Next Generation MRAM for Energy-Efficient AI Applications

W. Hwang[1], F. Xue[1], M.A. Di Perna[1], W. Tsai[2], D. Fan[3], S. X. Wang[1,2]

[1]*Department of Electrical Engineering, Stanford University, Stanford, California 94305, USA*

[2]*Department of Materials Science & Engineering, Stanford University, Stanford, California 94305, USA*

[3]*School of Electrical, Computer and Energy Engineering, Arizona State University, Tempe, Arizona 85281*

Energy-efficient computing, particularly in exploding data-intensive AI applications spurred by large foundation models, is essential for addressing the rising power demands and ensuring sustainable technology advancement. Next Generation Magnetoresistive Random Access Memory (MRAM) has emerged as a promising technology in this domain, offering nonvolatile memory solutions that combine low power consumption with high performance. Spin-Orbit Torque MRAM (SOT-MRAM) and its variants stand out for its potential to deliver SRAM-like performance at a higher bit-cell density. In this talk, we present a novel high-density STT-assisted SOT-MRAM (SAS-MRAM) technology designed for energy-efficient Edge AI applications [1]. SAS-MRAM capitalizes on the advantages of both spin-transfer torque (STT) and spin-orbit torque (SOT) mechanisms to achieve field-free deterministic magnetization switching while preserving high density cell array layout. Importantly, SAS-MRAM utilizes a multi-bit-shared SOT write line to achieve high-speed, high-density, and high-endurance performance. Our experimental results validate the potential of SAS-MRAM, including multi-bit programmability with a single SOT write current pulse [2] and nanosecond timescale writability without being assisted by an external magnetic field, which are desirable to address limitations of current STT-MRAM and SOT-MRAM technologies. An AI accelerator emulation chip with on-device continual learning and SAS-MRAM based In-Memory Computing (IMC) shows an impressive performance of ~3277 GOPS and an energy efficiency of 131 TOPS/W, which is better than similar implementations with CMOS SRAM or Resistive Random Access Memory (RRAM), showing great promises for low power training and inference of Edge AI models [1,3].

This work was supported in part by the National Science Foundation and industry partners under Grant No. 2328804 in the Future of Semiconductors (FuSe) program, Grant No. 2345655 in the Partnerships for Innovation (PFI) program, Grant No. 2314591 in the Advanced Chip Engineering Design and Fabrication (ACED Fab) program, and in part by the Stanford Nonvolatile Memory Technology Research Initiative (NMTRI) Affiliates program. SXW also acknowledges support from the Semiconductor Research Corporation (SRC) and TSMC, and collaboration from Sandia National Laboratory and Honeywell Aerospace.

References

[1] F. Xue, W. Hwang, F. Zhang, W. Tsai, D. Fan, and S. X. Wang, IEEE Trans. Mag., 61(4), 3400508, 2025.

[2] W. Hwang, F. Xue, M.-Y. Song, C.-F. Hsu, T. C. Chen, W. Tsai, X. Bao, and S. X. Wang, Electronics Device Lett., 45(10), 1800-3, 2024.

[3] F. Zhang, A. Sridharan, W. Hwang, F. Xue, W. Tsai, S. X. Wang, and D. Fan, IEEE TCAD), vol. 43, no. 8, pp. 2393-2404, Aug. 2024,

* Corresponding author: email: sxwang@stanford.edu

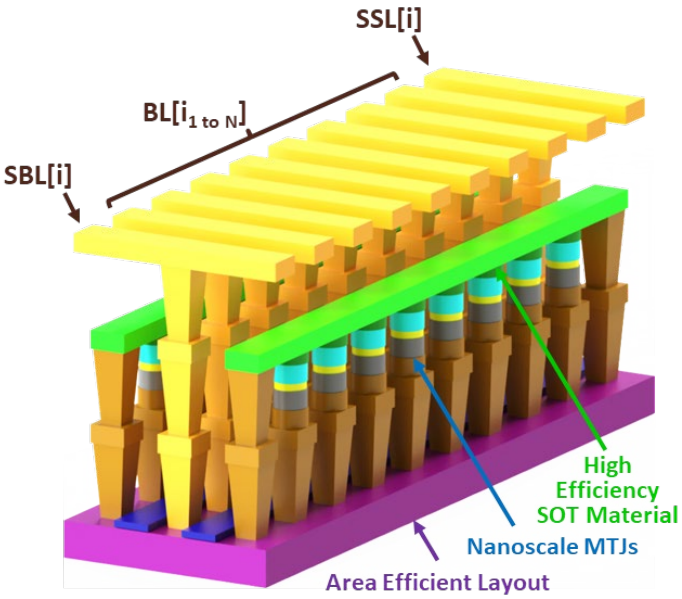

Figure 1: SAS-MRAM layout with 8 MTJs (bits) sharing one SOT write line. It achieves: 1) field-free, high-speed switching (~1 ns), 2) high cell density (~1T1MTJ), and 3) minimal current through tunnel barrier (high endurance).

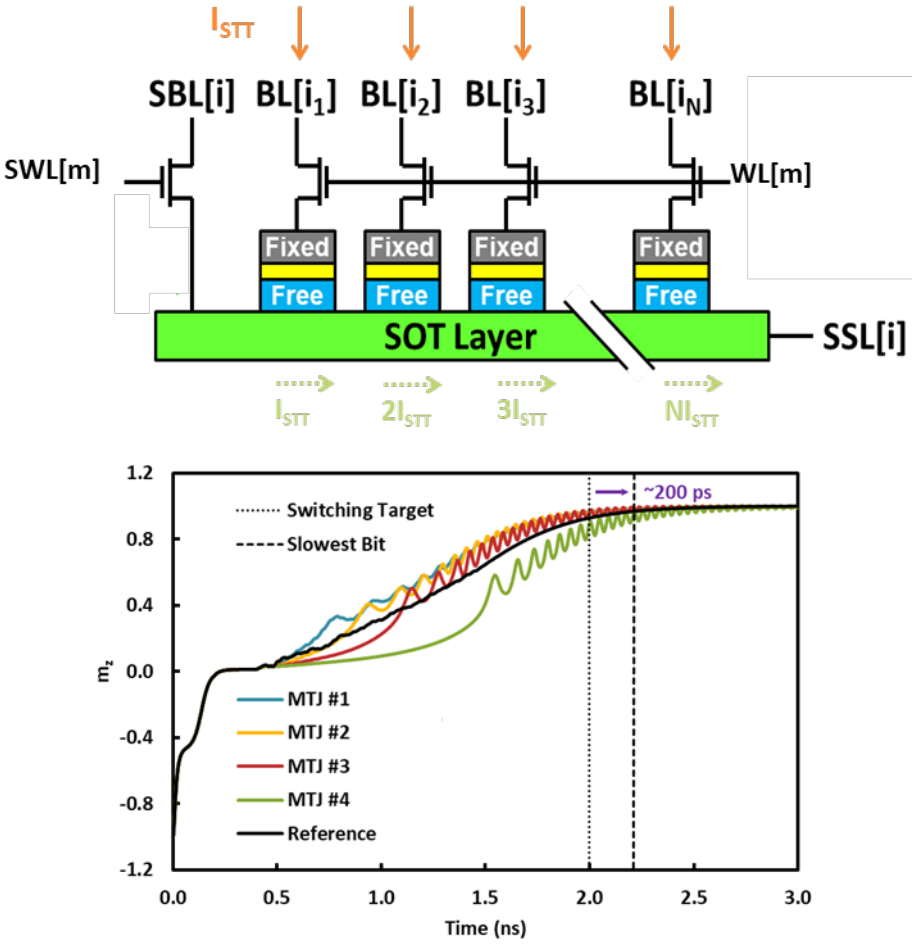

Figure 2: 1) **Top:** SAS-MRAM writing scheme 4 bits sharing one SOT write line. 2) **Bottom:** Micromagnetic simulation of worst-case scenario on 4-bit z-type SAS-MRAM device with a switching time of < 3ns for all 4 bits; SOT disturbance accelerates the switching speed of MTJ #1-#3, and delays MTJ #4 by only 200 ps, keeping deterministic SAS switching for all, including 4-16 bits per SOT line.