Tunable Fine Learning Rate controlled by pulse width modulation in Charge Trap Flash (CTF) for Synaptic Application

Shalini Shrivastava, Udayan Ganguly
[1]*Indian Institute of Technology Bombay, Mumbai, India*
Email: shalinishrivastava@iitb.ac.in

Abstract: The brain-inspired neuromorphic computation is on high demand for the next generation computational systems due to its high performance, low-power and high energy efficiency. The highly mature technology of today, Flash memory, is the first and has been a promising electronic synaptic device since 1989. The linear, gradual and symmetric learning rate are the basic requirements for a high performance synaptic device. In this paper, we demonstrate a fine-controlled learning rate in Charge Trap Flash (CTF) by pulse width modulation of input gate pulse. We further study the effect of cycle to cycle (C2C) and device to device (D2D) variability, and limits of charge fluctuation with scaling on the learning rate. The comparison of CTF as synapse with other state-of-the-art devices is carried out. The learning rate with CTF can be tuned from 0.2% to 100%, which is remarkable for a single device. Further, the C2C variability does not affect the conductance however it is limited by D2D variability only for learning levels > 8000. We also show that the CTF synapse has a lower sensitivity to charge fluctuation even with scaled devices. The designable learning rate, and lower sensitivity to variability and charge fluctuation in CTF synapse is significant compared to the state-of-the-art. The tunable learning rate of CTF is very promising and of great interest for brain-inspired computing systems.

Motivation: Brain-inspired computation promises complex cognitive tasks at biological energy efficiencies fig.1.

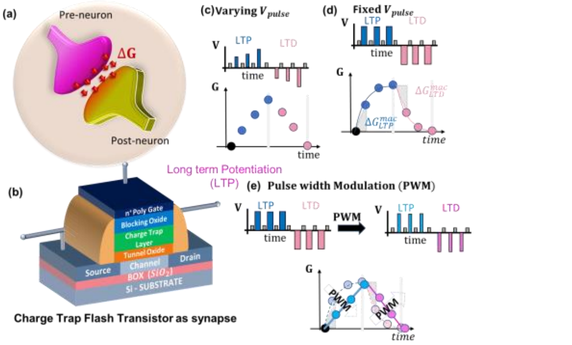
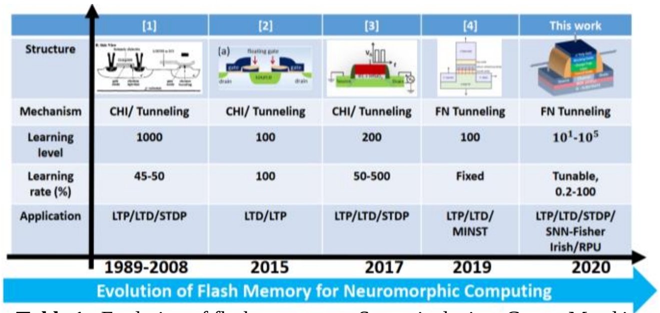


Fig. 1 **a)** A synapse connects a pre-neuron to a post-neuron. **b)** An artificial synapse is implemented in a Charge Trap Flash (CTF) enables LTP & LTD by the threshold voltage (V_T) shift to replicate the learning rate behavior at fixed gate pulses. **c)** G changes with the V_{pulse} increase is typical memristors behavior. **d)** G change and then saturates when fixed V_{pulse} is applied repeatedly. The ΔG^{max} occurs initially. This is the critical requirement for SNNs **e)** pulse width modulation (PWM): the input gate pulse width is reduces to get linear conductance results in tunable learning rate.

The evolution of flash memory as a synaptic device shown in table 1. The channel hot injection (CHI) for programming during LTP uses high current to inject electrons in the floating gate, resulting in a high energy loss. Negative bias temperature instability (NBTI) based trap generation in high κ dielectric MOSFETs is shown to mimic synaptic activity, but it requires large electrical stressing to prepare the device [2-3]. Thus, a highly energy-efficient, scalable synaptic device with gradual LTP and LTD is recently demonstrated [4]. Recently, instead of large digital circuits to represent analog weights, various nanoscale memristive synapses have been proposed [5]. These memristive devices store the weight as an analog conductance (G) value to provide excellent areal density improvement. The synapse "learns" by conductance change (ΔG) which depends upon the spike time difference (Δt) of pre- and post-neurons (Fig. 1b), known as spike time dependent plasticity (STDP). For memristive synapses, the application of a Δt-dependent pulse voltage (V_p) at fixed pulse-width (t_p) causes ΔG. The ΔG per pulse depends upon both the instantaneous conductance G and V_{pulse}.



	[1]	[2]	[3]	[4]	This work
Structure					
Mechanism	CHI/ Tunneling	CHI/ Tunneling	CHI/ Tunneling	FN Tunneling	FN Tunneling
Learning level	1000	100	200	100	10^1-10^5
Learning rate (%)	45-50	100	50-500	Fixed	Tunable, 0.2-100
Application	LTP/LTD/STDP	LTD/LTP	LTP/LTD/STDP	LTP/LTD/ MINST	LTP/LTD/STDP/ SNN-Fisher Irish/RPU
	1989-2008	2015	2017	2019	2020

Evolution of Flash Memory for Neuromorphic Computing

Table 1: Evolution of flash memory as Synaptic device- Carver Mead in 1989, first propose the flash memory for neuromorphic computing. The table shows the evolution of Flash memory over time with technological benefit. Programing and erasing mechanism plays important role in energy consumptions.

Synapses have two key challenges. First, both LTP and LTD need to be gradual SNN algorithms in software [6]. In fact, the learning rate i.e. the $max(\Delta G)$ for repeated *maximum* V_{pulse} needs to be low (<2% of the total range of G) for stable weight evolution in a network during training [7-8]. In other words, the G change should saturate after larger number of identical V_{pulses} i.e. in excess of 256 identical pulses for analog-valued datasets like Fischer Iris. Secondly, low energy operation (especially write-energy) in the analog synapse is a critical challenge for memristors [9]. Thus, a highly energy-efficient, scalable, gradual and designable learning rate with synaptic device is desirable. Stochastic learning in deep neural networks based on nanoscale is demonstrated in PCMO RRAM [10]. Earlier we have demonstrated the bulk Si technology based CTF capacitors exhibit gradual V_T shift by FN tunneling to enable symmetric LTP & LTD with high energy efficiency. A mathematical model of the experimental LTP/LTD was developed [15]. An SNN for Iris classification was implemented with CTF synapse and STDP was demonstrated. Recently with this device we have demonstrated the software-level accuracy using stochastic computing [16]. In this paper, (1) the gradual and fine-tuned learning rate is demonstrated by pulse width modulation. (2) To evaluate the reliable V_T shift C2C variation within device (noise) and D2D variation (variability) as a standard deviation (σ) on a ΔV_T shift per spike is experimentally demonstrated. (3) A comparison of $\sigma/\ Range,$ essentially noise-limited resolution vs. range of scaled devices– as well as fundamental limits of charge fluctuation, is carried out. Finally, a benchmarking of this device with the state-of-the-art is presented.

Tunable learning rate: Effected by Variability and Scalability: The gradual V_T shift with pulse number is designed by varying the pulse-width Fig. 2 shows V_T shift becomes more gradual with decreasing pulse-width for a fixed pulse amplitude (12.5 V for LTD & -12.5 V for LTP). The results from Fig. 2a is tunable by modulating the input gate pulse width from 10 ms to (8, 5, 3, 0.8) ms for LTP and 5 μs to $(2.5, 1, 0.5, 0.3)\ \mu s$ for LTD with a fixed voltage. Experimentally, $\sim 10^3 - 10^5$ states for LTP and LTD are demonstrated by pulse-width reduction, which is $10^2\times$ improved than Flash [1-4] or memristor devices [5]. As shown in Fig. 2c, none of the other devices demonstrated as the synapses in the literature have achieved designable and lower learning rate specification.

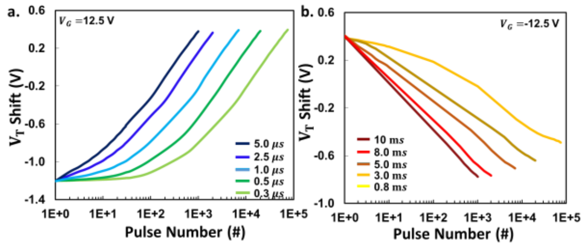
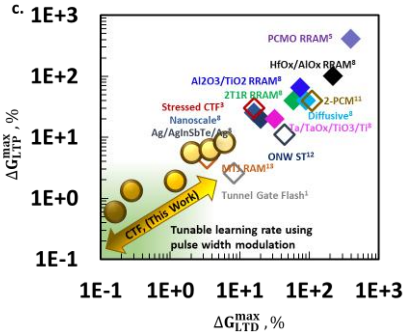


Fig. 2. Experimental **a)** LTD behaviour can be tuned by the pulse-width arbitrarily, for fixed V_G 12.5 V (LTD). b**)** LTP behaviour can be tuned by the pulse-width arbitrarily, for fixed V_G -12.5 V (LTP).



c) The tunable (shed of yellow circles) learning rate i.e. ΔG_{LTP}^{max} and ΔG_{LTD}^{max} normalized by range $(G_{max} - G_{min})$ is $10\times$ lower for CTF compared to the state-of-the-art.

Variability: Tunable learning rate with pulse width variation has fixed ΔV_T window (fig 3a). The variability (C2C &D2D) each for 10 cycles are demonstrated in fig.3a. LTD has fixed window with varing pulse width. The deep trap in the nitrite layer of CTF results in not fully erased in LTP, results in $reduced$ ΔV_T for lower pulse width. The C2C, σ/μ in LTP/LTD is < 0.05 and fixed. C2C variation does not limit the learning rate shown in fig3b. However, D2D σ/μ – increases with decreases in pulse width (fig3b), hence the tunability of learning rate is limited by D2D variability for > 8000 learning level.

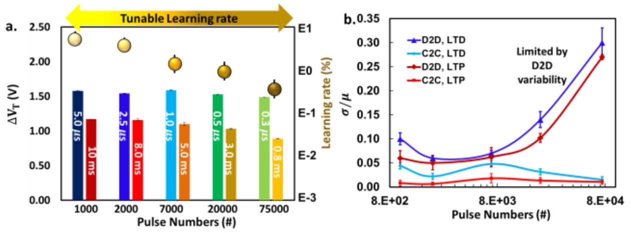


Fig. 3. Tunability with pulse width variation and Variability (C2C &D2D) each for 10 cycles and devices. **a)** LTD- Programing, has fixed window with pulse width, whereas LTP-erasing, ΔV_T $slightly$ decreases with decrease in pulse width. Max of learning rate of LTD/LTP (from fig. 2c) on secondary axis, with variability in learning rate is as bar error. **b)** σ/μ of V_T shift per spike for C2C in LTP & LTD is < 0.05 and constant. Whereas, D2D σ/μ of V_T shift per spike – increase with decreases in pulse width for both LTP/LTD, hence limit the learning rate by D2D variability.

Scalability: The noise-limited resolution vs range is critical for scaled device. The $\sigma/Range$ of scaled devices, as well as fundamental limits of charge fluctuation, the flash synapse (this device), has lower sensitivity even with scaling compared to PCMO-RRAM[5] & PCM[11] synapse variability (fig 4.)

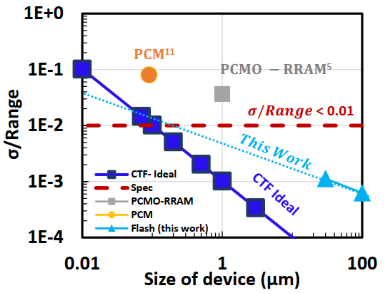


Fig. 4. A comparison of $\sigma/$ $Range$, of scaled devices PCM and RRAM with our work – as well as fundamental limits of charge fluctuation. Flash synapse has lower sensitivity to charge fluctuation even with scaled device. The standard deviation (σ) on a ΔV_T and $range$ $(= V_{Tmax} - V_{Tmin} \approx 1.6V)$, for this device

Benchmarking & Conclusion: The benchmark with the state-of-the-art artificial synapses presented in table 2.

Synapse Technology	Energy (fJ)	Area (F^2)	Timing (ns)	CMOS Compatible	LTP/ LTD cycles	Learning level	Gradual & Symmetric
Human Brain[7]	10^4	-	10^6	-	∞	∞	yes
SRAM[14]	0.5	120	1	High	∞	2	No
STTRAM[14]	$>10^4$	20	>2	Mid	10^{12}	~100	No
PCM[11]	$>10^3$	4	50	High	10^{12}	>100	No
RRAM[5,8,9]	$>10^4$	4	>10	High	10^{11}	~100	No
ONW ST[12]	1	8	10^4	Low	10^4	20	Yes
Tunneling gate Flash[1]	$>10^5$	75	10^4	High	10^4	1000	Yes
SST's ESF[2]	$>10^3$	20	$>10^4$	High	NR	100	yes
Stressed MOSFET[3]	10^5	8	$>10^4$	High	NR	~200	No
FB ST[15]	$>10^3$	8	10^3	High	NR	20	Yes
SONOS CTF[4]	10	8	10^6	Highly	–	fixed	yes
This work	2.5	8	10^5	Highly	$>10^6$	$10\text{-}10^5$	Yes

Table 2: Comparison with state of art. The energy for this work is estimated at 180 nm technology*

The CTF is among the lowest energy possible after SRAM technology, which is volatile and expensive in terms of area (>120 F^2), binary and not amenable to cross-bar implementation. The timescale of CTF is comparable to biology (~1 ms), which can be interesting for some real-time learning applications from natural data – as opposed to accelerated applications in software. This technology is highly manufacturable in the CMOS silicon industry and shows the gradually $\sim 10^5$ level of learning, which is a record improvement by >2 orders. Thus, the energy of write, CMOS compatibility, technological maturity, gradual and symmetric LTP and LTD, and LTP/LTD cycling reliability are the significant advantages. The tunable learning rate (0.2%-100%) from a single synaptic device is remarkable compare to the other devices in state of art. The fundamental limits of charge fluctuation in CTF synapse has lower sensitivity even with scaling compare to the state-of-the-art. The CTF is very promising and of great interest to the community.

References: [1] Diorio. et al., *IEEE TED*,1972-1980. [2] F.M. Bayat, et al., 2015, ISCAS. [3] Gu, X., et al., 2017, EDL. [4] Agarwal, et al., 2019. JESSCDC. [5] Panwar, N., et al., 2014, *DRC*. [6] Yu, S., 2018.*Proceedings of the IEEE*. [7] Abbott, L.et al., 2000, *Nature neuroscience*. [8] Aditya S, et al., IJCNN 2018. [9] Rajendran, et al., 2013, TED. [10] AV Babu, et al., 2018 *Neurocomputing*. [11] Bichler, O. et al., (2012). *TED*. [12] Xu, W. et al.,2016, S*cience advances*. [13] Krzysteczko, P., et al., 2012. *Advanced Materials*. [14] Kim, H., et.al., 2017. *Nanotechnology*. [15] S. Shrivastava, et al., (Feb. 2019). https://arxiv.org/abs/1902.09417. [16] V. Bhatt, et al., IJCNN 2020.