LOW-POWER DUAL-PORT ASYNCHRONOUS CMOS SRAM DESIGN TECHNIQUES

Tan Soon-Hwei, Loh Poh-Yee, Mohd-Shahiman Sulaiman, Zubaida Yusoff

Multimedia University, Cyberjaya, Malaysia

Key words: SRAM, Low-Power, CMOS, Dual-Port, Asynchronous, Non-volatile

Abstract: This paper describes the review and short tutorial on design techniques for low-power SRAM, focusing on the design of a 1-Mb CMOS SRAM on CMOS 0.25-μm process. The building blocks of the SRAM are individually discussed and various techniques are described, with the most appropriate one chosen for the block. SRAM power saving techniques are also described and implemented in the 1-Mb memory. The designed SRAM is simulated across different Process, Voltage, and Temperature (PVT) corners under the presence of parasitics. The performance of the 1-Mb SRAM is then compared with that of the previously published work. It is found that a minimum read access time of 4.26ns is achieved. The SRAM can operate at maximum frequency of 220MHz in dual-port mode and dissipates minimum active power of 31mW and is able to retain data at 0.1V supply voltage and consumes a standby power of 80nW. The SRAM occupies an area of 115mm^2.

Tehnike načrtovanja asinhronih dvovhodnih CMOS SRAM vezij z nizko porabo

Kjučne besede: SRAM, majhna poraba, CMOS, dvovhodni, asinhroni

Izvleček: V prispevku podamo pregled tehnik načrtovanja asinhronih dvovhodnih CMOS SRAM vezij z nizko porabo s poudarkom na načrtovanju 1-Mb CMOS SRAM vezja v 0,25um CMOS tehnologiji. Vsakega posebej opišemo sestavne bloke vezja SRAM, kakor tudi najbolj primerno tehniko načrtovanja. Ravno tako obravnavamo tehnike za zniževanje porabe in opišemo konkretni primer pri 1-Mb vezju. Načrtano vezje SRAM simuliramo pri različnih parametrih procesa ter vrednostih napetosti in temperature (PVT) ob prisotnosti parazitnih dejavnikov. Te rezultate primerjamo z drugimi predhodno objavljenimi rezultati. Ugotovimo, da dosežemo minimalni bralni čas 4.26ns. Vezje SRAM lahko deluje z največjo frekvenco 220MHz v dvovhodnem načinu, pri čemer porabi najmanj 31mW ter je zmožno ohraniti podatke tudi pri napajalni napetosti 0.1V. Poraba moči v stanju pripravljenosti je 80nW. Površina vezja je 115mm^2.

1 Introduction

Large portion of modern digital chips are occupied by memory and its capacity is forecasted to further increase in the new era of System on Chip (SoC). Hence high density while maintaining high-speed memory design is urgently needed by the semiconductor industry especially due to a great demand for cache applications in very fast processors. Concurrently, VLSI circuit designers also have to take power consumption problem into consideration due to the increased integration and operating frequency. In addition, portable equipment such as laptop computers, PDAs and cellular phones are more widely used nowadays and this raises the importance of low power design for longer battery operation /1/,/2/. This paper aims at exploring and implementing high speed and power savings techniques into memory design to overcome speed degradation and high power dissipation issues caused by large memory capacity.

This paper describes a 1-Mb SRAM with 64K words x 16-bit organization. The SRAM operates properly with the supply voltage of 1.5V in order to support portable equipment running on 1.5V batteries. A minimum 4.6ns access time and 31mW active power have been achieved by Hierarchical Word Decoding architecture, current-mode technique and pulse-mode technique. The proposed SRAM has four operation modes - active read mode, active write mode, active dual-port mode and standby mode.

Section 2 of this paper discusses circuit design and circuit techniques used for the SRAM. Characteristics of the designed SRAM and brief performance comparisons with other published works are presented in Section 3. This paper is concluded in Section 4.

2 Circuit Design & Techniques

2.1 Chip Architecture & Hierarchical Word Decoding

HWD architecture reduces latency by accessing smaller memory blocks. The principle in this technique is to partition memory array into several portions and to map these portions to different physical memory banks that can be selected or deselected independently. Also, it reduces power dissipation by shutting down portions that are not accessed potentially /3/.

The overall block diagram of the SRAM is illustrated in Figure 1. A 1-Mb memory array is partitioned into four quadrants and each quadrant contains eight 32Kb local blocks. Each local block contains 256 rows and 128 columns of memory cells. Row select word-line has a three-level hierarchical structure, which are global word-lines, sub-global word-lines and local word-lines. All quadrants are connected to an I/O block that contains data, signal and address input buffers, data output multiplexers and data output buffers. Tapered buffers with large current driving capabil-

ity are used. Control block contains Address Transition Detection (ATD) circuit and global control circuitry. The pre-decoder, global row decoder, sub-global row decoder and local row decoder works together to generate row select signals for the local blocks. Global row decoder contains a 2-to-4 decoder that selects one out of the four quadrants. Sub-global row decoder contains a 3-to-8 decoder that selects one out of the eight local blocks. Local row decoder contains buffers that drive the large word-line capacitive load. A pre-decoding stage is used to reduce transistors count and decoding delay. Two sets of decoder were placed to support dual-port design.

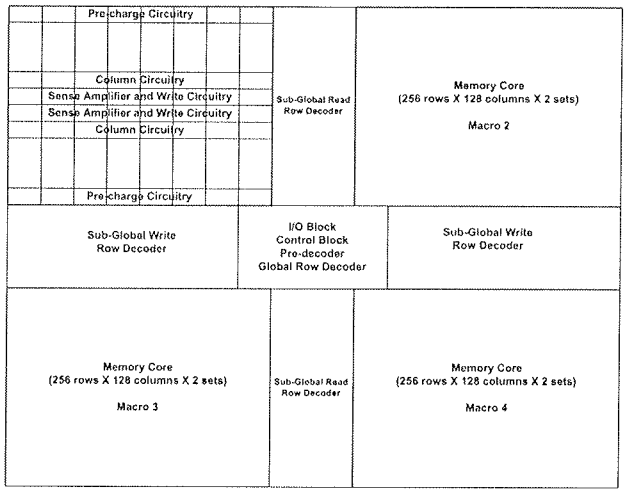


Fig. 1: Overall Block Diagram of 1Mb SRAM

Row select signals are distributed to 32 local blocks. These signals can be ensured to arrive at different blocks with similar delay by using Balanced Path H- Tree or X-Tree Signaling technique. H-Tree signaling technique was chosen for this paper. It is particularly useful for regular array networks as the loading of each node leaf is equal.

Figure 2 shows the HWD decoding stages for read and write decoders. Read/write enable signal is ANDed with ATD signal at the very beginning of decoding stage to avoid instantiating internal operation when SRAM is disabled and thus save power. The SRAM is also implemented advanced read function implicitly. When read memory instruction with same address is assigned to the SRAM subsequently, the ATD circuit stops generating control pulse and thus no internal operation will occur. Instead, previous data outputs are used and this contributes to more power savings.

2.2 Current-mode Sensing & Writing Technique

Voltage-mode sense amplifier senses differential voltage in the bit-lines and convert it to full-rail swing output. Voltage-mode write circuitry pulls down bit-lines voltage to certain voltage level that can overpower the memory cell. These voltage-mode circuitries can cause large bit-line voltage swing and hence power dissipation problem. Also, speed degradation is another problem as the discharging delay for pulling down bit-lines voltage is long especially when the bit-line capacitance is large. Other than that, a long period is required to restore bit-lines voltage after every operation.

To avoid these problems, current-mode technique is implemented for the SRAM in this paper. It was first proposed by /4/. The main idea is to use a low-resistance current-signal circuit to reduce voltage swing on the long transmission line and to avoid speed degradation problem caused by signal delay in the high capacitive interconnect. Delay of current-mode circuitry is insensitive to bit-line and data-

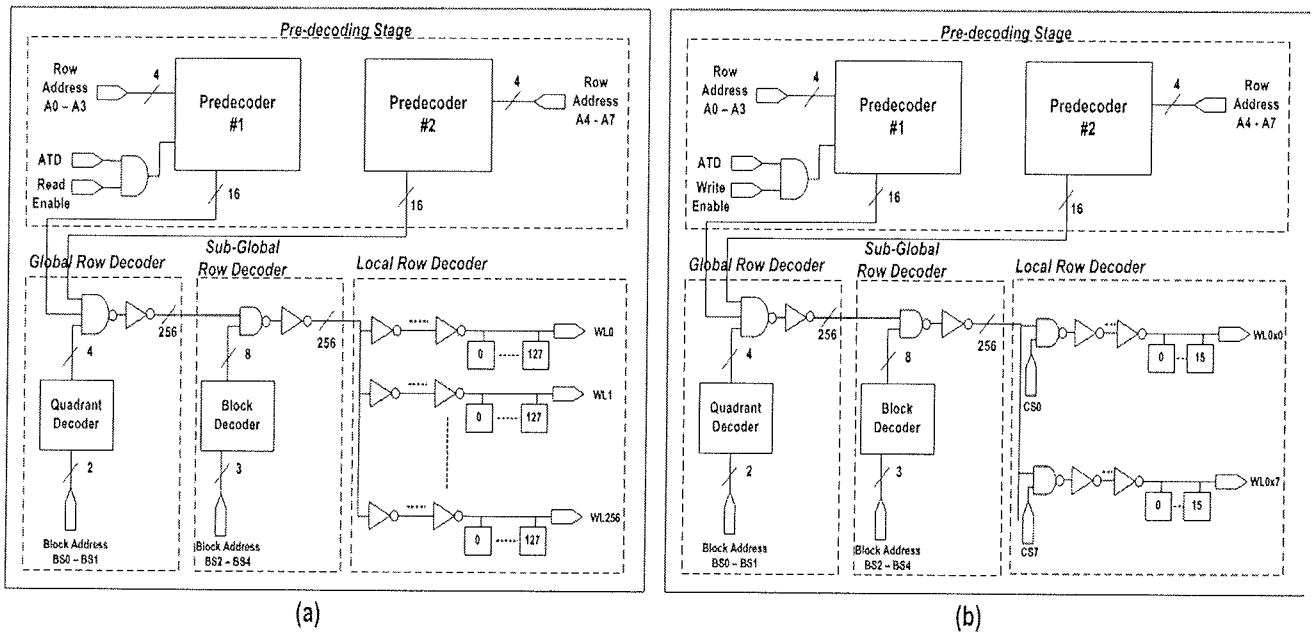


Fig. 2: (a) HWD Read Decoding Stages (b) HWD Write Decoding Stages

line capacitance /3/,/5/. In addition, it helps to reduce coupling noise to the adjacent bit-lines as bit-lines voltage swing is minimized.

Figure 3 shows a block diagram that describes the internal structure of the local 32Kb memory block that implements current-mode technique. Two sets of local row decoder and column decoder have been placed to support the dual-port feature. Column selector and current conveyor pass data in electrical forms to the local output circuitry. Column selector determines which column is selected for operation. Current conveyor acts like a current signals transmitter and is part of the current-mode circuitry. Local data output circuitry has a current-mode sense amplifier, output latch and tapered buffers. Local data input circuitry has a current-mode write circuitry and tapered buffers. Write circuitry translates input data into recognized electrical signals to modify memory cell's data.

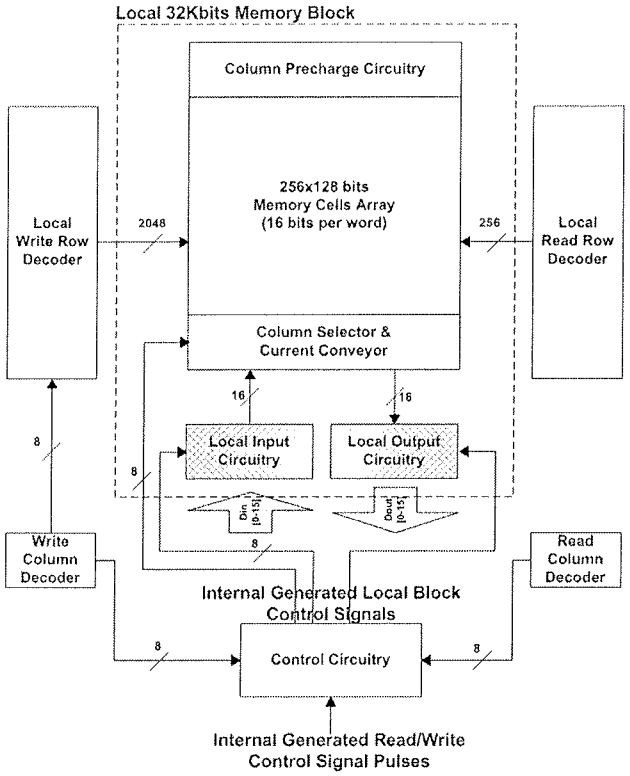


Fig. 3: Block Diagram of Local Memory Block

Figure 4 shows the bit-line peripheral circuits. It includes 7T memory cell, per-charge circuitry, current conveyor and column selector. 7T memory cell is used to implement current-mode writing technique. An additional equalization transistor is added and it forms a write port whereas the access transistors form a read port. The Source terminals of the PMOS transistors are connected to two common voltage lines that are called write bit-lines. These bit-lines connect pre-charge circuitry and other memory cells of the same column. PMOS-based pre-charge circuitry is used and is biased in linear region. No external control signal is supplied to the pre-charge circuitry as bit-lines capacitance is not needed to be discharged. Two sets of pre-charge circuitry are placed to support dual-port design. A two-input two-output PMOS-based current conveyor based on Caprio's bipolar cross-coupled quad circuit is used to implement current-mode technique /4/. This current conveyor forms a virtual short-circuit to the bit-lines and transports currents to the inputs of Clamped Data Line Sense Amplifier. In addition, it also serves as a column selector circuit. Current conveyor consists of four PMOSs and all transistors must be of equal size to create similar voltage level at input nodes of the circuit. Also, it contains an additional equalization transistor to solve the pattern-dependant problem /5/. Column equalization signal enable this transistor after every read operation. All the PMOS transistors in the current conveyor circuit operate in saturation region.

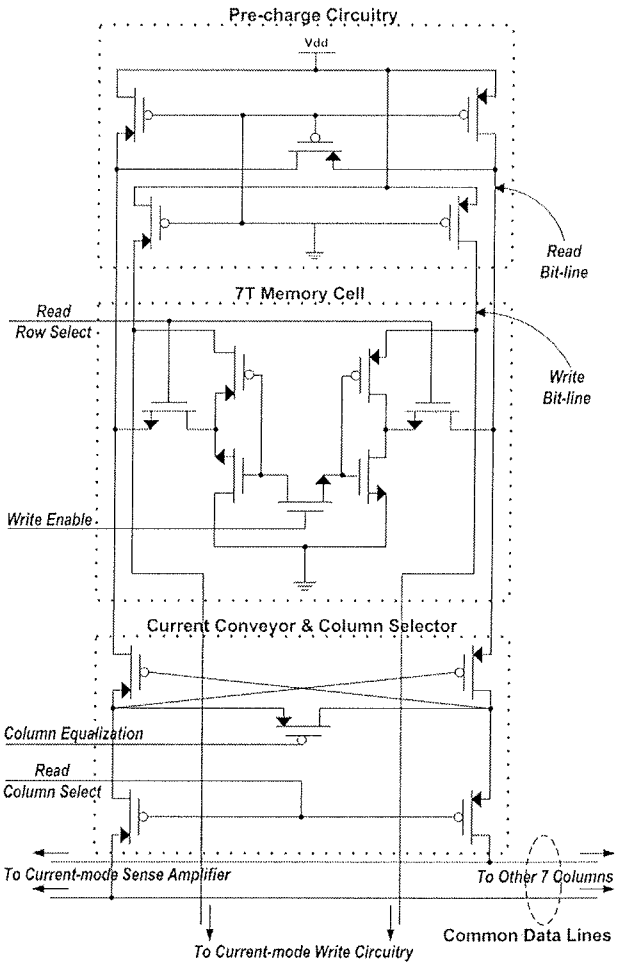


Fig. 4: Bit-line Peripheral Circuits

In general, current-mode sensing circuit is composed of current transporting circuit and current-to-voltage converter. A current transporting circuit or current conveyor is required to have two characteristics - low input resistance and unity current gain. Its main function is to transport differential currents from bit-lines to common data-lines when a memory column is selected. A current-to-voltage converter senses the differential input currents and convert these differential signals to a full swing CMOS voltage. Clamped Data Line Sense Amplifier (CDLSA) circuit topol-

ogy is chosen in this paper /6/. Figure 5 shows the transistor-level circuit of CDLSA. It clamps the data-lines using two NMOS transistors (M5 & M6) that are tied to ground and forces it to a voltage level closed to ground to provide virtual short-circuit feature. It also contains a cross-coupled inverter M1-M4 that provides complementary outputs. It has an equalization NMOS, Meq, that drives the cross-coupled inverter to meta-stable state during equalization phase.

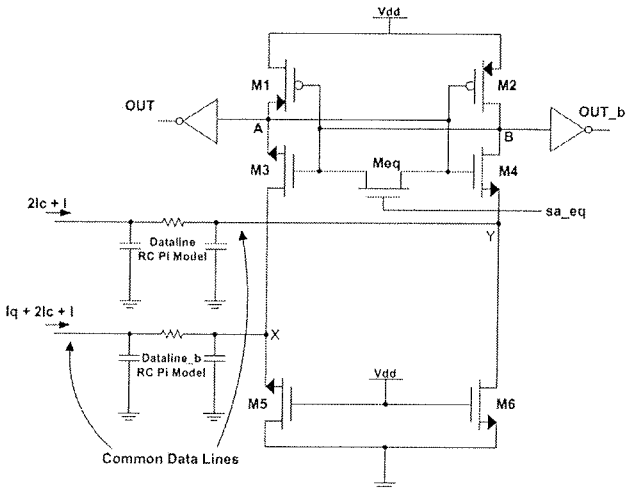


Fig. 5: Clamped Data Line Sense Amplifier

The circuit operates in two phases: equalization phase and sensing phase. During equalization phase, sense amplifier equalization signal (sa_eq) is pulled high to equalize voltage at node A and node B. The signal is disabled to start the sensing phase after sufficient differential currents are built up. M3 and M4 sources the differential currents into sense amplifier output nodes. Then, the small parasitic capacitance at output nodes are charged to either higher or lower voltage level depends on the differential currents flow in. Cross-coupled inverter forms a positive feedback amplifier to amplify the differential nodes voltage to corresponding CMOS voltage level.

The SRAM implements current-mode writing technique that was proposed by /7/. The proposed technique has a limitation where size of the row of a memory block must equal to SRAM word size. In this paper, row decoder and local control circuit have been redesigned so that the write circuitry can perform operation to the chosen columns only. The column select signal is ANDed with row select signal before reaching memory cell's write port. Figure 6 shows the transistor-level circuit of current-mode write circuitry and its circuit operation. Current-mode write operation can be achieved by loading differential currents into storage nodes of memory cell through the write bit-lines. This circuit operates in two phases: equalization phase and evaluation phase. Before equalization phase begins, write enable signal arrives earlier and either M8 or M9 is turned ON. If M9 is turned ON, current flows through M9 to ground and causes current flowing into the storage node becomes lesser. During equalization phase, write equalization signal turns ON M7 and the storage nodes are forced to a voltage level closed to V_{dd} /2. As M3 and M5 sources different currents, nodes capacitance is charged to different voltage levels. During evaluation phase, M7 is disabled and the differential voltage between two nodes is amplified to full swing CMOS voltage level. M10-M13 are placed as shown in Figure 6 to pass both high and low input voltages effectively and hence M8-M9 can be biased to operate in either saturation region or cut-off region.

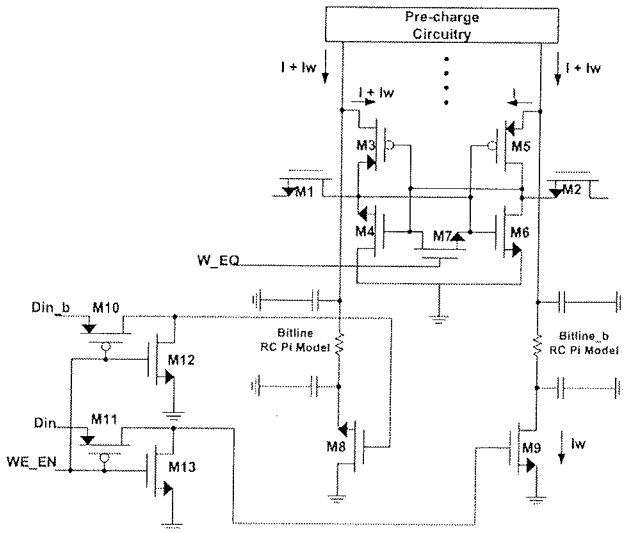


Fig. 6: Current-mode Write Circuitry

2.3 Pulse-mode Technique & ATD Circuit

Pulse-mode technique is one of the two most commonly used techniques in reducing dc current /8/. When external clock signal is used for generating internal control signals, the circuit is turned on every cycle yet no operation is requested. Besides that, the clock cycle period might be too long for an operation to complete. As a result, asynchronous SRAM design using Address Transition Detection (ATD) circuitry is invented to reduce power consumption by supplying pulse signals to the internal circuitries.

A simple ATD circuit together with the ATD pulse signal waveform is shown in Figure 7. The main function of ATD is to generate a pulse whenever a transition of address signal is triggered. Duration of the ATD pulse is determined by the delay element in ATD circuit. The delay element circuit is designed such that pulse duration will be just enough for an SRAM operation to complete. In order to generate pulses of similar pulse width under various conditions, modified Schmitt Trigger delay circuit is used as the delay element. It uses capacitors to control the delay time. This design can generate different delay by altering the size of capacitor /9/. The transistor-level circuit of delay circuit is shown in Figure 8.

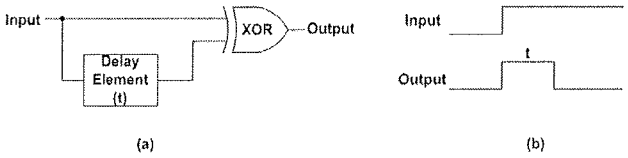


Fig. 7: ATD Circuit and Output Waveform

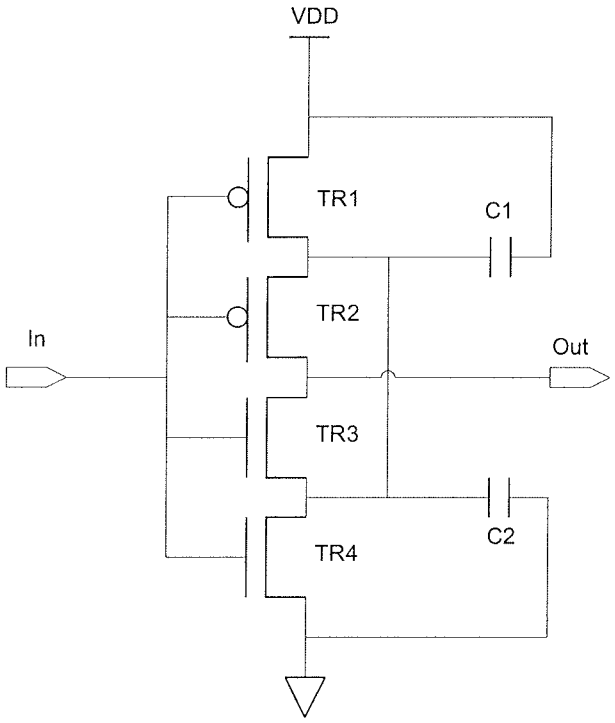


Fig. 8: Modified Schmitt Trigger Delay Circuit

Pseudo-NMOS is implemented to perform NOR operation to the 17 inputs (Refer to Figure 9). Pseudo-NMOS also avoids the series connection of PMOS transistors and reduces transistor count. However, this design suffers from reduced noise margin and high static power dissipation.

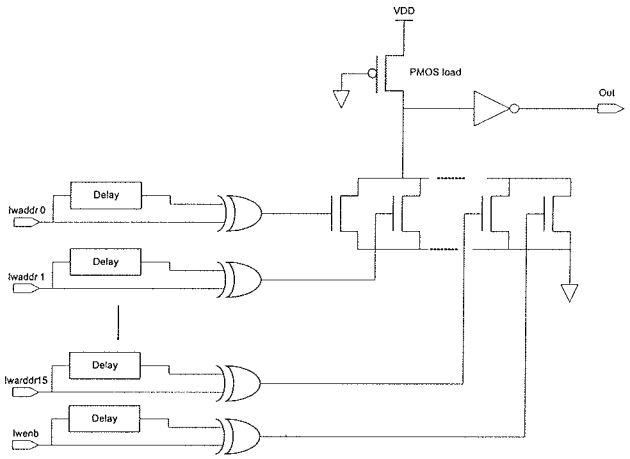


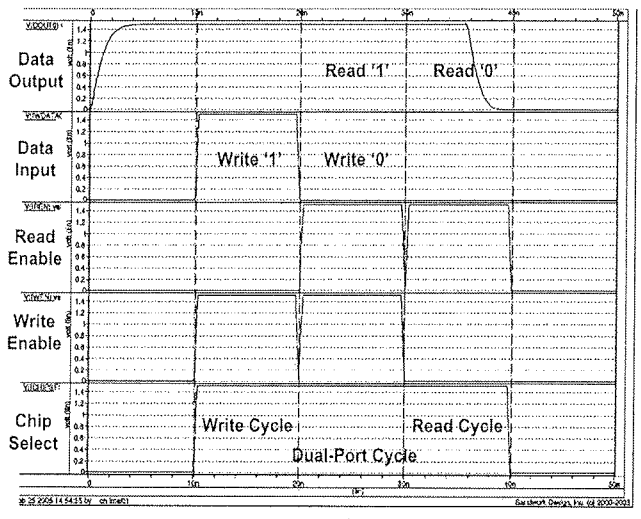
Fig. 9: Pseudo NMOS OR gate ATD

3 RAM performance

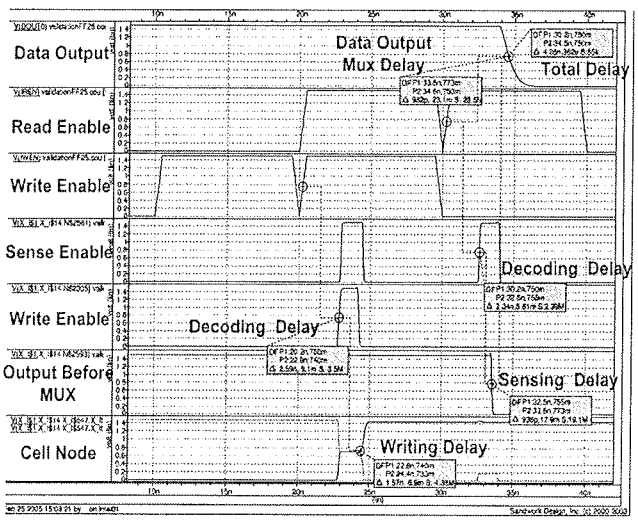
The SRAM was validated under various process, voltage and temperature (PVT) corners as listed in Table 1. Stimulus was setup so that a write operation was performed first followed by a dual-port operation and then a read operation.

Table 1: Simulation Conditions for 1Mb SRAM

Parameter	*Range*
Supply Voltage	1.4V - 1.6V
Temperature	25°C - 85°C
Process Corner	TT,FF,FS,SF,SS
Input Signal Slope (10% - 90%)	80ps - 400ps



(a)



(b)

Fig. 10: (a) Simulation Waveform of 1Mb SRAM (b) Delay Measurement

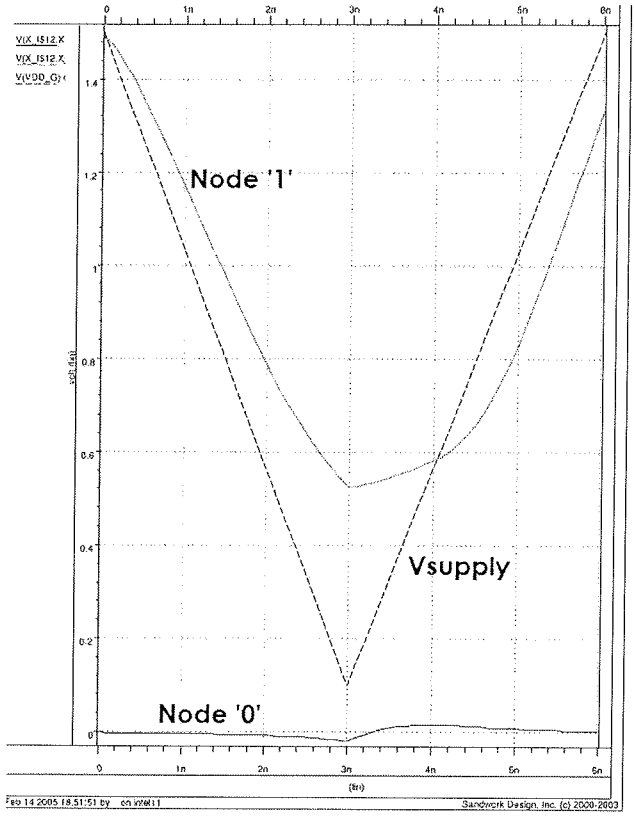


Fig. 11: Data Retention Analysis

Simulation waveform is shown in Figure 10(a) together with the minimum delay measurement shown in Figure 10b. Minimum read access delay, decoding delay, sensing delay and output MUX delay were measured as 4.26ns, 2.34ns, 0.93ns and 0.98ns, respectively. The maximum operating frequency is 220MHz@FF 25°C, 1.5V, and the typical operating frequency is 160MHz@TT, 1.5V. In contrast to the voltage-mode SRAMs, the designed SRAM has omitted two delay stages. They are bit-line/circuit recovery delay and discharging delay. Minimum average active current for read cycle is 13.35mA, write cycle is 13.52mA and dual-port cycle is 20.75mA.

Data retention analysis was performed and the simulation waveform is shown in Figure 11. It is used to measure the minimum supply voltage for which the designed memory cell can still retain the stored data. From simulations, it is found that the memory cell is able to retain data even the supply voltage is ramped from 1.5V to 0.1V across all skews with varied temperature. Minimum standby current at 0.1V supply voltage is 0.8µA. Therefore, standby power can be reduced significantly by lowering the supply voltage during suspend mode. However, additional voltage down converter (VDC) circuit is needed to downgrade the supply voltage to 0.1V.

Compared to other published papers on SRAM design, this project is projected to achieve smaller read access delay and lower active power. The read access delay comparison is shown in Table 2 and active power consumption comparison is shown in Table 3. The major characteristics of the SRAM are summarized in Table 4.

Table 2: Read Access Delay Comparison

	This Project	*/9/*	*/10/*	*/11/*
CMOS Process	0.25µm	0.3µm	0.35µm	0.5µm
Vsupply	1.5V	3V	3V	3.3V
Sensing Technique	Current Mode	Current Mode	Current Mode	Voltage Mode
Memory Size	1Mb	1Mb	256Kb	4Mb
Read Access Delay	4.26ns	7ns	9ns	25ns

Table 3: Active Power Consumption Comparison

	This Project	*/9/*	*/12/*	*/10/*
CMOS Process	0.25µm	0.3µm	0.5µm	0.35µm
Vsupply	1.5V	3V	5V	3V
Operating Freq(MHz)	150	100	100	100
Memory Size	1Mb	1Mb	1Mb	256Kb
Active Power	31.13mW (Dual-Port)	140mW	260mW	84mW

Table 4: Characteristics of SRAM

Organization	64K words x 16-bit
Power Supply Voltage	1.5 ± 0.1V
Operating Frequency	Max: 220MHz Typical: 160MHz
Read Access Time	Min: 4.26ns Typical: 5.94ns
Average Active Current	Min : Read: 13.35mA Write: 13.52mA Dual-Port: 20.75mA Typical : Read: 21.13mA Write: 23.18mA Dual-Port: 29.75mA
Min Data Retention Voltage	0.1V
Standby Current@0.1V	Min: 0.8µA Typical: 6.2µA
Die Size	≈ 11.5mm x 10mm
Tested Input Signal Slope	100ps - 500ps

4 Conclusions

An example for a design of an asynchronous dual-port 1-Mb CMOS SRAM with 64K words x 16-bit organization has been presented. The SRAM employs various low-power SRAM design techniques and is capable of performing read and write operations simultaneously. Hierarchical Word Decoding has been implemented at architecture-level to reduce decoding delay needed for large memory array. At circuit-level, the proposed SRAM implements current-mode techniques for both read and write operation. Power consumption can be further reduced by using pulse-mode technique together with the ATD circuit.

5 References

/1/ Kiyoo Itoh et. al, "Trends in Low Power RAM Circuit Technologies", *IEEE Symp. Low Power Elec. Dig. of Tech. Papers*, pp. 84-87, Oct. 1994.

/2/ Hiroki Morimura et. al, "A 1-V 1-Mb SRAM for Portable Equipment", *IEEE In. Symp. on Low Power Electronics and Design*, pp. 61-66, Aug. 1996.

/3/ P. Y. Chee et. al, "A High-speed Current-mode Sense Amplifier for CMOS SRAMs", *Proc. of the 35th Mid. Symp. on Circuits and Sys.*, pp. 620-622, Aug. 1992.

/4/ Evert Seevinck et. al, "Current-mode Techniques for High-speed VLSI Circuits with Application to Current Sense Amplifier for CMOS SRAMs", *IEEE J. Solid-State Circuits*, vol. 26, pp. 525-536, April 1991.

/5/ H. Wang et. al, "A Low Power Current Sensing Scheme for CMOS SRAM", *IEEE In. Workshop on Memory Tech. Design & Testing*, pp. 37-45, Aug. 1996.

/6/ Jinn-Shyan Wang et. al, "Low-Power Embedded SRAM Macros with Current-Mode Read/Write Operations", *Proc. IEEE Symp. Low Power Elec. and Design*, pp. 282-287, Aug. 1998.

/7/ Muhammad M. Khellah et. al, "Circuit Techniques For High Speed And Low Power Multi-Port SRAMs", *Proc. 11th Annual IEEE In. ASIC Conference*, pp. 157-161, Sept. 1998.

/8/ Martin Margala, "Low-Power SRAM Circuit Design", *IEEE In. Workshop on Memory Tech., Design and Testing*, pp. 115-122, Aug. 1999.

/9/ Katsuro Sasaki et. al, "A 7-ns 140-mW 1-Mb CMOS SRAM with Current Sense Amplifier", *IEEE J. Solid-State Circuits*, vol. 27, pp. 1511-1518, Nov. 1992.

/10/ S. M. Wang et. al, "Full Current-Mode Techniques for High-Speed CMOS SRAMs", *IEEE In. Symp. on Circuits and Sys*, pp/ 580-582, May 2002.

/11/ Fumio Miyaji et. al, "A 25-ns 4-Mbit CMOS SRAM with Dynamic Bit-Line Loads", *IEEE J. Solid-State Circuits*, vol. 24, pp. 1213-1217, Oct. 1989.

/12/ Teruo Seki et. al, "A 6-ns 1-Mb CMOS SRAM with Latched Sense Amplifier", *IEEE J. Solid-State Circuits*, vol. 28, pp. 478-483, April 1993.

/13/ Toshihiko Hirose et. al, "A 20-ns 4-Mb CMOS SRAM with Hierarchical Word Decoding Architecture", *IEEE J. Solid-State Circuits*, vol. 25, pp. 1068-1074, Oct. 1990.

/14/ Akinori Sekiyama et. al, "A 1-V Operating 256-kb Full CMOS SRAM", *IEEE J. of Solid-State Circuits*, Vol. 27, May 1992.

/15/ Jinn-Shyan Wang et. al, "A New Current-Mode Sense Amplifier for Low-Voltage Low-Power SRAM Design", *11th Annual IEEE In. ASIC Conference*, pp. 163-167, Oct. 1998.

Tan Soon-Hwei, Loh Poh-Yee, Mohd-Shahiman Sulaiman, Zubaida Yusoff
Multimedia University, Cyberjaya, Malaysia
zubaida@mmu.edu.my

Prispelo (Arrived): 20.07.2007 *Sprejeto (Accepted):* 15.06.2007