LOW-POWER, HIGH-SPEED SRAM DESIGN: A REVIEW

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

Faculty of Engineering, Multimedia University, 63100 Cyberjaya, Selangor, Malaysia

Key words: CMOS, SRAM, Low-Power, High-Speed, Dual-Port, Asynchronous

Abstract: This article describes the challenges in the design of low-power SRAM and difficulties in designing a high-speed static RAM. Following an overview of general issues, approaches/techniques in achieving low-power SRAM, as well as techniques to increase SRAM operating frequency are described. Following the overview of each approach, the challenges and trade-offs are outlined.

Načrtovanje hitrih vezij SRAM z majhno porabo - pregled

Kjučne besede: CMOS, SRAM, nizka poraba, velika hitrost, dvojni vhod, asinhronski

Izvleček: V prispevku opišemo izzive pri načrtovanju vezij SRAM z nizko porabo in težave pri načrtovanju hitrih pomnilnikov RAM. Po uvodnem pregledu splošnih pojmov opišemo tehnike in pristope k načrtovanju vezij SRAM z nizko porabo, kakor tudi tehnike za povečanje frekvence delovanja vezij SRAM. Po opisu obeh pristopov naštejemo njune izzive in kompromise.

1 Introduction

Static Random Access Memory (SRAM) is mainly used as an embedded block (EBB) memory circuitry such as level 1 and level 2 caches in a microprocessor operating based on the Principle of Locality or used as a data buffer at a chip's interface. SRAM occupies a large portion of the modern digital chips and its capacity is forecasted to further increase in the new era of System on Chip (SoC). Increased SRAM density results in larger internal capacitance and thus limits the operating frequency and power consumption. Besides that, leakage current is now one of the major contributors to chip's overall power consumption. As static power continues to dominate, appropriate measures have to be taken to minimize the effects of leakage currents, especially in short channel devices.

This article summarizes power savings and high-speed techniques that can be implemented into SRAM design to overcome speed degradation and high power dissipation issues. The next section of this article addresses general issues in designing SRAM with large memory capacity, with example referring to the implementation of a 1-Mbit SRAM.

2 General Considerations

The general power consumption equation for SRAM is /1/:

$$P = V_{dd}[mI_{active} + m(n-1)I_{hold}] + V_{dd}[(n+m)C_{de}V_{\text{int}}f] + V_{dd}[C_{pt}V_{\text{int}}f + I_{dcp}] \quad (1)$$

First term of the equation dominates SRAM power consumption nowadays. I_{active} is active current of m selected memory cells while I_{hold} is the data retention current of others unselected memory cells. Second term represents power consumption of decoders and is the power needed for switching internal node capacitance. Third term represents power consumption of peripheral circuitries. It consists of dc current and current needed for switching capacitance. In general, active power depends on the switching capacitance and direct-path current while data retention power depends on the leakage current (sub-threshold leakage current - main contributor and reverse-biased current) and size of memory.

SRAM's memory operation delay usually consists of word-line decoding delay, data sensing or data writing delay and delay for resetting circuitry to initial condition. These delays are mostly proportional to the size of memory and also depend on the techniques that adopted into the design.

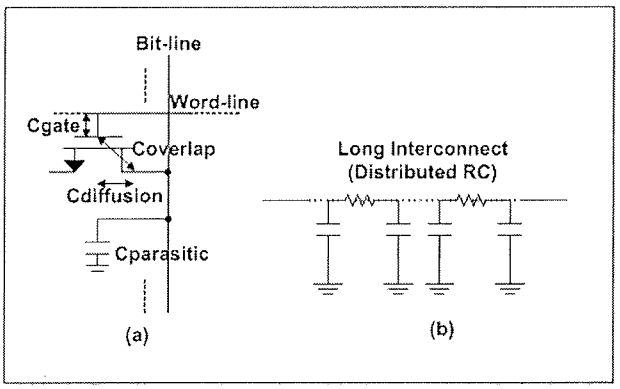


Figure 1 (a) Transistor's Parasitic
(b) Transmission line's Parasitic

Technology scaling and larger memory capacity contributes to larger bit-line capacitance and resistance as longer interconnect is required and more transistors are connected to a common line (Refer to Figure 1). Therefore, charging or discharging the line capacitance to transmit voltage signal causes huge active power consumption and signal propagation delay. In addition, larger number of memory cells raises the leakage power and decoding de-

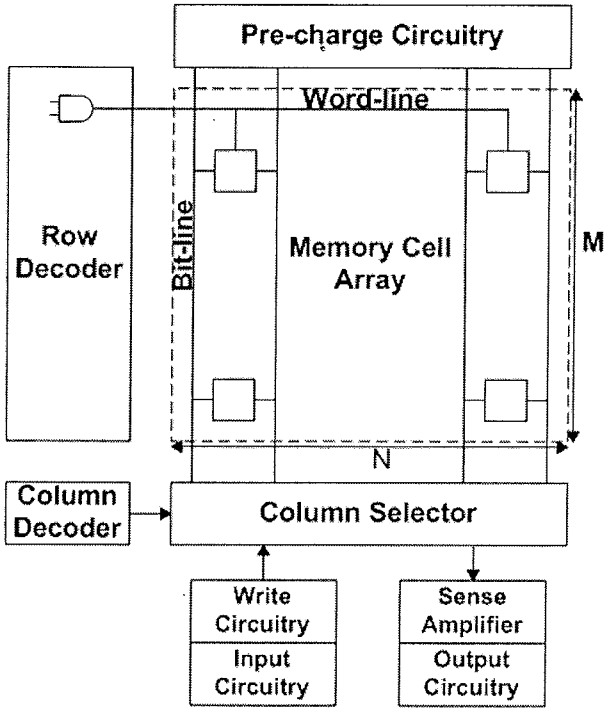


Figure 2 General SRAM Architecture

lay as well. The leakage power has been foreseen to dominate SRAM power in the Nano-range process technology as threshold voltage is scaled down along with the transistor's channel length. Other than that, designers have to make sure that the power race conditions in circuit are handled well during design phase to diminish the crowbar current.

2.1 Low-Power and High-Speed SRAM Architecture

Figure 2 is the general architecture that utilizes two way decoding method. If a 1-Mbit SRAM is going to be designed with 32-bit word size and 16:1 column muxing, it will have physical size of 2048 rows x 512 columns. The huge capacitive load on bit-lines and word-lines (large C_{de} and C_{pt}) causes high power and speed degradation problem. Instead, memory partitioning and bit-line partitioning techniques can be employed during the architecture-level design phase.

The principle in memory partitioning 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. The word-lines are divided into several sub levels and only certain sub word-lines are activated during operation. As a result, the number of transistors that are connected to the bit-lines and word-lines is lesser and hence smaller capacitance switched during operation. Also, it reduces active current I_{active} by shutting down portions that are not accessed potentially. Indirectly, it reduces the RC delay of decoding stage and word-line as well. Divided Word Line (DWL) architecture (Refer to Figure 3a) and Hierarchical Word Decoding (HWD) architecture (Refer to Figure 3b) are two of the most famous memory partitioning techniques that are proposed by Masahiko Yoshimoto et al. /2/ and Toshihiko Hirose et al. /3/ respectively.

DWL architecture requires lesser number of decoding stages and hence lesser number of transistors and smaller decoding delay. Furthermore, it is easier to be adopted into the design especially for layout design as routing and placement is simpler than the HWD architecture. However, the active power and decoding delay of DWL architecture exceeds HWD architecture when the memory capacity is above 256-Kbit due to larger number of multi-divided blocks that raises the load capacitance on global word-lines while keeping the number of selected memory cells small /3/. When the size of SRAM exceeds 1-Mbit, HWD architecture should be adopted into the design instead /4/. However, the number of portions should be constrained as excessively large number of portions imposes area overhead and wiring overhead that tends to increase power dissipation. Hence, HWD architecture might not be able to help in the near future as memory capacity is kept increasing. Divides the word-line into higher number of levels is no more a feasible solution as the decoding delay is increased as well.

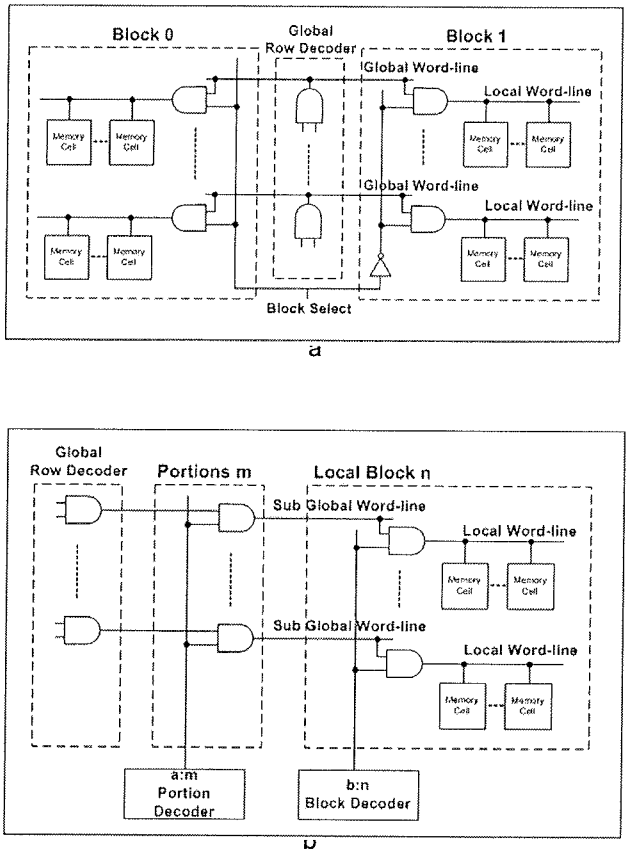


Figure 3 (a) Divided Word-Line (DWL) Architecture (b) Hierarchical Word Decoding (HWD) Architecture

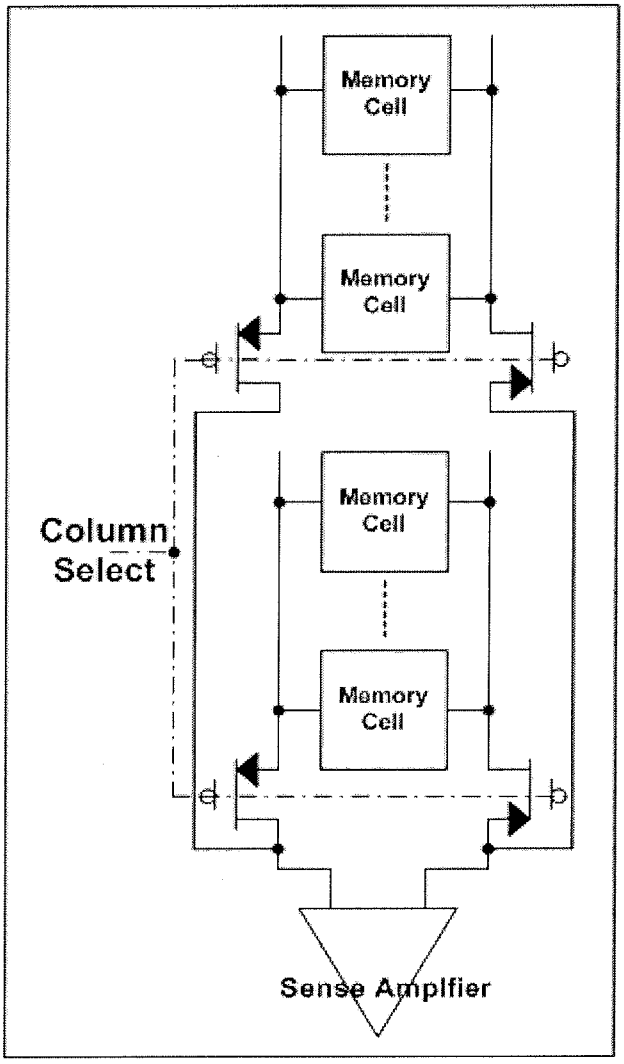
Bit-line partitioning technique is used to reduce bit-line capacitance and hence lower active power and access delay at the expenses of slightly larger number of transistors. It can be implemented in two ways, either using multiple-level muxing (Refer to Figure 4a) or divided bit-line approach (Refer to Figure 4b) proposed by Karandikar Ashish et al. /5/. Multiple-level muxing method increases node capacitance at the output of pass-transistors and hence number of the bit-line partition level should be limited. Besides, the load on column select signal line becomes larger due to longer wire routing and more transistors are connected. Divided bit-line approach reduces the bit-line capacitance to achieve lower active power and smaller RC signal delay. Besides that, decoding stage can be made smaller for this approach and thus reduces the decoding delay. Although the total current driving force of memory cell becomes smaller due to additional pass transistor in between the global bit-line and local bit-line, it can achieve better performance as long as the number of divided blocks and number of memory cells per a divided block are constrained. Theoretically, the optimal number of memory cells per a divided block is around 8. Sizing of the global bit-line pass transistor has to be large enough to pull down the bit-line voltage. It also has to be smaller than the size when the drain capacitance added to global bit-line exceeds the current driving capability.

3 Low Power and High Speed SRAM Circuit Design Techniques

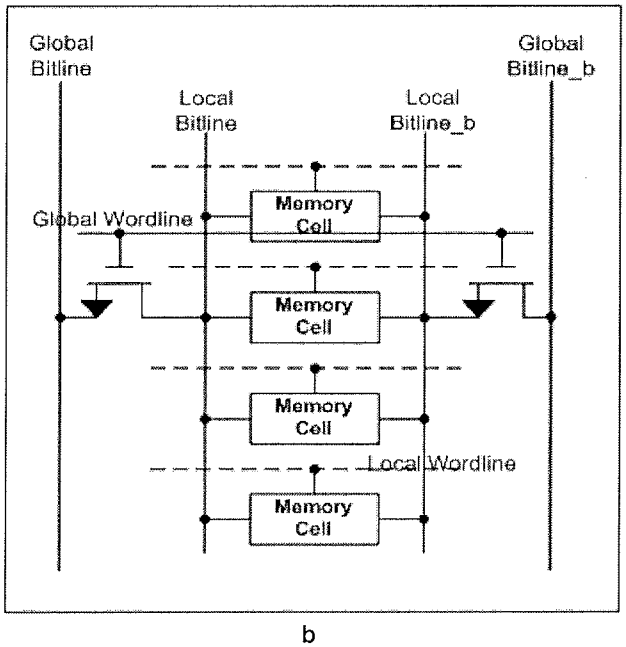
ISSCC2005 mentioned that semiconductor industry begins to shift from synchronous SRAM design to asynchronous design to solve power consumption problem. If external clock signal is used to derive internal control signals, pulse width of the derived signals might be too large for an operation to complete. For instance, bit-lines voltage can be pulled down to a level more than sufficient for a memory operation (Refer to Figure 5a). Besides, some internal circuits are turned on every cycle yet no operation is requested. As a result, asynchronous SRAM design using Address Transition Detection (ATD) circuitry is invented to reduce dc current, active current and switching capacitance by supplying pulse signals to the internal circuitries /6/. A simple ATD circuit together with the pulsed-signal waveform is shown in Figure 5b.

A pulse is generated whenever there is a transition in the input signals. The delay element circuit is designed such that pulse duration will be just enough for an SRAM operation to complete. Modified Schmitt trigger delay circuit that has lesser delay variation and hot-electron effect can be used /7/. However, this technique is still not widely used by the industry due to large variation of generated pulse across process, voltage and temperature (PVT) corners.

Voltage-mode sense amplifier and write circuitry usually cause large bit-line voltage swing and hence power dissipation problem as $P_{discharge} = C_{bl}\left[dV_{swing} / dt_{operation}\right] \cdot V_{dd}$.



a



b

Figure 4 (a) Multiple-level MUXING Method (b) Divided Bit-line Approach

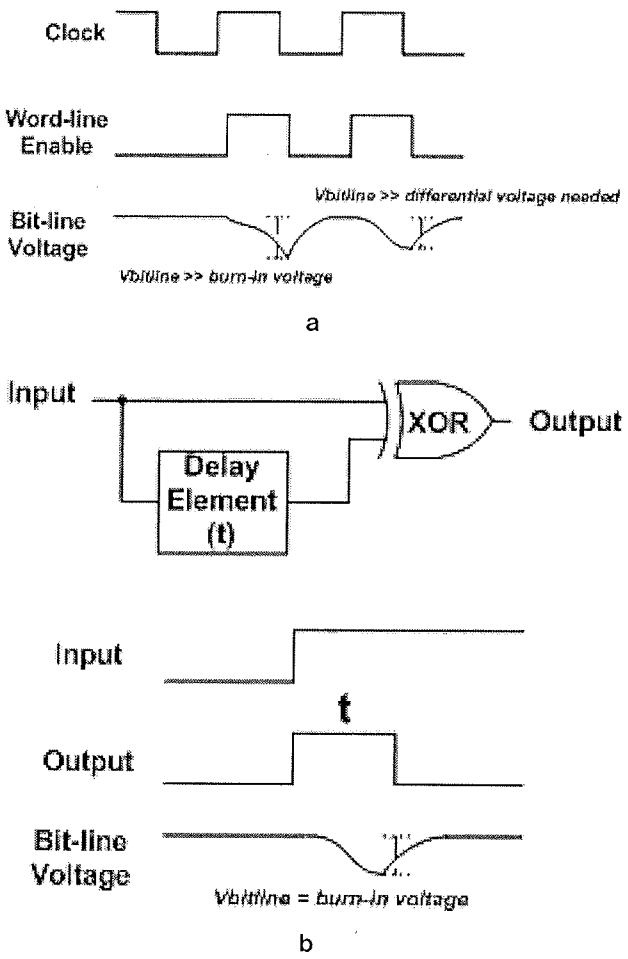


Figure 5 (a) Bit-line Over-discharged
(b) ATD and Output Pulsed-Signal

Also, speed degradation is another problem when the bit-line parasitic is large. Usually SRAM cell current is small as it is designed to have minimum-sized transistors and thus bit-line development time is long. Other than that, another long period is required to restore bit-lines voltage level after every memory operation. Current-mode technique can be implemented to avoid these problems. It was first proposed by Evert Seevinck et al. in 1991 /8/. Its 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. Current-mode technique can be implemented for both sense amplifier and write circuitry. Figure 6a shows the bit-line peripheral circuit that has implemented full current-mode technique. Figure 6b and 6c illustrate the current-mode read (Clamped Data-line Sense Amplifier) and write circuitry /9/. It is found that the delay of sensing and writing process is insensitive to the bit-line or data-line capacitance /4/,/10/. In addition, it helps to reduce coupling noise to the adjacent bit-lines as bit-lines voltage swing is minimized. However, the current-mode circuit itself causes large crowbar current and direct-current path also exist between pre-charge circuit and sense amplifier during operation. Thus the circuit has to

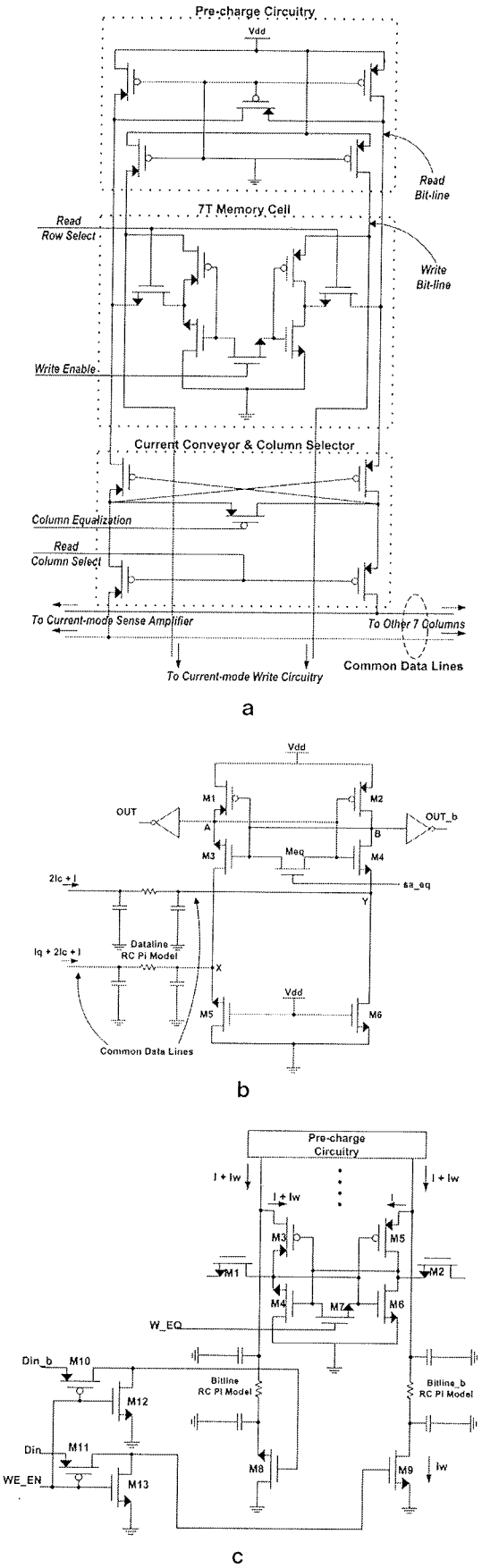


Figure 6 (a) Bit-line Peripheral Circuit
(b) Clamped Data-line Sense Amplifier
(c) Current-mode Write Circuitry

be properly designed or else it will offset the advantage of reducing capacitance switched during operation. It can be minimized by using dynamic load controlling circuit which has lesser effects to the speed of current-mode circuit. The load circuit is turned when there's no operation to reduce the current flowing. Besides that, it also requires additional transistors for memory cells that increase the die size.

Charge transfer sense amplifier (CTSA) allows low power sensing without sacrificing the speed (Refer to Figure 7). It operates based on charge re-distribution mechanism between very high bit-line capacitance and low output capacitance of the sense amplifier /11/. According to the charge conservation equation: $\Delta V_{bitline} \cdot C_{bitline} = C_{out} \cdot V_{out}$, bit-line swing depends on the ratio between bit-line and output capacitance where larger bit-line capacitance leads to lower bit-line swing and hence lower power. Besides, this circuit diminishes the direct-path current as the pre-charge circuit can be turned off after the word-line is enabled. Unlike current-mode sense amplifier that requires the pre-charge circuit to be turned on.

There're also several techniques proposed purposely to reduce writing delay and write power. There is a low power write technique proposed by Alowersson that biases the bit-lines voltage closed to low voltage level /12/. It first builds up a small differential voltage for the bit-lines and utilizes the regenerative effect of cross-coupled PMOSs to amplify small differential storage nodes voltage of the memory cell. However, this technique causes slow read operation as word-lines voltage are required to be reduced to prevent spurious discharge of memory cell storage nodes. Another low swing write technique proposed by Amrutur /13/ uses boosted voltage for word-line signal and virtual ground line (Refer to Figure 8). After the virtual ground line is driven high to clear memory cell's content, small differential voltage between the bit-lines is built up and this small differential signal is transferred into the storage node. The memory cell amplifies this small differential signal after virtual ground line is driven low. This technique has several problems such as voltage drop of virtual ground line, boosted voltage generation circuit is required and memory block width is required to be the same as input data size.

In the decoder design, designer is required to choose suitable gate logic style and then finds the optimal sizing for each gate and adding appropriate buffers to reduce decoding delay. Logical effort is one of the popular techniques in finding out optimal sizing for a chain of logic gates. Either NMOSs or PMOSs in the logic gates can be sized such that only one of the signal's transitions is enhanced and the opposite transition is weakened /14/. This helps to reduce the input capacitance of logic gates and thus overall address decoding delay. However, such implementation needs separate reset devices to prevent slow reset transition. These reset devices uses self-resetting logic (SR-CMOS) technique and delayed reset logic (DRCMOS) technique.

Another low power and high speed approach - half-swing pulse-mode technique uses half-swing signals to reduce the power in decoder and write circuitry /15/-/16/. Half swing technique limits the signal swing on the high capacitance bit-lines and I/O lines. In addition, the charge cycling between positive and negative pulses further reduces the power. The power dissipation is $P_{half-swing} = C_{bitline} \cdot (V_{dd}^2/2) \cdot f$ compares to the power of full swing voltage-mode circuits which is $P_{full-swing} = 2C_{bitline} \cdot V_{dd}^2 \cdot f$ and hence it can save 75% power theoretically. The problem with this technique is it uses half-swing pulse-mode gate family that requires dual-threshold or multi-threshold voltage transistors which might not be supported by certain process technology.

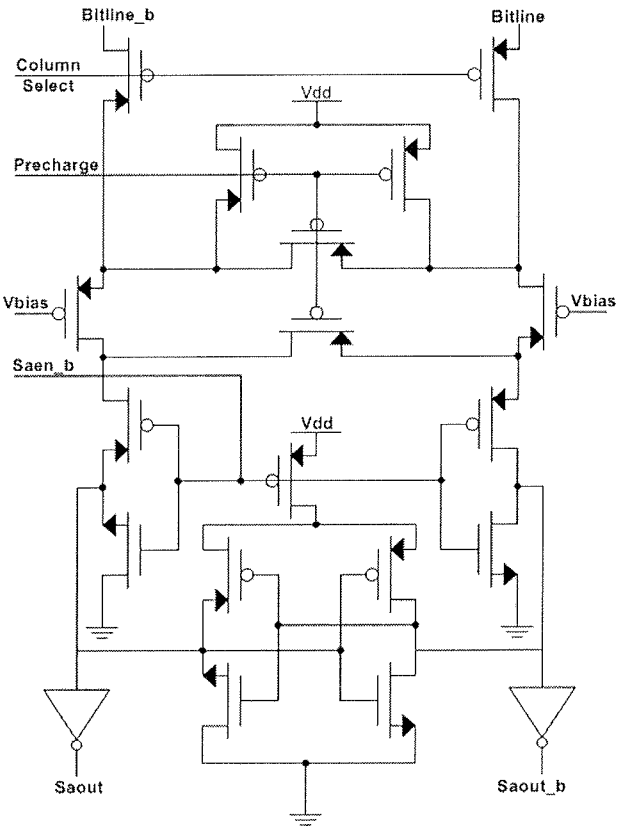


Figure 7 Charge Transfer Sense Amplifier (CTSA)

Supply voltage and threshold voltage are both lowered along with the technology scaling. The immediate effect is higher static power dissipation as the main contributor of leakage current - sub-threshold leakage is increased exponentially (Refer to the sub-threshold drain current I_{Dsub} equation in BSIM3v3 MOSFET model: $I_{Dsub} = I_{s0} \cdot \left[1 - e^{-V_{ds}/V_t}\right] \cdot \left[e^{\left[V_{gs} - V_T - V_{off}\right]/nV_t}\right]$). A new driving scheme which lowers the word-line voltage and raises the ground line voltage is proposed to reduce leakage current in standby mode /17/. Also, Gated-Vdd/Gated-Ground circuit techniques are proposed by Michael Powell et al. and Amit Agarwal et al. to reduce leakage power during standby mode /18/-/19/. The main idea is to disconnect

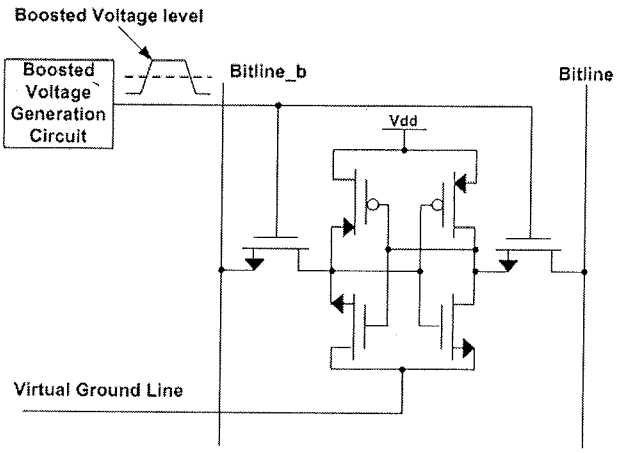


Figure 8 Low Swing Write Technique Proposed by Amrutur

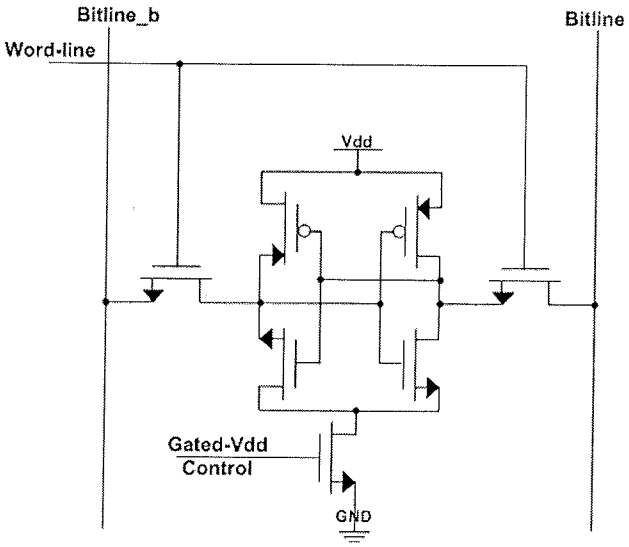


Figure 9 Gated-Vdd/Gated-Gnd Technique for Reducing Leakage

the SRAM cell circuit from supply voltage during standby mode and thus diminish the leakage current. However, an additional NMOS is needed to be put in the leakage path and this increases the layout of memory cell (Refer to Figure 9). Although gated-Vdd NMOS can be shared by multiple memory cells, but the transistor size has to be large enough to sink the current during active mode. Larger size improves the operation speed but worsen the power and area. Other than that, this technique needs no extra circuitry as the decoder itself can be used to control the gated-Vdd NMOS. On-chip Voltage Down Converter (VDC) circuit can be adopted to step down the supply voltage during standby mode /20/. This can greatly reduce the data retention power at the expenses of extra efforts to design the additional circuits and current consumption of the VDC.

4 Conclusion

Low-power high-speed SRAM design must deal with many circuit and architecture issues, including process limitations. The article has focused on difficulties in the design of low-power SRAM as well as the design of high-speed SRAM, presenting a number of power reduction techniques design approaches to increase SRAM's operating frequency.

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/ Masahiko Yoshimoto et al., "A Divided Word-Line Structure in the Static RAM and Its Application to a 64K Full CMOS RAM", *IEEE J. Solid-State Circuits*, vol. 18, pp. 479-484, Oct 1983.

/3/ 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.

/4/ Tan Soon Hwei, "A 160-MHz, 45-mW, Asynchronous Dual-Port 1-Mb CMOS SRAM", *Final Year Project Report - FOE Multimedia University*, April 2005.

/5/ Karandikar Ashish et al., "Low Power SRAM Design using Hierarchical Divided Bit-Line Approach", *Proc. International Conference on Computer Design*, pp. 82-88, Oct 1998.

/6/ Martin Margala, "Low-Power SRAM Circuit Design", *IEEE In. Workshop on Memory Tech., Design and Testing*, pp. 115-122, Aug. 1999.

/7/ Akinori Sekiyama et. al, "A 1-V Operating 256-kb Full CMOS SRAM", *IEEE J. Solid-State Circuits*, vol. 27, pp. 776782, May 1992.

/8/ 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.

/9/ Jinn-Shyan Wang et al., "Low-Power Embedded SRAM with the Current-Mode Write Technique", *IEEE J. Solid-State Circuits*, vol. 35, pp. 119-124, Jan 2000.

/10/ 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.

/11/ Manoj Sinha et al., "High-Performance and Low-Voltage Sense-Amplifier Techniques for sub-90nm SRAM", *Proc. International Conference on SOC*, pp. 113-116, Sept. 2003.

/12/ Jonas Alowersson et al., "SRAM Cells for Low-Power Write in Buffer Memories", *IEEE Symposium on Low Power Electronics*, pp. 60-61, Oct. 1995.

/13/ Bharadwaj S. Amrutur, "Design and Analysis of Fast Low Power SRAM", *Thesis for the Degree of Doctor of Philosophy - Stanford University*, Aug. 1999.

/14/ Bharadwaj S. Amrutur et al., "Fast Low-Power Decoders for RAMs", *IEEE J. Solid-State Circuits*, vol. 36, pp. 1506-1515, Oct 2001.

/15/ Kenneth W.Mai et al. "Low Power SRAM Design Using Half-Swing Pulse-Mode Techniques", *IEEE J. Solid-State Circuits*, vol. 33, pp. 1659-1670, Nov 1998.

/16/ T. Mori et al., "A 1V 0.9mW at 100Mhz 2kx16b SRAM utilizing a Half-Swing Pulsed-Decoder and Write Bus Architecture in 0.25um Dual Vt CMOS", *IEEE International Solid-State Circuit Conference*, pp. 354-355, 1998.

/17/ K. Osada et al., "16.7fA/Cell Tunnel-Leakage-Suppressed 16-Mbit SRAM Based on Electric-Field-Relaxed Scheme and Alter-

nate ECC for Handling Cosmic-Ray-Induced Multi-Errors", *ISSCC Digest of Technical Papers*, pp. 302-303, Feb 2003.

/18/ Michael Powell et al., "Gated-Vdd: A Circuit Technique to Reduce Leakage in Deep-Submicron Cache Memories", *Proc. of the 2000 Int. Symp. On LPED*, pp. 90-95, 2000.

/19/ Amit Agarwal et al., "DRG-Cache: A Data Retention Gated-Ground Cache for Low Power", *Proc. Conference on Design Automation*, pp. 473-478, June 2002.

/20/ Koichiro Ishibashi et al., "A Voltage Down Converter with Submicroampere Standby Current for Low-Power Static RAMs", *IEEE J. Solid-State Circuits*, pp. 920-926, June 1992.

Tan Soon-Hwei, Loh Poh-Yee, Mohd-Shahiman Sulaiman, Zubaida Yusoff
Faculty of Engineering, Multimedia University,
63100 Cyberjaya, Selangor, Malaysia

Prispelo (Arrived): 20.07.2006 *Sprejeto (Accepted):* 30.03.2007