Technology trends and near-future applications of embedded STT-MRAM

Shinobu Fujita, Hiroki Noguchi, Kazutaka Ikegami, Susumu Takeda , Kumiko Nomura, Keiko Abe
Toshiba Corporation, R&D Center
Advanced LSI technology laboratory
1 Komukai-Toshiba, Kawasaki, Kanagawa, Japan 2128582
Tel : +81-44-549-2315
Fax : +81-44-549-2318
e-mail : shinobu.fujita@toshiba.co.jp

Abstract - This paper presents fast and low-power embedded nonvolatile memory technologies and circuit designs based on perpendicular STT-MRAM. Future prospects of applications are also discussed.

I Introduction

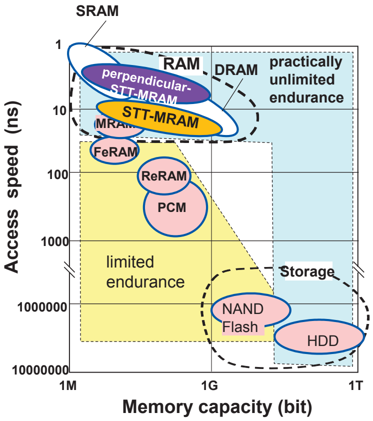


Fig.1. Memory capacity and access speed of various nonvolatile memories.

As shown in Fig.1, spin torque transfer (STT)-MRAM has potentials to cover working memory applications due to its high access speed and novel endurance. Furthermore, since perpendicular (p-) STT-MRAM has potentials to have higher access speed, p-STT-MRAM is expected to replace SRAM used as embedded working memory. Recently, static leakage power of SRAM has been increased with CMOS scaling and becomes a major part of power consumed in various semiconductor chips. "Power gating (PG)" technique enables reduction in the leakage power of SRAM effectively during long standby time (>0.1 ms) *when the application is not running (standby state)*. However, the power gating technique cannot be used *while application is running (active state)*, even though there are frequent short standby states (~several 10 ns) for SRAM.

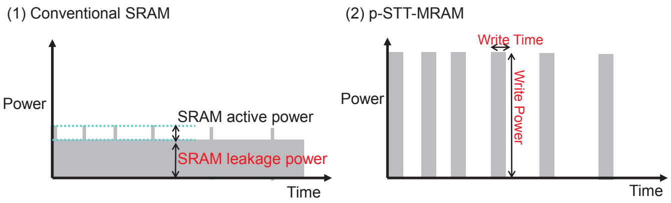


Fig.2. Comparison of power change with time for SRAM and STT-MRAM during active state.

Figure 2(1) shows typical power transition of SRAM during active state, indicating that power of SRAM is dominated by its leakage power. As shown in Fig.2(2), idealistically, there is no leakage power for STT-MRAM. However, to reduce total power (=active + leakage in Figs.2) by replacing SRAM with STT-MRAM, the most important point is reduction in write power of STT-MRAM and/or reduction in write time, as suggested in Figs.2.

II Towards write power reduction of MTJs

Figure 3 shows write current (write power = write current x V_{dd}(CMOS)) and write time of various MTJs. Though there was strong trade-off between write current and time (plot 1 to 7), modified MTJs have exhibited short write time and small current presented by Toshiba[13,18]. Total power of these MTJ based STT-MRAM is smaller than that of high-end SRAM.

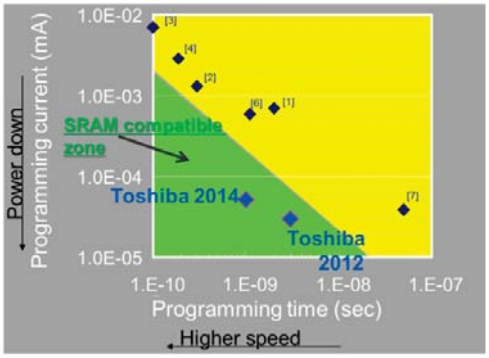


Fig.3. Write programming current and time of various MTJs. Green zone denotes SRAM compatible zone considering both active and leakage power shown in Fig.2. Ref. [1] Sony corp. IEDM (2005). [2] New York univ. Appl. Phys. Let. 97, 242510 (2010). [3] Cornel Univ. Appl. Phys. Let. 95, 012506 (2009). [4] Minnesota univ. J. Phys. D: Appl. Phys. 45, 025001 (2012). [6] IBM corp. Appl Phys Lett 98, 022501 (2011). [7] TDK-Headway Applied Physics Express 5 093008 (2012) .

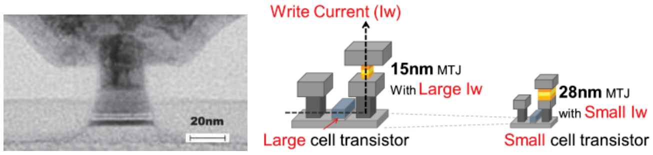


Fig.4. Cross sectional view of 28nm MTJ developed for embedded memory and illustration to show relationship between memory cell size and write current.

978-1-4673-6933-6/15/$31.00 ©2015 IEEE

As the write current is decreased, memory cell size is reduced, since the cell size is determined by not MTJ size but cell transistor size as illustrated in Fig.4. Therefore, for example, cell size using 28nm MTJ with small write current is smaller than that using "15nm MTJ with large write current". Thus, write current reduction contributes power saving, increase in write speed and decrease in memory area.

III Towards high-speed read for embedded memory

Although write speed is increased by improving MTJ features, read speed cannot be easily increased, since MTJ has fundamentally small on-off resistance ratio (typically x2 ~ 3). To improve read access speed, some of the authors previously proposed a "nonvolatile SRAM" (NV-SRAM) [1-3] using cross coupled inverters based latches combined with two p-MTJs, as shown in Fig.5. Since NV-SRAM works as an SRAM, read speed is fast. Just before PG, data in SRAM cells are stored complementally in MTJs. However, these NV-SRAMs include "leakage current paths" like SRAM. These leaky memory cells can be identified by "normally-on" type memory cells. Although other kinds of NV-SRAM combined with SRAM cells and MTJs were recently proposed [4,5], all of them include normally-on type memory cells.

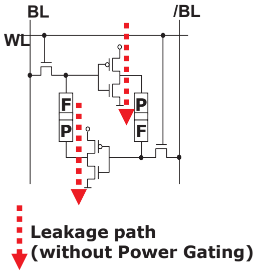


Fig.5. Nonvolatile-SRAM cell having leakage path like SRAM.

For normally-on type, highly frequent PG is needed for reducing leakage power during active state. The highly frequent PG with NV-SRAMs, however, increases circuit area largely and degrade computing performance based on our analysis [6]. Also, as for power and energy saving, there is limited effect on power reduction, since leakage energy is not decreased as shown in Fig.6.

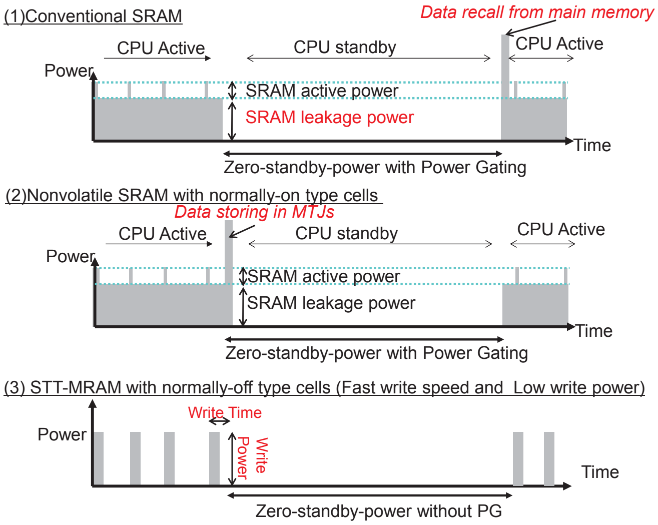
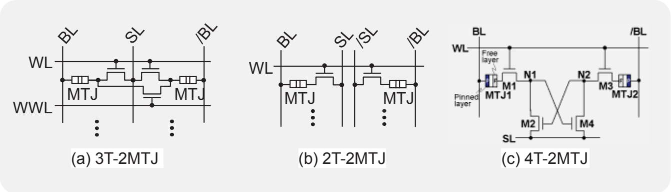

(1)Conventional SRAM
Data recall from main memory
CPU Active
CPU standby
CPU Active
Power
SRAM active power
SRAM leakage power
Time
Zero-standby-power with Power Gating
(2)Nonvolatile SRAM with normally-on type cells
Data storing in MTJs
CPU Active
CPU standby
CPU Active
Power
SRAM active power
SRAM leakage power
Time
Zero-standby-power with Power Gating
(3) STT-MRAM with normally-off type cells (Fast write speed and Low write power)
Write Time
Power
Write Power
Time
Zero-standby-power without PG

Fig.6 Power transition of SRAM, nonvolatile SRAM with normally-on type cells and fast and low power p-STT-MRAM with normally-off type cells.

The authors proposed to change the memory design concept from normally-on type with PG to "normally-off" type without PG that enable to reduce power effectively even for just one clock-cycle standby state[6]. Figures 7 show three kinds of the novel normally-off type memory cells without leakage paths based on advanced p-STT-MRAM. No leakage paths can decrease in leakage power even for a very short standby state during active state of CPU. The key point to increase the read access speed is that we use two MTJs with complementary resistance state, where one MTJ has high resistance state and another has low resistance state. As a result, write time was decreased to less than 5ns[7-9], which has confirmed that degradation of CPU performance is as small as less than 3 %.

Furthermore, the memory cell area of normally-off type shown in Figs. 6 is about x1.5 to x4 smaller than that of SRAM. On the contrary, the memory cell area of "normally-on type" is x1.2 to x2.5 "larger" than that of SRAM[6], which is severe disadvantage for L2 or LLC. For using dual MTJ cells, although the write energy becomes double of single MTJ, its read speed is much faster than that of single MTJ cell.

A test chip for 1Mb normally-off type (2T-2MTJ) STT-MRAM was fabricated in 65nm CMOS technology and 4ns access time was confirmed[10] as shown in Figs.8.



(a) 3T-2MTJ (b) 2T-2MTJ (c) 4T-2MTJ

Fig.7 Three kinds of normally-off type memory cell designs using (a)3T-2MTJ[7], (b)2T-2MTJ[8] and (c)4T-2MTJ[9]. Note that there are no leakage paths in these cells.

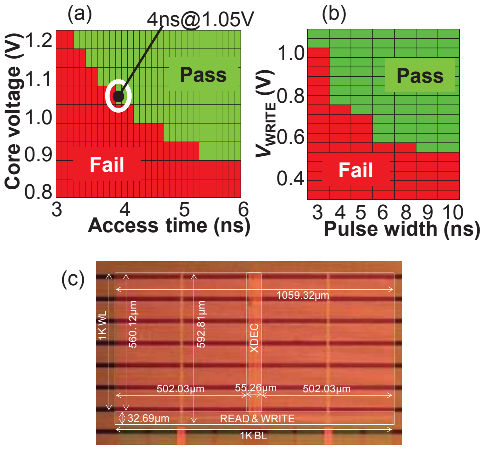


Fig8. Measurement data for read access (a) and write access (b) of fabricated 1Mb STT-MRAM with 65nm CMOS technology and photo of micrograph on the test chip surface[10].

IV High-performance processor applications

In advanced mobile processors, thanks to PG technique, low-power and high-performance operation (500mW~ 2 W) has been achieved[11]. For conventional processor architecture, volatile memories based on SRAM are used for high frequency access memories like resister files, L1 cache memory, and L2 cache memory. Based on conventional PG technique[11], power supply is cut off to CPU cores but power supply is not cut off to L2 cache to keep data in it (transition to "CPU core sleep state") during relatively short standby state, where the average power is x2 to x3 lower than that in active state. When standby state time is long, power supply is also cut off to L2 cache except for small SRAM blocks to retain resister state. This is a critical transition to "deep power down state (**DPS**)" that has over x10 lower power than that in active state. Clearly, frequent state transition to DPS plays a vital role to reduce processor power effectively. However, since L2 cache cannot retain data during DPS, after "wake-up" the processor system must recover the lost data from main memory. From this reason, the "wake-up time", time needed for recovery from DPS, is so long, which can degrade performance largely. Therefore, there is severe trade off between power and performance based on conventional PG. Although speed and frequency for PG has been improved, ratio of leakage power for L2 cache in total consumed power for CPU is increased with increasing speed of PG, since it is fundamentally difficult to perform PG for volatile cache memory.

It is, hence, considered that average CPU power determining battery life depends strongly on L2 or L3 cache memory capacity. Cache memory capacity has been dramatically increased as shown in Fig. 9. This is because cache memory capacity has been simply increased for improving the performance. Also, for multi-core processors based on conventional architectures, shared cache memory capacity, L2 or L3, is increased with increasing number of core. As a result, power consumed by cache memory occupies major part in the recent processor chips.

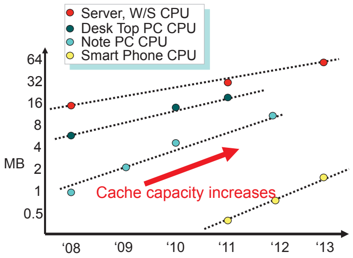


Fig.9. Increment of L2, L3 cache memory capacity.

Considering such situation of HP-mobile processors, the authors have proposed STT-MRAM/SRAM hybrid memory hierarchy, [3,12]. Here, SRAM-based cache memory is applied to level-1 (L1) cache, and STT-MRAM-based cache is applied to level-2 (L2) cache. This is because, whereas active power is dominant for L1 cache due to frequent access to L1, L2 cache infrequently operates only in the case of L1 cache miss. Further, L2 cache capacity is 10 times or much larger than that of L1. Leakage power in all cache memory is, therefore, dominated by L2 cache. To reduce the leakage power in CPU, STT-MRAM should be applied to L2 cache.

However, general STT-MRAM has large active power especially for write operation and/or large latency. Total energy consumption is, hence, largely increased, although leakage power can be largely decreased by nonvolatile cache memory for high-performance processors. In general, nonvolatile memories have little static power but large active power. This issue has been recognized as "dilemma of nonvolatile memory"[12]. Therefore, reduction in write power and time for STT-MRAM is inevitable to solve this issue. Figures 10 shows how fast write is needed to reduce the total average power including active and static power compared with that of SRAM based on preliminary case studies. The results suggest that 5 ns is a target access time. Recently, p-MTJ[13] has demonstrated 3ns fast write with write current 50uA per bit. This advanced p-STT-MRAM has, thus, a strong potential to replace SRAM for the lower power cache memory. Although write time of 3 ns per one-bit cell is short, it is still longer than that of one-bit SRAM cell. However, the difference in the write time decreases as memory capacity is increased to MB order. Also, access frequency for L2 or L3 cache is more than x10 to 100 smaller than that for CPU core memories like resisters.

Therefore, influence of write time (delay) of STT-MRAM on CPU performance is very limited. This fact was confirmed

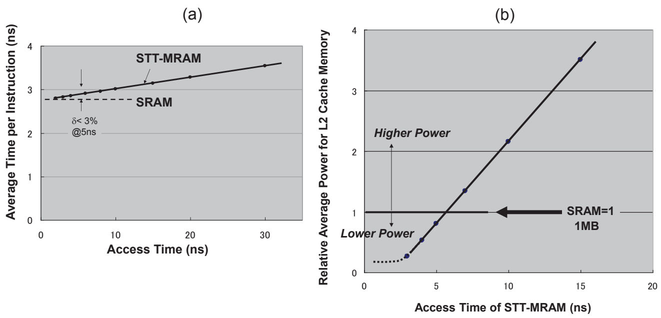


Fig.10 (a): Average time per instruction of CPU with STT-MRAM based L2 cache as a function of access time of STT-MRAM. In this case study, L1, L2 and main memory access latency are 1ns, 10ns and 100ns, respectively.

Fig.10 (b): Respective cache miss rate of L1 and L2 are 90% and 80%.Relative average power for STT-MRAM based L2 cache memory (1MB) compared with SRAM as a function of access time of STT-MRAM. Write current is assumed to be 50uA.

by simulations on this study, as shown in Fig.10(a). This indicates that L2 cache latency does not degrade CPU performance much, if STT-MRAM access time is shorter than 5ns.

It should be noted that decreasing write time of STT-MRAM leads effective saving of total average power, as shown from simulation results in Fig.10 (b). This is because the largest part of total power for STT-MRAM is “write time x write power”, whereas “read time x read power” is 10x or more lower than that of the write. On the other hand, the largest part of total power for SRAM is leakage power, as shown in Fig.2.

Using the measured data on 1Mb STT-MRAM, power and performance in processor with 2T-2MTJ STT-MRAM based cache memory for LLC with was evaluated with customized CPU simulator using gem5 [18]. The CPU is ARMv7- single core using 3-way issue and out-of-order. Processor simulations were conducted with SPEC CPU2006 benchmarks. Compared with the typical 0.8 V operating SRAM design, our proposed normally-off type STT-MRAM, can reduce the energy per instruction (EPI) of the total cache memory reduced by 64%. It has an instruction per cycle (IPC) performance degradation within 6 %, as shown in Fig. 11. On the other hand, the EPI for other STT-MRAMs previously reported is largely increased. It should be noted that decrease in power for CPU active state is attributable to decrease in leakage power of L2 cache memory.

p-STT-MRAM based nonvolatile cache is effective for power reduction in not only mobile processors but also high-end processors for servers. In this application, multi-core CPU is used and processors include large capacity of last level cache (LLC). Due to large memory capacity, eDRAM is also used for area reduction, although eDRAM consumes much large refresh power. CPU performance (execution time) and LLC energy was simulated using reported data and our measurement data, as shown in Figs.12. There was no performance degradation, and consumed energy of LLC was decreased with p-STT-MRAM by 59.6% compared with SRAM based SRAM[10]. Further, LLC energy has been decreased by about 80% with fast PG for peripheral circuits[19].

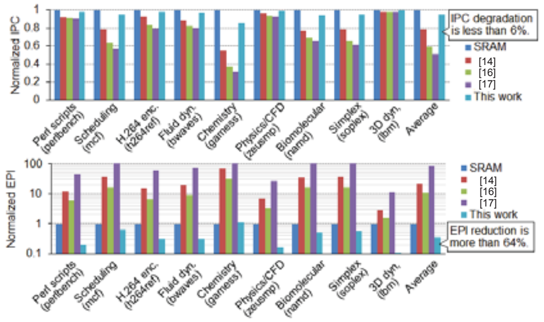


Fig. 11. Simulated IPC (Instruction per Clock Cycle) and EPI (Energy per Instruction) performances of 1MB L2cache using 2T-2MTJ design.

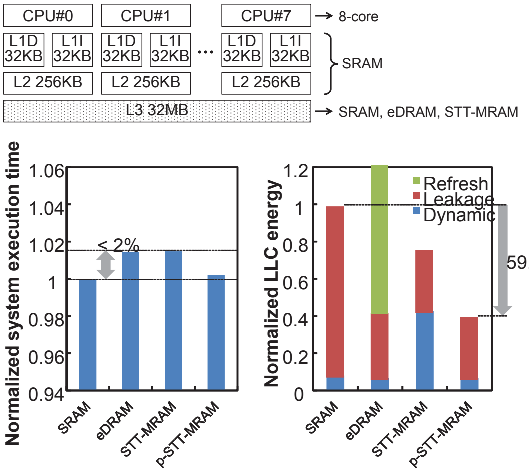


Fig. 12. Block diagram of an 8-core processor for high-end server and simulation results of normalized execution time of CPU and LLC energy.[10]

V New applications : Big Data, IoT and security.

For another applications, the authors proposed reconfigurable memory based computation based on embedded p-STT-MRAM[20], as shown in Fig.13. This concept can be applied to a computing system for Big Data applications[21]. Recently, FPGA is used for some executions in SQL to enhance parallel data process between big data storage and memory. Power reduction can be expected by replacing FPGA with p-STT-MRAM based computation, as FPGA include massive number of normally-on type SRAM. Also, resource can be also saved, as LLC with p-STT-MRAM can be used for the memory computation.

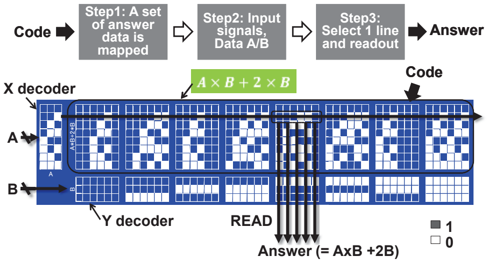

Code
Step1: A set of answer data is mapped
Step2: Input signals, Data A/B
Step3: Select 1 line and readout
Answer
$A \times B + 2 \times B$
Code
X decoder
A
B
Y decoder
READ
Answer (= AxB +2B)
1
0

Fig. 13. An example of memory computing process and memory mapping for execution of arithmetic[20].

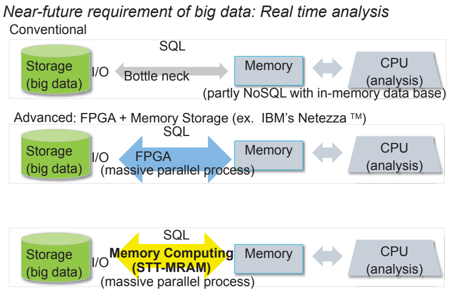

Near-future requirement of big data: Real time analysis
Conventional
SQL
Storage (big data)
I/O
Bottle neck
Memory
CPU (analysis)
(partly NoSQL with in-memory data base)
Advanced: FPGA + Memory Storage (ex. IBM's Netezza ™)
SQL
Storage (big data)
I/O
FPGA
(massive parallel process)
Memory
CPU (analysis)
SQL
Storage (big data)
I/O
Memory Computing (STT-MRAM)
(massive parallel process)
Memory
CPU (analysis)

Fig.14. Embedded p-STT-MRAM based memory computing for Big Data applications[20].

For IoT (internet of things), high-end SoC uses large capacity of SRAM and/or embedded NOR-flash. It has been suggested that p-STT-MRAM has also potential to replace these memories[21]. Furthermore, STT-MRAM can be used for truly random number generators that plays an important role of information security [22]. Thus p-STT-MRAM is expected for various kinds of ICT applications in future.

Acknowledgments

This work was partly supported by the Normally-Off Computing Project of the New Energy and Industrial Technology Development Organization (NEDO) in Japan. Normally-off processor architecture has been studied based on collaboration with Prof. Hiroshi Nakamura and his group of University of Tokyo.

References

[1] K. Abe, S. Fujita and T. H. Lee, European Micro and Nano Systems (EMN04), pp. 225-229, Oct. 2004.
[2] K. Abe, S. Fujita and T. H. Lee, Proceedings of the 2005 NSTI Nanotechnology Conference, vol. 3, pp 203–206, May 2005.
[3]K. Abe, S. Fujita et al., International Conference on Solid State Devices and Materials (SSDM), pp. 1144-1145, Sep. 2010.
[4] S. Yamamoto et al, Japn. J. Appl. Phys. 48 (2009) 043001.
[5] T. Ohsawa et al., Symp. VLSI Circuits, pp.46-47, July 2012.
[6] K. Abe et al., IEDM Tech. Dig., pp.243-246, 2012.
[7]A. Kawasumi et. al, IMW, p.p. 76 - 79, 2013.
[8] H. Noguchi et. al, VLSI Technology Symposium, p.108, 2013.
[9] C. Tanaka et. al, International Conference on Solid State Devices and Materials (SSDM), p.p. 1092-1093, 2013.
[10] H. Noguchi et. al, VLSI Technology Symposium, p.97, 2014.
[11] G. Gerosa et al., ISSCC Technical Digest, 2008, p.p. 256-25.
[12] K. Ando, "A Normally-off Computer", FED Journal, **12**, 89, 2001(in Japanese). K. Ando, et al., "Normally-Off Computer: new roles of nonvolatile devices in future computer systems", in *Sustainable Green Computing*, IGI press, June, 2012.
[13] H. Yoda, Session 11.3, IEDM Technical Digest, 259, 2012. E. Kitagawa, et al., Session 29.4, IEDM Technical Digest, P.677, 2012.
[14] R. Nebashi et al., ISSCC Technical Digest, 2009, p.p. 462-463.
[15] K. Tsuchida et al., 2010 ISSCC Technical Digest, 258, 2010.
[16]J. P. Kim et al., Symposium on VLSI Circuits, pp.296-297, 2011.
[17] T. Ohsawa et al., Symp. VLSI Circuits, pp.46-47, July 2012.
[18] D. Saida Int'l Magnetics Conf., p1162, 2014.
[19] Noguchi et. al, ISSCC Technical Digest, p.136, 2015.
[20]Noguchi et. al, IEDM Technical Digest, 2013.
[21] S. Fujita et al., ISSCC 2015, Forum 2: Memory Trends: From Big Data to Wearable Devices.
[22] K. Lee et al, ISLPDE, p.131, 2014 .
[23]S. Yuasa et. al, IEDM Technical Digest 3.1.1, 2013.