CBRAM devices as binary synapses for low-power stochastic neuromorphic systems: auditory (cochlea) and visual (retina) cognitive processing applications

M. Suri, O. Bichler, D. Querlioz, G. Palma, E. Vianello, D. Vuillaume, C. Gamrat, B. de Salvo

CBRAM Devices as Binary Synapses for Low-Power Stochastic Neuromorphic Systems: Auditory (*Cochlea*) and Visual (*Retina*) Cognitive Processing Applications

M. Suri[1], O. Bichler[2], D. Querlioz[4], G. Palma[1], E. Vianello[1], D. Vuillaume[3], C. Gamrat[2], and B. DeSalvo[1]

[1]CEA-LETI-MINATEC, 38054, Grenoble, France, [2]CEA-LIST, Gif-sur-Yvette, [3]CNRS-IEMN, Lille, [4]IEF-Paris
Contact: (manan.suri@cea.fr , barbara.desalvo@cea.fr) +33-438781086

Introduction

Aggressive device-scaling and low-power operation trends have improved the silicon economy, but at the cost of intrinsic variability. Thus, future computing systems have to be designed to be immune to, or even exploit, the technology variability and intrinsic stochasticity. Although neuromorphic hardware is ascribed to be tolerant to stochasticity, it has rarely been shown how.

Abstract

In this work, we demonstrate an original methodology to use Conductive-Bridge RAM (CBRAM) devices as binary synapses in low-power stochastic neuromorphic systems. A new circuit architecture, programming strategy and probabilistic STDP learning rule are proposed. We show, for the first time, how the intrinsic CBRAM device switching probability at ultra-low power can be exploited to implement probabilistic learning rule. Two complex applications are demonstrated: real-time auditory (from 64-channel human cochlea) and visual (from mammalian visual cortex) pattern extraction. A high accuracy (audio pattern sensitivity >2, video detection rate >95%) and ultra-low synaptic-power dissipation (audio 0.55µW, video 74.2µW) are obtained.

CBRAM technology

1T-1R CBRAM devices (both isolated and in 8x8 matrix), integrated in standard CMOS platform [1], were tested (Fig.1). CBRAM operating relies on an electrochemically active electrode metal (Ag), drift of highly mobile Ag^+ cations in the conducting layer (30nm-thick GeS_2), and their discharge at the (inert) counter electrode (W), leading to the growth of Ag dendrites (i.e. a highly conductive filament) in the ON (set) state. Upon reversal of voltage polarity, an electrochemical dissolution of the conductive bridge happens, resetting the system to the OFF (reset) state (Fig. 2). Ease of fabrication, CMOS compatibility, scalability and low operating-voltages make CBRAM an ideal choice for the design of low-power bio-inspired systems.

Limitations of Multi-level CBRAM Synapses

In literature [2], CBRAM multi-level programming was proposed to emulate biological synaptic-plasticity (Long Term Potentiation-LTP and Long Term Depression-LTD). LTP behavior (i.e. ON-state resistance decrease) is demonstrated in our samples by applying a positive bias at the anode and gradually increasing the select transistor gate voltage (Vg) (Fig.3a). This phenomenon can be explained with our physical model [3] assuming a gradual increase in the radius of the conductive filament formed during the set-process. Nevertheless, this approach implies that each neuron must generate pulses with increasing amplitude while keeping a history of the previous state of the synaptic device, thus leading to additional overhead in the neuron circuitry. Moreover, we found it very difficult to emulate a gradual LTD-like effect using CBRAM. Fig.3b shows the abrupt nature of the set-to-reset transition in CBRAM devices, due to the difficulty in dissolving the conductive filament in a controlled way. To overcome these issues, we propose hereafter a new methodology based on binary CBRAM synapses with a probabilistic STDP (spike-time-dependent-plasticity) learning rule.

Experiments and Probabilistic Switching

Fig.4 shows the On/Off resistance distributions of an isolated 1T-1R CBRAM (during repeated cycles with strong set/reset conditions). The OFF state presents a larger dispersion. This can be interpreted in terms of stochastic breaking of the filament during the reset process, due to the unavoidable defects [4-6] close to the filament which act as preferential sites for dissolution. By fitting this data with our physical-model [3], the distribution of the left-over filament-height was computed (Fig.5a). Note that this also implies a spread on the voltage (V_{SET}) and time (T_{SET}) needed for the consecutive set operations (Figs.5b,c). In other words, when weak SET programming conditions are used immediately after a RESET, a probabilistic switching of the device appears (Fig.6). To take into account the device-to-device variability, we performed similar analysis on the matrix devices. Fig.7 shows the 'On/Off' resistance distributions for all devices cycled 20 times with strong conditions. In Fig.8, we note that switching probability (criterion for successful switch: Roff/Ron>10) increases for stronger programming conditions. We thus argue that CBRAM device switching probability can be tuned by using the right combination of programming conditions.

Stochastic STDP and Programming Methodology

At the system level, a functional equivalence [14] exists between multi-level deterministic synapses and binary probabilistic synapses (Fig.9). Note also that in real biological systems synaptic transmission is probabilistic [15]. Based on these assumptions, we performed system level simulations with our "Xnet" tool [7, 8]. The synapses were defined by fitting data of Fig.7 with a lognormal distribution (Fig.10). An original stochastic and simplified STDP learning rule, inspired from biological one [9] and optimized by genetic-evolution algorithms [8], was adopted (Fig.11). Fig.12 shows the core circuit of our architecture with CBRAM synapses connected to Leaky-Integrate and Fire (LIF) input- and output- neurons. When an output neuron is

active (i.e. fires), if the input neuron was active recently (in the "T_{LTP}" time window) the CBRAM has a given probability to switch into the ON state (probabilistic LTP), if not, the CBRAM has a given probability to switch OFF (probabilistic LTD). In a real circuit, the switching probability of CBRAM synapses can be implemented in two ways (Fig.14): (i) Externally, by multiplying the LTP/LTD signal of the input spiking neuron with pseudo-random number generator (PRGN) output (Figs.12, 13), whose signal probability can be tuned by customizing the shift registers cascade sequence [10]; (ii) Internally, by utilizing the intrinsic CBRAM switching probability with weak programming conditions (Fig.6,8). *Note that exploiting the intrinsic CBRAM switching probability avoids the presence of PRGN circuits, thus saving silicon footprint, and reduces the programming power.* Fig.15 describes our generic neuromorphic processing core.

Auditory and Visual Processing (Cochlea and Retina Application)

Fig.16b shows the network designed to learn, extract, and recognize hidden patterns in auditory data. Temporally encoded auditory data are filtered and processed using a 64-channel silicon cochlea emulator [11] (implemented in 'Xnet'). The processed data are then presented to a single layer feed forward spiking neural network (SNN) with 192-CBRAM synapses. Initially (0s-400s), pure noise is used as input to the system, and the firing pattern of the output neuron is completely random (Fig.18a). Then (400s-600s), an arbitrarily created pattern is embedded in the input noise data and repeated at random intervals. In this period, the output neuron starts to spike predominantly when the pattern occurs; then, the system becomes entirely selective to it. At the end of the test case (600s-800s), pure noise is represented to the system. As expected, the output neuron doesn't activate at all (Figs.18b, 19). The system attained sensitivity higher (>2) than the human ear (Fig.19a) [12], with a very low false-spike rate (Fig.19b) and extremely low synaptic power consumption of 0.55µW (Table 1). This example acts as a prototype for applications such as speech recognition and sound-source localization. Fig.17b shows the network simulated to process temporally encoded video data, recorded directly from an artificial silicon retina [13]. Video of cars passing on a freeway recorded in AER format is presented to a 2-layered SNN consisting about 2-million CBRAM synapses. We implemented a similar network in [7] exploiting multi-level Phase-Change Memory synapses. The CBRAM-based system learns to recognize the driving lanes, extract car-shapes (Fig.20), with more than 95% average detection-rate and a total synaptic-power dissipation of just 74.2µW (lower than [7]) (Table 1). Applications such as image classification and target tracking can be realized with the same network.

Conclusions

We proposed for the very first time a bio-inspired system with binary CBRAM synapses and stochastic STDP learning rule able to process asynchronous analog data streams for recognition and extraction of repetitive patterns in a fully unsupervised way. The demonstrated applications exhibit very high performance (auditory pattern sensitivity >2, video detection rate >95%) and ultra-low synaptic power dissipation (audio 0.55µW, video 74.2µW) in the learning mode. Such systems are extremely promising in two possible fields of application: low-power neuromorphic computing tasks or bio-medical devices in future neural-processing prosthetics.

Acknowledgements

The authors would like to thank Altis Semiconductors for providing the CBRAM devices for this study. The PhD scholarship of Manan Suri is partially funded by DGA-France.

References

[1] C.Gopalan, Solid State Elec., 58, p.54, '11.
[2] S.Yu, IEDM, '10.
[3] G.Palma, IMW, 2012.
[4] S.Choi, IMW, '12.
[5] R.Soni, JAP, 107, 024517, '10.
[6] D.Ielmini, APL, 96, 053503, '10.
[7] M.Suri, IEDM, '11.
[8] O.Bichler, Neural Nets, 32, p. 339, '12.
[9] G.Q.Bi, J. Neurosci. 18, 24, 10464, '98.
[10] F.Brglez, Int. Test Conf. p. 264, '89.
[11] V.Chan, Circ.& Sys., IEEE Trans., 54, p.48, '07.
[12] T.R.Agus, Neuron, 66, n.4, p. 610, '10.
[13] P.Lichtsteiner, IEEE J. Solid-State Circuits, 43, '08.
[14] D.H. Goldberg, Neural Nets, 14, p.781, '01.
[15] B.Walmsley, J. of Neurosci., p.1037, '87.

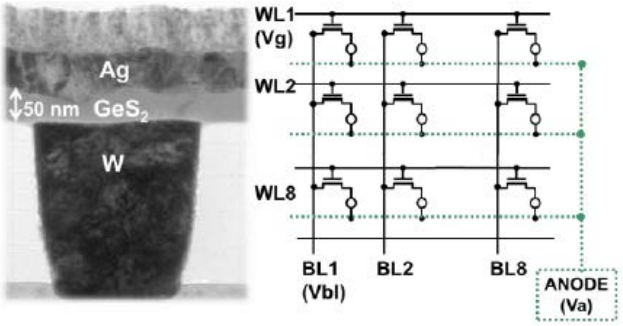


Fig.1 (Left) TEM of the CBRAM resistor element. (Right) Circuit schematic of the 8 X 8 1T-1R CBRAM matrix.

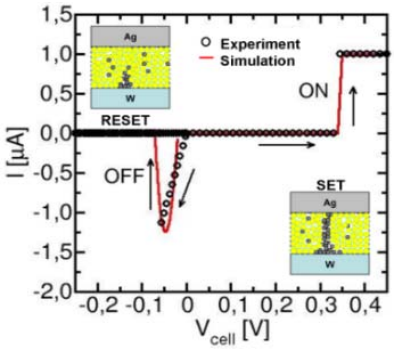


Fig.2 Quasi-static IV curve for the CBRAM device showing the bipolar operation. Model [3] is also shown.

Experiments

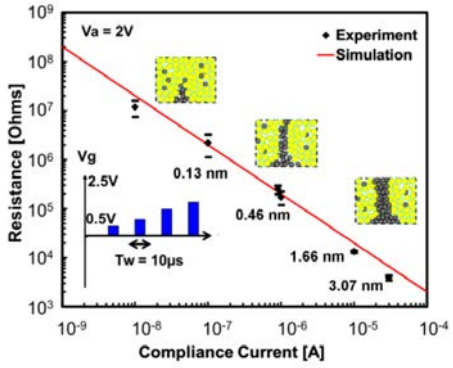


Fig.3a On-state resistance modulation using current compliance. Simulations using model [3] are also shown (extracted filament radius are indicated).

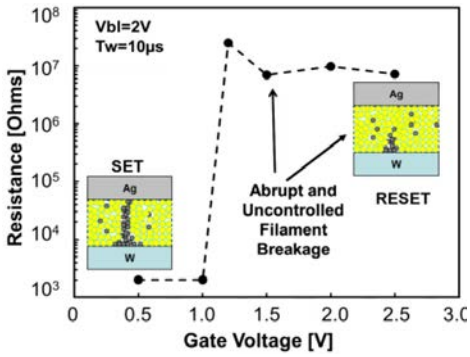


Fig.3b Resistance dependence on gate voltage during the SET-to-RESET transition.

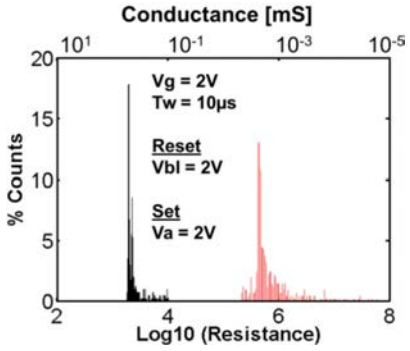


Fig.4 On/Off resistance distribution of an isolated 1T-1R device during 400 cycles when strong programming is used.

Physical Modeling

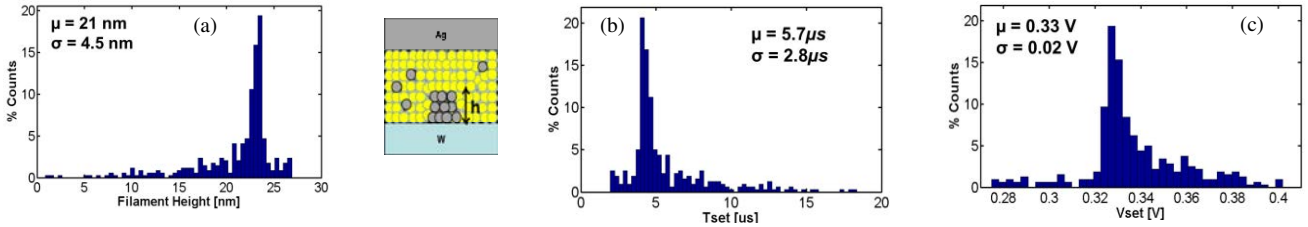


Fig.5 Computed distributions (generated using *Roff* data from Fig.4 and model [3]) of: (a) left over filament-height after RESET; needed (b) Tset and (c)Vset, values for consecutive successful SET operation (mean value µ and sigma σ are indicated).

Stochastic Switching

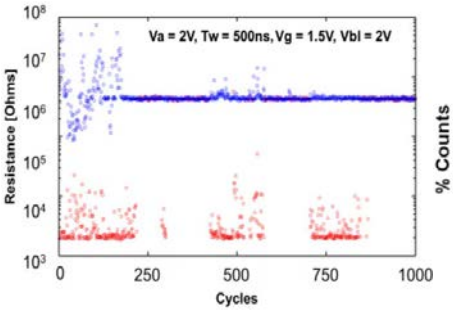


Fig.6 Stochastic switching of 1T-1R device during 1000 cycles using 'weak'-conditions (switch-probability=0.49).

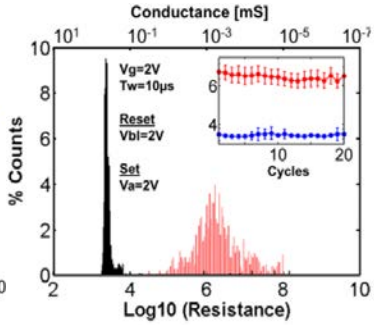


Fig.7 On/Off resistance distributions of the 64 devices of the 8x8 matrix cycled 20 times.

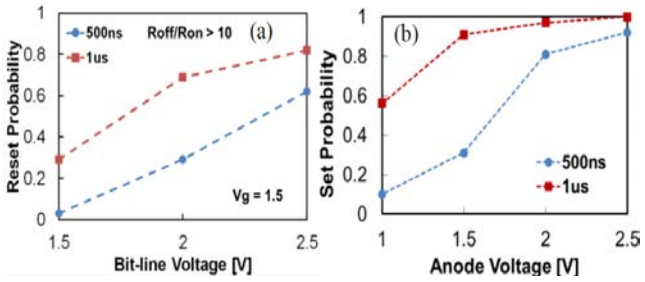


Fig.8 Switching probability for the 64 devices of the matrix (switching being considered successful if Roff/Ron>10) using (a) weak-reset conditions and (b) weak-set conditions.

Probabilistic Neural Learning

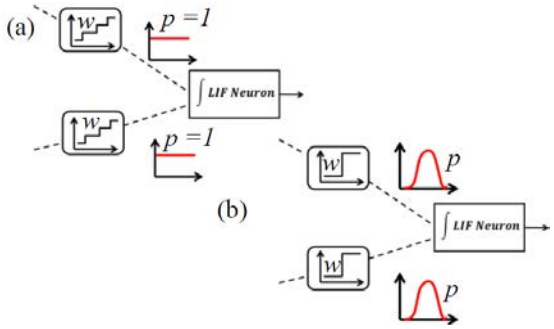
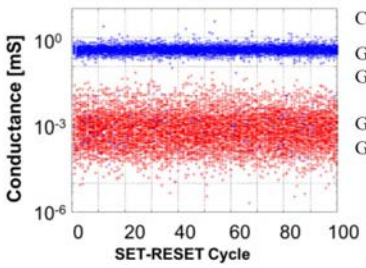


Fig.9 Schematic illustrating (a) multi-level deterministic- and (b) binary probabilistic-synapses connected to a LIF neuron (*W*: weight, *p*: probability).



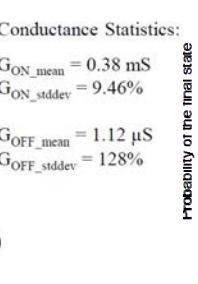


Fig.10 Binary synapses simulated in XNET by fitting Fig.7 data using a log-normal distribution (Statistics are indicated).

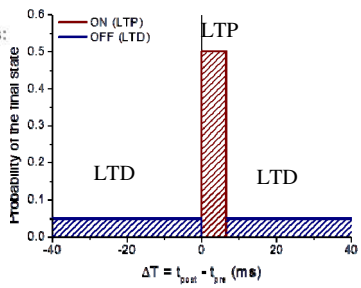


Fig.11 Probabilistic STDP learning rule (used for audio application). X-axis shows the time difference of post-and pre-neuron spike.

Programming Methodology

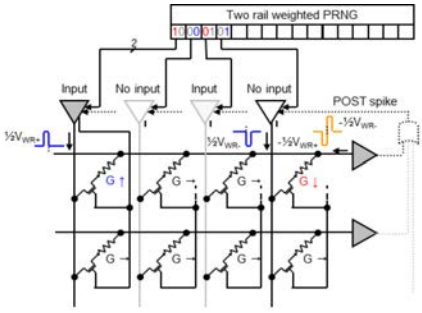


Fig.12 Circuit schematic with CBRAM synapses, LIF neurons, and program pulses (conductance G↑ indicates Switch-On, G→: No-Switch and G↓: Switch-Off).

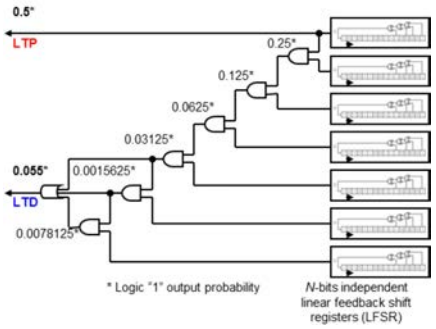


Fig.13 Tunable Pseudo-random-number generator (PRGN) circuit [10], the output being tuned according to STDP in Fig.11.

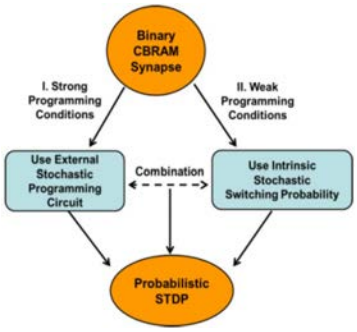


Fig.14 Schematic for the two different approaches possible for using CBRAM device as stochastic binary synapse.

Neuromorphic Processing and Models of Human Cochlea and Retina

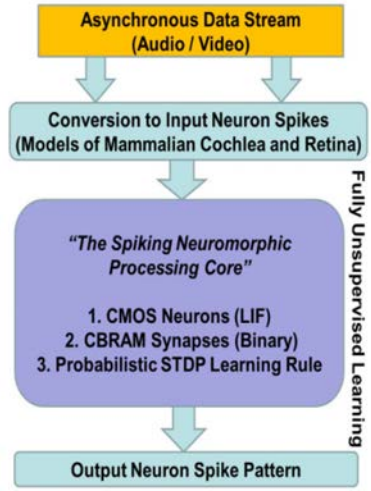


Fig.15 Concept and data flow of the proposed simulated spiking-neuromorphic processing core.

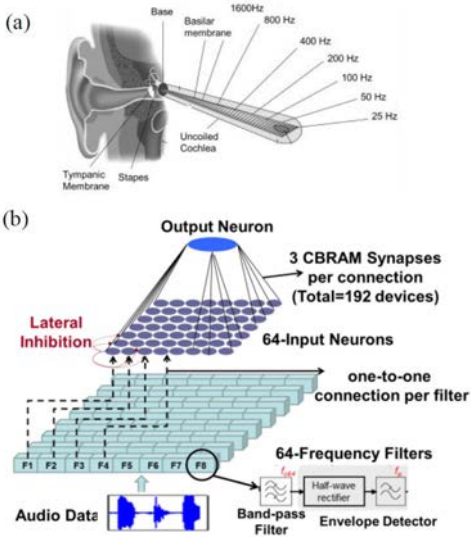


Fig.16 (a) Picture of the uncoiled human cochlea. (b) Our single layer spiking neural network simulated for auditory processing.

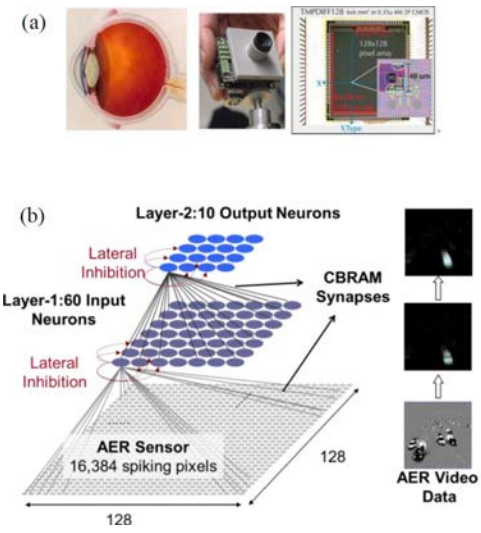


Fig.17 (a) Picture of the human (left) and artificial silicon retina [13] (right). (b) Our 2-layer spiking neural network simulated for processing video data.

Learning Results

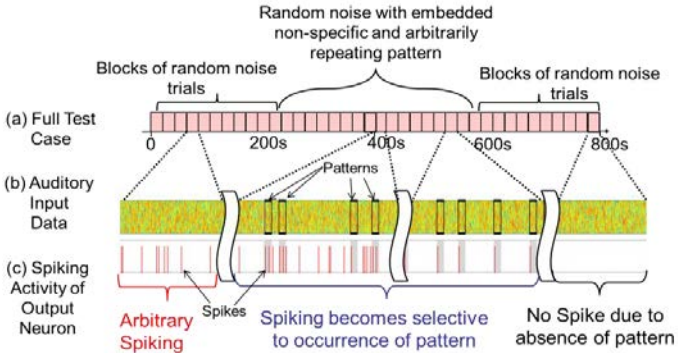


Fig.18 (a) Full auditory-data test case with noise and embedded repeated patterns. (b) Auditory input data and (c) spiking activity for selected time intervals of the full test case of the output neuron (shown in Fig.16b).

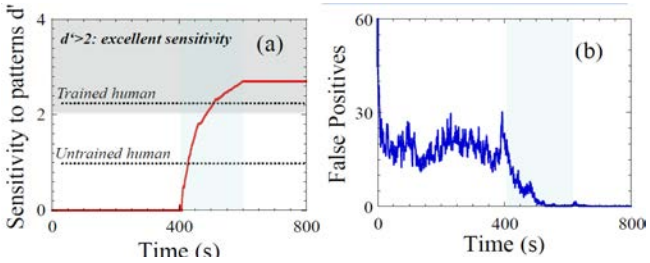


Fig.19 (a) Pattern Sensitivity index (d') for the test case shown in Fig.18a. The system reaches a very high sensitivity (d'>2). (b) Number of false detections by the output neuron during the learning case of Fig.18.

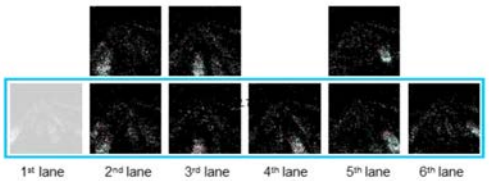


Fig.20 Final sensitivity map of 9 output neurons from the 1st layer of the neural network shown in Fig.17b. Average detection rate for 5 lanes was 95%.

Network Statistics		Energy/Power Statistics	
Audio Test Case		*Audio Test Case*	
Total Set Events:	102646	Total duration: 800 s	
Total Reset Events:	41810	Total Energy dissipated: 436 μJ	
Total Read Events:	2.10×10^7	**Synaptic Programming Power:**	
Total CBRAM Synapses: 192		**0.55 μW**	
Video Test Case			
Total Set Events:	449725	*Video Test Case*	
Total Reset Events:	26837412	Total duration: 680 s	
Total Read Events:	2.49×10^9	Total Energy dissipated: 50.4 mJ	
Total CBRAM Synapses: ~ 2 million		**Synaptic Programming Power:**	
(Iset = 155μA, Ireset = 90μA)		**74.2μW**	

Table 1 Network statistics and power dissipation for the two applications.