Extended Abstracts of the 2020 International Conference on Solid State Devices and Materials, VIRTUAL conference, 2020, pp81-82

Impact of PCM Flicker Noise and Weight Drift on Analog Hardware Inference for state-of-the-art Deep Learning Networks

J.-P. Han[1], M. J. Rasch[1], Z. Liu[2], P. Solomon[1], K. Brew[2], K. Cheng[2], I. Ok[2], V. Chan[2], M. Longstreet[3], W. Kim[1], R. Bruce[1], C. Cheng[1], N. Gong[1], P. Adusumilli[2], S. Pinkett[2], H. Li[3], N. Saulnier[2], T. Yamashita[2], M. Brightsky[1], and V. Narayanan[1]
[1]IBM T. J. Watson Research Center, Yorktown, USA, phone: +1-914-945-1778, email: hanjp@us.ibm.com
[2]IBM Research - Albany Nano Tech, USA, [3]IBM System, Hopewell Junction, NY, USA

Abstract **We have characterized flicker noise of $Ge_2Sb_2Te_5$ (GST) based phase change memory (PCM) cells and found agreement with potential trap activation and defect annihilation in asymmetric cell structure for reset and set, respectively. We assessed the impact of flicker noise and drift on analog deep learning (DL) hardware for state-of-the-art networks at scale. We show the effect of accumulated flicker noise $\propto \sqrt{A_r * wait\ time}$ and conductance drift $\propto (wait\ time)^{-\upsilon}$, where A_r is the flicker noise figure of merit (FOM) and υ is the drift coefficient, on inference accuracy. We find that hardware aware (HWA) retraining is essential and a tight control of A_r and υ is vital for DL inference.**
Keywords: **flicker noise, drift, PCM, analog DL Inference.**

Introduction Analog crossbar arrays have been broadly investigated for DL acceleration [1]. The effect of conductance drifting has been addressed [2,3], yet the impact of low frequency flicker noise (incl. *1/f* [4] and random telegraph noise (RTN)) has received little attention, even though such noise might affect inference and training accuracy considerably. Here, we systematically characterize flicker noise of typical GST based PCM cell [5] processed in advanced technology back end of line (BEOL). We describe read voltage polarity dependence of the power spectral density (PSD) in terms of asymmetric cell structure and trap-assisted mechanism and discuss the dependency of A_r and υ on device resistance (R). We evaluate their impact on analog DL inference by including these phenomena in a hardware simulator for state-of-the-art DL networks at scale [6]. This is the first study of such networks to provide insight into the importance of flicker noise and weight drift.

Characterization of flicker noise The GST device is pre-programmed with full reset, partial reset, and set pulses prior to measuring flicker noise (Fig. 1a), then a read voltage is applied at the top electrode (TE) or the bottom electrode (BE) of the asymmetric cell structure. The normalized PSD ($S_{id}/i^2(f)$; Figs.1de) shows that the PSD gap between reset and set in the BE case is much bigger than in the TE case, even though that device I-V is symmetric in the reset state. The asymmetric structure and the existence of traps may contribute to the read voltage polarity dependence. As shown in Figs. 2a-f, crystalline GST (set state) has mainly acceptor-type traps, while amorphous GST (reset state) has, in addition, donor-type and lone pair traps [7]. In the reset state, when 0.2V is applied at BE, it may trigger the donor-type traps near the heater, resulting in increased *1/f* noise. In the set state, vacancy-type defect annihilation may occur causing the *1/f* noise to decrease, however, reduced set resistance may also play a role here. In the TE case, the phase segregation due to the asymmetric cell structure prevents traps from being triggered or destroyed.

Flicker FOM for analog DL The empirically determined FOM for assessment of flicker noise on analog DL are defined in Figs. 3ab. For well-behaved *1/f* noise, it is $S_{id}/i^2(f) = A_r/f$, where A_r is scale independent. This was modeled, in our resistive processing unit (RPU) simulation tool [6], using a log-uniform spectrum of RTN traps, each trap causing a fractional change in weight value (Figs. 3ac). For flicker noise with a prominent shoulder, we extrapolated A_r' (Figs. 3bd) which approximately integrates to the same amount as the A_r'/f spectrum. In a 3-layer FC-DNN on MNIST, we find that during the waiting time before inference the flicker noise will accumulate, causing an increase in test error (Figs. 3ef). Note that this toy-level network may underestimate HW requirements for realistic cases (see below), although it provides initial guidance for prototype development.

Resistance dependence of flicker and drift Flicker noise $S_{id}/i^2(f)$ (Fig. 4a) and resistance drift $\log(R/R_0) = \upsilon\log(t/t_0)$ [8] (4b) show clear resistance dependency, a positive correlation between A_r (or A_r') and υ is evident (4c). Double well potential (DWP; see Fig. 4d) have been used to illustrate defect fluctuation with voltage bias modulation, which may explain the correlation of flicker and drift [9,10], although mechanisms may differ. The asymmetry in the two barriers (Fig. 4e) results in different transition rates between energy potential minima for structural relaxation or crystallization.

Impact on state-of-the art DL networks To study the impact of flicker noise and drift at scale, we investigate their effects on inference for large vision networks (8M-143.7M free parameters, see Fig. 5) on the ImageNet dataset (1.2M images, [11]). First, we initialize with floating-point (FP) trained weights (from [12]) and used our standard RPU specification, which includes clipped weight ranges, DAC/ADC discretization and system noise etc. (see [1, 6]), for inference. Accumulated flicker noise is added to the initial weights as a waiting time t dependent Gaussian noise with variance $\sigma^2=A_r \ln(t/t_{read})$ (RPU array integration time t_{read} is 100ns). Figs. 6ad show the impact of flicker and drift on the top-1 error when system noise is turned off (vertical lines mark example values for A_r, υ, and t). In Fig. 6be, realistic RPU system noise is turned on and results in severe degradation of inference accuracy. In Fig. 6cf, we retrain the FP networks by including the RPU specification into the forward training pass only (similar to [13], but with distilling [14]). When retraining in this hardware-aware (HWA) fashion, inference becomes robust to all analog RPU imperfections. Moreover, flicker tolerance can be further improved when including additive weight noise (reminiscent of *1/f* noise) in HWA training (dashed lines in Fig. 6c). Our results show clearly that HWA training is necessary for analog inference at scale.

Conclusion We have established a methodology of characterizing flicker noise of analog synaptic devices (e.g. PCM) and evaluating inference performance on analog DL. Results indicate that inference may be prone to flicker noise, in addition to weight drift and system noise, especially for large-scale networks. Material and process optimization along with HWA retraining will be needed to mitigate the overall noise impact on analog inference DL hardware.

[1] Gokmen & Vlasov, Front. Neurosci. (2016) [2] Ambrogio et al., IEDM (2019) [3] Nandakumar et al., IEEE-ICECS (2019) [4] Fugazza et al., IEDM (2009) [5] Burr et al., IEEE Trans. Elec. Dev. (2015) [6] Rasch et al., IEEE Design & Test (2019) [7] Pirovano et al., IEEE Trans. Elec. Dev. (2004) [8] Karpov et al., J. Appl. Phys. (2007) [9] Ielmini et al., IEDM (2007) [10] Nardone et al., Phys. Rev. B (2009) [11] Deng et

al., CVPR09 (2009) [12] https://pytorch.org/docs/stable/torchvision [13] Gokmen et al., IEDM (2019) [14] Hinton et al., arXiv:1503.02531 (2015)

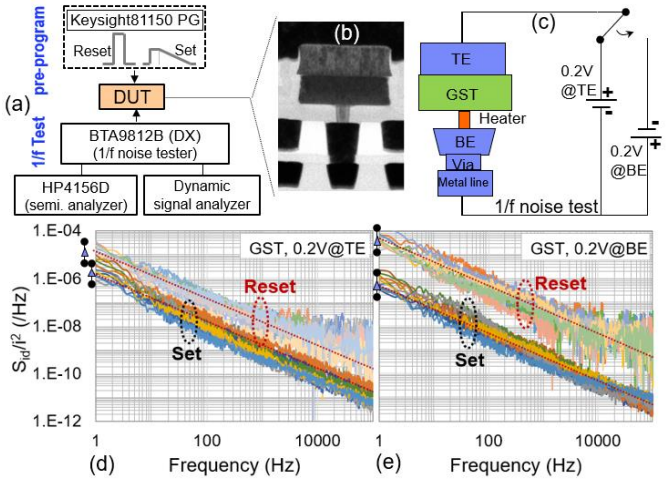

Fig. 1 (a) Pre-programing and testing *1/f* noise on GST device under test (DUT), asymmetric cell cross-section TEM (b) with 0.2V applied to top electrode (TE) or bottom electrode (BE) (c). Corresponding normalized PSD of reset / set when read from (d) TE or (e) BE.

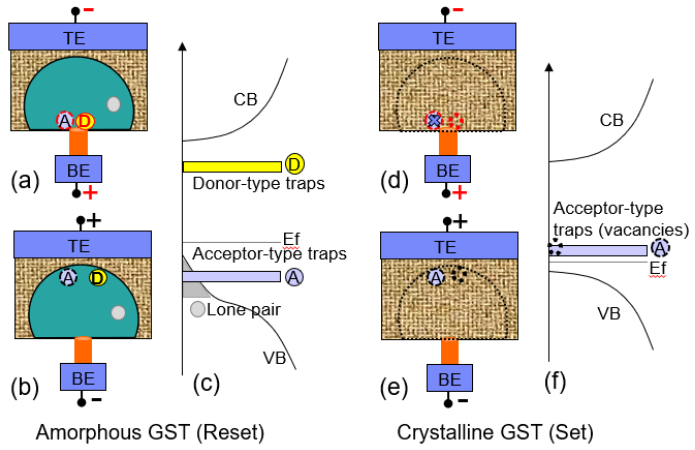

Fig. 2 Illustrations for explaining read voltage polarity dependence for amorphous (a-c) and poly-crystalline GST (d-f). Band diagrams of set (c) and reset (f) with various types of defects [7]. If read applied at BE, donor-type traps near the heater increase flicker in reset state, while possible vacancy-type defect annihilation decreases flicker noise in set state. In the TE case, phase segregation may prevent both.

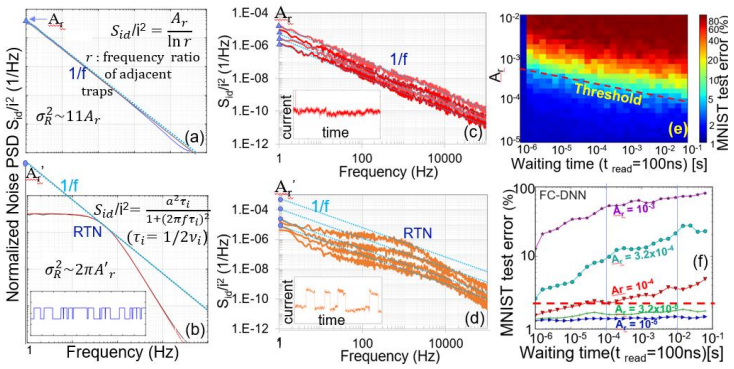

Fig. 3 Empirically determined FOM (a) A_r for simulated *1/f*, (b) A_r' for simulated RTN. (c) A_r and (d) A_r' obtained experimentally from PCM devices, extrapolated from normalized $S_{id}/i^2(f)$. (e-f) Inference impact for MNIST 3-layer fully connected (FC) DNN toy network with our analog RPU model [1] for different settings of A_r and waiting time t. Note that the test error degrades with increasing A_r or t.

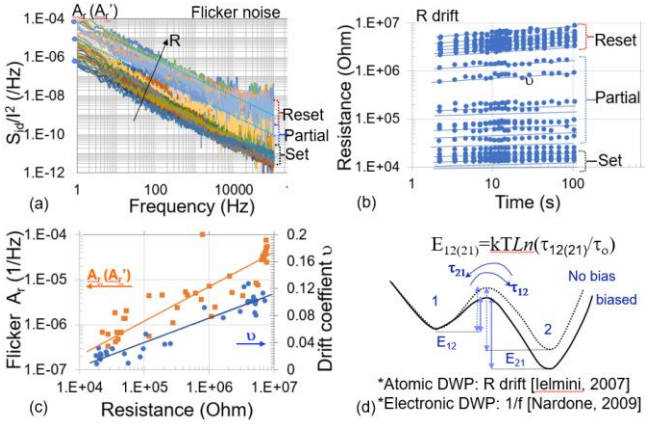

Fig. 4. PCM reset / partial set / set states (a) *1/f*: S_{id}/i^2 vs. f. (b) drift: R vs. t. (c) A_r and drift coefficient υ as a function of R showing clear correlation. (d) Double well potential modulation due to defect fluctuation in 1/f (electronic DWP) and structural relaxation in drift (atomic DWP) may explain the correlation between A_r and υ.

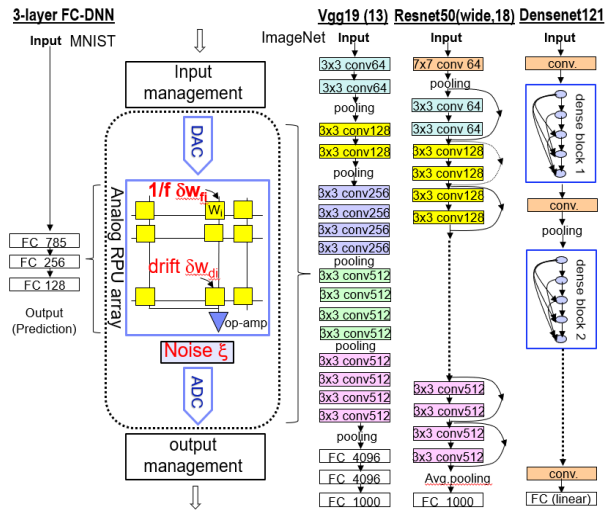

Fig. 5. To investigate the effect *1/f* noise and weight drift for analog DL inference, we performed RPU hardware simulation [1,6] of 6 large, state-of-the-art networks on ImageNet: VGG13, VGG19, ResNet18, ResNet50, WideResNet50, Densenet121. For a size comparison, a toy 3-layer FC-DNN for MNIST is shown on the left.

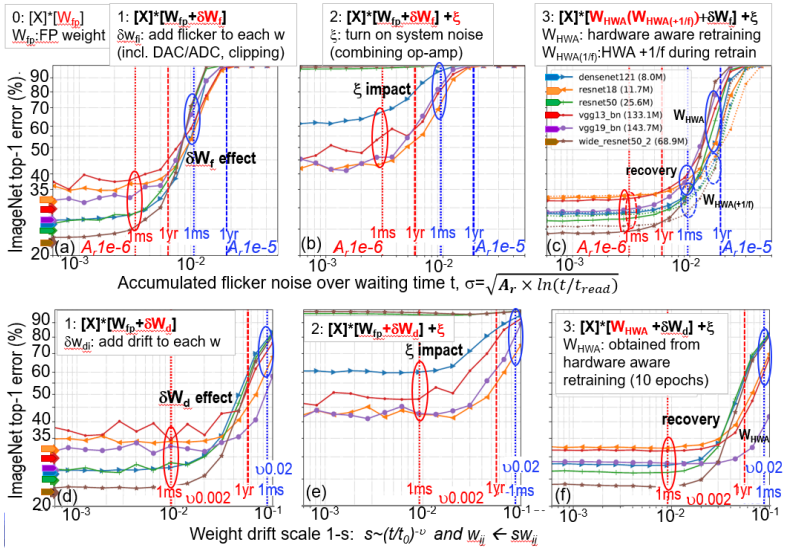

Fig. 6 (a) RPU model without system noise using FP trained weights. Larger A_r or longer waiting time t increases top-1 error. (b): as in (a) but with system noise on show failure. (c): HWA retraining recovers functionality. HWA (*+1/f*) retraining increases flicker tolerance (dotted lines). (d-f): as in (a-c) but for drift, showing degradation (d), ξ impact (e), HWA recovery (f). Left-hand color arrows: FP baseline references for 6 state-of-the-art networks as labeled in legends in (b).