Extended Abstracts of the 15th Conference on Solid State Devices and Materials, Tokyo, 1983, pp. 269-272

A-7-6

Double Word Line and Bit Line Structure for VLSI RAMs —Reduction of Word Line and Bit Line Delay—

Takayasu SAKURAI, and Tetsuya IIZUKA

Semiconductor Device Engineering Laboratory, Toshiba Corporation
1-Komukai Toshiba, Saiwai-ku, Kawasaki 210 Japan
81-44-511-2111 ex 2676

This paper describes new word and bit line structures especially suited for VLSI RAMs, namely double word/bit line (DWL/DBL) structures. The feasibility of the DWL structure is studied for 256Kbit high-resistive poly-Si load static RAM. It is shown that a word line delay reduction is possible even though relatively high-resistive poly-Si layer is used as a 2nd word line. Power dissipation is about a half of the conventional structure. The DBL structure is devised and bit line delay reduction by a factor of two is expected at a stage of 256Kbit static RAM. DWL/DBL structure becomes more important in future VLSI SRAMs where proportion of word/bit line delay in access time increases.

1. Introduction

As storage bit capacity of MOSRAMs becomes larger, capacitance of a word line (WL) or a bit line (BL) increases. This is because WL/BL line capacitance per one memory cell is scaled by a factor less than two, even though the number of cells per WL/BL is doubled as RAM storage capacity is quadrupled. Moreover, resistance of a WL also increases as device dimensions are scaled down. Then, WL/BL delay will not be scaled as geometries. On the other hand, peripheral circuit delay is roughly scaled down by a factor of $1/S^2$, where S is a scaling factor. Therefore, WL/BL delay reduction becomes vital to a fast access time. This situation is demonstrated in Fig. 1.

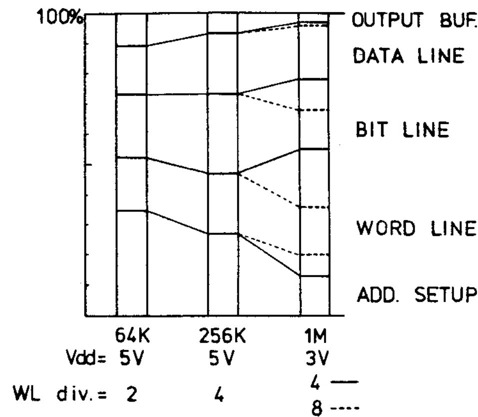

Fig.1 Delay component of SRAM. External output capacitance is set zero. WL and BL delay component increase in future SRAMs.

The DWL structure was proposed to reduce the word line capacitance[1,2]. In this paper, the feasibility of the DWL structure is studied for 256Kbit high-resistive poly-Si load static RAM. In section 2, word line delay reduction by a factor of 2 ~ 3 is shown by a computer simulation. The DBL structure is devised for a fast bit line response in section 3. Computer simulation is carried out assuming 256Kbit SRAM. Discussion on future VLSI SRAM application is also given. The last section is dedicated for conclusion.

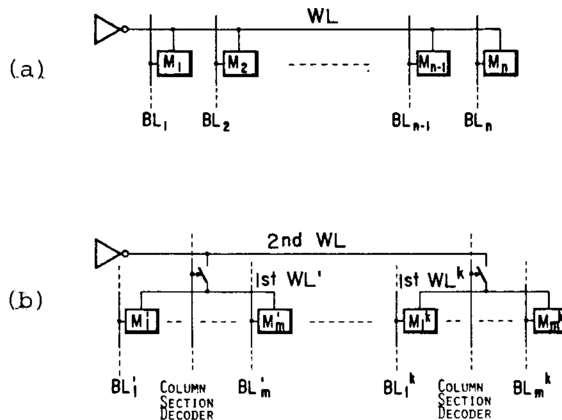

Fig.2 DWL structure. (a) Conventional WL structure. Many cells are connected to one WL. (b) DWL structure. Smaller number of cells are driven so as to lower the WL capacitance. A block belonging to one 1st WL is called a section.

2. Double Word Line (DWL) structure

A conventional word line structure is shown in Fig. 2a where many memory cells are directly con-

nected to one word line. On the other hand, in DWL structure shown in Fig. 2b, the 1st word line drives smaller number of memory cells. One of the 1st word lines is connected to a 2nd word line at a time through a switching circuit controlled by a column section decoder. The 2nd WL is driven by a row decoder. Total word line capacitance is reduced, because a word line has to drive smaller number of memory cell capacitance.

Some concrete circuits are devised for the switching circuit, namely drain input structure (Fig. 3a) and gate input structure (Fig. 3b). Schematic operation waveforms are shown in Fig. 4. ϕ_I is a word line inhibit pulse which is triggered by a row address transition detector. Whenever row address changes, the inhibit pulse is generated. When the inhibit pulse goes high, all 1st word lines are discharged through n-ch transistor in the switching circuit. Meanwhile 2nd word line goes low slowly and column section decoder completes decoding. When the inhibit pulse ends, one certain 1st word line is charged up through p-ch transistor in the switching circuit.

In order to calculate a behavior of these switching circuits by a circuit simulator (SPICE), good models for word lines are investigated[3]. L-ladder CR circuit model shown in Fig. 5a is widely used to simulate a word line which can be considered as a distributed CR line. 2-step L-ladder model is applied to a case of 256Kbit CMOS RAM. The voltage response is compared with an exact solution in Fig. 6. The 2-step L-ladder model gives about 30% slower response than a distributed CR line, which can not be said satisfactory.

Here, we propose pi-ladder model shown in Fig. 5b, whose voltage response is also shown in Fig. 6. In this figure, the exact solution is obtained by using 30-step pi-ladder model whose results coincide with 30-step L-ladder model. A good agreement is observed, either at the nearest point to or the farthest point from a word line driver. The computational cost of pi-ladder model is almost the same as that of a L-ladder model, so that 3-step pi-ladder model is adopted in the following calculation.

Computer simulation by SPICE is carried out assuming 256Kbit poly-Si load static RAM. The transistor models are fitted to measured data of 1.2um-gate MOSFETs. Two dimensional effects are

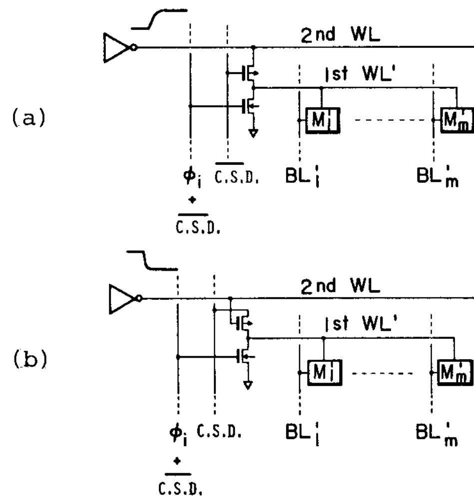

Fig.3 Switching circuits. (a) Drain-input structure where p-ch MOSFET operate as a transfer gate. (b) Gate-input structure in which p-ch MOSFET serves as an accelerator to generate fast 1st WL response.

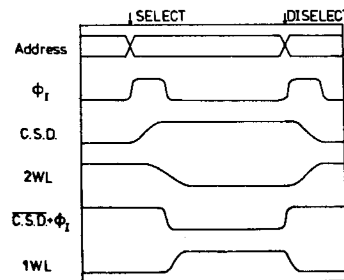

Fig.4 Schematic operation waveforms for DWL switching circuit. ϕ_I is an inhibit pulse triggered by address transition, and C.S.D. is column section decoder output, which goes high when the section is selected.

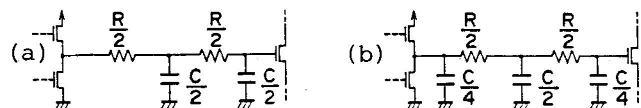

Fig.5 Word line models by lumped circuits. (a) 2-step L-ladder circuits which is widely used to simulate WLs. (b) Proposed 2-step pi-ladder model.

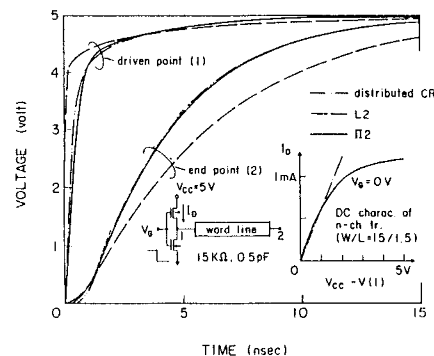

Fig.6 Simulated WL responses assuming 256Kbit SRAM. Widely used L-ladder model cannot give satisfactory results. Whereas the proposed pi-ladder model can reproduce the correct response either at the nearest point to and the farthest point from WL driver.

taken into account in capacitance estimation[4]. The results are shown in Fig. 7. Gate-input structure turns out to operate faster than drain-input structure. This is because p-ch MOSFET in switching circuit of gate-input structure converts slow 2nd WL waveform to fast 1st WL response.

Word line delay reduction by a factor of 2 ~ 3 is observed by using DWL, even though sheet resistance of 2nd WL is over 50 ohm. This level of resistivity is achieved by selective doping of high-resistive 2nd poly-Si layer, although it is difficult to lower it to a level of 1st poly-Si layer due to a lack of sufficient activation heat process. This means that DWL structure is possibly made without adding any extra process. Since 2nd WL is stacked on 1st WL, memory cell area remains unchanged.

Power consumption is also decreased in the DWL scheme because the number of activated cells is much smaller than conventional structure. About 60% of operation current flow as cell current in the conventional structure. This current component is decreased to 32/128 or 16/128 depending on the number of cells per one 1st WL. Therefore, total power consumption is cut down to about a half.

Another way to reduce the CR delay of a word line is the use of aluminum word line. However, in this method no decrease of power dissipation is possible. Cell current is almost proportional to drive capability of a memory cell transfer gate. As a MOSFET is scaled down and conductivity of transfer gate increases, more power consumption is expected at activated memory cells. In order to cut down the power dissipation, multi-block scheme is usually adopted. But more area for row decoders is required than DWL structure, because row decoder takes about 20 times as large area as a memory cell compared with only 2 memory cell area for section select switching circuits.

3. Double Bit Line (DBL) structure

Conceptional DBL structure is shown in Fig. 8. 1st BL is connected to fewer memory cells and discharges smaller capacitance so as to move faster. Figures 9a and 9b show drain- and gate- input DBL structure, respectively. Since BL is bi-directional bus, two amplifiers should be placed in gate-input structure. Figure 10 is BL voltage waveforms for different structures. The 2nd BL takes 4ns to arise 0.5V voltage difference, while conventional BL takes 9ns.

The DBL structure requires one more metal layer, but the design rule for this layer is not so severe since not every cell has contact holes. Since this metal layer can be separated from the under-layers with thick oxide, capacitance of 2nd WL is relatively small.

In further SRAMs, we suppose that the vertical dimensions such as poly-Si line thickness and field oxide thickness remain unchanged except for gate oxide thickness. Lateral dimensions are scaled down by a factor of 1.5. These conditions give the most optimistic value for bit line delay. Figure 11 is BL voltage waveforms for 1Mbit SRAM, assuming 3V for supply voltage. The DBL structure effectively reduces the BL delay in this case, too.

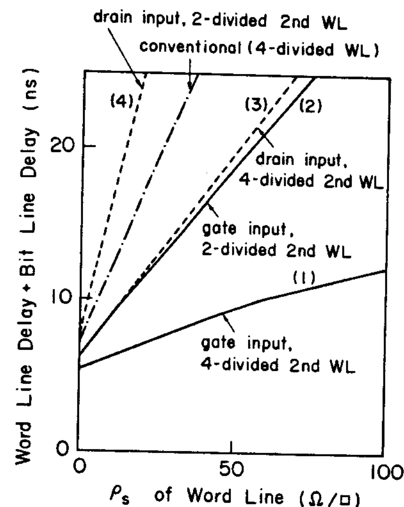

Fig.7 Word line delays for various structure. Delay reduction by a factor of 2 ~ 3 is possible using DWL structure. Gate-input structure operates faster than Drain-input structure. The time when BL is discharged to 4.5V is plotted as a function of WL sheet resistance. 1st poly-Si is about 20 ohm/□ and 2nd poly-Si is over 50 ohm/□.

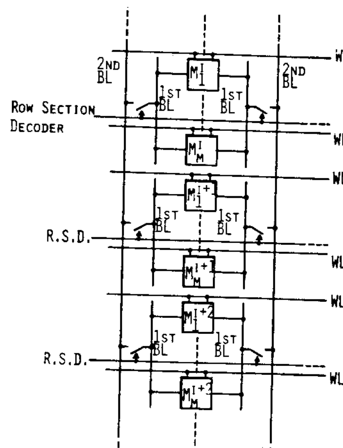

Fig.8 Conceptional DBL structure. Smaller number of cells are connected to 1st BL so that faster response is expected.

4. Conclusion

In conclusion new double word/bit line structures can be realized by present technology and are effective in reducing delays and power consumption at a stage of 256Kbit SRAMs. The importance of these structures increase in future SRAMs where word and bit delay become limiting factors for fast operation.

Acknowledgments

The authors would like to express their sincere appreciation to Y. Nishi, S. Kohyama, Y. Uchida, and K. Tamaru for encouragement throughout the work, and to their colleagues at Toshiba Semiconductor Device Engineering Laboratory for their cooperation.

References

1) T.Sakurai, T.Furuyama and T.Iizuka, "Methods for analyzing a word line delay and their applications", The community of solid state device of IECE of Japan, SSD82, P15, Oct. 1982.

2) M.Yoshimoto et al, "A 64Kb Full CMOS RAM with Divided Word Line Structure", ISSCC Dig. of Tech. Papers, P58, Feb. 1983.

3) T.Sakurai, "Approximation of wiring delay in MOSLSI", J. of Solid State Circuits, to be published.

4) T.Sakurai and K.Tamaru, "Simple formulas for two- and three-dimensional capacitances",IEEE ED-30, p183, Feb. 1983., FCAP2: Hewlett-Packard two-dimensional capacitance calculation program.

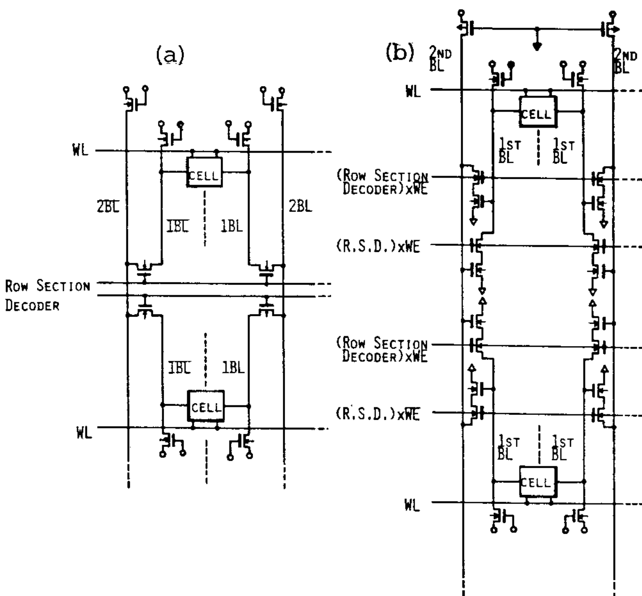

Fig.9 Switching circuits for DBL structure. (a) Drain-input structure. (b) Gate-input structure.

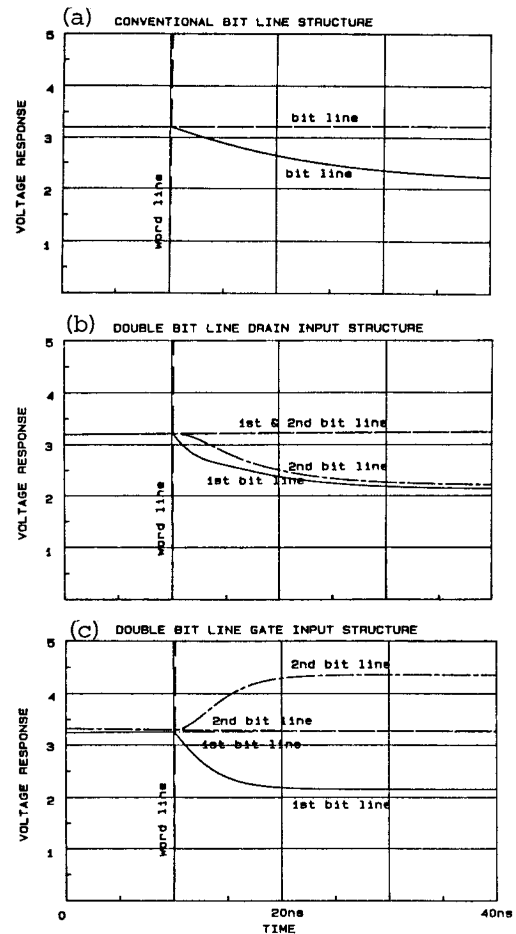

Fig.10 Bit line delay assuming 256Kbit SRAM. (a) Conventional structure. (b) DBL with drain-input structure. (c) DBL with gate-input structure which operates faster than drain-input structure. 2nd bit lines take 4ns to arise 0.5V voltage difference, compared to 9ns in conventional structure.

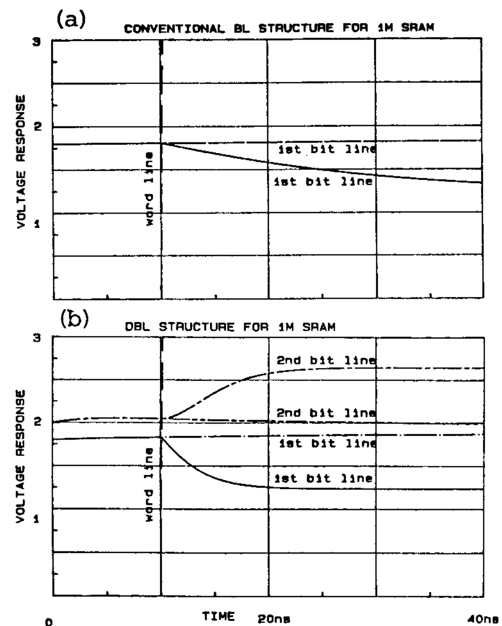

Fig.11 Bit line delay assuming 1Mbit SRAM. (a) Conventional structure. (b) DBL with gate-input structure.