

Low-Power SRAM Circuit Design

Martin Margala

Department of Electrical and Computer Engineering

University of Alberta

Edmonton, Alberta, Canada T6G 2G7

e-mail: margala@ee.ualberta.ca

Abstract

This paper presents an extensive summary of the latest developments in low-power circuit techniques and methods for Static Random Access Memories. The key techniques in power reduction in both active and standby modes are: capacitance reduction by using divided word-line structure or single-bitline cross-point cell activation, pulse operation by using ATD generator and reduced signal swings on high-capacitance predecode lines, write bus lines and datalines, AC current reduction by using multistage decoding, operating voltage reduction coupled with low-power sensing by using charge-transfer amplification, step-down boosted word-line scheme or full current-mode read/write operation and leakage current suppression by using dual-Vt, Auto-Backgate-Controlled multiple-Vt, or dynamic leakage cut-off techniques.

1. Introduction

In recent years, a rapid development in VLSI fabrication has led to decreased device geometries and increased transistor densities of integrated circuits and circuits with high complexities and very high frequencies have started to emerge. Such circuits consume an excessive amount of power and generate an increased amount of heat. Circuits with excessive power dissipation are more susceptible to run time failures and present serious reliability problems. Increased temperature from high-power processors tends to exacerbate several silicon failure mechanisms. Every 10°C increase in operating temperature approximately doubles a component's failure rate. Increasingly expensive packaging and cooling strategies are required as a chip power increases [1, 2]. Due to these concerns, circuit designers are realizing the importance of limiting power consumption and improving energy efficiency at all levels of the design.

The second driving force behind the low power design phenomenon is a growing class of personal computing devices, such as portable desktops, digital pens, audio- and video-based multimedia products, and wireless communications and imaging systems, such as personal digital assistants, personal communicators and smart cards. These devices and systems demand high-speed, high-throughput computations, complex functionalities and often real-time processing capabilities [3, 4]. The performance of these devices is limited by the size, weight and lifetime of batteries. *Serious reliability problems, increased design costs and battery-operated applications prompted the IC design community to look more aggressively for new approaches and methodologies that produce more power-efficient designs, which means significant reductions in power consumption for the same level of performance.*

Memory circuits form an integral part of every system design as Dynamic RAMs, Static RAMs, Ferroelectric RAMs, ROMs or Flash Memories, significantly contributing to the system level power consumption. Two examples of recently presented reduced-power processors show that 43% and 50.3% respectively of the total system power consumption is attributed to memory circuits [5, 6]. Therefore, reducing the power dissipation in memories can significantly improve the system power-efficiency, performance, reliability and overall costs.

SRAMs have experienced a very rapid development of low-power low-voltage memory design during recent years due to an increased demand for notebooks, laptops, hand-held communication devices and IC memory cards. Table 1 summarizes some of the latest experimental SRAMs for very low-voltage and low-power operation.

2. Sources of SRAM Power

There are different sources of active and standby (data retention) power present in SRAMs. The active power is the sum of the power consumed by the following components:

- decoders
- memory array

- sense amplifiers
- periphery (I/O circuitry, write circuitry, etc.) circuits

Table 1
Low-Power SRAMs Performance Comparison

Memory size (Ref.)	Power supply	CMOS technology	Access time	Power dissipation
4Kb [13]	0.9 V	0.6μm	39ns	18μW@1MHz
4Kb [13]	1.6 V	0.6μm	12ns	64μW@1MHz
32Kb [18]	1 V	0.35μm	17ns	5mW@50MHz
32Kb [20]	1 V	0.35μm	11.8ns	3mW@10MHz
32Kb [16]	1 V	0.25μm	7.3ns	0.9mW@100MHz
32Kb [15]	1 V	0.25 μm	-----	0.9mW@100MHz
32Kb [22]	1 V	0.25μm	7ns	3.9mW@100MHz
256Kb [24]	1.4 V	0.4μm	60ns	3.6mW@5MHz
1Mb [21]	1 V	0.5μm	74ns	1mW@10MHz
1Mb [23]	0.8 V	0.35μm	10ns	5mW@100MHz
4.5Mb [29]	1.8 V	0.25μm	1.8ns	2.8W@550MHz
7.5Mb [25]	3.3 V	0.6μm	6ns	8.42mW@50MHz
7.5Mb [30]	3.3 V	0.8μm	18ns	4.8mW@20MHz

The total active power of an SRAM with $m \times n$ array of cells can be summarized by the expression [8, 9, 10]:

$$P_{active} = (m i_{active} + m(n-1)i_{leak} + (n+m)fC_{DE}V_{INT} + m i_{DC}\Delta t f + C_{PT}V_{INT}f + I_{DCP}) * V_{dd}$$

where i_{active} is the effective current of selected cells, i_{leak} is the effective data retention current of the unselected memory cells, C_{DE} is the output node capacitance of each decoder, V_{INT} is the internal power supply voltage, i_{DC} is the DC current consumed during the read operation, Δt is the activation time of the DC current consuming parts (i.e. sense amplifiers), f is the operating frequency, C_{PT} is the total capacitance of the CMOS logic and the driving circuits in the periphery and the I_{DCP} is the total static (DC) or quasi-static current of the periphery. Major sources of I_{DCP} are column circuitry and differential amplifiers on the I/O lines.

The standby power of an SRAM has a major source represented by i_{leakmn} because the static current from other sources is negligibly small (sense amplifiers are disabled during this mode). Therefore, the total standby power can be expressed as:

$$P_{standby} = mni_{leak} * V_{dd} \quad (1)$$

3. Techniques for Low-Power Operation

In order to significantly reduce the power consumption in SRAMs all contributors to the total power must be targeted. The most efficient techniques used in recent memories are:

- Capacitance reduction of word-lines and the number of cells connected to them, data lines, I/O lines and decoders
- DC current reduction by using new pulse operation techniques for word-lines, periphery

- circuits and sense amplifiers
- AC current reduction by using new decoding techniques (i.e. multi-stage static CMOS decoding)
- Operating voltage reduction
- Leakage current reduction (in active and standby mode) by utilizing multiple threshold voltage (MT-CMOS) or variable threshold voltage technologies (VT-CMOS)

3.1. Capacitance reduction

The largest capacitive elements in a memory are word-line, bitlines and datalines each with a number of cells connected to them. Therefore, reducing the size of these lines can have a significant impact on power consumption reduction. A common technique used often in large memories is called Divided Word Line (DWL) which adopts a two-stage hierarchical row-decoder structure as shown in Figure 1 [10]. The number of sub-word lines connected to one main word line in the dataline direction is generally four, substituting the area of a main row decoder with the area of a local row decoder. DWL features two-step decoding for selecting one word-line, greatly reducing the capacitance of the address lines to a row decoder and the word-line RC delay.

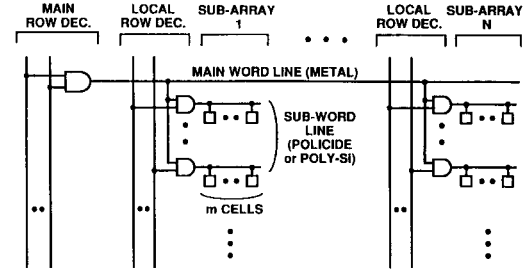


Figure 1. Divided Word-Line Structure (DWL) [10].

A single bitline cross-point cell activation (SCPA) architecture reduces the power further by improving the DWL technique [11]. The architecture enables the smallest column current possible without increasing the block division of the cell array thus reducing the decoder area and the memory core area. The cell architecture is shown in Figure 2. The Y-address controls the access transistors and the X-address. Since only one memory cell at the cross-point of X- and Y- is activated, a column current is drawn only by the accessed cell. As a result, the column current is minimized. In addition, SCPA allows the number of blocks to be reduced because the column current is independent of the number of block divisions in the SCPA. The disadvantage of this configuration is that during the write "high" cycle, both X- and Y- lines have to be boosted using a word-line boost circuit.

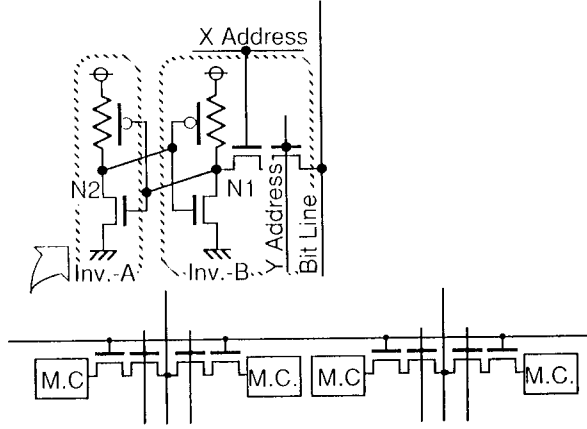


Figure 2. Memory cell used for SCPA architecture[11].

Caravella has proposed a similar sub-division technique as DWL which he demonstrated on 64×64 bit cell array [12, 13]. If C_j is a parasitic capacitance associated with a single bit cell load on a bitline (junction and metal) and if C_{ch} is a parasitic capacitance associated with a single bit cell on the word-line (gate, fringe and metal), then the total bitline capacitance is $64 * C_j$ and the total word capacitance is $64 * C_{ch}$. If the array is divided into four isolated sub-arrays of 32×32 bit cells, the total bitline and word-line capacitances would be halved (see Figure 3). The total capacitance per read/write that would need to be discharged or charged is given $1024 * C_j + 32 * C_{ch}$ for the sub-array architecture as opposed to $4096 * C_j + 64 * C_{ch}$ for the 64×64 array. This technique carries a penalty due to additional decode and control logic and routing.

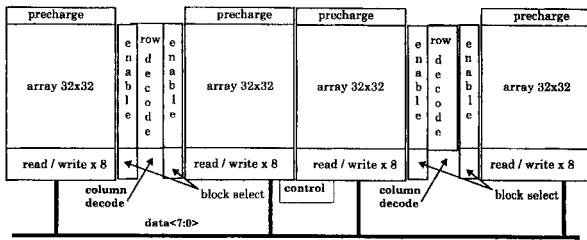


Figure 3. Memory architecture [13].

3.2. Pulse operation techniques

Pulsing the word-lines, equalization and sense lines can shorten the active duty cycle and thus reduce the power dissipation. In order to generate different pulse signals, an on-chip address transition detection (ATD) pulse generator is used [10]. This circuit, shown in Figure 4, is a key element for the active power reduction in memories.

An ATD generator consists of delay circuits (i. e. inverter chains) and an XOR circuit. The ATD circuit generates a $\phi(a_i)$ pulse every time it detects a "L" to "H" or "H" to "L" transition on the input address signal a_i . Then all ATD generated pulses from all address transitions are summed through an OR gate to a single pulse ϕ_{ATD} . This final pulse is usually stretched out with a delay circuit to generate different pulses needed in the SRAM and used to reduce power or speed up a signal propagation. Pulsed operation techniques are also used to reduce power consumption by reducing the signal swing on high-capacitance predecode lines, write-bus-lines and bit lines without sacrificing the performance[14, 15, 16]. These techniques target the power that is consumed during write and decode operations. Most of the power savings come from operating the bitlines from $V_{dd}/2$

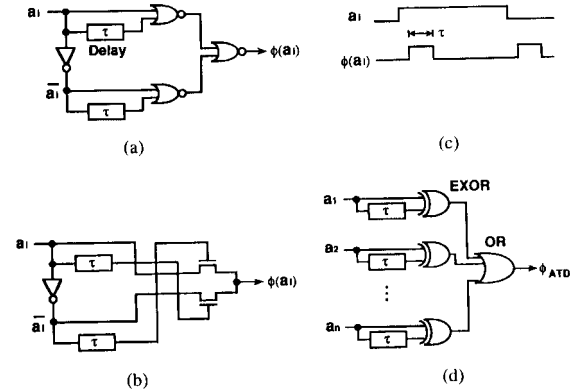


Figure 4. Address transition detection circuits; a) and b) ATD pulse generators, c) ATD pulse wave-forms, d) a summation circuit of all ATD pulses generated from all address transitions [10].

rather than V_{dd} . This approach is based on new half-swing pulse-mode gate family. Figure 5 shows a half-swing pulse-mode AND gate. The principle of the operation is in a merger of a voltage-level converter with a logical AND. A positive half-swing (transitions from a rest state $V_{dd}/2$ to V_{dd} and back to $V_{dd}/2$) and a negative half-swing (transitions from a rest state $V_{dd}/2$ to Gnd and back to $V_{dd}/2$) combined with the receiver-gate logic style result in a full gate overdrive with negligible effects of the low-swing inputs on the performance of the receiver. This structure is combined with a self-resetting circuitry and a PMOS leaker to improve the noise margin and the speed of the output reset transition (see Figure 6).

Both negative and positive half-swing pulses can reduce the power consumption further by using a charge recycling. The charge used to produce the assert transition of a positive pulse can also be used to produce the reset transition of a negative pulse. If the

capacitances of positive and negative pulses match, then no current would be drawn from the $V_{dd}/2$ power supply ($V_{dd}/2$ voltage is generated by an on-chip voltage converter). Combining the half-swing pulse-mode logic with the charge recycling techniques, 75% of the power on high-capacitance lines can be saved

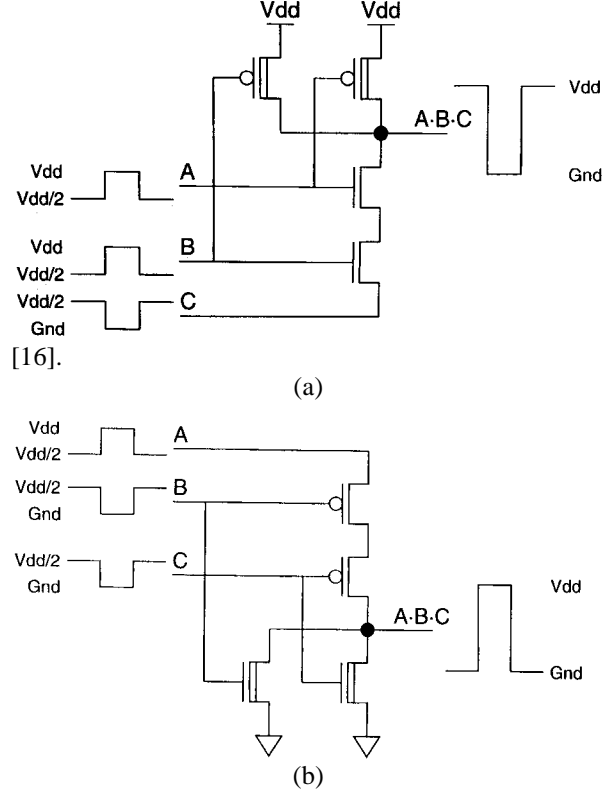


Figure 5. Half-swing pulse-mode AND gate; a) NMOS-style, b) PMOS-style.

3.3. AC current reduction

One of the circuit techniques that reduces AC current in memories is a multi-stage decoding. It is common that fast static CMOS decoders are based on decode architecture, the number of transistors, fanin and the loading on the address input buffers is reduced (see Figure 8). As a result, both speed and power are optimized. The signal ϕ_x , generated by the ATD pulse generator, enables the decoder and secures pulse activated word-line. OR/NOR and AND/NAND architectures. Figure 7 shows one example of a row decoder for a three-bit address. The input buffers drive the interconnect capacitance of the address line and also the input capacitance of the NAND gates. By using a two-stage decode architecture, the number of transistors, fanin and the loading on the address input buffers is reduced (see Figure 8). As a result, both speed and power are optimized. The signal ϕ_x , generated by the ATD pulse generator, enables the

decoder and secures pulse activated word-line.

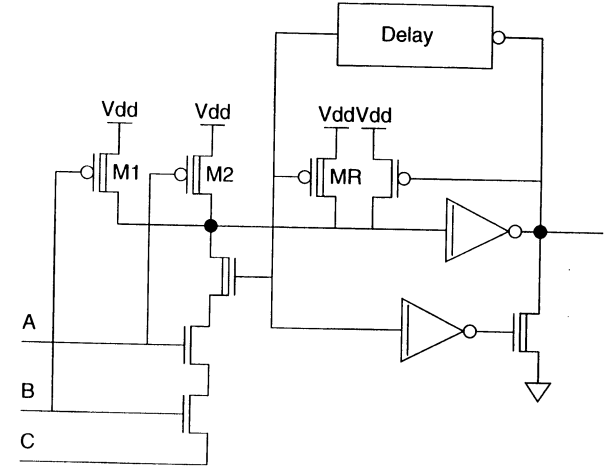


Figure 6. Self-resetting half-swing pulse-mode gate with a PMOS leaker.

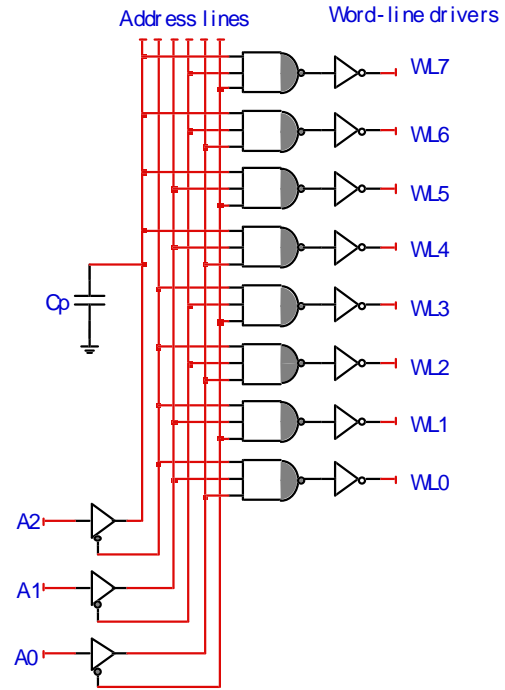
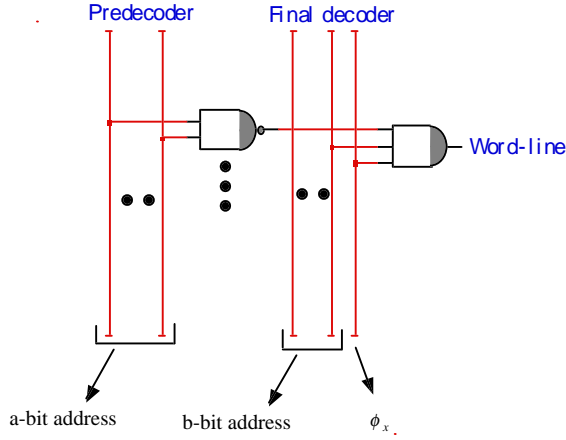


Figure 7. A row decoder for a three-bit address.

3.4. Low-power sensing techniques

An efficient method for reducing the AC power of bitlines and datalines is to use the current-mode read and write operations based on new current-based circuit techniques [25, 26, 27]. Wang et. al. proposed a new SRAM cell that supports current-mode operations with very small voltage swings on bitlines and datalines. A fully current-mode technique consumes



$a + b$: number of bits for row decoding
Figure 8. A two-stage decoder architecture.

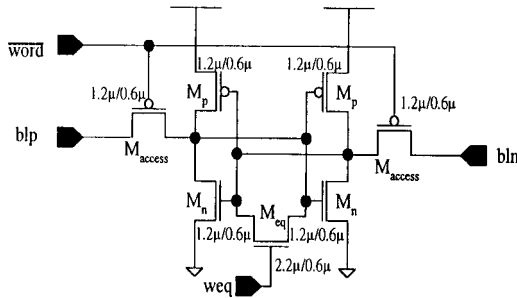


Figure 9. New 7-transistor SRAM memory cell.

only 30% of the power consumed by a previous current-read only design. Very small voltage swings on bitlines and datalines lead to a significant reduction of AC power. The new memory cell has seven transistors as shown in Figure 9. The additional transistor M_{eq} clears the content of the memory cell prior to the write operation. It performs the cell equalization. This transistor is turned off during the read operation so it does not disrupt the normal operation. An N-type current conveyor is inserted between the data input cell and the memory cell in order to perform a current-mode write operation which is a complementary way to read. The equalization transistor is sized to be as large as possible to improve fast equalization speed, but not to increase the cell size. After a suitable sizing, the new 7-transistor cell is 4.3% smaller than its 6-transistor counterpart, as illustrated in Figure 10.

Another new current-mode sense amplifier for 1.5V power supply was proposed by Wang and Lee [27]. The new circuit overcomes the problems of a conventional sense amplifier with pattern dependency by implementing a modified current conveyor. Pattern-dependency problem limits the scaling of the operating

voltage. Also, the circuit does not consume any DC power because it is constructed as a complementary device. As a result, the power consumption is reduced by 61-94% compared with a conventional design. The circuit structure of the modified current conveyor is similar to a conventional current conveyor design. However, an extra PMOS transistor Mp7, as seen in Figure 11, is used. The transistor is controlled by RX signal (a complement of CS). After every read cycle, transistor Mp7 is turned on and equalizes nodes RXP and RXN which eliminates any residual differential

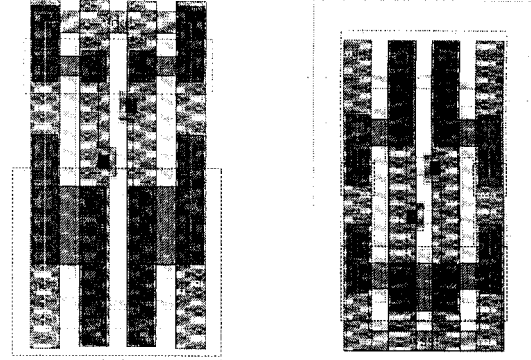


Figure 10. SRAM cell layout, a) 6T cell, b) new 7T cell.

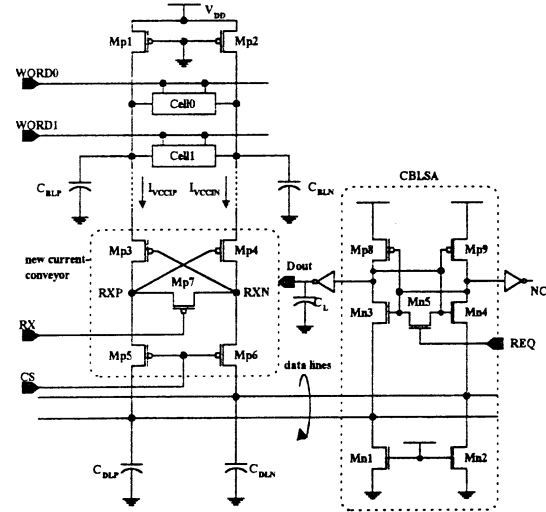


Figure 11. SRAM read circuitry with the new current-mode sense amplifier.

voltage between these two nodes (limitation in conventional designs).

3.5. Leakage current reduction

In order to effectively reduce the dynamic power consumption, the threshold voltage is reduced along with the operating voltage. However, low threshold voltages increase the leakage

current during both active and standby modes. The fundamental method for a leakage current reduction is a Dual- V_{th} or a Variable- V_{th} circuit technique. An example of one such technique is shown in Figure 12 [18, 22]. Here, high V_{th} MOS transistors are utilized to reduce the leakage current during standby-mode. As the supply voltage for the word decoder (g) is lowered

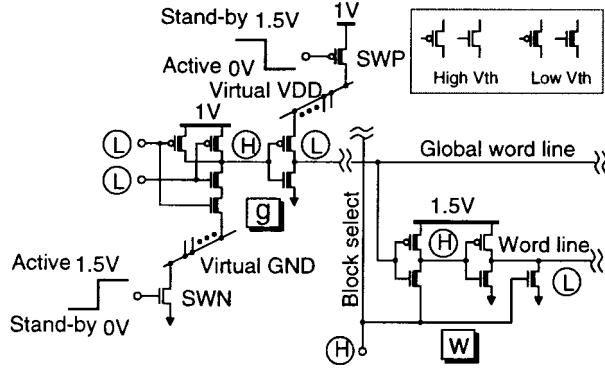


Figure 12. Dual V_{th} CMOS circuit scheme.

to 1V, all transistors forming the decoder are low V_{th} to retain high performance. The leakage currents during the stand-by mode are substantially reduced by a cut-off switch (SWP, SWN). SWN consists of a high V_{th} transistor and SWP consists of a low V_{th} transistor. Both switches are controlled by a 1.5 V signal. Hence, the SWN gains a considerable conductivity. SWP can be quickly cut off because of the reverse biasing. The operating voltage of the local decoder (w) is boosted to 1.5 V. The high operating voltage gives sufficient drivability even to high V_{th} transistors. This technique belongs to schemes that use dynamic boosting of the power supply voltage and the word lines. However, in these schemes the gate voltage of MOSFETs is raised often to more than 1.4 V even though the operating voltage is 0.8 V. This creates reliability problems.

Kawaguchi et. al. introduced a new technique, a dynamic leakage cut-off (DLC) scheme. Operation waveforms are shown in Figure 13 [28]. A dynamic change of n-well and p-well bias voltages to V_{DD} and V_{SS} respectively for selected memory cells is the key feature of this architecture. At the same time, the non-selected memory cells are biased with $\sim 2V_{DD}$ for V_{NWELL} and $\sim -V_{DD}$ for V_{PWELL} . After this, the V_{th} of the selected cells becomes low which aids in high drive, thus a fast operation is executed. On the other hand, the V_{th} of the unselected memory cells is high enough to achieve low subthreshold consumption. This technique is similar to the Variable Threshold CMOS (VT CMOS) technique; however, the difference is in the synchronization signal of the well bias. While in

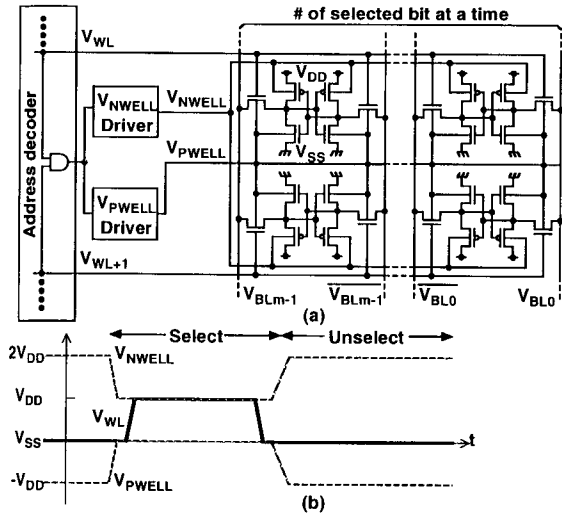


Figure 13. Dynamic Leakage Cut-Off Scheme: (a) Circuit Schematic (b) Its Operation.

current VT CMOS the well bias is synchronized with a standby signal, DLC technique is synchronized with the word-line signal.

Nii et. al. improved the MT-CMOS technique further and proposed the Auto-Backgate Controlled (ABC) MT-CMOS method [20]. The ABC MT-CMOS reduces significantly the leakage current during the

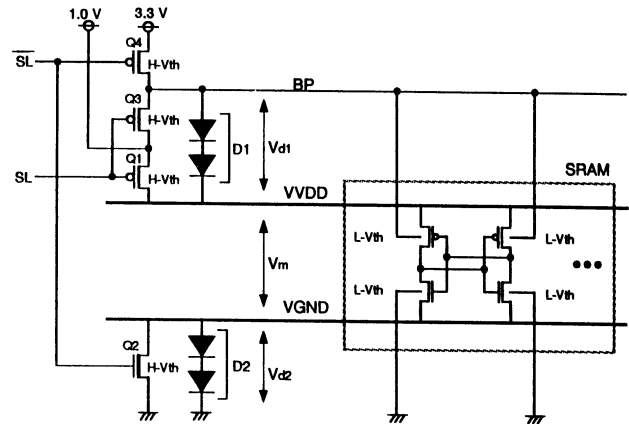


Figure 14. A schematic diagram of ABC-MT-CMOS circuit.

"sleep" mode. The circuit diagrams of this method is shown in Figure 14. Transistors Q1-Q4 are high-threshold devices that act as switches to cut-off the leakage current. The internal circuitry is designed with low- V_t devices. During the active mode, signal \overline{SL} is pulled low and SL is pulled high. Q1, Q2, and Q3 turn on and Q4 turns off and virtual power supply VVDD and the substrate bias BP become 1V. During the sleep mode, signal \overline{SL} is pulled high and SL is pulled low and Q1, Q2 and Q3 turn off whereas Q4 turns on and

BP becomes 3.3V. The leakage current that flows from V_{dd2} to ground through D1 and D2 determines voltages V_{d1}, V_{d2} and V_m. V_{d1} is a bias between the source and the substrate of the PMOS transistors, V_{d2} is a bias of the NMOS transistors and V_m is a voltage between the virtual power line V_{VDD} and the virtual ground V_{GND}. The leakage current is reduced to 20pA/cell.

4. Conclusion

In this paper, the latest developments in low-power circuit techniques and methods SRAMs were reviewed. All major sources of power dissipation in these memories were analyzed. Key techniques for drastic reduction of power consumption were identified. These are: capacitance reduction, very low operating voltages, DC and AC current reduction and suppression of leakage currents. Many of reviewed techniques are applicable to other applications such as ASICs, DSPs, etc. Battery and solar-cell operation requires an operating voltage environment in sub-1V area. These conditions demand new design approaches and more sophisticated concepts to retain high device reliability. Experimental circuits operating at these voltage levels slowly start to emerge in all types of memories. However, there is no universal solution for any of these designs and many challenges still await for memory designers.

References

- [1] D. Pivin, "Pick the Right Package for Your Next ASIC Design," *EDN*, vol.39, no.3, pp.91-108, February 3, 1994.
- [2] C. Small, "Shrinking Devices Put the Squeeze on System Packaging," *EDN*, vol.39, no.4, pp.41-46, February 17, 1994.
- [3] D. Manners, "Portables Prompt Low-Power Chips," *Electronics Weekly*, no.1574, p.22, November 13, 1991.
- [4] J. Mayer, "Designers Heed the Portable Mandate," *EDN*, vol.37, no.20, pp.65-68, November 5, 1992.
- [5] R. Stephany et. al., "A 200MHz 32b 0.5W CMOS RISC Microprocessor," in *ISSCC Digest of Technical Papers*, pp.15.5-1 - 15.5-2, February, 1998.
- [6] H. Igura et. al., "An 800MOPS 100mW 1.5V Parallel DSP for Mobile Multimedia Processing," in *ISSCC Digest of Technical Papers*, pp.18.3-1 - 18.3-2, February, 1998.
- [7] A. K. Sharma, "Semiconductor Memories - Technology, Testing and Reliability", *IEEE Press*, 1997.
- [8] M. Margala and N. G. Durdle, "Noncomplementary BiCMOS Logic and CMOS Logic Styles for Low-Voltage Low-Power Operation - A Comparative Study", *IEEE Journal of Solid-State Circuits*, vol. 33, no. 10, pp.1580-1585, October 1998.
- [9] A. Bellaouar and M. I. Elmasry, "Low-Power Digital VLSI Design, Circuits and Systems", *Kluwer Academic Publishers*, 1996.
- [10] K. Itoh et. al., "Trends in Low-Power RAM Circuit Technologies", *Proceedings of the IEEE*, pp.524-543, April 1995.
- [11] M. Ukita et. al., "A Single Bitline Cross-Point Cell Activation (SCPA) Architecture for Ultra Low Power SRAMs," in *ISSCC Digest of Technical Papers*, pp.252-253, February 1994.
- [12] J. S. Caravella, "A 0.9V, 4K SRAM for Embedded Applications," in *Proceedings of CICC*, pp.119-122, May 1996.
- [13] J. S. Caravella, "A Low Voltage SRAM For Embedded Applications," *IEEE Journal of Solid-State Circuits*, vol. 32, no. 3, pp.428-432, March 1997.
- [14] B. S. Amrutur and M. A. Horowitz, "Techniques to Reduce Power in Fast Wide Memories," in *Proceedings of SLPE'94*, pp.92-93, 1994.
- [15] T. Mori et. al., "A 1V 0.9mW at 100MHz 2kx16b SRAM utilizing a Half-Swing Pulsed-Decoder and Write-Bus Architecture in 0.25 μ m Dual-V_t CMOS", in *ISSCC Digest of Technical Papers*, pp.22.4-1 - 22.4-2, February 1998.
- [16] K. W. Mai et. al., "Low-Power SRAM Design Using Half-Swing Pulse-Mode Techniques," *IEEE Journal of Solid-State Circuits*, vol. 33, no. 11, pp.1659-1671, November 1998.
- [17] H. Morimura and N. Shibata, "A 1-V 1-Mb SRAM for Portable Equipment," in *Proceedings of ISLPED'96*, pp.61-66, August 1996.
- [18] S. Kawashima et. al., "A Charge-Transfer Amplifier and an Encoded-Bus Architecture for Low-Power SRAM's," *IEEE Journal of Solid-State Circuits*, vol. 33, no. 5, pp.793-799, May 1998.
- [19] H. Morimura and N. Shibata, "A Step-Down Boosted-Wordline Scheme for 1-V Battery Operated Fast SRAM's," *IEEE Journal of Solid-State Circuits*, vol. 33, no. 8, pp.1220-1227, August 1998.
- [20] K. Nii et. al., "A Low Power SRAM using Auto-Backgate-Controlled MT-CMOS," in *Proceedings of ISLPED*, pp.293-298, August 1998.
- [21] H. Sato et. al., "A 5-MHz, 3.6mW, 1.4-V SRAM with Nonboosted, Vertical Bipolar Bit-Line Contact Memory Cell," *IEEE Journal of Solid-State Circuits*, vol. 33, no. 11, pp.1672-1681, November 1998.
- [22] I. Fukushi et. al., "A Low-Power SRAM Using Improved Charge Transfer Sense Amplifiers and a Dual-V_{th} CMOS Circuit Scheme," in *Digest of Technical Papers of Symposium on VLSI Circuits*, pp.142-143, June 1998.
- [23] H. Yamauchi et. al., "A 0.8V/100MHz/sub-5mW-Operated Mega-bit SRAM Cell Architecture with Charge-Recycle Offset-Source Driving (OSD) Scheme", in *Digest of Technical Papers of Symposium on VLSI Circuits*, pp.126-127, June 1996.
- [24] K. Itoh et. al., "A Deep Sub-V, Single Power-Supply SRAM Cell with Multi-V_t Boosted Storage Node and Dynamic Load," in *Digest of Technical Papers of Symposium on VLSI Circuits*, pp.132-133, June 1996.
- [25] J.-S. Wang et. al., "Low-Power Embedded SRAM Macros with Current-Mode Read/Write Operations," in *Proceedings of ISLPED*, pp.282-287, August 1998.
- [26] M. Khellah and M. I. Elmasry, "Circuit Techniques for High-Speed and Low-Power Multi-Port SRAMs," in *Proceedings of ASIC*, pp.157-161, September 1998.
- [27] J.-S. Wang and H.Y. Lee, "A New Current-Mode Sense Amplifier for Low-Voltage Low-Power SRAM Design," in *Proceedings of ASIC*, pp.163-167, September 1998.
- [28] H. Kawaguchi et. al., "Dynamic Leakage Cut-off Scheme for Low-Voltage SRAM's," in *Digest of Technical*

Papers of Symposium on VLSI Circuits, pp.140-141, June 1998.

[29] H. Nambu et. al., "A 1.8-ns Access, 550-MHz, 4.5-Mb CMOS SRAM," *IEEE Journal of solid-State Circuits*, vol. 33, no. 11, pp.1650-1658, November 1998.

[30] K. J. Shultz et. Al.," Low-Supply-Noise Low-Power Embedded Modular SRAM," *IEE Circuits, Devices and Systems*, vol. 143, no. 2, pp.73-82, April 1996.