A Non-Volatile ReRAM and Subthreshold-FET-Based Kernel for Energy-Efficient Temporal Signal Processing Using Reservoir Computing

G. S. Chaitanya[1]*, A. Arora[1]

[1]*Department of Electronics Engineering, Indian Institute of Technology (BHU) Varanasi, Varanasi – 221005, India*

The escalating demand for energy-efficient artificial intelligence has catalysed interest in neuromorphic hardware capable of real-time temporal processing. Reservoir Computing (RC) offers a compelling framework by leveraging the intrinsic dynamics of nonlinear systems [1]. However, conventional memristive RC implementations are constrained by the devices' intrinsic dynamics, limiting their adaptability across diverse edge AI workloads, spanning timescales from milliseconds to seconds.

In this work, we propose a novel RC kernel architecture as shown in Fig. 1(a) that synergistically integrates subthreshold-operated FETs with non-volatile ReRAM devices, enabling post-deployment tunability of reservoir time constants. By operating FETs in the subthreshold regime, we exploit exponential $I_D - V_G$ characteristics as shown in Fig. 1(b), analogous to biological ion channels, thereby achieving rich nonlinear dynamics at low power, as shown in Fig. 1(c). Further, ReRAM elements configured as programmable gate resistances offer real-time control over the kernels' time constant across multiple orders of magnitude.

Our analysis reveals that tailoring the subthreshold behaviour of FETs, particularly through reductions in the subthreshold slope, can significantly enhance the reservoir's memory capacity as shown in Fig. 1(d), which is a key determinant for effective temporal processing. Emerging steep-slope transistor technologies, such as tunnel FETs (TFETs) and ferroelectric FETs (FeFETs), present promising avenues for achieving this by enabling sharper transitions, broader nonlinear regimes, and reduced energy-delay products [2].

The proposed architecture achieves a low normalized root mean squared error of 0.0513 on benchmark time-series tasks such as the Henon map with an average energy consumption of $0.152\ nJ$ per timestep, evaluated over 2000 input pulses. The interplay between the device subthreshold parameters and their impact on the performance of reservoir-based temporal processing will be discussed thoroughly.

These findings underscore a critical opportunity for device-architecture co-optimization, where advances at the materials and device levels directly translate into scalable, adaptive, and energy-efficient neuromorphic systems.

References

[1] M. Yan, C. Huang, P. Bienstman, et al., Nat. Commun. 15 (2024) 2056.

[2] Y. Zhai, Z. Feng, Y. Zhou, S.-T. Han, Mater. Horiz. 8 (2021) 1601–1617.

* gseshasai.chaitanya.cd.ece22@iitbhu.ac.in

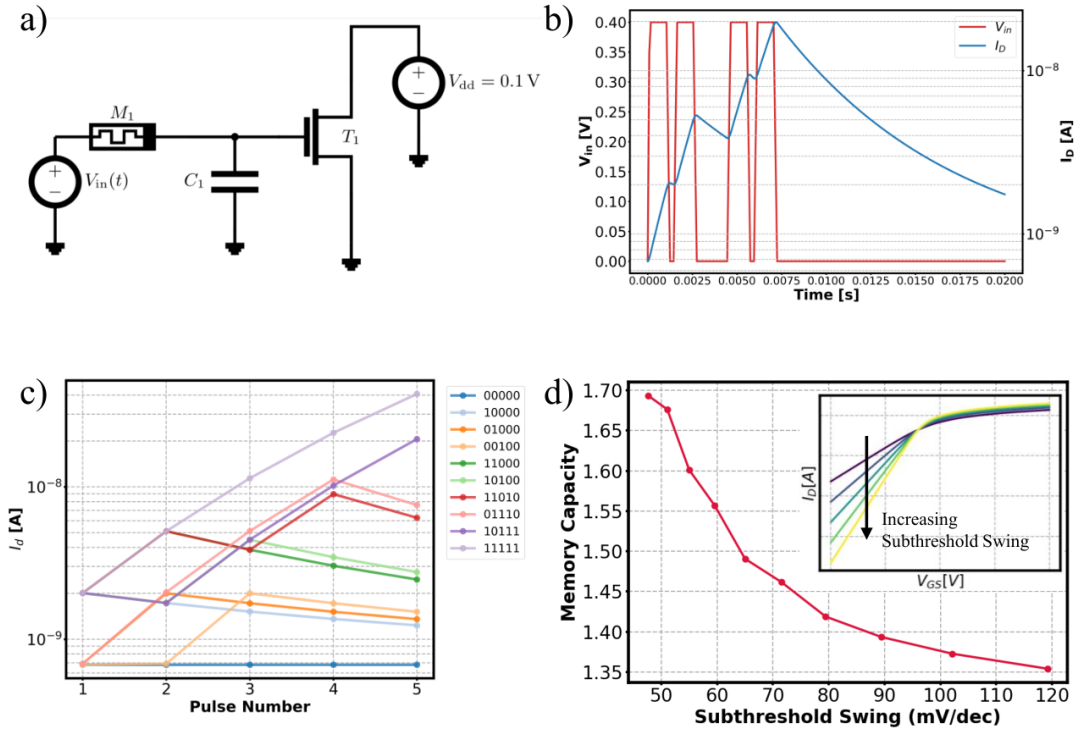

Fig.1 (a) Schematic of the proposed FET based kernel. The input voltage is fed to a memristor ($M1$) - capacitor ($C1$) network which is further connected to the gate of a DC-biased NMOS FET operating in the sub-threshold region. (b) Drain current response of the kernel in response to the input pulse voltages. Here we have chosen an input pulse train corresponding to the binary sequence '11011' (c) Drain current evolution across different binary sequqences of input pulses, demonstrating the separability of I_D values (d) Variation of memory capacity of the proposed kernel across different values of subthreshold swing. A lower subthrehshold swing results in a higher memory capacity of the kernel. Inset shows the simulated $I_D - V_G$ characteristics for different subthreshold swings, at a fixed threshold voltage.