4-4

Realistic Scaling Scenario for Sub-100nm Embedded SRAM Based on 3-Dimensional Interconnect Simulation

Yasumasa Tsukamoto, Tatsuya Kunikiyo*, Koji Nii, Hiroshi Makino, Shuhei Iwade,
Kiyoshi Ishikawa* and Yasuo Inoue*
System LSI Development Center, *ULSI Development center, Mitsubishi Electric Corporation
4-1 Mizuhara Itami Hyogo 664-8641, Japan
E-mail: tsukamoto.yasumasa@lsi.melco.co.jp

Abstract – It is still an open problem to elucidate the scaling merit of the embedded SRAM with the Low Operating Power (LOP) MOSFET's fabrication in 50, 70 and 100nm CMOS technology node. Taking into account the realistic SRAM cell layout, we evaluate the parasitic capacitance of Bit Line (BL) as well as Word Line (WL) in each generation. By means of 3-Dimensional (3D) interconnect simulator (Raphael), we focus on the scaling merit through the comparison of the simulated SRAM BL delay in each CMOS technology node. In this paper, we propose two kinds of original interconnect structures which add some modifications to ITRS, and clarify for the first time that the original interconnect structures guarantee the scaling merit of the SRAM cell fabricated with the LOP MOSFET's in 50 70 and 100nm CMOS technology node.

I. INTRODUCTION

Scaling studies on embedded SRAM and its parasitic capacitance due to interconnect are critical to CMOS technology node below 100nm. The problem of scaling merit is still open if the drain saturation current of Low Operating Power (LOP) MOSFET's keeps its value for sub-100nm CMOS technology node. The aim of this paper is to manifest the scaling merit of the embedded SRAM with LOP MOSFET's through the comparison of the simulated SRAM BL delay in 50, 70 and 100nm CMOS technology node. The SRAM BL delay is subjected to interconnect delay rather than gate delay. Therefore, the interconnect parasitic capacitance is a necessary ingredient to simulate the SRAM BL delay. Taking into account the realistic SRAM cell layout, we evaluate the parasitic capacitance of Bit Line (BL) as well as Word Line (WL) in the SRAM cell, using 3-Dimensional (3D) interconnect simulator (Raphael)[1-2]. In the interconnect capacitance simulation, we assume the original interconnect structures which add some modifications to ITRS (International Technology Roadmap for Semiconductors)[3].

The construction of this paper is as follows. In Sec. II, we precisely investigate the capacitance of BL and WL in the SRAM cell based on 100nm CMOS technology node. Then we show that the BL capacitance is proportional to the height of the plug connecting gate area (GA) to first metal-layer (M1). This result indicates the necessity of shrinking the vertical structure in the advanced technology node. In Sec. III, we introduce the scaled SRAM in 50 and 70nm CMOS technology nodes referring to the structure used in Sec. II, and evaluate the capacitance of BL and WL in each generation. Applying these simulated results to the circuit simulation, we propose for the first time the interconnect structure which guarantees the scaling merit of the SRAM cell with the LOP MOSFET's. Finally, in Sec. IV, we offer some conclusions.

II. INTERCONNECT CAPACITANCE OF THE SRAM CELL BASED ON 100NM CMOS TECHNOLOGY NODE

In this section, we investigate the BL and WL interconnect capacitance and show the importance of the vertical structure through the analysis of the SRAM cell based on 100nm technology node. Focusing on how the contact plug height affects the total BL capacitance, we make it clear that it is necessary to shrink the vertical structure, as well as to introduce new low-k dielectrics in the future technology node.

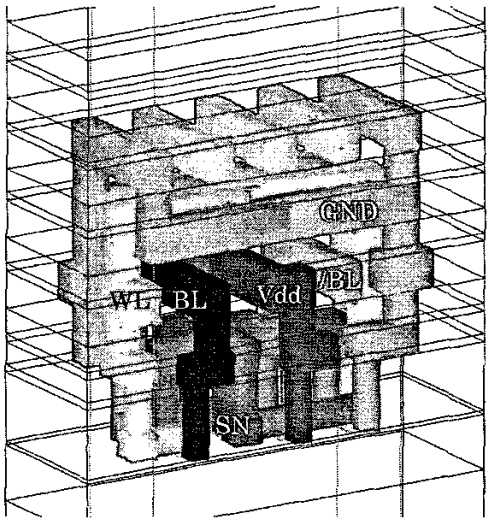

Fig.1. A bird's-eye view of 3D-SRAM memory cell structure based on 100nm CMOS technology node (unit cell). This structure is used for 3D-interconnect capacitance simulation [1-2].

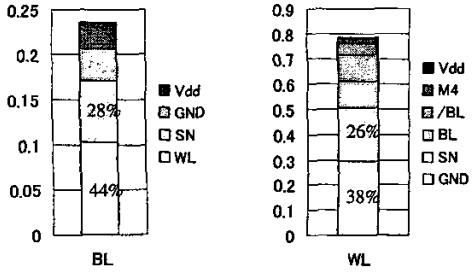

Fig.2. Simulated total capacitance per cell and the contribution rate of each component for BL and WL capacitance. The ordinate is capacitance in fF unit.

Figure 1 shows a bird's-eye view of a realistic 3D-SRAM cell structure for 3D-interconnect capacitance simulation. One of the most characteristic features of the present SRAM cell layout is that the BL is fabricated in second metal layer (M2) and the WL in third metal layer (M3), reducing the BL capacitance [4]. Information about the vertical structure is summarized in Table I.

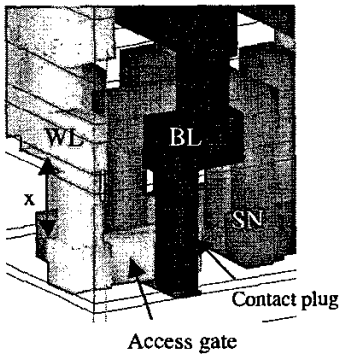

Fig.3. 3D graphics focusing on the contact plugs connecting GA (or AA) to M1. SN (Storage Node) expresses the node where the data ("H" or "L") is stored.

By means of 3-D interconnect simulator (Raphael), we calculate the capacitances of BL and WL. The simulated BL and WL capacitance is 0.237fF/cell and 0.792fF/cell, respectively. We also show the contribution rate of each component to the total capacitance in Fig. 2. The most dominant contribution in the total BL capacitance is from WL (44%), and that in total WL capacitance from ground interconnection (GND, 38%).

It is generally known that the higher operation speed of SRAM macro cell can be mainly achieved by the reduction of the BL capacitance. Therefore, we especially focus on the BL capacitance, taking into account the scaling merit discussed later. Figure 3 shows an enlargement of the SRAM cell around the BL structure. The contact plugs connecting gate (GA) or AA to first metal layer (M1) are fabricated in very fine pitch, so that the coupling capacitance between plugs connecting GA (or AA) to M1 is considered to be a major contributor to total BL capacitance. In order to examine the effect of the contact plug to the total BL capacitance quantitatively, we investigate the gate plug height (GA-M1 distance: shown as x in Fig. 3) dependences of the total BL capacitance (Figure 4). In Fig.4, solid line and dotted line indicate the structure having TEOS (k=4.2) and rather low-k material (k=3.4) for the interlayer dielectrics between GA and M1, respectively. The total BL capacitance has the linear dependence of the gate plug height of the SRAM cell. This result gives us some consequential features; the coupling capacitance between contact plugs is dominant over the total BL capacitance compared to the effect of the interconnection in the upper layer (M1), and the height of the vertical structure is also an important design matter for achieving high-speed SRAM cell. In other words, it is necessary to shrink the height of the memory cell in the future technology node.

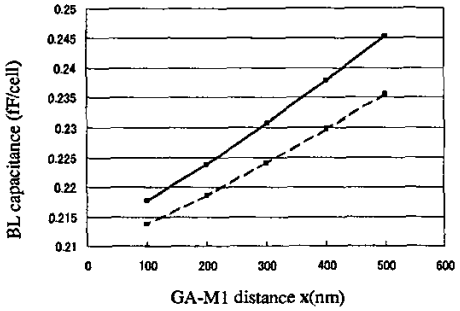

Fig.4. Plug height dependences of BL capacitance. Solid line and dotted line indicate the structure having TEOS (k=4.2) and "low-k" (k=3.4) material for interlayer dielectrics between GA and M1.

Table I. Interconnect structure and low operating power logic technology for 50, 70 and 100nm CMOS technology node.

Technology Node	0.1μm	0.07μm	0.05μm
Physical gate length(μm)	0.1	0.07	0.05
Saturation drive current:I_{dd}(μA/μm)	600	600	600
Power supply voltage:V_{dd}(V)	1	0.9	0.7
SRAM cell size(μm^2)	1.25	0.61	0.31
T_{ox} Equivalent (nm)	2	1.9	1.2
Gate height(μm)	0.16	0.12	0.08
GA-M1 distance(μm)	0.3	0.2	0.15
M1(M2-M5) L/S(μm)	0.12/0.12(0.14/0.14)	0.08/0.08(0.1/0.1)	0.06/0.06(0.08/0.08)
M1(M2-M5) height(μm)	0.192(0.238)	0.136(0.17)	0.108(0.144)
M1(M2-M5) A/R	1.6(1.7)	1.7(1.7)	1.8(1.8)
1T(Through hole) -4T height(μm)	0.21	0.16	0.128
1T-4T A/R	1.5	1.6	1.6

III. INVESTIGATION OF SCALING MERIT OF EMBEDDED SRAM IN 50 AND 70NM CMOS TECHNOLOGY NODE

In this section, we examine the BL delay of the scaled SRAM in 50 and 70nm CMOS technology node by employing the shrunk vertical structure as well as the shrunk layout of the SRAM cell shown in Fig. 1.

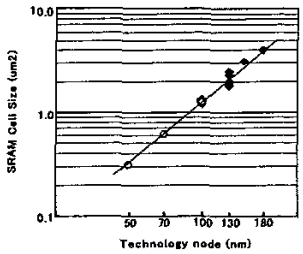

Fig.5. Scaling trend of the SRAM cell size. Dots represent cell sizes realized in the past, and circles those expected in the future, which we used for each simulation.

Figure 5 shows the cell size in each generation [5]. The logarithm of the cell size depends linearly on the feature size of the technology node. Therefore, in the present study, the cell size in 50 and 70nm technology node is determined in such a way as to extrapolating the present scaling trend. Table I shows our original scaling scenario, which adds some modifications to ITRS on interconnect structures. As shown in Fig.6, we assume two types of interconnect structures by referring to the results obtained in the previous section. In Structure A (Fig.6(b)), low-k dielectrics (low-k1) is used only between intralayer interconnect and TEOS is used for interlayer dielectrics. In Structure B (Fig.6 (b)), second low-k dielectrics (low-k2) replaces TEOS layer in Structure A. Low-k2 is assumed to be used for etch stopper for Cu damascene process. Based on Table I and Fig.6, we precisely estimate the capacitances of BL and WL in each generation. Figure 7 shows the total capacitance and its component in each generation. By applying our original structures, it is achieved that each capacitance reduces by approximately 30% compared to that of the previous technology node.

In order to estimate the scaling merit more precisely, we calculate the "BL delay time" of the SRAM cell applying the simulated BL and WL capacitance to the circuit simulation in each generation. Here we define BL delay time as shown in Figure 8, which is related to the SRAM read time. In the simulation, we deal with 512-row and 256-column memory cell array. Moreover, we assume that the transistor in each generation has the same saturation drive current (see Table I). Figure 9 shows the BL delay time at each CMOS technology node. This result indicates that the SRAM cell having Structure A keeps the scaling tendency except for 50nm CMOS technology node. Moreover, the reduction of the BL delay is accelerated by 4.6% in 70nm and by 11.6% in 50nm CMOS technology node for the SRAM cell having Structure B in comparison to Structure A. The simulation results in the present study suggest that the coupling capacitance between intralayer interconnect is dominant over the total BL capacitance and the low-k dielectrics having dielectric constant 1.2 is indispensable to achieve the scaling merit of the SRAM fabricated with the LOP MOSFET's.

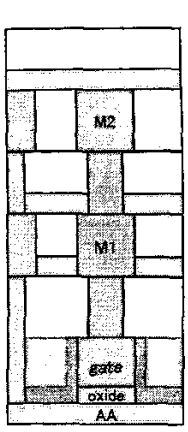

Fig.6-(a)

Structure A	Structure B
TEOS k=4.2	low-k2
SiC(k=4.8)	SiC(k=4.8)
low-k1	low-k1
TEOS k=4.2	low-k2
SiC(k=4.8)	SiC(k=4.8)
low-k1	low-k1
SiC(k=4.8)	SiC(k=4.8)
SiO_2 k=4.2	low-k2
<Side Wall> SiN(k=7.4)	<Side Wall> SiN(k=7.4)
SiO2	SiO2

Fig.6-(b)

	100nm	70nm	50nm
low-k1	2.8	2	1.2
low-k2	*	2.8	2

Fig.6-(c)

Fig.6. Schematic illustration of the interconnect structure (a), and low-k dielectrics in each generation suggested in this paper ((b) and (c)).

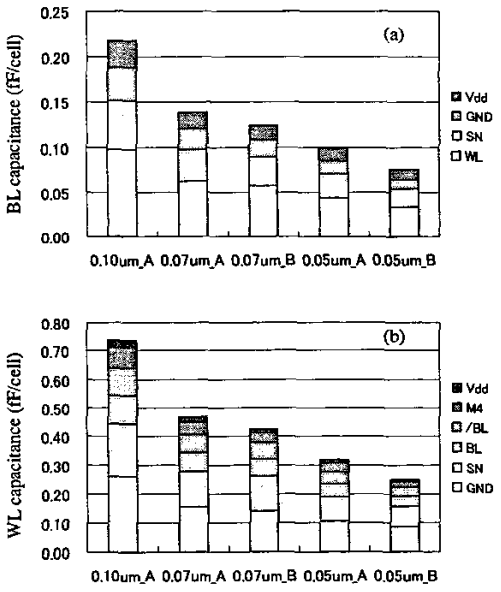

Fig.7. BL and WL capacitance vs CMOS technology node; (a) BL capacitance and (b) WL capacitance. Note that A and B represent the structure A and B explained in Fig.6, respectively.

Before concluding this section, we briefly mention the importance of the Gate Overlap Capacitance (GOC) in 50nm technology node. As shown in Fig.9, there is no scaling merit on the BL delay time compared to 70nm technology node with structure A. In order to estimate the contribution of the GOC over the BL delay time, the BL delay time is simulated by changing the transistor model parameters of circuit simulation in such a way as to reduce the GOC and keep the saturation current. Figure 10 shows the BL delay dependence of the GOC reduction with structure A. Reducing the GOC by 50% compared to that used for simulation shown in Fig.9, the BL delay time is almost the same as that of the structure B. Therefore, we conclude that GOC is also an important transistor-design matter to achieve high-speed SRAM in the 50nm technology node. Even though the scaling merit on the BL delay time is small in 50nm technology node, the scaling merit on power consumption may be large because the power supply voltage is reduced from 0.9 to 0.7V.

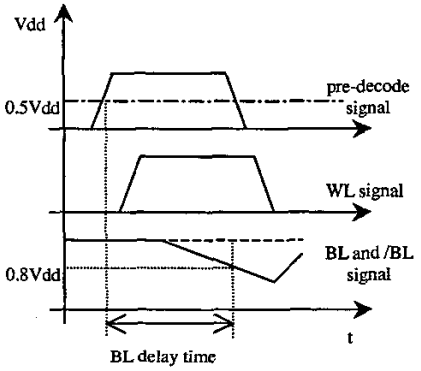

Fig.8. Read operation diagram. The pre-decoded signal activates WL signal. Then a memory cell discharges one or other of bit line pair BL or /BL. The BL delay time indicates an interval from the pre-decoded signal rising to a half Vdd to the BL voltage falling to 0.8Vdd. We calculate this BL delay time in each technology node.

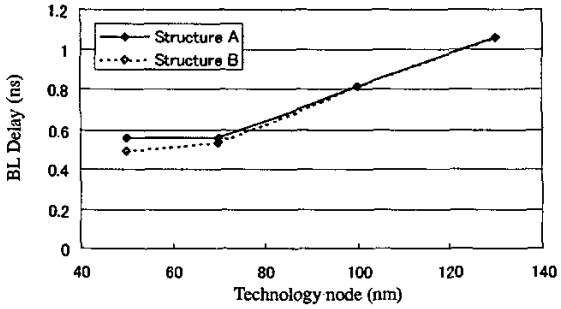

Fig.9. Simulated BL delay against each CMOS technology node. For the purpose of comparison, the simulated result in 130nm technology node is also plotted.

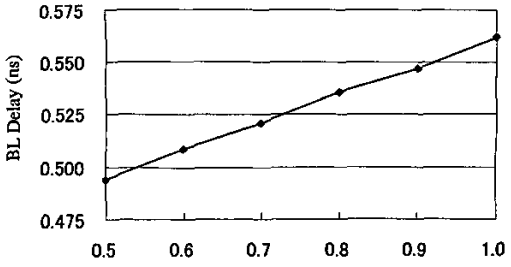

Fig.10. Gate Overlap Capacitance (GOC) dependence of BL delay in 50nm technology node with Structure A. Horizontal axis expresses the GOC ratio: GOC of transistor with Structure An over that employed in the simulation shown in Fig. 9. The BL delay time in Structure A is close to that in Structure B by reducing the GOC by 50%.

IV. CONCLUSIONS

The in-depth study of the scaling merit of the embedded SRAM cell has allowed us to unequivocally demonstrate that both low-k material between intralayer interconnects and the height of contact plugs connecting AA (or GA) to M1 play an important role to determine the BL delay of the SRAM cell. Showing the fact that the structure between GA and M1 affects the BL capacitance, it is found that BL capacitance depends on the plug height linearly. This result indicates that it is important to shrink the height of the vertical structure in order to reduce the BL capacitance. Moreover, by proposing two types of interconnect structures referring to the ITRS, we have for the first time investigated the SRAM cell based on 50, 70 and 100nm CMOS technology node, respectively. Precise simulation concludes that our original scenario for the interconnect structure as well as reduction of the gate overlap capacitance is indispensable for keeping the scaling trend of the SRAM cells fabricated with LOP MOSFET's.

ACKNOWLEDGMENT

The authors would like to thank to Tetsuya Watanabe, Tomoaki Yoshizawa, Susumu Imaoka and Yoshio Matsuda for their valuable discussion.

REFERENCES

[1] S. Yoon et al., Proc. of SISPAD-2000, p.94, Sep., 2000.
[2] http://www.avanticorp.com
[3] http://public.itrs.net/Files/2001ITRS/Home.htm
[4] K. Tomita et al., Symp. on VLSI Tech., p.14, 2002.
[5] K. Miyashita et al., Symp. on VLSI Tech., p.11, 2001.