Progress of STT-MRAM Technology and the Effect on Normally-off Computing Systems

H.Yoda[1], S.Fujita[2], N.Shimomura[2], E.Kitagawa[2], K.Abe[2], K.Nomura[2], H.Noguchi[2] and J.Ito[2],
[1] Center For Semiconductor Research & Development, Toshiba Corporation, Kawasaki 212-8520
[2]Corporate R&D Center, Toshiba Corporation, Kawasaki 212-8582, Japan
Tel: +81-44-548-2522, Fax: +81-44-548-8324, Email: hk.yoda@toshiba.co.jp

1. Introduction:

Figure 1-1 shows an access-speed vs. density map of existing memories. The ideal memory, a fast and dense non-volatile memory with unlimited endurance, does not exist. Consequently, most systems use a combination of working memories such as SRAM and DRAM and storage memories such as NAND Flash and HDD as shown in Fig. 1-2.

The working memories have fast read/write access speed and unlimited endurance but do not have nonvolatility. The storage memories have nonvolatility and density but do not have fast read/write access speed. They work cooperatively with each other to attain both fast accessibility and nonvolatility of data.

However, there are drawbacks. When such a system is turned on, data, which is located in storage, should be copied to the working memories to set up memory systems for usage. The setup is called booting and it takes about a minute in the case of personal computers. When such a system is in use, the working memories consume a lot of energy to keep those data because SRAM and DRAM are volatile. The battery power consumption is a major problem for mobile systems such as cellular phones.

Fast nonvolatile memories with unlimited endurance solve these problems. MRAM is the only nonvolatile memory that has relatively fast read/write accessibility and unlimited endurance. However, an innovation was needed. The commercialized MRAM is not as fast as SRAM or not as dense as DRAM.

Much work has been done on MTJ materials and MRAMs. In particular, spin transfer torque (STT)-writing on MTJs with perpendicular magnetization (simply expressed as P-MTJs) solved most of the problems [1],[2],[3],[4],[5],[6]. As a result, we are very close to solving the above-mentioned drawbacks of the present SRAM, DRAM/NAND, HDD memory hierarchy by designing a normally-off memory hierarchy.

In this paper, the progress of P-MTJs is reviewed and prospects for the normally-off memory hierarchy based on new results are discussed.

2. Progress of MTJ development

2.1 Issues:

Since field-writing MRAM seems to have a scalability limit of around a hundred megabits because of its large milliampere-order write current, STT-writing MRAM (simply expressed as STT-MRAM) has been intensively developed.

The unit cell consists of a selection transistor and an MTJ and the cell size is as small as 6-8 F^2, the same as that of DRAM where F is design node.

In order to shrink F, write current should be reduced below the small selection transistor's drivability.

Commercialized field-writing MRAMs use MTJs with in-plane magnetization (simply expressed as I-MTJs), because of their maturity. I-MTJs have been used for STT-MRAMs development. However, the write current was of the order of hundreds of microamperes, far beyond the small selection transistor's drivability. The write current should have been greatly reduced.

The typical target of the write current is of the order of several tens of microamperes, and 30 microamperes for below 20 nm node is desirable [7]. To read data, large MR over 150% is also required [7].

2.2 Breakthrough Technology, P-MTJs:

P-MTJs were proposed by Toshiba and intensively developed by the Spin-RAM working group in New Energy and industrial technology Development Organization (NEDO) spintronics non-volatile devices project team [1]. To solve mismatch between MgO tunnel barrier and perpendicular magnetic materials, insertion of perpendicular CoFeB layer was proposed [2].

2.3 Ics reduction trend by P-MTJs:

Figure 2-1 shows the Ics (critical switching current) reduction trends. Clearly, P-MTJs contributed greatly to the Ics reduction and finally reached 7-11microamperes with 5ms current pulse width and 15 microamperes with 30ns current pulse width [4], [6]. Thus, P-MTJs have solved the biggest obstacle, large write current that had prevent MRAM densification for a long time.

In addition to the write efficiency improvement by P-MTJs themselves, reduction in damping constant of MTJ storage layer and an increase in MR contributed to the Ics reduction as shown in Fig. 2-2.

2.4 MR increase trend for P-MTJs:

Figure 2-3 shows the MR increase trend. The insertion of perpendicular CoFeB between perpendicular magnetization layer and MgO tunnel barrier contributed to an MR increase and finally achieved MR of over 200% [5]. Toshiba released a news that Toshiba had developed P-MTJs having MR ratio of 200% and small Ics [6].

2.5 Switching speed trend:

Figure 2-4 shows the switching speed trend. Cornell Univ. proved by using in-plane GMR elements that writing with 0.05 ns pulse was possible. Osaka Univ. and Toshiba proved by using perpendicular GMR elements that writing with 0.5 ns pulse was possible [8]. IBM and Toshiba proved by using P-MTJs that writing with 1 ns current pulse was possible [9]. Because of small Iw of P-MTJs, heating by the current was greatly reduced and the MTJ endured the fast switching tests.

It is thought that the switching speed does not depend on whether the magnetization is in-plane or perpendicular. Switching faster than 1 ns will be easily realized with MTJs if the Ics decreases.

2.6 Energy consumption trend:

Figure 2-5 shows the switching energy trend. Although the demonstrated switching speed of the I-MTJs is faster than that of P-MTJs, the energy consumption of the I-MTJs is quite large. Toshiba showed 0.05 pJ/bit energy consumption of the P-MTJ that was much smaller than that of I-MTJ [9]. This drastic reduction of the switching energy is due to the drastic Ics reduction by the P-MTJs.

3 Progress of MRAM development

3.1. Denser Application

The main issue for denser applications is large write current that increases the size of select transistor in the memory cell. We have decreased Ics using low-power P-MTJs whose write current is much lower than that of other "general" P-MTJs and in-plane-MTJs, as shown in Fig. 2-1. Using these P-MTJs, we fabricated and demonstrated 64Mbit MRAMs. Figure 3-1 shows the history of increase in the storage density of the STT MRAM.

3.2. Faster Application

The main issue for faster applications is slow write speed of MTJ that limits the write time of memory arrays. Figure 3-2 shows recently reported data on write speed of MTJs. Although this figure shows that high speed of even sub- 1 ns is feasible in some cases, write current is very large for those cases, as shown in Fig. 3-2, indicating serious tradeoff between write speed and write current. In the fastest applications such as replacement of SRAM, large increase in power is unacceptable. Therefore, decrease in write current and increase in write speed have to be achieved simultaneously by resolving the tradeoff. Only our P-MTJs with the lowest write energy (0.09 pJ/bit) overcome this tradeoff, as shown clearly in Fig.3-2.

3.3. Key Technologies for Low Power Applications

Furthermore, MRAMs are expected to be used to decrease operation energy for high-performance (hp-) mobile SoC (smart phone, tablet PC). As memory access speed is more increased with increasing clock frequency, power consumption of the memory has to be greatly decreased to save battery power. Although a nonvolatile memory such as MRAM is a low-power memory to store a large amount of data with zero standby power, it should be noted that active power, especially write power, is much larger than that of conventional volatile memories. This issue called the "dilemma of nonvolatile memory"[10] means that *"zero-standby power circuit" is not "low power circuit"*. Also, the power gating technique enables SRAM to have zero-standby power of during long standby state when the application software is not running, as shown in Fig. 3-3 (1). To reduce the power effectively in this situation, leakage power of the memory has to be decreased for a ***short standby state*** (SSS) during running application, as shown in Fig.3-3(2). For that purpose, the break even time (BET) for the energy reduction should be shorter than the time of SSS, as shown in Fig.3-3(3).BET is expressed by:

$$\mathrm{BET} = \boldsymbol{P}_W \times \boldsymbol{t}_w / P_S \times \boldsymbol{N}_M/\boldsymbol{N}_S \quad \ldots (1)$$

If the BET is longer than average SSS time, the total energy of MRAM-based circuit is increased rather than that of conventional SRAM-based one, as shown in Fig. 3-3(3). Clearly, reduction in both $\boldsymbol{P}_w$ and $\boldsymbol{t}_w$ (shift to the lower left in Fig.3-2) is a key factor to shorten BET for low-power applications. Also, the smaller the leakage current of SRAM becomes, the longer BET is. Furthermore, memory number ratio, $\boldsymbol{N}_M/\boldsymbol{N}_S$, is another key factor to shorten the BET, as described in the next section and shown in Table 1.

4. The effect on Normally-off computing systems

4.1. History towards Normally-off computing

A new concept was proposed for dramatic change from a conventional "Normally-on" computer to "Normally-off (N-off) computer" with nonvolatile memories to achieve an ultimate low-power-computing in 2001[10]. To realize N-off computer we started from "entirely nonvolatile memory hierarchy (ENV-MH)", as shown in Fig.4-1. Since the access speed of 1T-1MTJ STT-MRAM is slower than those of SRAM or flip flops in the processor core (fc $\geq$2GHz), we designed nonvolatile cross-coupled inverters (NVC) using STT-MRAM as shown in Figs. 4-2[11]. The NVC is used as a NV-flip flop (NV-FF) and NV-SRAM. The unique feature of these NV-circuits is that while CPU is in active state, these circuits act as conventional CMOS-based flip flop or SRAM with very high speed (>2GHz). While CPU is in standby state, data are stored in MTJs and zero-standby power is achieved by the power gating. After power supply returns, the data saved in MTJs are automatically recalled in the SRAM or flip flop, then the processor core quickly becomes ready to start arithmetic operation [11]. In our first ENV-MH, NV-FFs are used for registers in CPU core, NV-SRAM are used for L1, L2 cache memory. We checked effectiveness of power reduction by calculating BETs with equation (3), as shown in Table 1. For this calculation, we use two kinds of P-MTJs, P-MTJ with the lowest E_w of 0.09 pJ and general P-MTJs having E_w of 3 pJ. Even using lowest-E_w P-MTJs, table 1 suggests that, if the NV-FF and NV-SRAM are used in the hp-CPU core, *the processor consumed ultra high power* for high frequency operation. On the other hand, the BET can be effectively shorten by applying NV-memory to L2 or L3 cache memory, since the write power is consumed by only 64B for MRAM, whereas standby power is consumed by 512 kB or 4MB for SRAM. As a result, BET is reduced to as short as 57ns, which enables power saving even for the SSS time. Similarly, it was found that L3 cache is also an available target even with 1T-1MTJ perpendicular STT-MRAM.

It should also be noted that the higher the level in memory hierarchy is, the more average power becomes dominated by active power, as shown in Fig.4-1. This means that there is little opportunity for power saving by reducing standby power using nonvolatile memory in the higher-level memories. Based on our rethink, we proposed more suitable and effective memory hierarchy composed of volatile memories in the CPU core and nonvolatile memories for L2 or L3 cache memories (T2010-MH)[16], as shown in Fig. 4-1.

After our proposal of NVCs, similar NV-SRAM with 8T(transistors) [12] or 4T [13], or NV-FF with 19T[14] or 14T[15] was presented. However, our NVCs are superior to those considering power, area and performance, as shown in Fig. 4-4 for comparison examples. It should be noted that the BET cannot be shorten by modifying NV-circuit designs by changing T-number, since BET is determined only by the balance between P_w of single MTJ and P_s of one SRAM cell.

4.2. Power/performance evaluation of Normally-off computing

Since estimation of BET is not enough for analyzing the power for real applications, we evaluated CPU performance using a CPU core/cache simulator, where the conventional power gating scheme for SRAM-based cache memory is applied [17]. Fig. 4-3 shows that P-MTJ-based cache memory is only the case in which power can be reduced without degrading performance. It is because general MTJs have large E_w, as shown in Fig.3-2, the BET is longer than the SSS time. Further, to approximate real mobile processors, we evaluated CPU power and performance using a processor emulator for ARM-core based CPU on Linux OS while running two kinds of applications (MPEG, video game). These results indicate that the power consumed in the cache memory can be reduced by over 80% without any performance penalty, as shown in Fig. 4-5. These facts mean that this new architecture based on P-STT-MRAM can realize N-off computing.

Conclusion: Intensive work on P-MTJs overcame the major obstacle of large write current. It was proved to have very fast switching speed of 1 ns and very small switching energy of 0.05 pJ. These data cleared the specification of 22 nm node STT-MRAM. In addition, a new memory hierarchy with volatile/nonvolatile hybrid using P-MTJ was presented and effective power reduction by over 80% without performance degradation was confirmed for real applications running on mobile CPUs for the first time. This is a big step toward normally-off computing systems.

Acknowledgements: We would like to thank all the members of the Spin-RAM working group of the NEDO Spintronics Nonvolatile Devices project for their contributions, and Prof. Hiroshi Nakamura for fruitful discussion on processor architecture. This study was partly done in the NEDO normally-off computing project.

References:

[1] H.Yoda, et al., presented at 7th IWFIPT, Session IIIc
[2] M. Nakayama et al., J. Appl. Phys., 103, 07A710 (2008).
[3]T. Kishi et al., IEDM Tech. Dig. p309 (2008).
[4] T. Daibou et al., MMM-Intermag 2010, DA-08, Jan. 2010.
[5] K. Nishiyama et al., MMM-Intermag 2010, EF-07, Jan. 2010.
[6] Toshiba RDC HP, Jun. 2011.
[7] ITRS roadmap 2011 version
[8] H. Tomita et al.,IEEE Trans. Magn., 47, 1599 (2011)
[9] E. Kitagawa et al., Session 29.4, IEDM 2012.
[10] K. Ando, "*Roles of Non-Volatile Devices in Future Computer System: Normally-off Computer*", Energy-Aware Systems and Networking for Sustainable Initiatives, edited by N. Kaabouch and W.-C. Hu, published by IGI Global, June, 2012.
[11] K. Abe, et al. European Micro and Nano Systems, Session : ThuAmOR1 H: MEMS components I 2004, Paris. ; K. Abe, et al. NSTI Nanotech. Conf. Anaheim, P. 203, 2005.
[12] S. Yamamoto et al. J.J. Appl. Phys. Vol. 48, 4, 2009.
[13] T. Ohsawa et al., Japan. J. Appl. Phys. 51 (2012) 02BD01. T. Ohsawa et al., Symposium on VLSI Circuits , 2012, p46.
[14] N. Sakimura et. al, , IEEE JSSC, Vol. 44, No.8, 2009.
[15] T. Endoh, IEDM 2011, Technical Digests, p75.
[16] K. Abe et al., SSDM2010, P.P. 1144-1145, 2010.
[17] K. Nomura et al., MMM 2011, GD-03, Nov. 2011.
[18] K. Nomura et al., The 25th Workshop on Circuit and Systems, Japan. As2-3-1, July, 2012.
[19] K Abe et al., Session 10.5, IEDM 2012.

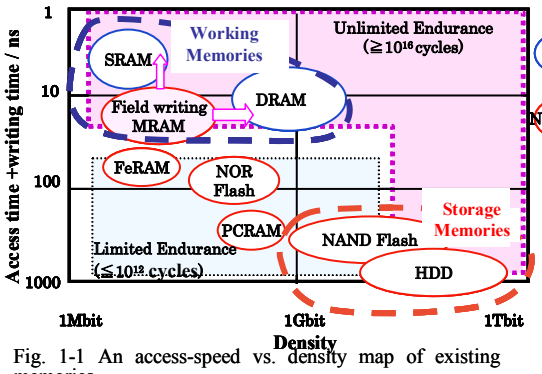


Fig. 1-1 An access-speed vs. density map of existing memories
There are two major categories, working memories and storage memories. MRAM is the only nonvolatile working memory but commercialized MRAMs are neither very dense nor very fast. Challenges for MRAM are indicated by the two arrows, namely, a faster memory such as SRAM and a denser memory such as DRAM.

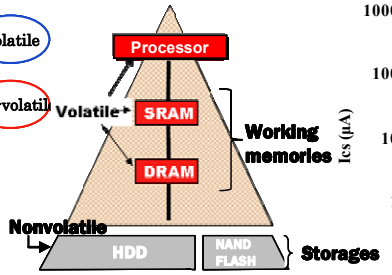


Fig. 1-2 Typical memory hierarchy used in the present systems
Since working memories are volatile and storage memories are slow, a combination of working memories and storage memories is used.

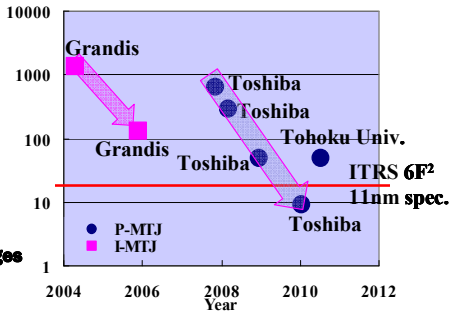


Fig. 2-1 Ics reduction trends
P-MTJs overcame the obstacle that prohibited MRAM's densification. They cleared the specification for 11nm node in ITRS STT-MRAM roadmap.

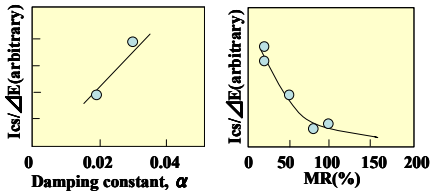


Fig. 2-2 Further reduction in Ics by reduction in α and an increase in MR
Ics/⊿E is a metrics for 1/(write efficiency), critical switching current per data retention energy. Both reduction in α and an increase in MR contributed to an increase in write efficiency, i.e. the reduction in Isc for given retention energy.

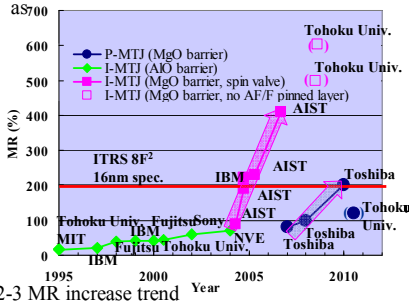


Fig. 2-3 MR increase trend
Insertion of perpendicular CoFeB or Fe between perpendicular magnetic layer and MgO tunnel barrier contributed an MR increase of over 200%. It cleared the specification of ITRS 16 nm node. All the MTJs used to get above data have stable reference layer except data in parentheses. Data in parentheses can not be compared with other data because the MTJ did not have pinning layers which stabilized the reference layers. Adding pinning layer degrades MR a lot.

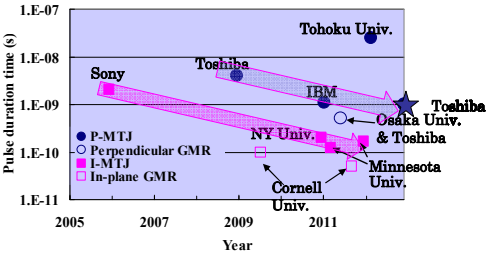


Fig. 2-4 Switching speed trend
STT-writing was proved to have fast switching. The potential was demonstrated by in-plane GMR. The fastest switching ever demonstrated is 0.05 ns. The star indicates the data to be presented in ref. [9].

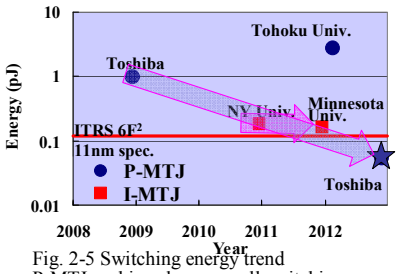


Fig. 2-5 Switching energy trend
P-MTJs achieved very small switching energy of 0.05 pJ/bit and cleared the specification for 11nm node in ITRS STT-MRAM roadmap.

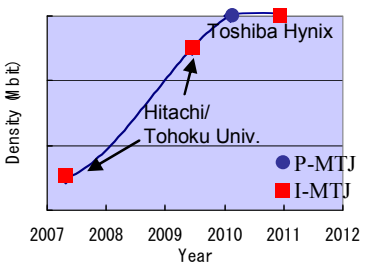


Fig.3-1 History of increase in the storage density of the STT MRAM.

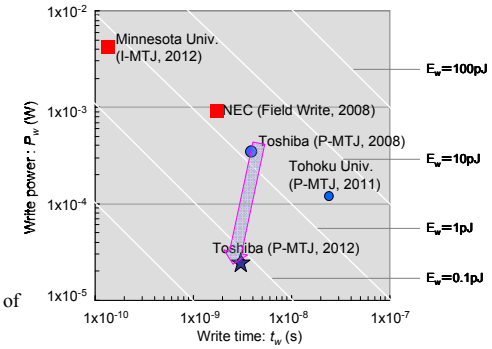
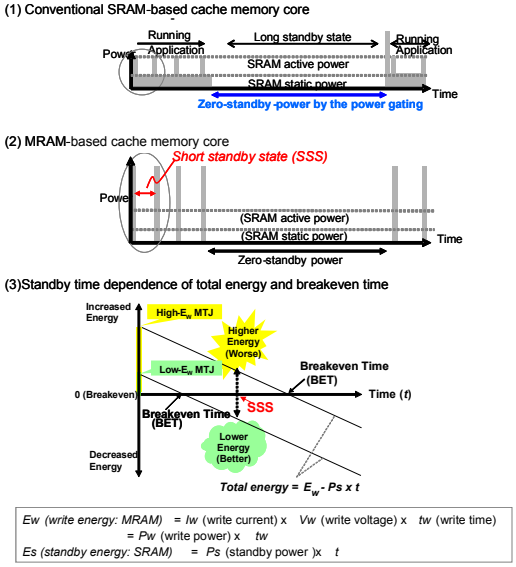


Fig.3-2: Recently reported data on write speed and power of MTJs.



Ew (write energy: MRAM) = *Iw* (write current) x *Vw* (write voltage) x *tw* (write time)
= *Pw* (write power) x *tw*
Es (standby energy: SRAM) = *Ps* (standby power)x *t*

Fig.3-3: (1)(2)Memory power reduction during running application and (3)energy reduction condition considering Ew, Ps and BET[9].

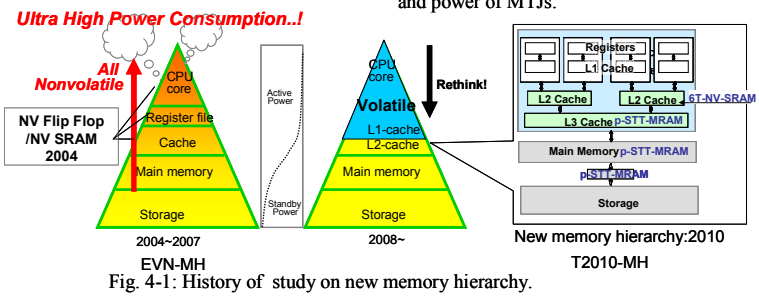
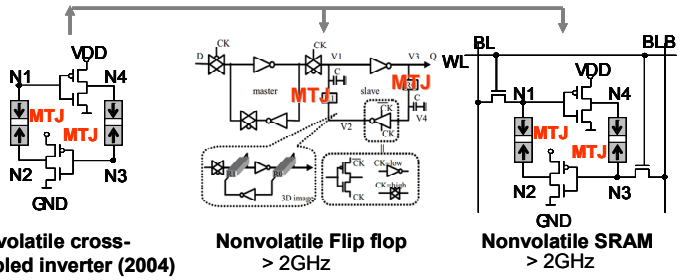


Fig. 4-1: History of study on new memory hierarchy.



Nonvolatile cross-coupled inverter (2004) | **Nonvolatile Flip flop** > 2GHz | **Nonvolatile SRAM** > 2GHz

Fig. 4-2 Nonvolatile (NV-) latch using MTJs for STT-MRAM and NV-flip flop and NV-SRAM[11].

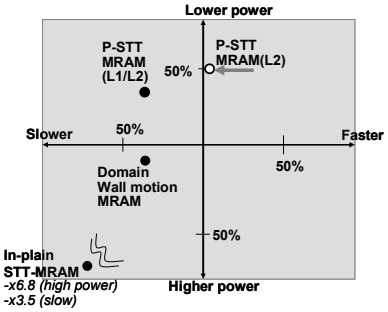


Fig. 4-3 CPU performance and cache power with various MTJ-based cache memories [17].

Table 1 : Breakeven time for various memories in CPU in a case study.

Memory Hierarchy	Volatile Circuit	Standby Power (P_s)	NV-Circuit	N_s/N_M	BET case1 (Advanced p-STT-MRAM)	BET case1 (General p-STT-MRAM)	Typical Time of Short Standby State(SSS)
Registers In CPU core	Flip flop	0.75 nW	NV-Flip flop	1	120us	4 ms	<1~2ns
Register File	SRAM	0.75nW	NV-SRAM	~1	120us	4 ms	<1~2ns
L1 Cache	SRAM	0.75nW	NV-SRAM	0.1~1	12 ~ 120us	0.4~ 4 ms	3~5ns
L2 Cache (512kB)	SRAM (Low power)	0.2nW	NV-SRAM	$6.3x10^{-5}$	57ns	1.9us	30~ 100ns
L3 Cache (4MB)	SRAM (Low standby power)	0.02nW	STT-RAM	$8x10^{-6}$	72ns	2.4us	100~ 300ns

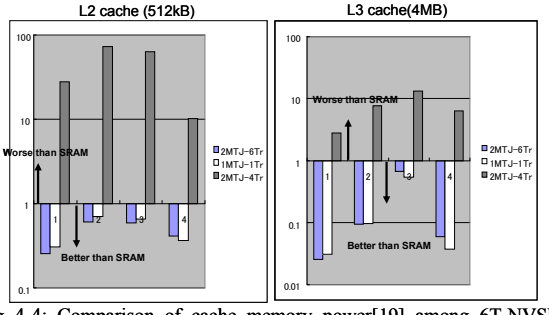


Fig 4-4: Comparison of cache memory power[19] among 6T-NVSRAM (6T-2MTJ), 1T-1MTJ-STTMRAM[9], and 4T-NVSRAM (4T-2MTJ)[13]. The operation power of 4T-2MTJ is much higher than that of SRAM, since the active power of 4T-2MTJ circuit is high due to large write/read current and the write energy of MTJ is also high.

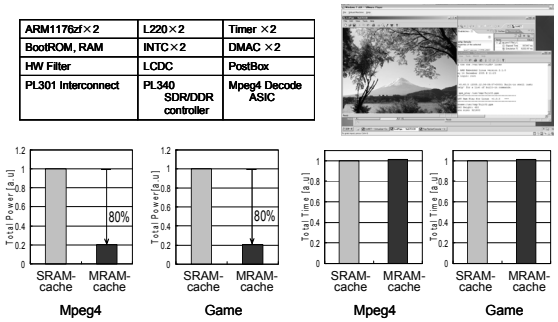


Fig4-5. CPU power and performance based on SRAM-cache and MRAM-cache evaluated on ARM-core based CPU with Linux OS while running two kinds of applications (MPEG, video game) [18]