A Folded Bit-Line Architecture For High-Speed CMOS SRAM

Sejun Kim Ilkwon Chang Seoungyoung Seo Kaedal Kwack
Semiconductor Lab, Department of Electronics.
Hanyang University
Hangdang-dong , Sungdong-gu ,Seoul 133-791
Korea

Abstract: - This paper describes a new architecture and schemes for high speed SRAM. It summarized as follows:1)a Folded Bit-Line Architecture(FBLA) to reduce the delay time of bit-line by decreasing the parastic capacitance, to reduce the area. 2) a Double Word-Line Activation(DWLA) technique to increase the data-rate twice and minimize row path delay, and 3) a High speed sensing scheme to decrease the delay time of sense amplifier. To verify above these, a 8kb SRAM was designed using 0.6μm technology. it realized a 600Mbyte/s(300M×8×2) data-rate and the die size is 2.8mm × 0.85mm.

Key words: - high-speed SRAM, Folded Bit-Line Architecture, Double Word-Line Activation,
High speed sensing scheme, Data-rate, area, Burst length *IMACS/IEEE CSCC'99 Proceedings*, Pages:5041-5046

1 Introduction

In recent years, the speed of the microprocessor has been increased very high. Therefore, it is very urgent problem for on-chip cache SRAM's to keep up with the increased speed of microprocessors. For the while, the studies to reduce the access time have been reported. On one approach, many critical blocks such as decoder, sense amplifier, output buffer, and etc. have been improved.[1,4] On the other approach, many cache SRAM's adopting variable pipeline schemes have been proposed.[3] As a result, recently, the access time has been improved to 1.8ns ~ 2.5ns.[2,5]. However, as memory capacity goes large, the improvement of speed in these schemes induced the excessively increased area. To overcome this limitation, we propose a new architecture and schemes for high-speed SRAM.

2 Folded Bit-line Architecture

Fig.1 shows a block diagram of typical CMOS SRAM. The access time for the SRAM consists of the sum of the delay times of the four main circuits, i.e., address buffer, decoder, memory-cell array, and sense amplifier circuits. Among these circuits, the delay times of decoder, memory-cell array, and sense amplifier circuits occupy very largely in total delay time. Fig.2(a) shows typical bit-line architecture. In the memory cell array, the parastic bit line capacitance is very large. As memory capacity is increased, the data in cell cannot be transfered speedily to sense amplifier owing to much a larger bit-line capacitance. Therefore, today, almost all large-scale SRAM's have multi-memory cell arrays as shown in Fig.2(b). But it increase the area of SRAM largely. Fig.3 shows our proposed memory cell array with Folded Bit-Line Architecture(FBLA). Because the number of pass transistor in the cell causing the parastic bit- line capacitance is decreased by half, FBLA reduces the parastic capacitance by half. Therefore, FBLA transfers larger voltage swing to sense amplifier as shown in Fig.4. Eventually, the access time can be reduced. Also, SRAM with FBLA occupies much smaller area than conventional SRAM having same memory capacity with multi cell arrays. Fig.5 shows the comparison of area according to memory capacity.

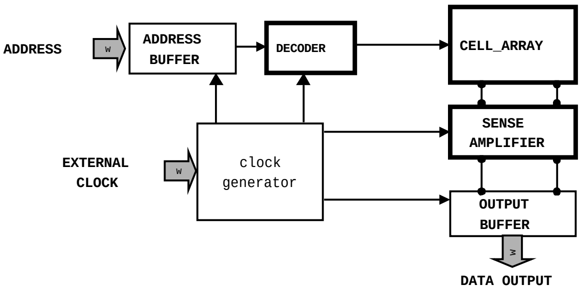


Fig.1 Block diagram of a typical CMOS SRAM

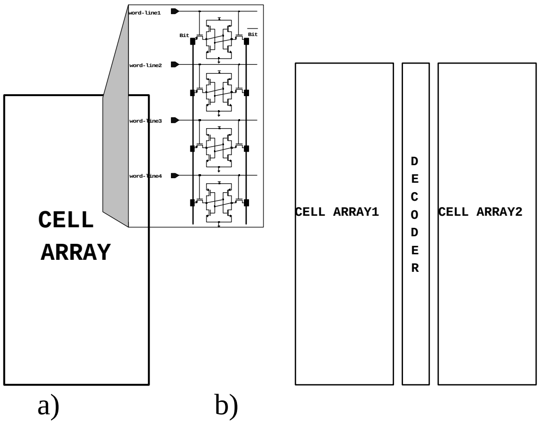


Fig.2 Typical bit-line and multi-memory cell arrays

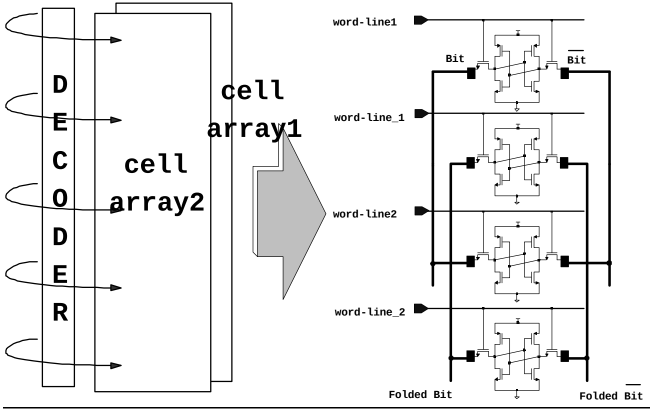


Fig.3 Folded Bit-Line Architecture

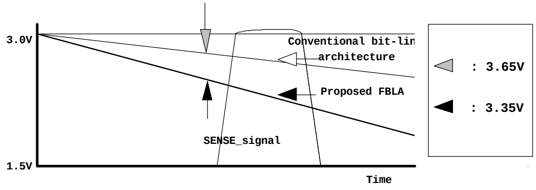


Fig.4 bit-line voltage Swing

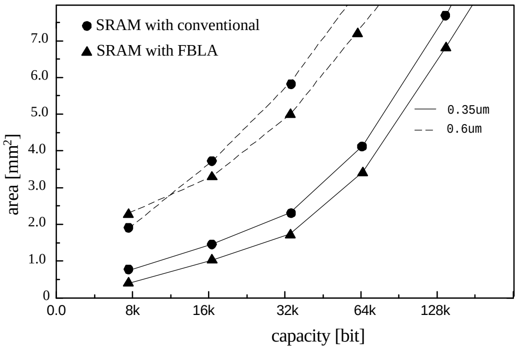


Fig.5 the comparison of area

3 High speed schemes

3.1 Double word-line activation scheme

Above all, FBLA has another important advantage. This structure has two separated bit-lines as shown Fig.3. Therefore, it enables two word-lines to be driven almost simultaneously. In the end, this increases a data-rate twice. Fig 6 shows Double Word-Line Activation(DWLA)scheme.The operation of DWLA is as follows. At rising edge of clock, The 'word_drv' signal is asserted to word-line driver and it drives first word-line. And, at falling edge of same clock, it drives second word-line through the delay chain with exact delay time. Fig.7 shows the comparison of proposed SRAM and conventional burst pipelined SRAM. The other advantage is that it simplifies the decoder. Generally, the decoder is composed of logic gate, such as, NAND NOR.. the delay time of decoder accounts for about 40~50% of the overall access time.[2] In recent years, the decoder has multistage structure for minimize gate delay. But it is very complex and increases the number of logic gate. But, as our proposed DWLA scheme drives two the different word-line almost simultaneously, the necessary row address is reduced by half. And, the reduced row address simplifies the structure of decoder. Eventually, In comparison to the decoder in the SRAM having same capacity, It has an effect that reduces the delay time of decoder.

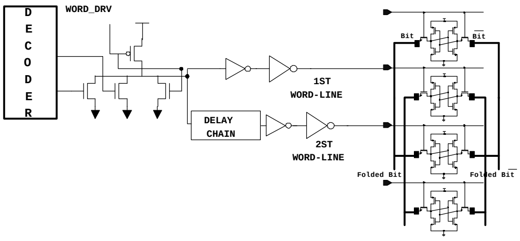
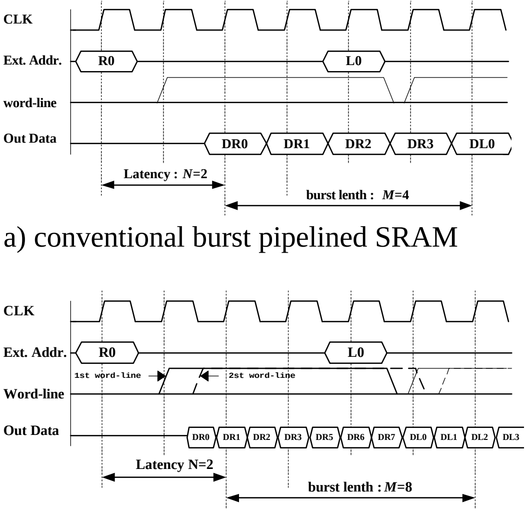


Fig.6 Double Word-Line Activation scheme



a) conventional burst pipelined SRAM

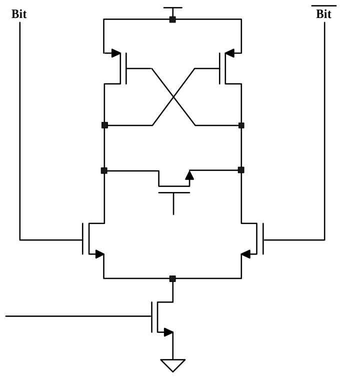
b) proposed burst pipelined SRAM

Fig.7 The comparison of time diagram

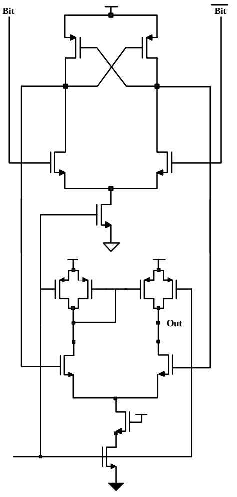
3.2 High speed sensing scheme

Generally, the delay time of sense amplifier account for about 20~30% of the overall access time.[2] Until now, conventional SRAM has used sense amplifier as shown Fig.8(a). To reduce the delay time of sense amplifier, the multi-stage sense amplifier in Fig.8(b) was proposed. But it occupies the large area and power dissipation. Fig.8(c) shows our the proposed

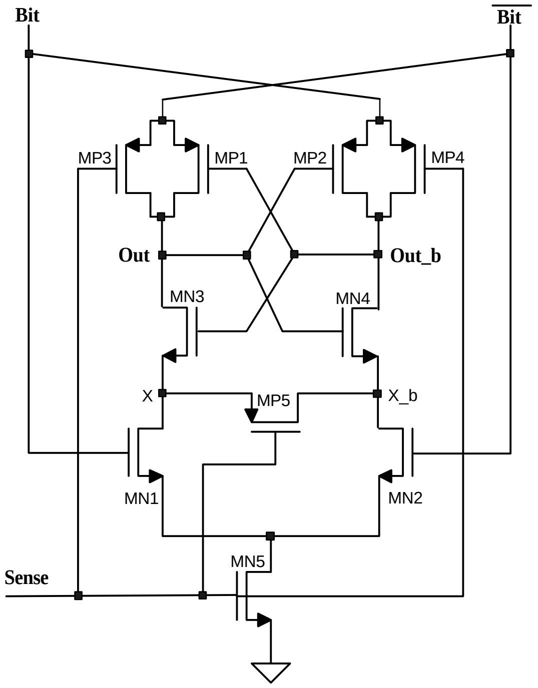
Multi Point Sense Amplifier(MPSA). MPSA simply consists of CMOS-coupled pair, equalizing transistors and sensing input transistors. The operation of MPSA is as follows. Before the Sense signal is asserted, the MP3 and MP4 transistor transfer the bit-line voltage difference from both bit lines to the nodes out and out_b, which causes the splitting voltage level on output nodes. And X and X_b voltage is equalized by MP5. This equalizing scheme accelerates the sensing speed. As the sense signal goes from the low level to high level, the current that is flowing through MN5 causes lowering the voltage on the drain node of MN5. As the voltage difference between gate and source of MN1 or MN2 is large enough to turn on the transistor, the voltage difference of bit-lines is amplified. The remarkable feature of MPSA is that bit-line voltage difference transfered in advance reduces the delay time of sense amplifier with relatively simple structure. Fig.9 shows the sensing delay of conventional sense amplifier and proposed sense amplifier.



a)conventional sense amplifier



b)multi-stage sense amplifier



c) proposed multi point sense amlifier

Fig.8 Sense amplifier

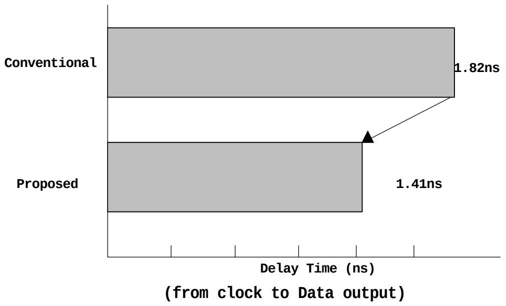


Fig.9 Delay of sense amplifiers

4 Simulation Results

To evaluate these new schemes, we designed a 8kb(128word × 64 column) burst pipelined SRAM macro on 0.6-μm CMOS technology. Fig.10 shows the simulation results of both read and write operation. The supply voltage is 3v. The clock frequency is 300MHz. The access time is 1.41ns. The clock latency is 2. And the burst length is 8. Results show that our SRAM has the same performance as operating at 600MHz. The main features of the SRAM are listed in Table 1. Fig.11 shows the layout of SRAM.

Table 1 Features of the SRAM

Technology	0.6 μm 2 poly 3 metal
Ext. Clock	300MHz
Clock access time	1.41ns
Supply voltage	3V
Capacity	8kb(128word×64 column)
Die size	2.8mm×0.85 mm

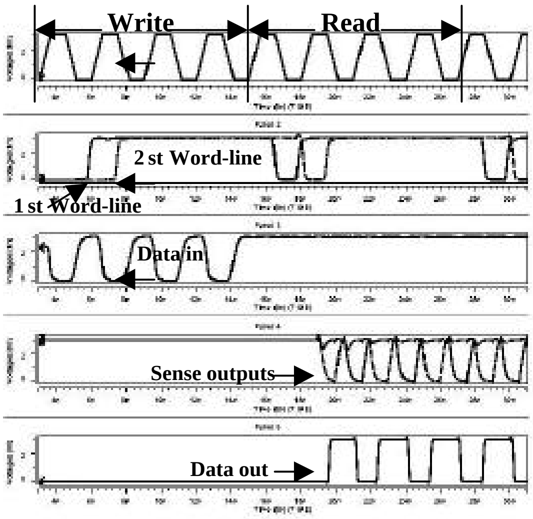


Fig.10 Simulation results

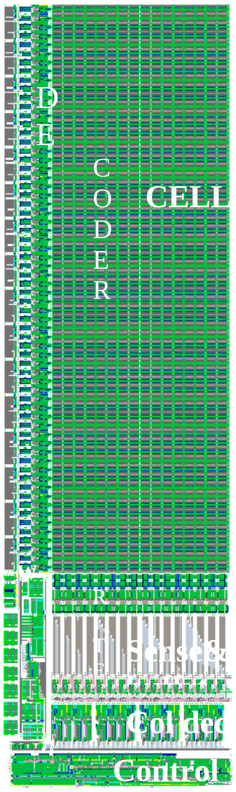


Fig.11 layout of SRAM

5 Conclusion

For high-speed operation, we adopted the effective architecture and schemes as follows: 1)Folded Bit-line Architecture, 2)Double Word-Line Activation, 3)High speed sensing scheme. On these architecture and schemes, We designed 8kb SRAM using 0.6µm technology. Our SRAM showed the very excellent performance. Our proposed architecture and schemes will be a scalable solution for high speed SRAM. Specially, Folded Bit-Line Architecture and Double Word-Line Activation scheme are expected to solve the many problems of the large-scale SRAM.

Reference:

[1] Tsugro kobayashi, et al.., “A Current-mode Latch Sense Amplifier and a static Power Saving Input Buffer for Low-Power Architecture”,IEEE J.Solid-State Circuits, vol 28, pp. 523-527, Apr.1993

[2]H. Nambu, et al.., “A 1.8ns Access, 550Mhz 4.5Mb CMOS SRAM”,ISSCC , pp 360-361 Feb.1998

[3]Hoi-jun Yoo“A Study of Pipeline Architectures for High-speed Synchronous DRAM’s”,IEEE J.Solid-State Circuits, pp. 1597-1603,Oct 1997

[4]Harold Pilo, et al.., “A 300Mhz, 3.3v 1Mb SRAM Fabricated in a 0.5µm CMOS Process”, ISSCC ,pp 148- 149, 1996

[5]Ken-ichi,et al..,A 2ns Access,285Mhz,Two-Port Cache Macro using Double Global Bit-Line Pairs”,ISSCC,pp.402,1997