HfO2-based FeFET and FTJ for Ferroelectric-Memory Centric 3D LSI towards Low-Power and High-Density Storage and AI Applications

Masumi Saitoh, Reika Ichihara, Marina Yamaguchi, Kunifumi Suzuki, Keisuke Takano, Keisuke Akari, Kota Takahashi, Yuta Kamiya, Kazuhiro Matsuo, Yuuichi Kamimuta, Kiwamu Sakuma, Kensuke Ota, and Shosuke Fujii   
Institute of Memory Technology R&D, Kioxia Corporation, Yokkaichi, Japan, E-mail: masumi.saitoh@kioxia.com

***Abstract*—** We present the recent progress in HfO2-based ferroelectric FET (FeFET) and ferroelectric tunnel junction (FTJ) memory towards low-power and high-density storage and AI applications. A huge amount of interface trap charges coupled to spontaneous polarization significantly alters the operating model and improvement guideline of HfO2 FeFET irrespective of elements doped into HfO2. Performance and reliability of in-memory reinforcement learning (RL) with HfO2 FTJ array are enhanced by improving the characteristics of FTJ memory cells.

The measured device in this study is a poly-Si channel TFT with FE-HfSiO. The thickness of interfacial SiO2 is around 1.5nm. Figs. 3 and 4 show Id-Vg after program/erase and Vth as a function of program voltage (Vpgm) or erase voltage (Vers). The maximum MW is around 1.2V for Vpgm/Vers of +/-5V. It should be noted that Id-Vg measurement is performed at 100sec after program both in Figs. 3 and 4. As shown in Fig. 5, Vth shifts to the positive direction (opposite to the expected direction) just after program, then rapidly decreases in about 10sec and falls below the erased Vth (Vth

|  |  |  |
| --- | --- | --- |
| **I.** | **INTRODUCTION** | shift inherent to FeFET is obtained). |
| Time evolution of ΔPs, ΔQt, and ΔQt-ΔPs (corresponding |

Ferroelectric-HfO2 (FE-HfO2) based memory has attracted much attention because of the CMOS compatibility and scalability. Among various ferroelectric memory cell structures, FeFET and FTJ are especially promising in terms of non-destructive nature and compactness. HfO2 FeFET is suitable for high-speed and low-power storage applications [1-3], while HfO2 FTJ can be utilized for high-density, high-

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| performance, | and | low-power | in-memory | computing |

applications because of its analog switching property, high non-linearity, and excellent uniformity [4-6].

Fig. 1 illustrates the concept of ferroelectric memory centric 3D-LSI. In the BEOL layer of high-performance logic CMOS, HfO2 FeFET is fabricated in the form of poly-Si channel TFT (ferroelectric TFT). Above FeFET layer, cross-point HfO2 FTJ arrays are stacked. Low-power/high-speed in-memory computing and storage applications are executed in a single chip. In this paper, we present the improvement guideline of the performance and reliability of both HfO2 FeFET and HfO2 FTJ for storage and in-memory applications on the basis of the physical understanding of both devices.

**II.HFO2FEFET**

In spite of attractive features of HfO2 FeFET, challenges such as small memory window (MW), window reduction during retention, and limited endurance remain to be solved. So far, it has been reported that charge trapping at the interface between the ferroelectric layer and the interfacial layer in addition to polarization itself strongly affects FeFET operation [1,7,8]. However, the quantitative extraction of both spontaneous polarization (Ps) and trap charges (Qt) during memory operation has not been performed.

Very recently, we have developed an analysis scheme to extract Ps and Qt during memory operation in HfO2 FeFET and clarified the dominant factors of MW and reliability [9]. Fig. 2 explains the developed analysis scheme. By monitoring Qm (the integration of Ig), Vth and Cg during memory operation, ΔPs and ΔQt can be extracted.

to ΔVth) after program extracted using the developed scheme is shown in Fig. 6. ΔQt is larger than ΔPs just after program, but a part of ΔQt rapidly disappears. Rapid Vth reduction just after program in Fig. 5 is caused by unstable charge detrapping (unrelated to FE). ΔVth after 100sec inherent to FeFET is determined by stable components of ΔPs and ΔQt.

Fig. 7 shows Vth as a function of the program pulse width for various Vpgm. Correlation between stable ΔPs and ΔQt for various program conditions in Fig. 7 is clearly shown in Fig. 8. ΔQt is almost 90% of ΔPs irrespective of program conditions. Moreover, ΔQt is 90% of ΔPs also at 107ksec after program. This means that 90% of ΔPs is always compensated by trap charges except for the short-term domain just after program. Fig. 9 shows the relationship between MW and ΔPs. Since only 10% of ΔPs contributes to MW because of the charge compensation, MW is proportional to ΔPs for a wide range as long as MW is lower than the coercive voltage (Vc) limit. As a result, larger ΔPs is desirable for larger MW. Fig. 10 shows ΔQt/ΔPs for HfSiO and HfO2 with another dopant. ΔQt/ΔPs is 90% irrespective of dopants.

In the long-term domain of Fig. 6, both ΔPs and ΔQt decrease with keeping ΔQt/ΔPs. As depicted in Fig. 11, depolarization field across FE-HfO2 decreases ΔPs during retention, which then decreases ΔQt. Since ΔVth corresponds to ΔQt-ΔPs, ΔVth diminishes. Therefore, ΔVth reduction in long-term domain is triggered by ΔPs reduction, not by charge injection. Stabilization of ΔPs is required to improve retention.

Fig. 12(a) shows the cycling characteristics. MW starts to decrease after 1000 cycles. The change of ΔPs, ΔQt, and ΔQt/ΔPs during cycling is shown in Fig. 12(b). There are two origins for MW reduction. One is ΔPs decrease (fatigue), and the other is ΔQt/ΔPs increase. ΔQt/ΔPs increase with cycling is caused by the increase in the trap sites which can compensate Ps as illustrated in Fig. 13. Fig. 14 shows ΔQt/ΔPs during cycling for HfSiO and HfO2 with another dopant. ΔQt/ΔPs increases with cycling irrespective of dopants. Fig. 15 shows the impact of excessive program stress. Excessive stress

|  |  |  |
| --- | --- | --- |
| 978-1-7281-8888-1/20/$31.00 ©2020 IEEE | 18.1.1 | IEDM20-375 |

degrades cycling as a result of faster ΔQt/ΔPs increase. Since excessive stress causes a larger amount of the unstable electron trapping, the repetition of electron trap/detrap via IL-SiO2 is one possible origin for the additional trap sites at the interface. Suppression of charge injection/ejection via IL-SiO2 is a key to higher endurance.

Fig. 16 shows the effective electron mobility against the inversion carrier density in poly-Si channel TFTs with SiO2 and FE-HfSiO (fresh state). While mobility of the HfSiO TFT is lower than the SiO2 TFT due to Coulomb scattering by interface states or trap charges, the peak mobility of the HfSiO TFT is still higher than 70cm2/Vs thanks to advanced solid-phase crystallization process [10].

Improvement guideline of FeFET based on the revised model is summarized in Fig. 17.

**III.HFO2FTJ FOR IN-MEMORY COMPUTING**

We have recently proposed and demonstrated in-memory RL using cross-point HfO2 FTJ array to solve path-finding and control problems [4]. Fig. 18 illustrates the concept of FTJ-based RL system [11]. An agent performs an action according to the policy and receives a positive or a negative reward, thereby making the policy updated. The action is determined in a winner-take-all manner by selecting the highest-current bit-line under parallel word-line read. Conductance of the FTJs in the array is modulated when the policy is updated.

Fig. 19 shows switching I-V and demonstration of analog switching by positive and negative voltage pulses in HfO2 FTJ. Conductance can be modulated continuously depending on the number of voltage pulses and the pulse amplitude.

In this section, we clarify the dominant factors and improvement guideline of performance and reliability of the FTJ-based RL system (Fig. 20(a)). For benchmarking, we simulated the standard problem of training to balance a pole on a cart (Fig. 20(b)). The RL system learns a policy to keep a pole on moving cart for a target duration. The number of episodes required to reach the target is considered as a performance index of the system. In our simulation, the conductance change of FTJ was calculated using a behavior model considering the switching stochasticity of HfO2 FTJ [4].

Learning performance is strongly affected by relaxation and variability. Fig. 21 shows the distribution of the read

thickness is useful for minimization of variability and improvement of learning efficiency.

Since the cross-point RL system does not have access transistors for each cell in the array, breakdown of an FTJ could result in a fatal failure of the system. Therefore, the design space where none of the devices in the array falls into breakdown failure is required for the reliable RL system. In the following, we assess whether the FTJ satisfies endurance requirement for practical RL array operation based on the TDDB (time-dependent dielectric breakdown) data.

The system typically achieves the target with several hundreds of episodes (Fig. 20(b)). We calculated the number of pulses which every device receives in 162x100 array during 500 episodes [11]. The maximum number of pulses is less than 100. We calculated the lifetime of the RL system, that is, the time-to-breakdown at which one of 162x100 devices falls into breakdown by the extrapolation using TDDB data, as shown in Fig. 23. By improving FE-HfO2 and increasing the breakdown voltage, the operation voltage can be increased as compared to conventional FE-HfO2 [11] and reaches 5V even for the extreme case. Thus, design space sufficient for pole-cart problem can be obtained with the improved HfO2 FTJ.

On the other hand, when the device area is reduced, the lifetime becomes even longer as shown in Fig. 24. This means that scaled FTJ (e.g. several tens of nm in cell diameter) can be applied to the tasks more complicated than pole-cart. In addition to scaling, FE layer optimization is useful to enlarge the breakdown voltage margin and improve endurance [5].

As well as RL, low-power vector-matrix multiplication using HfO2 FTJ array was demonstrated very recently [13].

**IV.CONCLUSION**

Improvement guidelines of FeFET were established on the basis of the revised operating model. Increase and stabilization of Ps and suppression of charge injection via the interface layer during cycling are essential. Performance of the RL system with the FTJ array is enhanced by improving relaxation and variability of FTJ cells. Scaled FTJ with improved reliability has high potential to be utilized in highly-functional RL system. FE-memory centric 3D-LSI composed of FeFET and FTJ is greatly promising for low-power and high-density storage and AI applications.

|  |  |
| --- | --- |
| current just after program [12]. Read operation was performed twice. In FTJ with conventional FE-HfO2, the distribution of | REFERENCES |

second read shows a deviation from that of first read in the range below -1σ. By optimizing the FE deposition process, the quick relaxation is much suppressed, and the first and second read currents are almost identical until -3σ. The simulation results shown in Fig. 22 reveal that learning efficiency is improved by suppressing quick relaxation, because loss of the updated policy between each episode can be suppressed by minimizing the quick relaxation.

Cycle-to-cycle variability of the conductance change by voltage pulse also has an influence on the learning efficiency. While moderate level of variability is tolerable by optimizing the pulse voltage, too large variability results in too long learning time [5]. Optimization of FE layer property such as

[1] K. Ni et al., IEEE Trans. Electron Devices 65, 2461 (2018).

[2] T. Mikolajick et al., IEDM Tech. Dig., p. 354 (2019).

[3] T. Ali et al., IEDM Tech. Dig., p. 665 (2019).

[4] R. Berdan et al., Dig. Symp. VLSI Tech., p. 22 (2019).

[5] K. Ota et al., IEDM Tech. Dig., p. 114 (2019).

[6] T.-Y. Wu et al., IEDM Tech. Dig., p. 118 (2019).

[7] M. Pesic et al., IEDM Tech. Dig., p. 588 (2018).

[8] K. Toprasertpong et al., IEDM Tech. Dig., p. 570 (2019).

[9] R. Ichihara et al., Dig. Symp. VLSI Tech., TF1.2 (2020).

[10] M. Oda, K. Sakuma, Y. Kamimuta, and M. Saitoh, IEDM Tech. Dig., p. 125 (2015).

[11] M. Yamaguchi et al., Proc. IRPS, 8B.4 (2020).

[12] S. Fujii et al., Dig. Symp. VLSI Tech., TF2.3 (2020).

[13] R. Berdan, T. Marukame, K. Ota, M. Yamaguchi, M. Saitoh, S. Fujii, J. Deguchi, and Y. Nishi, Nature Electron. 3, p. 259 (2020).

IEDM20-376 18.1.2

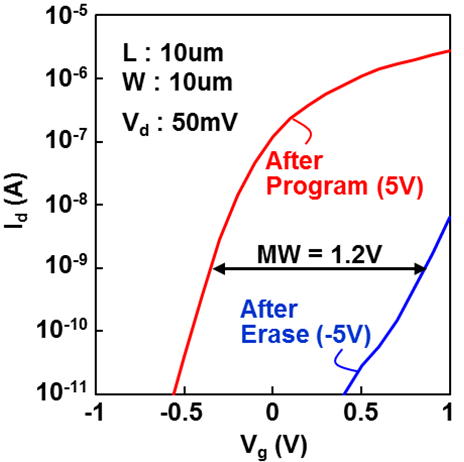
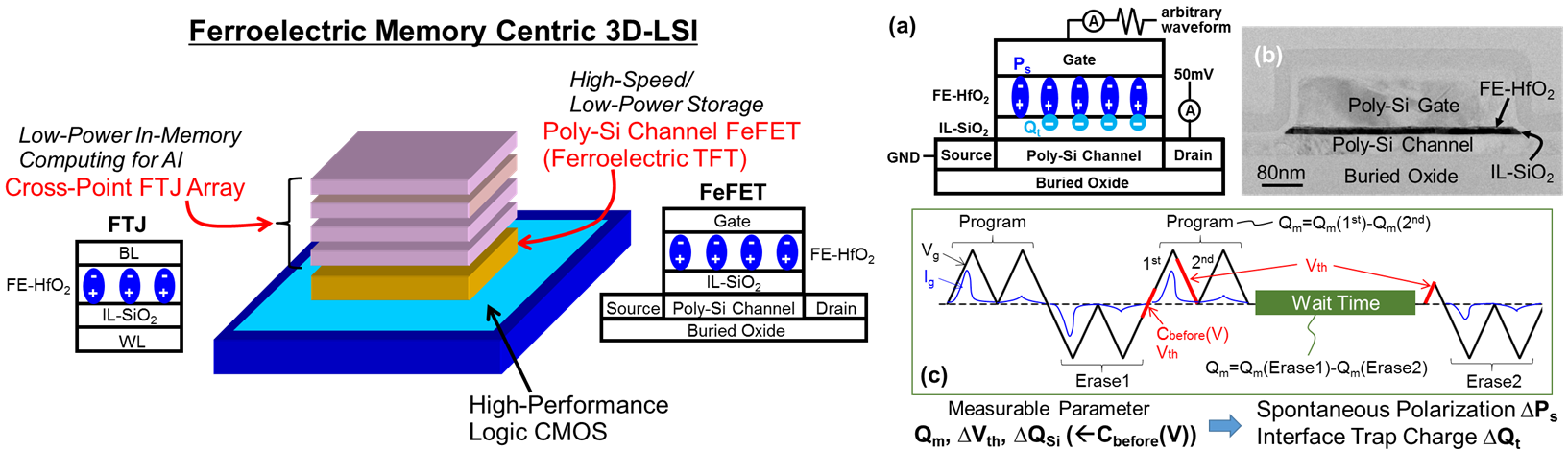


Fig. 1. Concept of ferroelectric memory centric 3D-LSI. Fig. 2. (a) Schematic of FeFET under Fig. 3. Id-Vg after program at Cross-point FTJ arrays for low-power/high-speed in- measurement. (b) Cross-sectional TEM of the 5V and erase at -5V. memory computing and poly-Si channel FeFETs for measured FeFET. (c) Measurement sequence Measurement is performed at storage applications are stacked in a single chip. to extract ΔPs and ΔQt. 100sec after program.

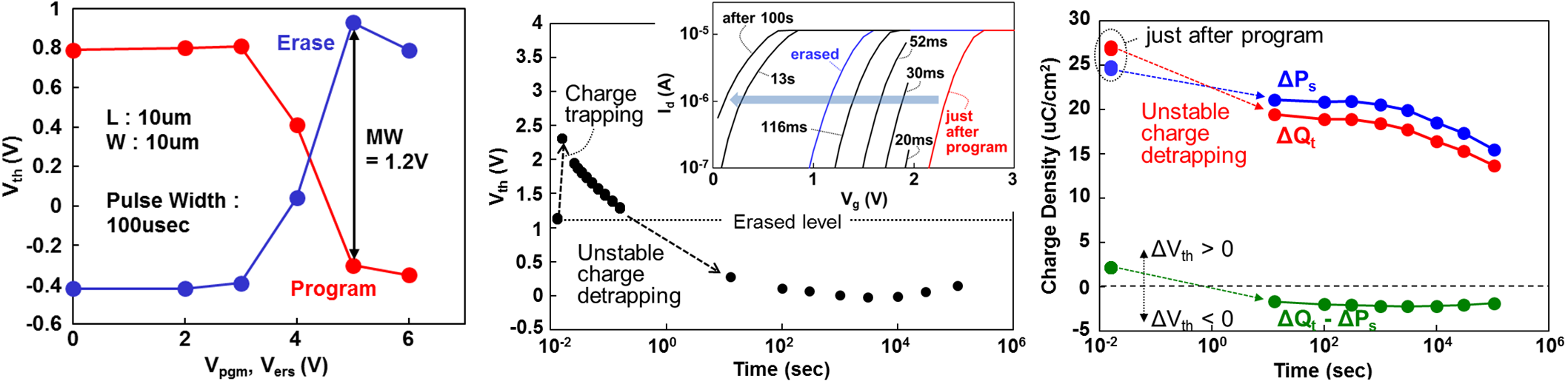


Fig. 4. Vth as a function of Vpgm and Vers. Fig. 5. Time evolution of Vth after program. Fig. 6. Time evolution of ΔPs, ΔQt, and ΔQt-ΔPs Voltage pulse width is 100usec. Vth shifts to positive just after program, then (corresponding to ΔVth) after program extracted Measurement is performed at 100sec rapidly decreases in 10sec (Revised from [9]). using the developed scheme (Revised from [9]). after program. Inset shows Id-Vg at each point of time.













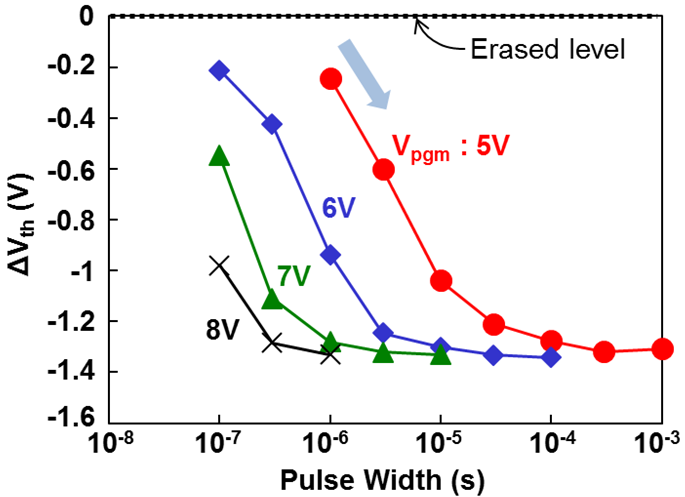
 

























































































































































































































































































































































































































































































































































































































































































































































































































































































































































































































Fig. 7. ΔVth as a function of the program Fig. 8. Relationship between stable ΔPs and Fig. 9. Relationship between MW and ΔPs pulse width for various Vpgm. ΔVth of 1V is ΔQt for various program conditions in Fig. 7 (Revised from [9]). MW is proportional to obtained within 100nsec for Vpgm of 8V. and 107ksec after program (Revised from   
 ΔPs for a wide range as long as MW is lower [9]). ΔQt is always ~90% of ΔPs. than 2Vc limit.

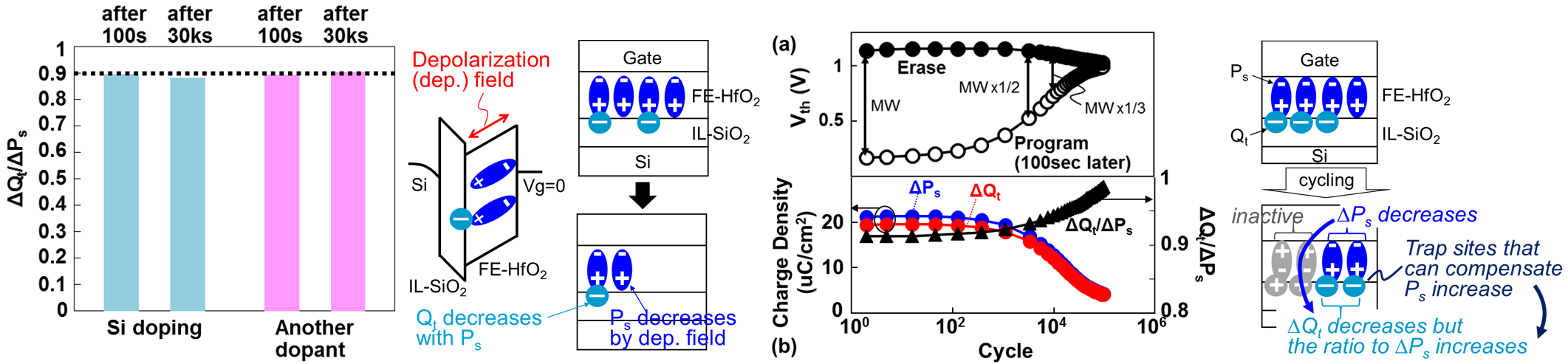
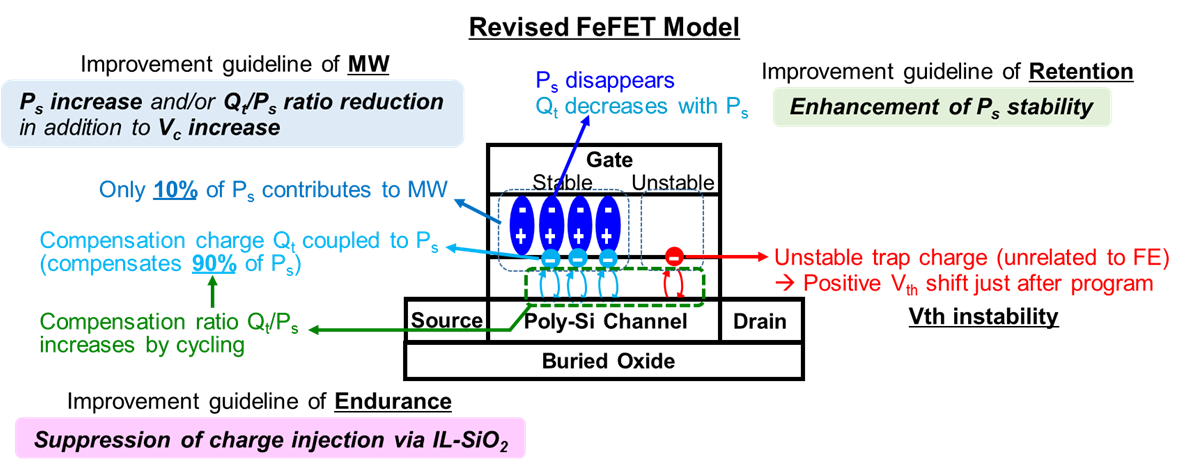


Fig. 10. ΔQt/ΔPs for HfSiO and Fig. 11. Schematic of Ps and Fig. 12. (a) Cycling characteristics and Fig. 13. Schematic of Ps and HfO2 with another dopant for 100s Qt behavior during retention (b) change of ΔPs, ΔQt, and ΔQt/ΔPs Qt behavior during cycling and 30ks after program. ΔQt/ΔPs is process (Revised from [9]). during cycling (Revised from [9]). process (Revised from [9]). 90% irrespective of dopants.

18.1.3 IEDM20-377

|  |  |  |
| --- | --- | --- |
|  |  |  |
| Fig. 14. Change of ΔQt/ΔPs during cycling for | Fig. 15. Impact of excessive program stress | Fig. 16. Effective electron mobility against |
| HfSiO and HfO2 with another dopant. ΔQt/ΔPs | on cycling. Excessive stress degrades | the inversion carrier density (Ninv) in poly-Si |
| after 2cycles and the cycles where MW is | cycling as a result of faster ΔQt/ΔPs | channel TFTs with SiO2 and FE-HfSiO |

reduced to 1/2 and 1/3 are shown. increase (Revised from [9]).



(fresh state).

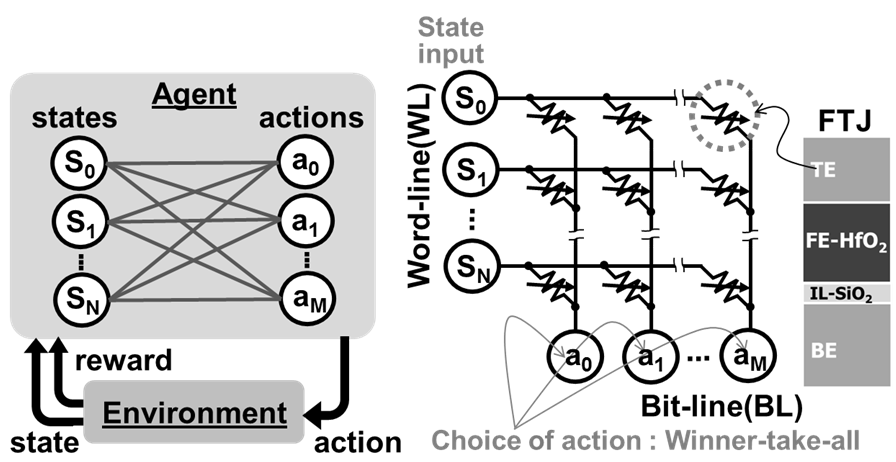


Fig. 17. Improvement guideline of FeFET performance (MW) and Fig. 18. Concept of FTJ-based RL system (Revised from

reliability (retention and endurance) based on the revised model (Revised from [9]). Increase and stabilization of Ps and suppression of charge

[11]). Agent takes actions determined by the highest current in the bit-lines under word-line parallel read in the cross-point

injection via IL-SiO2 during cycling are keys to the improvement. FTJ array.

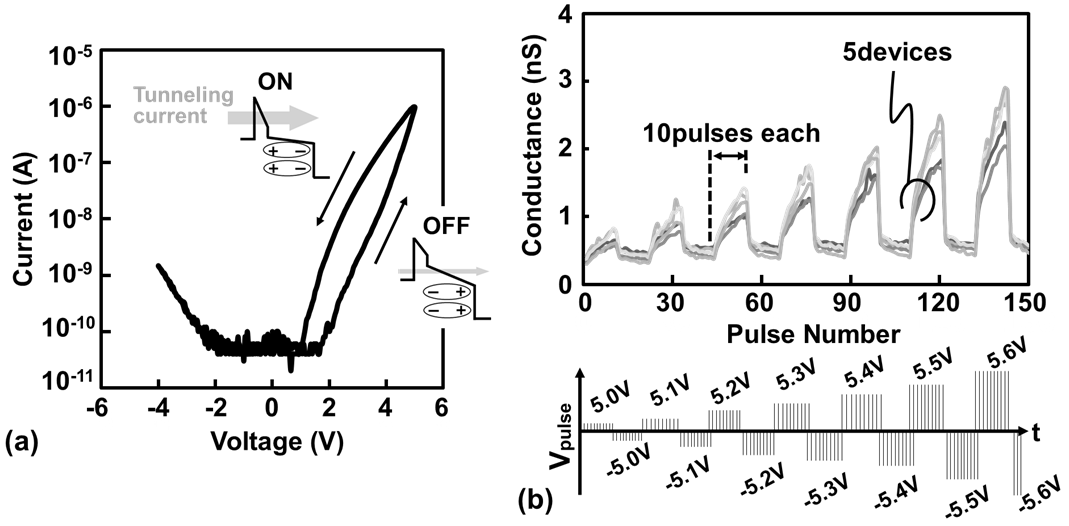


Fig. 19. (a) Switching I-V and (b) demonstration of analog switching by positive and negative voltage pulses in HfO2 FTJ. Conductance can be modulated continuously depending on the number of voltage pulses and the

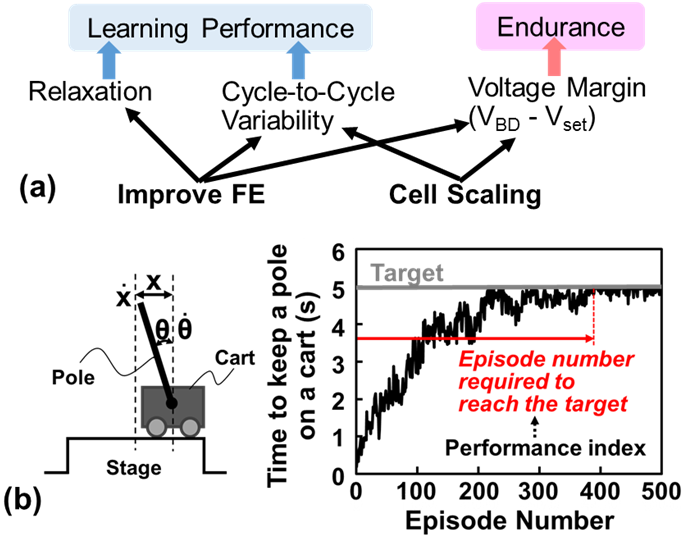


Fig. 20. (a) Improvement guideline of the performance and endurance of the FTJ-based RL system. (b) Schematic of the Pole-cart balancing problem and the example of

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| pulse amplitude.    Fig. 21. Read current distribution just after program for conv. and improved FE (Revised from [12]). Read operation was performed twice (inset).  IEDM20-378 | learning history to keep a pole on a cart. | | | | | | | | | |
|  | | |  | |  | | | | |
| Fig. 22. Pole-cart simulation | | | Fig. 23. Time-to-failure of the | Fig. 24. Time-to-failure of the RL | | | | | |
| for conv. FE, improved FE | | | RL system as a function of the | system as a function of the cell area | | | | | |
| and | ideal | (no-relaxation) | pulse voltage for conv. FE | for | | improved | FE. | Scaled | FTJ |
| devices (Revised from [12]). | | | [11] and improved FE. | improves the lifetime and can be | | | | | |
| applied to more complicated tasks.  18.1.4 | | | | | | | | | |