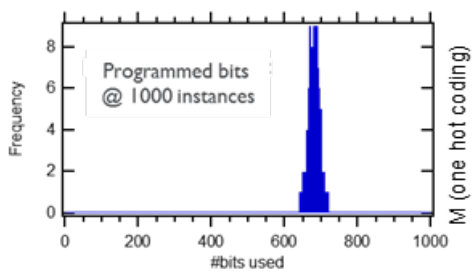
- Periodic function is translated as loop of connections
- Two periodic functions = 2 loops with weaker interconnects
- Noisy function adds several unwanted connections

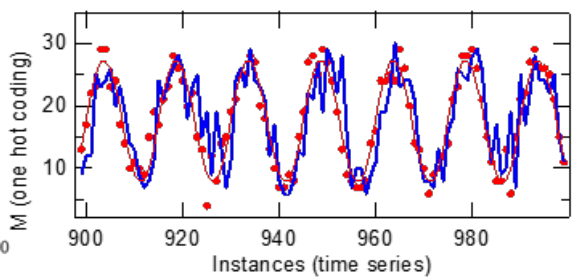
SUPPRESSION OF INFREQUENT CONNECTIONS BY READ DISTURB

SOLUTION FOR COPING WITH NOISE IN DATA

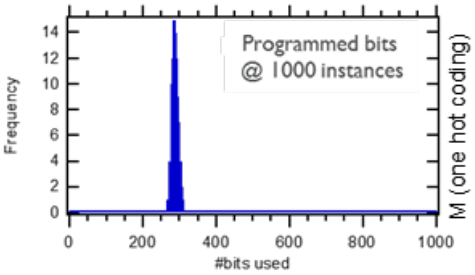


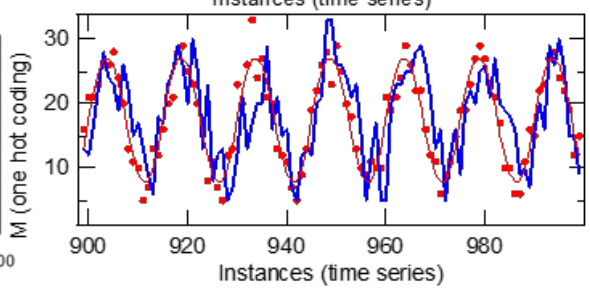
Read during predict weakens connections
Only the trained connection is confirmed again





Without reset



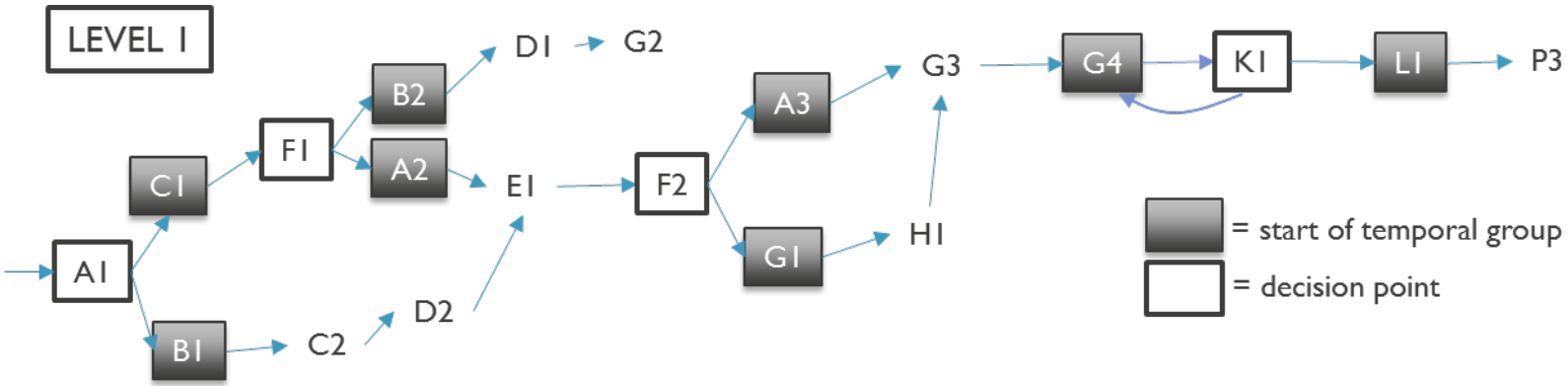


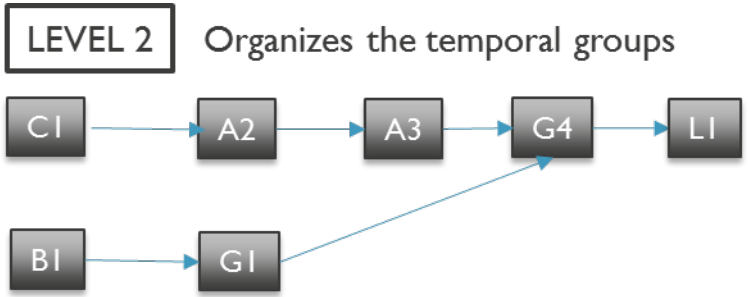
With reset

- Over-programming avoided by using **read disturb** mechanism as forget mechanism
- V_{read} = -0.5 V → causes a small (non-abrupt) reset to high resistive state

HIERARCHICAL STRUCTURE FOR MORE COMPLEX BEHAVIOR

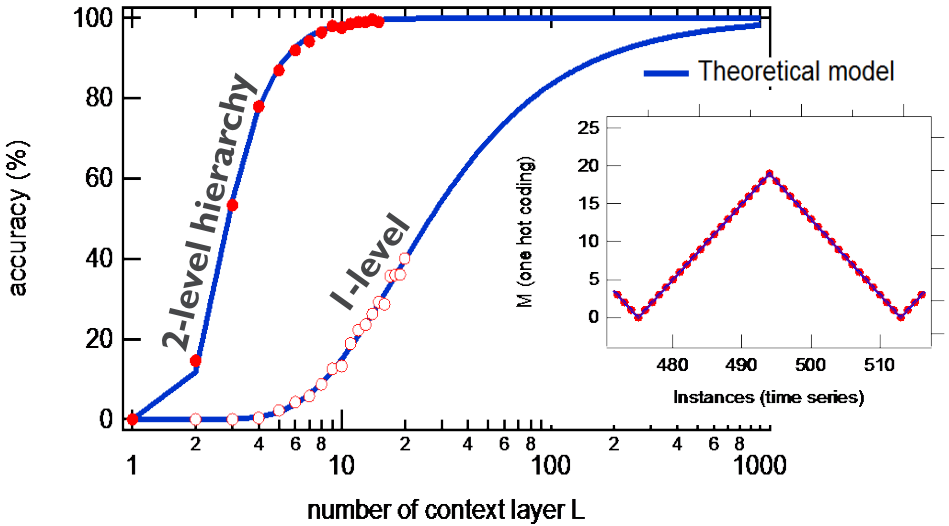
TIME SEQUENCES ARE DIVIDED OVER SHORT AND LONGER TERM STRUCTURE





- At each decision point in level 1, we address level 2 both read and program
- Level 1 = organization of states in temporal groups
- Level 2 = organization of temporal groups in larger groups

CONTROL ILLUSTRATES OPERATION AND EFFECTIVENESS



- Symmetric saw tooth has each point in two contexts
  - Indistinguishable with connection strength
  - Hierarchical approach is effective

SUMMARY & CONCLUSIONS

- Find a way to exploit filamentary RRAM properties in a learning algorithm
- Our answer
  - Continuous unsupervised learning
  - Temporal data
  - Hardware implementable
  - Filamentary RRAM used as stochastic memory technology
  - Read and program are at constant voltage and time
  - Read disturb as forget mechanism
- Sequence of data is stored as a structured 2D network of hardwired connections with statistical properties