A 28nm Integrated True Random Number Generator Harvesting Entropy from MRAM

Kaiyuan Yang[1,2], Qing Dong[1,3], Zhehong Wang[1], Yi-Chun Shih[3], Yu-Der Chih[3], Jonathan Chang[3], David Blaauw[1], and Dennis Sylvester[1]
[1]University of Michigan, Ann Arbor, MI, [2]Rice University, Houston, TX, [3]TSMC, Hsinchu, Taiwan (Email: kyang@rice.edu)

Abstract

This paper presents an integrated True Random Number Generator (TRNG) based on the random switching behavior of Magnetic Tunnel Junctions (MTJs) under low write current. A complete TRNG is designed with minimal overhead to an existing embedded MRAM in 28nm CMOS. To the best of our knowledge, this is the first experimental study of this random process and the first TRNG implemented with commercial STT-MRAM technology. The prototype adds only 180μm^2 to a standard MRAM array for TRNG operation. It passes all NIST randomness tests across −25 to 100°C, while consuming 18pJ/bit with 66Mbps throughput at the nominal condition.

Introduction

Random Number Generators (RNGs) are key security primitives in secure systems for generating random keys and cryptographic nounces. A compromised RNG can be exploited to implement system attacks. In recent years, hardware TRNGs harvesting entropy from physical noise have been increasingly studied. Conventional CMOS TRNGs amplify device thermal noise and usually require long integration time, extensive calibration, and post-processing to ensure a high level of randomness [1-4]. Therefore, the use of other random processes, such as MOS oxide breakdown [5] or the stochastic switching of Magnetic Tunnel Junctions (MTJs) [6] and Resistive RAMs (RRAMs) [7], have been proposed. Compared to CMOS designs, MTJ and RRAM can potentially offer higher density and throughput. MTJs are more suitable for TRNGs because of their significantly better endurance. Furthermore, MTJ-based Spin Torque Transfer Magnetic RAM (STT-MRAM) is widely considered a promising CMOS-compatible, non-volatile solution for cache and main memory. Therefore, this work presents a compact and high-speed MTJ-based TRNG, which supplements a normal embedded MRAM with minimal overhead.

MTJ Switching Time-Based TRNG

The switching of MTJs from parallel (P) to anti-parallel (AP) and AP to P has been shown (both theoretically [8] and experimentally [6]) to be statistical due to thermal fluctuations. Previous work made use of the write success rate as an entropy source [6]. Specifically, certain combinations of write pulse voltage and width result in a 50% write success rate. Under such conditions, the MTJs final state can be considered a random bit. However, this design requires picosecond resolution of pulse width and an accurate write voltage to ensure the MTJ is biased at the exact 50% point, making it too expensive and complex to implement on chip. Moreover, even when accurate external instruments are used to generate the write pulse, the generated random bits still require post processing or continuous probability-based feedback to pass NIST tests. Alternatively, this work explores the switching time of MTJs under low write currents (2−3× smaller than nominal values) as an entropy source. This avoids the need for accurate control of write pulse and requires only standard MRAM peripherals to use MTJs as TRNGs, significantly reducing the complexity and area overhead of the TRNG. Only an 180μm^2 block, consisting of read/write peripherals, counters, and controllers, is added to each column. In practice, this area can be further amortized because read/write circuits are necessary for MRAM (~70μm^2) and this block is shared by a column of MTJs.

Fig. 2 (left) shows an MRAM array including the random number generation. A configurable on-chip β-multiplier current reference generates read/write currents. AP to P (write "1") and P to AP (write "0") writing require different currents in opposite directions, which are implemented by multiplexers controlled by the input data. To detect write completion, a continuous comparator monitors the voltage on bit line (BL). Because of the opposite writing current and resistance change direction, BL voltage always reduces when a successful write occurs. A compact comparator is implemented within the existing sense amplifier by reconfiguring its topologies (Fig. 2, top right). Write completion time is recorded using a ring oscillator and an asynchronous counter. A faster counter could provide higher resolution and more random bits using the same writing process. The latched comparator output (STOP signal in Fig. 2) stops the oscillator and the write current after write completion to save energy. As shown in the waveform (Fig. 2, lower right), the counter starts after word line is asserted and BL/SL are connected, while the comparator is enabled at a slightly later time in order to avoid false triggering during BL build-up phase. Since writing "0" is slower than "1" and requires higher current, the design uses only the time to write "1" as the entropy source for improved throughput and energy efficiency. The TRNG performs a normal (fast, high current) write "0" after each run to reset the MTJ for the next cycle.

MTJ Behavior and Measurement Results

Lacking accurate MTJ models, the switching behavior of MTJs is best studied using prototypes. The TRNG design was implemented together with a commercial MRAM fabricated in 28nm CMOS. To better study the write time, an 18GHz counter was used, while a slower oscillator can be used in practice. Measurement in Fig. 3a shows that the time to write "1" under a low current follows a skewed distribution, which can be best fitted with a generalized extreme function. The autocorrelation function of consecutive write times in Fig. 4a verifies that the write time is an independent variable. It is known that the counter LSBs of such a distribution can be directly used as random bits [3-4]. The latest NIST 800-90B test is employed to verify the i.i.d. assumption of the generated bit stream and estimate its min-entropy (Fig. 4b). Moreover, NIST 800-22 random test suite is used to verify the statistical behavior of the generated random bits (Fig. 4c). Fig. 3b also shows that MTJ write times are longer at lower temperatures. As a result, the distribution is wider and more high-quality random bits are harvested. Fig. 5 shows the estimated Shannon entropy and min-entropy of each LSB from measurement at three temperatures.

MTJs exhibit variations in their switching behaviors. Fifty MTJs under fixed write current yield 5 to 10 high-quality random bits (Fig. 5a), indicating the feasibility to use the same write configuration across all cells. To reconfigure an MRAM array for TRNG operation (flowchart in Fig. 5b), the write "1" current is lowered and write-1 reference is reduced to accommodate the BL voltage drop. Calibration is based on the mean and variance of the write time distribution, which decides the number and quality of generated bits. Lower currents generate wider distributions and vice versa.

We also explore optimizing the allowed TRNG write time. Fig. 3 shows that the write time distribution has a long tail. Allowing sufficient time for all writes to finish with high certainty penalizes overall throughput and energy efficiency. If the allowed time is reduced, the system clock becomes faster, but a portion of writes become invalid and the quality of higher LSBs can be reduced (Fig. 6). Optimal results occur when around 12% of total writes are discarded as tail bits. In addition,

20ns is allocated for writing "0" to reset the MTJ for the next cycle. Maximum throughput is 66Mbps, and best energy efficiency is 18pJ/b, based on the measured write time distribution. The current reference is shared by many TRNGs within an array to amortize its overhead. Assuming current reference power can be neglected via sharing, energy efficiency is 11pJ/b. The actual amortized energy efficiency will thus lie between 11 and 18pJ/b. Given a large MTJ array, throughput is easily improved in demanding applications, by accessing many MTJs in parallel. Table 1 summarizes the work and compares to state-of-the-art TRNGs. Fig. 7 provides a die micrograph.

References

[1] S. K. Mathew, et al., *JSSC*, pp. 1695-1704, 2016.
[2] S. K. Matthew, et al., *JSSC*, pp. 2807-2821, 2012.
[3] K. Yang, et al., *JSSC*, pp. 1022–1031, 2016.
[4] E. Kim, et al., *ISSCC*, 2017. [5] N. Liu, et al., *VLSI*, 2011.
[6] W. H. Choi, et al., *IEDM*, 2014.
[7] S. Balatti, et al., *JETCAS*, pp. 214-221, 2015.
[8] Z. Diao, et al., *J. Phys.: Condens. Matter*, 2007.

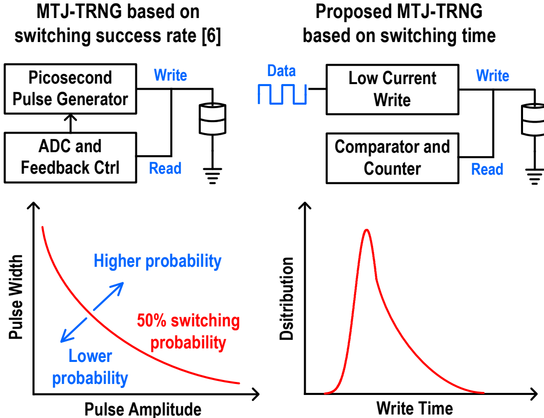


Fig. 1 Working principles of MTJ-based TRNGs using switching success rate [6] and switching time in this work.

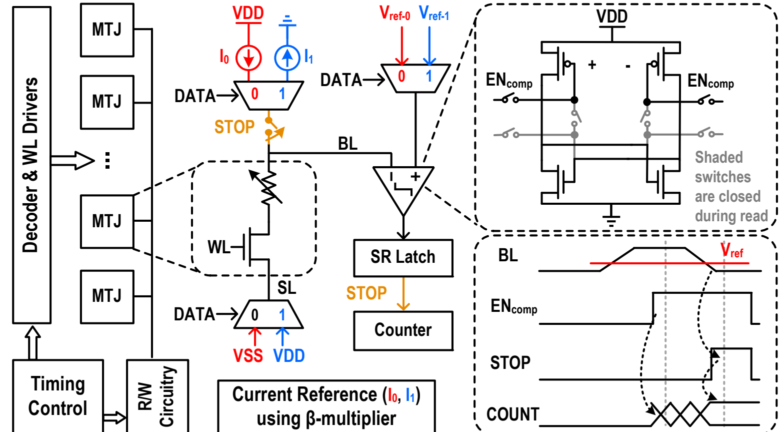


Fig. 2 Schematics and waveforms of the MTJ TRNG using a sense amplifier as continuous comparator during write. The blue and red paths indicate writing "1" and "0", respectively.

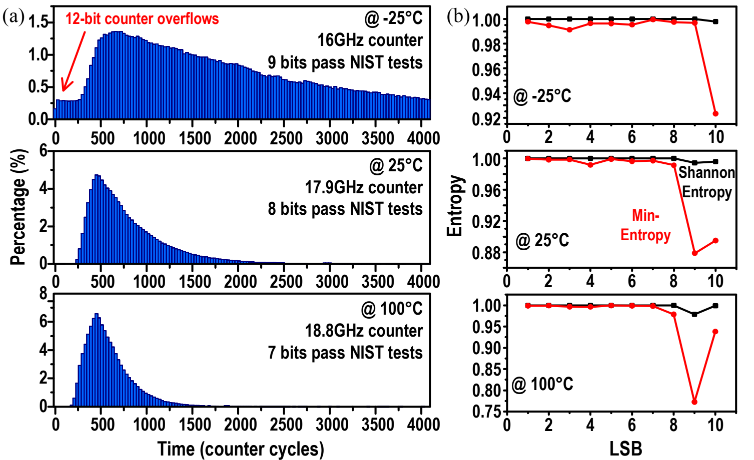
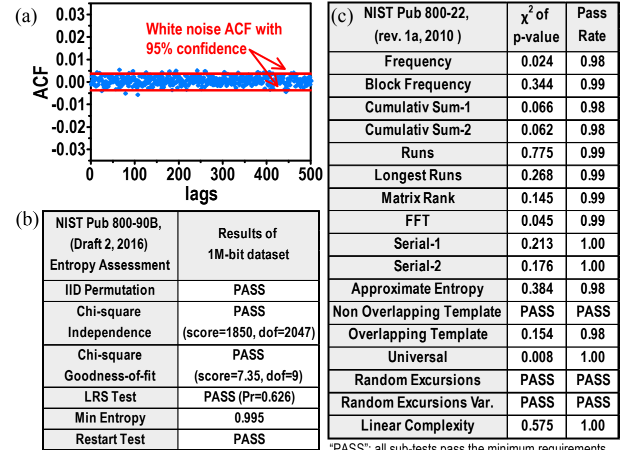


Fig. 3 Measured (a) distribution of write "1" time and (b) estimated Shannon entropy and min-entropy of the LSBs of write time at three temperatures with the same on-chip write current reference.



(b)

NIST Pub 800-90B, (Draft 2, 2016) Entropy Assessment	Results of 1M-bit dataset
IID Permutation	PASS
Chi-square Independence	PASS (score=1850, dof=2047)
Chi-square Goodness-of-fit	PASS (score=7.35, dof=9)
LRS Test	PASS (Pr=0.626)
Min Entropy	0.995
Restart Test	PASS

(c)

NIST Pub 800-22, (rev. 1a, 2010)	χ^2 of p-value	Pass Rate
Frequency	0.024	0.98
Block Frequency	0.344	0.99
Cumulativ Sum-1	0.066	0.98
Cumulativ Sum-2	0.062	0.98
Runs	0.775	0.99
Longest Runs	0.268	0.99
Matrix Rank	0.145	0.99
FFT	0.045	0.99
Serial-1	0.213	1.00
Serial-2	0.176	1.00
Approximate Entropy	0.384	0.98
Non Overlapping Template	PASS	PASS
Overlapping Template	0.154	0.98
Universal	0.008	1.00
Random Excursions	PASS	PASS
Random Excursions Var.	PASS	PASS
Linear Complexity	0.575	1.00

"PASS": all sub-tests pass the minimum requirements

Fig. 4 Randomness evaluation results: (a) Autocorrelation function of consecutive write times (AP to P); (b) NIST 800-90B test result; (c) NIST 800-22 randomness test results.

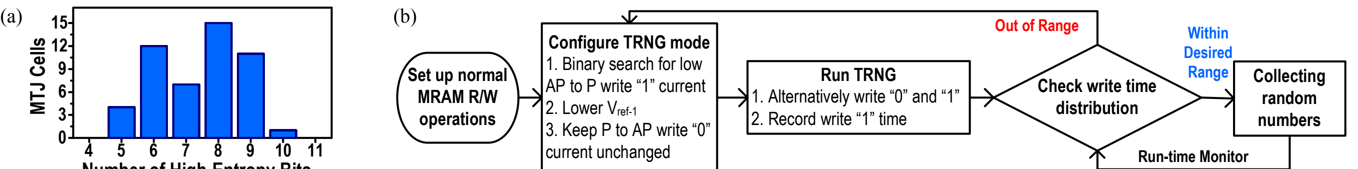


Fig. 5 (a) Histogram of the number of high-entropy bits generated by different MTJs and (b) a flow chart for calibrating MTJs for TRNG mode.

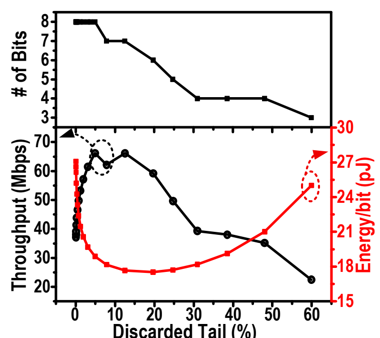


Fig. 6 Number of high-entropy bits, overall throughput and energy efficiency as a function of dropped tail bits.

	This Work	IEDM' 14 [6]	ISSCC' 17 [4]	JSSC' 16 [1]	JSSC' 16 [3]	VLSI' 11 [5]
Technology	28nm	N/A	65nm	14nm	40nm	65nm
Entropy Source	MTJ switching	MTJ switching	CMOS jitter	CMOS metastability	CMOS jitter	Soft oxide breakdown
Bit Rate (Mb/s)	66	N/A	9.9	162.5	2	0.011
NIST 800-22 Test Passed	All	10	All	All	All	All
NIST 800-90B Test Passed	All	N/A	N/A	N/A	N/A	N/A
Area (μm^2)	180	0.0085 (MTJ only)	920	1088	836	1200
Efficiency (pJ/bit)	18 (11*)	N/A	42	9	23	181810
Post Processing	No	Yes**	No	Yes	No	No

* Neglecting the static power of references shared by many MTJs and required by normal MRAM operation.
** Von Neumann corrector or off-chip output probability based feedback

Table 1 Summary of measurement results and comparison to state-of-the-art.

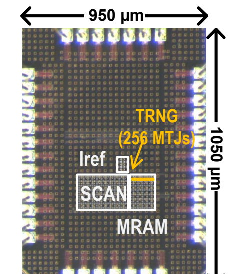


Fig. 7 Die Micrograph.