Memory-Logic Hybrid Gate with 3D-Stackable Complementary Latches for FinFET-based Neural Networks

Chieh Lee[1], Yue-Der Chih[2], Jonathan Chang[2], Chrong Jung Lin[1] and Ya-Chin King[1]
[1]Institute of Electronics Engineering, National Tsing Hua University, Hsinchu, Taiwan
[2]Design Technology Division, Taiwan Semiconductor Manufacturing Company, Hsinchu, Taiwan
Phone/Fax: +886-3-5162219, E-mail: ycking@ee.nthu.edu.tw

Abstract

A memory-logic hybrid gate with complementary resistive switching pairs on vias in BEOL FinFET technologies with an area-efficient, 3D-stackable structures is proposed. Stable output logic stages enabled by the complementary states on the RRAM pair have been demonstrated. Through stacked-vias architectures, logic operations based on multiple non-volatile states are achieved.

I. Introduction

Power consumption and data transmission in advanced computing units have become the critical bottlenecks, in recent years. Hence, researches and developments start to seek alternative approaches such as computing-in-memory (CIM) [1] and neural network (NN) based architectures [2] to further boost computational power and efficiency. Resistive random access memory (RRAM), featuring high-density and high-compatibility to CMOS technologies, has been one of the promising candidates for realizing future neuromorphic computing systems [3]. The "near-data processing" architecture places logic circuits inside memory to avoid the power and speed limitations as a result of the von-Neumann bottleneck [4]. However, RRAM-based NN subjects to stability and variability problems which can leads to faulty functions, computational glitches, and reliability concerns. The memristor-based nonvolatile logics has been reported in many prior studies [5][6], which typically required special RRAM processes which are not fully compatible to advance CMOS process. In this work, we investigate a non-volatile (NV) logic gate with complementary resistive switches through BEOL Via structures implemented in 16nm FinFET processes. By incorporating the twin-bit RRAM cell reported in previously in [7], a new memory-logic hybrid gate with full-compatibility to FinFET logic processed, superior endurance and stable logic output is demonstrated. Through the unique shared-via arrangement, the twin-pair RRAM forms a complementary latch (CL) to increase immunity to variations and ensure well-controlled logic states during evalutions.

II. Single-layer CL Hybrid Gate and Operation Principle

The most basic single-layer complementary latch hybrid logic gate is illustrated in the circuit schematic shown **Fig.1 (a)**. The non-volatile logic gate is composed of one FinFET and one twin-bit via RRAM cells which stored the non-volatile state, Q, in the R_1, R_2 level, where the R_1 and R_2 are always in the opposite states, i.e. one in the HRS, the other in LRS and vice versa. The common via in between the two RRAM is connected to the transistor which controls the logic output node "Y". The metal gate of the FinFET becomes the SET terminal, labelled as "S" and the input data terminal is label as "D". The corresponding circuit symbol and the truth table of the nonvolatile logic gate are shown in **Fig. 1(b)** and **Fig. 1(c)**, respectively. As shown on the truth table, when S is low, the transistor is off, the output data (Y) remains unchanged. On the other hand, when S is high, the transistor will turn on. The output level then changes in response to the input data, D.

Fig. 2(a) shows the layout configuration of a hybrid logic gate with the complementary latch placed above a standard FinFET transistor. The cross-sectional views of this device with single layer CL along the AA' plane and BB' planes are demonstrated, respectively, in **Fig. 2(b)** and **(c)**. The first layer of CL is formed with a designed clearance (ranging from14~18nm) between the two mask layers, Via1 and metal 1. As shown in the TEM picture, two RRAM devices are formed in the opposite sides of a single Via, where TMO layer consists of TaON and SiO_2 between the sidewalls Cu Via and Metal electrodes. By setting the two RRAMs in opposite states, one layer of a complementary latch can be combined with a FinFET, which controls the writing of the non-voltage state, Q, as well as gating the input, D, during logic operations. The basic flows of writing Q onto the NV latch are summarized in **Fig. 3**, while the biasing conditions at each stage are listed in **Table 1**.

Fig. 4 shows the 3D illustration of the hybrid logic gate with CL. Contrary with conventional D latch, the device stores the data in states of complementary pairs, (R_1, R_2), namely, NV data, Q, can remained when the power is off, as illustrated in **Fig. 5**. This combination forms the key unit for composing a state machine instead of conventional programmable logic arrays. For updating the data Q, one needs first to "Clear" the two RRAM into a balanced state initially. During "Clear" step, both RRAMs are set to high resistance state (HRS). For Write "0", D is raised high, so that R_2 is set to low resistance state (LRS) and R_1 remains in HRS. In Set "1", D is low, set R_1 to HRS. During Read, the S is low, and the output is determined by the complementary state in Q, the non-volatile data stored. The complementary settings of each RRAM pair can ensure better stability of the data, higher static noise margins and increase immunity to device-to-device variations.

III. Characteristics of single-layer CRS logic

The measured voltage waveform of the hybrid gate is arranged in **Fig. 6**, where the output responses as a results of logic operation of different logic states are demonstrated. With different input (S, D), the output (Y) will change accordingly. For Q=0, when the input (S, D) are (0, 0), (0, 1), and (1, 0), the output will be logic 0. When input is (1, 1), the output will be logic 1. When transistor is off, the output will then be control by state, Q. When S is high, the output will change with input data D. Furthermore, the restoration of Q on the hybrid gate is shown in **Fig. 7**. Measurement data demonstrated that Q are kept after power is off and can be retrieved directly for logic operations once the power is turn back on. To further evaluation the reliability of the NV data, Q, **Fig. 8** shows a stable output signal V_Y for over 100 hours under 150□ bake tests. This results shows that through the complementary states on the two RRAMs, the non-volatile latch exhibits superior data stability without supply power. In order to ensure no error during writing data Q, the over-set and over-reset tests are also performed to address data reliability, as shown in **Fig. 9**. The output voltage is found to remain stable for over 10K cycles. This result indicates the non-volatile logic gate is immune to disturbance from the operations of the RRAM on the other side. This will ensure the reliable Q state can be used provide correct logic state at the output terminal.

The variability of resistance levels from cycle-to-cycle and device-to-device are known as one of the most critical challenges for the RRAM based logics. To address the variation effect, we first investigate the distributions of the resistance of RRAMs between Via1 and M1 in its both states, see **Fig. 10(a)**. As expected, the resistance level in its LRS and HRS do exhibit fair large difference from device-to-device. However, as a result of the resistance ratio between the two states, output levels, V_Y, in response to the state Q can be well controlled, as summarized in **Fig.10 (b)**. Since, the TMO layers in the complementary latches are defined by the spacing between metal and via, they are susceptible to misalignments between the two mask layers. This is expected to lead to significant shifts in the characteristics of the twin-bit RRAMs. As show in **Fig. 11**, the initial characteristics of R_1 and R_2 are found to quiet diverse. This mismatching between the two RRAMs on both sides might still lead to non-ideal Q states in sizable circuits which large number of hybrid gates are used. In order to improve the matching between the two RRAMs on the opposite side of the same via, a hybrid logic gate with mirror structure is proposed. The circuit schematic of such the mirror gates shown in **Fig. 12** is composed of pair of two vias. **Fig. 13** is the layout of mirror structure non-volatile latch. Two via RRAMs on the left hand side formed a pair. In case of significant misalignments, both vias are expected shift in one-side. That is to say, twin-bit RRAM pair in the same side tend to shift in the same directions, increase the matching between RRAM pairs. Data in **Fig. 14** show that there is a strong correlation between the initial resistances from a RRAM pair These results suggest that the mirror structure can effectively increase the matching in individual pairs. Correspondingly, the output response in accordance to the Q state can be more precisely controlled, as compared, in **Fig. 15**.

III. 3D-Stackable CL for Multi-State Q

To further increase the number of available NV states, a hybrid gate with 3D stackable latches is purposed. First, we investigate the general feature of the twin-bit RRAM at each layer, where the set and reset characteristics are summarized in **Fig. 16**. These RRAM at all three metal layers can be set and reset with operation voltage lower than 3V. Using the complementary latches formed between different metal layers and the corresponding vias, an example and the multi-state NV gate is composed of one transistor, two pairs of M1-via1 RRAM, and one pair of M3-via3 RRAM, as explained in **Fig. 17**. The corresponding TEM pictures obtained from different cut sections in the 2-layer hybrid gate demonstrated the stackable via connections. The resistance distribution at two states are summarized in **Fig. 18**, revealing a resistance different over 20X. The circuit schematic this 2-layer hybrid gate example, its circuit symbol and truth table with 3 different CL storing Q_1, Q_2, Q_3 states are shown in **Fig. 19**. The detail operation flow of setting Q_1, Q_2, Q_3 states are outlined in **Fig. 20**, where the highest voltage required is $2V_{DD}$, which remains relatively low enough to be compatible to advanced FinFET circuits. During logic operation, terminal A and B are common to V_{DD}, to first simplify the logic input to 2. For operation with addition logic inputs, A, B terminals can be used to expanded the possible logic operation as needed. Finally, the distribution of the logic output voltage under all 8 logic stages in Qs' are compared in **Fig. 21**. Through the complementary configuration, tight V_Y distributions can still be maintained in the multi-state hybrid gate. **Fig. 22** shows the V_Y response during data restoration in Q_1, Q_2, Q_3, where the accurate output response are found at the output terminal.

IV. Conclusion

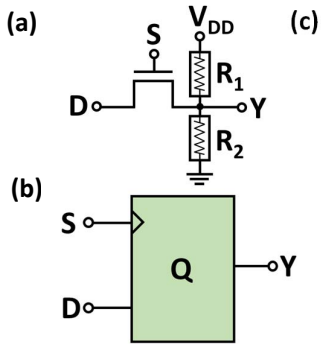
A hybrid logic gate with complementary latches from on the two sides of the Via is proposed and successfully demonstrated in a 16nm FinFET CMOS logic process. By measurement, all data clear, write, and read functions are demonstrated with stable and accurate response. This memory-logic hybrid gate with 3D-stackable CL features great area efficiency and immunity to variability, which can be the key unit for building FinFET based neural networks.

Acknowledgement

The authors would like to thank the support from the Ministry of Science and Technology, Taiwan, and Taiwan Semiconductor Manufacturing Company, Hsinchu, Taiwan.

References

[1] P. C, et al, ISCA, 2016
[2] T.M, et al, ISCA, 2010
[3] G. Indiveri, et al, Nanotechnology, Vol. 24, NO 38
[4] J. B, ACM Turing Award Lecture, 1977
[5] P. W. C. Ho, et al, TENCON, pp.1-4, 2015
[6] R. R, et al, IEEE T MAGN, Vol.51, No.6, 2015
[7] Y. W. C, et al, IEDM, 2014
[8] Y. H. S, et al, JJAP, Vol. 57, no. 4S, Mar. 2018
[9] Y. C. L, et al, JEDS, Vol. 2, No.6, pp. 149-153, Nov. 2014
[10] W. H. C, et. Al, ISSCC,2018



(c)

S	D	Q	Y
0	0	0	0
0	0	1	1
0	1	0	0
0	1	1	1
1	0	0	0
1	0	1	0
1	1	0	1
1	1	1	1

$Y=S\times D+Q$

Figure 1 (a) The schematic of the non-volatile logic gate, (b) its circuit symbol, logic expression and (c) the corresponding truth table for single-layer NV-latch.

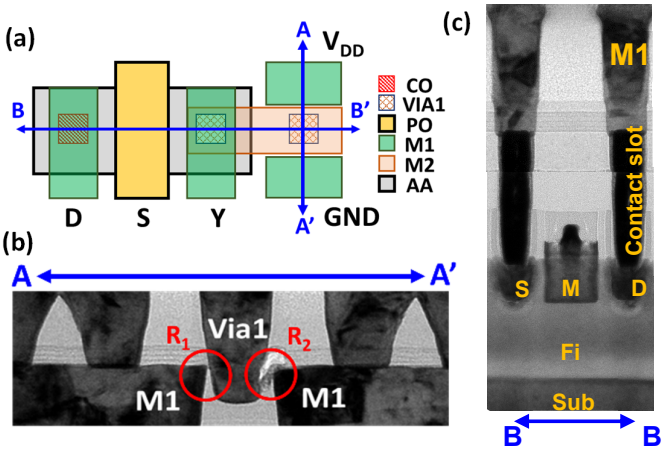


Figure 2 (a) Layout of twin-bit RRAM based non-volatile latch implemented by 16nm FinFET CMOS process and the cross-sectional TEM pictures showing the (b) twin-bit RRAM cut and (c) along FinFET channel directions.

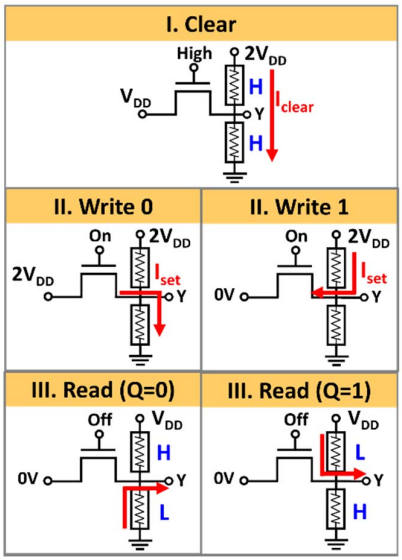


Figure 3 Operation steps of complementary latch. Every cycle begins with a clear function, followed by setting the Q level.

	Clear	Set		Read
		"0"	"1"	
V_{SE}	$2V_{DD}$	V_{DD}	V_{DD}	0
V_{DT}	V_{DD}	$2V_{DD}$	0	X
V_{DD}	$2V_{DD}$	$2V_{DD}$	$2V_{DD}$	V_{DD}

Table 1 List of bias conditions for clear/set/access phases, V_{DD}=1.3V.

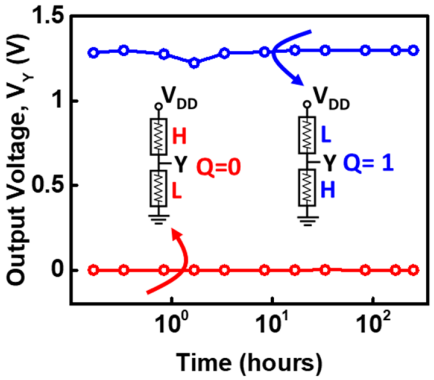


Figure 8 Stability of the latched data tested on 150C bake, showing stable output signal without alternations.

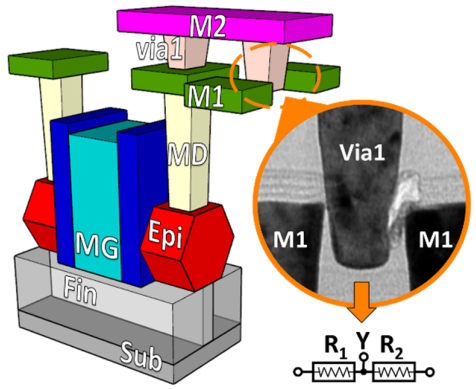


Figure 4 3D illustrations of the single-layer CL with FinFET and the close-up TEM of the twin-bit RRAM in CL.

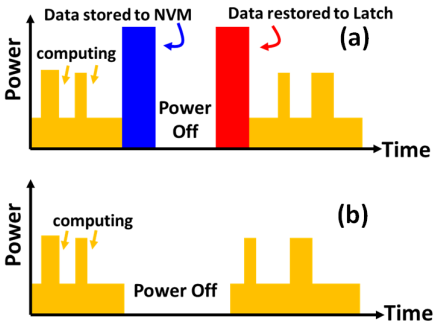


Figure 5 Illustration of power consumption in (a) a normal NV latch and (b) that in a RRAM based non-volatile logic circuits.

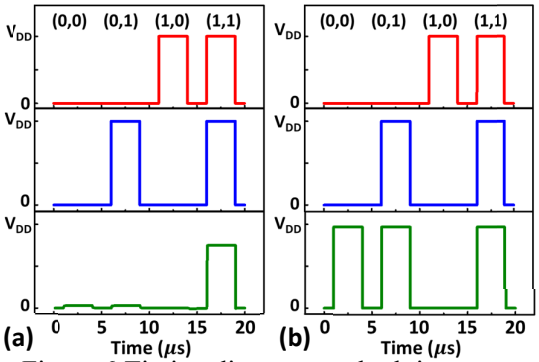


Figure 6 Timing diagrams on both input and that of the responding output during set stage for setting the non-volatile data, at (a) Q=0 and (b) Q=1, respectively.

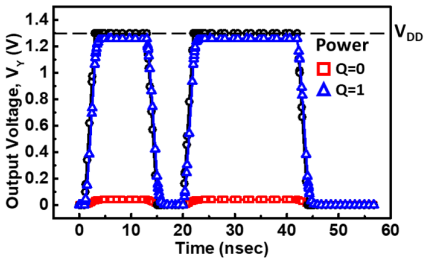


Figure 7 Restoration of NV data, Q, once power is back on, the Q, stored in complementary pair is successfully restored.

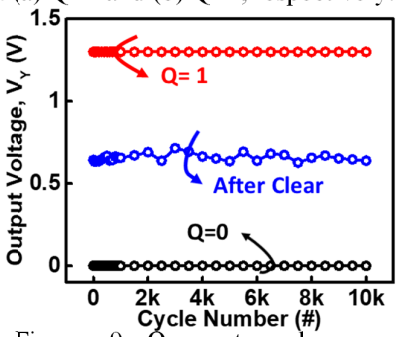


Figure 9 Over-set and over-reset testing results. Both set-0 disturb, set-1 are appeared to be disturb-free to Q.

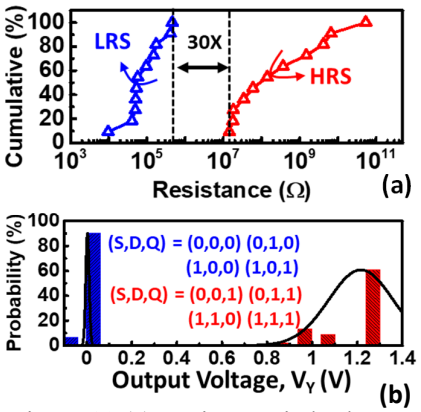


Figure 10 (a) Resistance in both states and (b) its effect of on its output levels under different logic states.

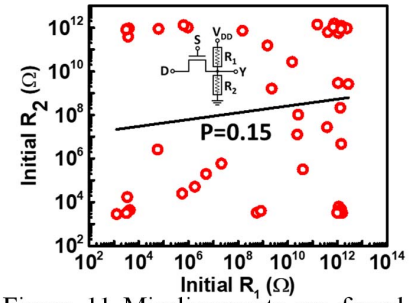


Figure 11 Misalignments are found to causes random shift in the RRAM pairs.

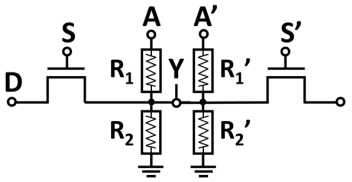


Figure 12 The circuit schematic of the proposed mirror complementary latch structure for the enhancement of the characteristic matching in the twin-bit RRAM pairs in sizable circuits.

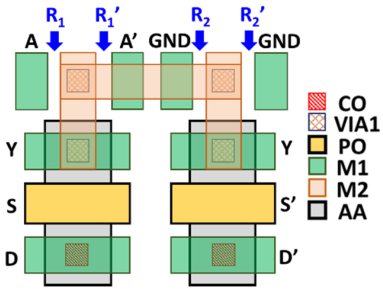


Figure 13 Layout of the mirror complementary NV latch gates.

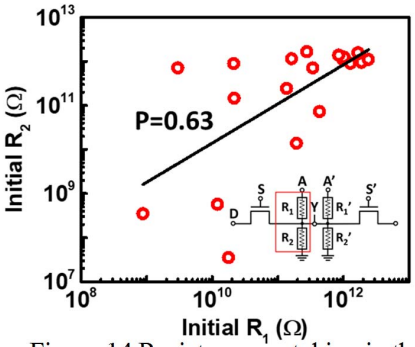


Figure 14 Resistance matching in the complementary latch significantly improves in the mirror structures.

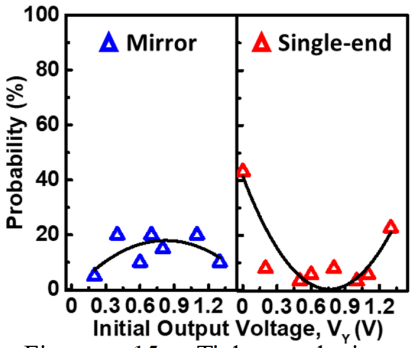


Figure 15 Tighten logic state distributions are found in the symmetric latch structures.

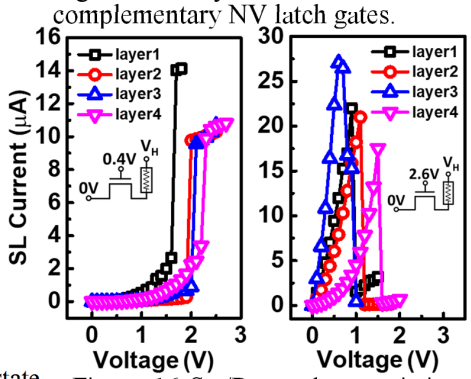


Figure 16 Set/Reset characteristics of the twin-bit RRAMs at each layer.

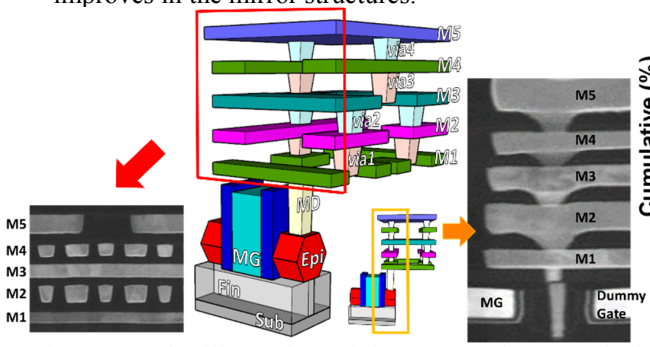


Figure 17 The illustration of the proposed 3D stacked latches and TEMs from cuts on different directions

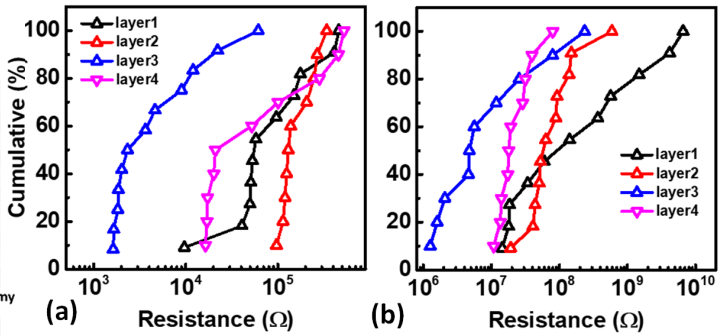
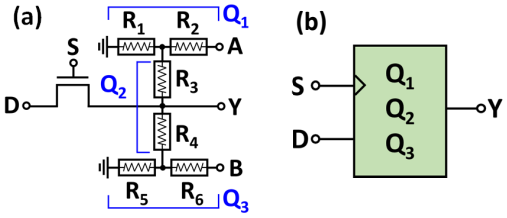


Figure 18 Distributions of resistance level in its (a) low and (b) high states are compared from RRAM from each layers.



S	D	Y
0	0	$Q_1Q_2+Q_3Q_4$
0	1	$Q_1Q_2+Q_3Q_4$
1	0	0
1	1	1

Figure 19 (a) Circuit Schematic of a 2 layers hybrid gate, (b) its circuit symbol and logic expression and (c) its complete truth table.

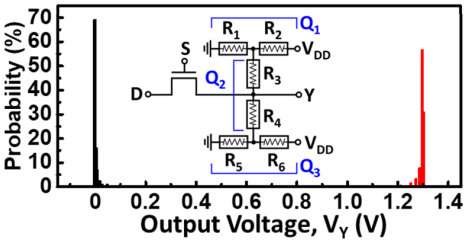


Figure 20 Output voltage distributions in response to the different non-volatile states.

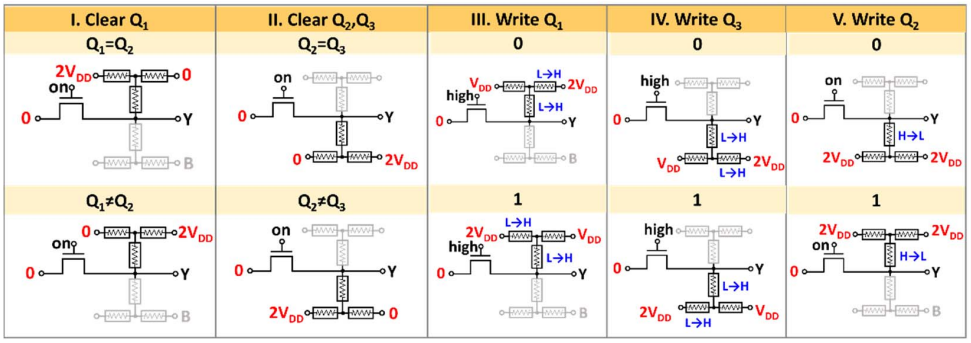


Figure 21 Illustrations on the writing procedure of non-volatile data, Q_1, Q_2, Q_3 onto different layer of stackable complementary latch pairs.

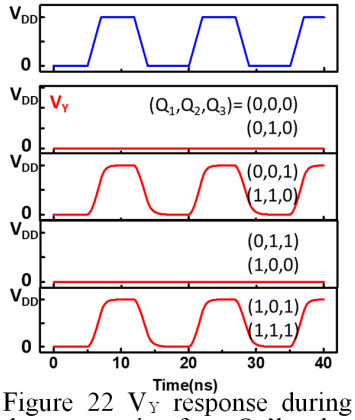


Figure 22 V_Y response during data restoration from Qs' levels.