TWO ORDERS OF MAGNITUDE LEAKAGE POWER REDUCTION OF LOW VOLTAGE SRAM'S BY ROW-BY-ROW DYNAMIC V_{DD} CONTROL (RRDV) SCHEME

Kouichi Kanda, Takayuki Miyazaki, Min Kyeong Sik, Hiroshi Kawaguchi, and Takayasu Sakurai

Institute of Industrial Science, University of Tokyo, 4-6-1 Komaba, Meguro-ku, Tokyo, 153-8505 Japan

ABSTRACT

A novel SRAM scheme is proposed that can reduce the active leakage power by two orders of magnitude. In low voltage region less than 1V, the VTH, V_{TH}, is lowered to less than 0.2V and leakage power of memory cells becomes a dominant issue. By dynamically dropping the supply voltage of un-accessed cells row by row, the cell leakage can be reduced exponentially through the Drain Induced Barrier Lowering (DIBL) effect. Additionally, to lower the leakage from bit-line through transfer gates of memory cells, un-accessed word line is applied negative voltage together with reduced swing write technique. The basic advantage is verified by measurement and the effectiveness in future generations is discussed by simulations.

I. INTRODUCTION

Low-power LSI design is strongly required not only for portable electronic systems but also for high-speed systems. SRAM's, which have been widely used as an important fabric of such systems, are getting dominant power consumer because of the large capacity and area. Moreover, it is estimated that 90% of the chip area will be occupied by memory in future system LSI's and the power reduction of SRAM's is increasingly important. The current operation voltage of SRAM's is rather high being more than 1.2V, and the threshold voltage, V_{TH}, of cell transistors are 0.6V. As a result, it is easy to assure the standby leakage power on the order of μW now.

When the channel length becomes less than 100nm, however, the supply voltage, V_{DD}, should be decreased down to 1V or less to assure sufficient reliability of scaled devices. In that case, V_{TH} should be also decreased in order to maintain the performance. Figure 1 shows an SRAM read delay comparison in sub-1V region. As depicted in the figure, if V_{DD} is less than 0.8V and V_{TH} is increased from 0.4V to 0.6V, the delay increases to more than twice. If V_{DD} is less than 0.5V and V_{TH} is increased from 0.2V to 0.4V, the delay increases to more than four times. The figure clearly shows that using high-V_{TH} device is prohibitive due to performance degradation.

The drawback of reducing V_{TH} is the explosion of leakage power. Assuming V_{TH} is 0.3V, leakage current goes up by a factor of 1000 compared with the case of 0.6V V_{TH}. The standby power of a 1M-bit SRAM will be of the order of mW. If V_{TH} is 0.1V, the standby power will be of the order of 1W. Thus SRAM designers should manage the large leakage power caused by low V_{TH} of 0.2V in future SRAM's.

The leakage power is a headache not only in the standby mode but also in an active mode. This is because even in the active mode, almost all the memory cells are in the standby mode and they are consuming the leakage power all the time. Techniques have been reported to suppress the leakage current in logic circuits through the use of power switches and/or substrate biasing. None of them, however, can be directly applied to SRAM cells. For example, since the memory cells should keep the stored data, inserting the power switch is impossible. Moreover, the area overhead of additional circuits should be very small in memories.

Figure 2 shows four leakage current paths in un-accessed memory cell. It should be noted that there are two types of leakage: cell leakage and bit-line leakage. The dominant leakage current is the cell leakage, I_{cell}, which is the sum of leakage current denoted as LP2 and LP3. The leakage current from bit-lines, I_{bit}, shown as LP1 and LP4, however, is also important because the ratio of I_{cell}/I_{bit} is around 10. In order to reduce the total leakage current by two orders of magnitude, both types of leakage should be well suppressed. Therefore, in the proposed scheme, different remedy is taken for each leakage. The following two sections are describing the suppressing methods for the two leakage types in detail.

II. REDUCING CELL LEAKAGE

First, how to suppress the cell leakage power is explained. Let's assume that all transistors in a memory cell have low V_{TH} of 0.1V~0.2V. The straightforward approach to reduce the cell leakage without degrading speed is to increase V_{TH} of transistors of only un-accessed

0-7803-7494-0/02/$17.00 © 2002 IEEE

cells. One method to realize such variable threshold is to take advantage of body bias effects and control the substrate/n-well bias as proposed in [1]. When this technique is applied, however, separation of wells causes relatively large area penalty and the speed degradation by changing the substrate/n-well potential is also an issue. Therefore, the SRAM proposed here utilize another phenomenon specific to deep sub-micron region, called Drain Induced Barrier Lowering (DIBL) effect.

The DIBL effect can be explained by the diagram of surface potential of MOSFET as shown in Fig.3. Two different potential diagrams of an off-state transistor for two different V_{DS} values are shown. When V_{DS} is low, the height of the potential barrier near the source node is increased, and the amount of leakage current decreases compared with high V_{DS} case. It is reported that the leakage current can be reduced to 10% or less by making use of the DIBL effects [2]. Assuming that the leakage current is reduced to 10% when cell V_{DD} is lowered to from V_{DD} to $0.2V_{DD}$, 98% of the leakage power can be saved, since the power is the product of the leakage current and the voltage. It is also reported in [2] that the leakage reduction through the DIBL effect will become larger as device size scales, which makes this technique more attractive.

Low voltage of $0.2V_{DD}$ is generated using a high efficiency DC-DC converter, which is already used in a microprocessor [3]. By supplying low V_{DD} from the DC-DC converter to un-accessed cells, leakage power of SRAM can be suppressed to 2% compared to that of conventional SRAM. When one row is accessed, both word line and cell V_{DD} line are activated. There is one important design consideration. If pass gates of a cell are turned on before the cell V_{DD} is not high enough, the two nodes in the cell storing data are charged from bit-lines through pass gates. In this case, the memory cell operates like a resistor-load type SRAM cell and its noise margin becomes small. Therefore, β ratio of a cell should be a little larger than the 6T cell. The stability of an SRAM cell in very low V_{DD} region is also an important issue, which will be treated in a later section.

III. REDUCING BIT-LINE LEAKAGE

Here, the bit-line leakage reduction techniques are introduced. It has been reported for DRAM's that applying negative voltage to inactive word lines successfully reduces bit-line leakage, since the gate-source bias voltage of pass gates (MT1 and MT2 in Fig.1) are both negative. If a word line voltage is –0.2V, the bit-line leakage is reduced approximately to 1% because it is virtually equivalent to increasing the V_{TH} of the pass gates by 0.2V. This method, however, has not been used in SRAM's because it suffers from degradation of device reliability since the oxide of the pass gate is overstressed. This becomes more and more important in short-channel transistors having ultra-thin gate oxide. In order to avoid the problem, two different techniques are used in the proposed scheme.

First technique is to build a bit-line load with an NMOS transistor as shown in Fig. 4. By the V_{TH} drop of the NMOS, bit-lines are precharged to V_{DD}-V_{TH}, and the oxide of pass gates is never overstressed. When cell V_{DD} is lowered to $0.2V_{DD}$, the gate-drain bias of the pass gate (MT2 in Fig.2) is also assured to be less than V_{DD}. Second technique is to make a write driver with a NMOS pull-up instead of normal PMOS pull-up, which is also shown in Fig.4. In this case, one of the bit lines is driven to V_{DD}-V_{TH} when writing data to a cell. As a result, pass gates connected to inactive cells are not overstressed and thus protected.

In Fig.4, it is assumed that neighboring two rows do not share one cell V_{DD} line. Such a cell layout can be drawn by using two metal layers for cell V_{DD} lines. If a cell V_{DD} line and also PMOS source junctions are shared by adjacent two rows, word line voltage should have triple states. That is because when one row is activated, the pass gates of cells in the neighboring row sharing a cell V_{DD} line should not be negative in order to avoid gate overstress. In that case, bit-line leakage through that row increases, but since there is only one row in the whole memory array, leakage power overhead is eventually negligible.

Though pulling up a capacitive bit line by NMOS require longer time compared with PMOS case, the time overhead is not serious because usually bit lines are precharged to high by the write recovery process. The driving signals for the write buffer in Fig.4 can be tuned such that short-circuit current through the buffer can be minimized.

The negative word line technique is also effective in solving asymmetrical bit-line problem in read cycles reported in [4]. Let's assume that 1024 cells are connected to one bit-line pair, and also that only one cell is storing data "0" and other 1023 cells are storing data "1". If that "0" cell is accessed in a read cycle, cell current should be far larger than the bit-line leakage of 1023 cells to avoid read error. When V_{TH} is low, it is difficult to avoid such a situation without high overhead compensation circuits like the one proposed in [4]. The negative word line scheme requires a negative voltage generator, but the area overhead is less than 0.1% of a chip.

IV. STABLE DATA RETENTION IN LOW-VOLTAGE REGION

Since inactive SRAM cells are operating under very low V_{DD} like 0.2V, V_{TH} fluctuation is to be handled seriously. Above all, V_{TH} unbalance of PMOS and NMOS in a memory cell may ruin the storing function. A V_{TH} mismatch between two pass gates is also important because if there is 0.1V mismatch, it is possible that the stored data is corrupted during reading operation. An SRAM cell, however, has a well-balanced structure and the V_{TH} mismatch on two closely-placed pass gates are less than 30mV. It is verified by SPICE simulation that it does not give rise to an erroneous flip of a cell even in the worst case. On the contrary, even though PMOS and NMOS in a cell are closely placed, V_{TH} mismatch can reach 0.1V if both transistors cause 50mV V_{TH} shift from the target value to opposite directions. In that case, the V_{TH} mismatch amounts up to $0.5V_{DD}$ when V_{DD} is 0.2V. Some V_{TH} compensation scheme is thus needed.

Figure 5 shows the circuit for compensating V_{TH} fluctuations that controls the n-well bias of SRAM cells. The n-well potential is monitored and adjusted so that the leakage of PMOS and NMOS is kept balanced. The concept of this scheme is similar to that of the Dynamic Leakage Control (DLC) SRAM in [1], but in the DLC SRAM, both n-well and p-well are to be biased dynamically, which leads to high area overhead. The n-well monitor circuit is based on sense-amplifying flip-flop topology, and periodically monitors the n-well potential. The power overhead of this additional circuit can be kept 1/10 of the cell leakage using devices with long channel device since speed is not an issue here.

Figure 6 shows a SPICE simulation result of a transfer curve of a CMOS inverter with the proposed n-well biasing scheme just described. It is seen that the trip point remains at the center even when V_{TH} of PMOS and NMOS oppositely shifts by more than 50mV. Without the V_{TH} compensation circuit, the CMOS inverter can no longer operate correctly, but with the proposed n-well biasing, the inverter can still work as a data storing element in very low V_{DD} of 0.2V.

V. MEASUREMENT AND SIMULATION RESULT

The effectiveness of the proposed leakage reduction scheme is demonstrated through SPICE simulation using future MOS models and also verified through measurement. Figure 7 shows the SPICE simulation results of leakage power consumption of one SRAM cell at room temperature at three different technology nodes. The high V_{DD} values used here are 1.0V for 70nm, 0.8V for 50nm and 0.6V for 30nm as are indicated in the International Technology Roadmap for Semiconductors. The low V_{DD} value is set to 0.2V in all cases. In the figure, three different types of leakage power are compared at each technology node. (A) is the conventional SRAM with high cell V_{DD} and non-negative word line. In (B), only cell V_{DD} is lowered to $0.2V_{DD}$, but the word lines are still non-negative. Finally, in (C), both low cell V_{DD} and negative word line are applied.

As shown in Fig.7, the proposed scheme successfully reduces the total leakage power by two orders of magnitude in all technology generations. The importance of reducing bit-line leakage in future SRAM's is also clearly shown from the figure. Figure 8 shows leakage power reduction at 100°C, where V_{TH} of transistors is reduced by approximately 0.2V from that at room temperature. In this case, the DIBL effect still works well to cut off the cell leakage.

Figure 9 shows the dependence of leakage power on cell V_{DD}. It is understood that in order to obtain sufficient V_{TH} increase through the DIBL effect, V_{DDL} should be around $0.2V_{DDH}$.

Figure 10 shows the measurement result of a commodity SRAM chip. V_{DD} is reduced from 5.0V to 1.0V and the leakage power reduction is measured. The leakage current reduction due to the DIBL effects is successfully measured, and the total leakage power at 1V is 99.2% smaller compared with that at 5.0V.

VI. DISCUSSIONS

The area overhead of the proposed SRAM scheme is mainly due to the drivers for word lines and cell V_{DD} lines. Compared with the conventional word line driver, the area for the driver should be doubled. For a 32-bit simultaneous read-out SRAM, the total chip area penalty is estimated to be 6%, and for a 64-bit SRAM, it will be 3%.

Bit-line leakage can also be reduced by increasing ground level (V_{SS}) of a cell. The ground lines, however, are usually structured as a mesh in SRAM's to assure sufficient reliability against electro-migration of the ground lines. Hence, it is difficult to control the ground voltage row by row. Another approach is to insert one NMOS into each cell as is proposed in [5]. In that case, however, the area overhead can be more than 10%. As for the scheme proposed in this paper, there is no change in memory cell itself and hence there is no area penalty in the memory cell itself. The overhead in access time in the proposed scheme is less than 7%, since the pass gates can be turned on without malfunction before the cell V_{DD} is fully restored to

V_{DDH} because the precharged bit-line plays a role of power supply line to the cell.

By dynamically controlling cell V_{DD} row by row (RRDV), charging and discharging power of V_{DD} line itself is consumed. This additional power is estimated to be 3.5 times larger compared with that of a word line. Since power consumed by the word line is typically 2% of the total power in a read cycle for 32-bit read-out case, the power overhead is about 5%. Since by the introduction of the RRDV, the dominating leakage power is reduced by two orders of magnitude, this overhead is negligible.

In summary, Row-by-Row Dynamic V_{DD} Control (RRDV) scheme is proposed and is shown to be effective in reducing SRAM cell leakage by two orders of magnitude. The effectiveness is verified through simulations and measurements. The proposed scheme is promising in future low voltage SRAM's.

REFERENCES

[1]H. Kawaguchi, Y. Itaka, and T. Sakurai, "Dynamic Leakage Cut-off Scheme for Low-Voltage SRAM's," Symposium on VLSI Circuits, pp.140-141, June 1998.

[2]S. Narendra, A. Chandrakasan, D. Antoniadis, S. Bokar, and V. De, "Scaling of Stack Effect and its Application for Leakage Reduction," International Symposium on Low Power Electronics and Design, pp.195-200, August 2001.

[3]K. Suzuki, S. Mita, T. Fujita, F. Yamane, F. Sano, A. Chiba, Y. Watanabe, K. Matsuda, T. Maeda, and T. Kuroda, "A 300MIPS/W RISC Core Processor with Variable Supply-Voltage Scheme in Variable VTH CMOS," Custom Integrated Circuits Conference, pp.587-590, May 1997.

[4]K. Agawa, H. Hara, T. Takayanagi, and T. Kuroda, "A Bit-Line Leakage Compensation Scheme for Low-Voltage SRAM's," Symposium on VLSI circuits, pp.70-71, June 2000.

[5]S. Hattori, and T. Sakurai, "90% Write Power Saving SRAM Using Sense-Amplifying Memory Cell," Symposium on VLSI circuits, pp.46-47, June 2002.

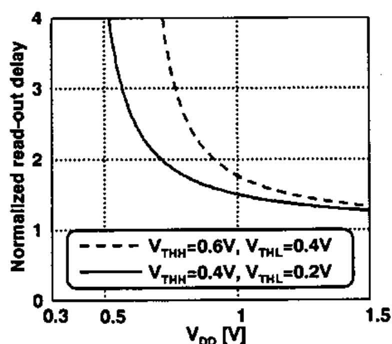

Fig.1 Degradation of read-out delay, D for two different VTHs, V_{THH} and V_{THL} in a low-voltage environment. The vertical axis represents the ratio of $D(V_{THH})/D(V_{THL})$. This graph clearly demonstrates the necessity of use of low VTH to maintain performance in sub-1V region.

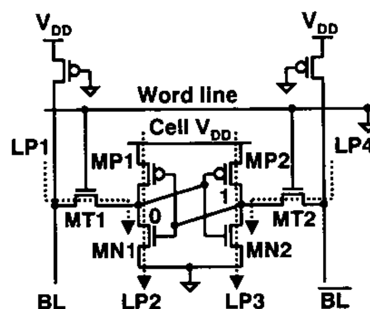

Fig.2 SRAM cell circuit and leakage current paths in standby mode. PMOS, MP1 and NMOS MN2 are assumed to be cut off. LP1 and LP4 are bit-line leakage current, and LP2 and LP3 are cell leakage current.

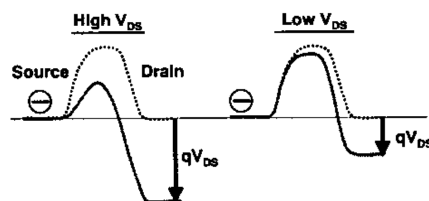

Fig.3 DIBL(Drain Induced Barrier Lowering) effect. Potential diagrams of a transistor with high V_{DS} and low V_{DS} are shown. Applying low V_{DS} reduces the lowering of the potential barrier, which keeps the leakage current smaller.

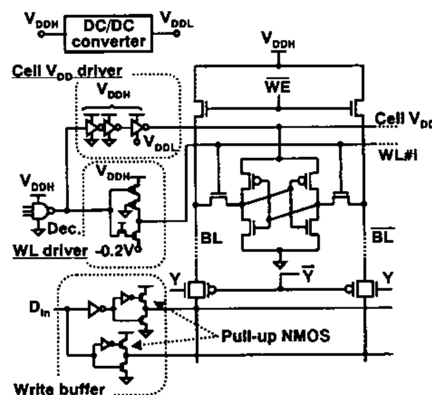

Fig.4 Overall structure of the proposed SRAM. PMOS's in bit-line load and write buffers are replaced by NMOS. A cell V_{DD} line is driven to V_{DDH} when the corresponding

word line is activated, while it stays V_{DDL} when the word line is low.

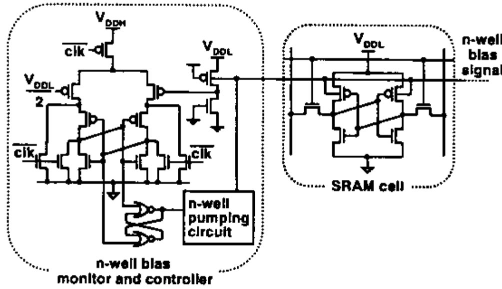

Fig.5 Circuits for VTH mismatch adjustment. N-wells of the whole memory cell array is monitored and controlled so that the threshold mismatch between PMOS and NMOS is minimized.

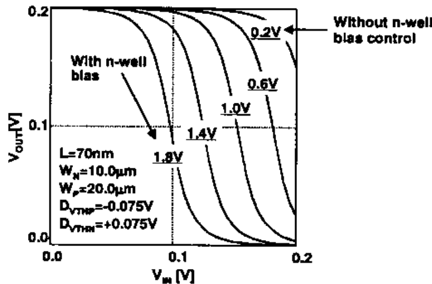

Fig.6 Simulated inverter transfer curve at 0.2V. Without n-well control, normal inverter operation can not be obtained when NMOS's V_{TH} shifts +75mV, and PMOS's V_{TH} shifts −75mV. It is shown that by controlling the n-well bias, balance of PMOS and NMOS is far improved.

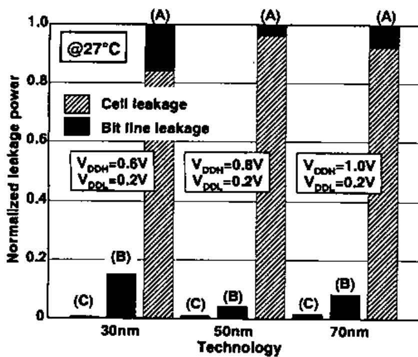

Fig.7 Simulated leakage power reduction at room temperature on three technology nodes. Gate length is 30nm, 50nm and 70nm, and supply voltage is 0.5V, 0.6V and 0.8V respectively. (A) Conventional SRAM, (B) low V_{DD} is applied to all un-accessed cells, (C) negative word line is also applied to (B).

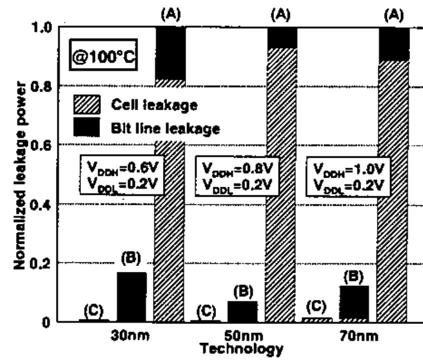

Fig.8 Simulated leakage power reduction at 100°C on three technology nodes. All the parameters are the same as those used in Fig. 7.

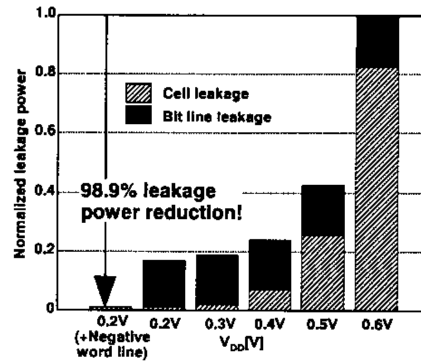

Fig.9 Simulation results showing the relation between cell V_{DD} and total leakage power, when cell V_{DD} is gradually reduced from 0.6V by 0.1V step.

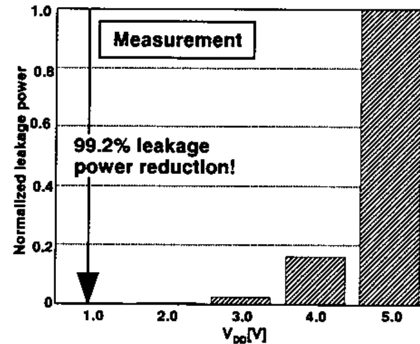

Fig.10 Measured leakage power reduction of a commodity SRAM chip by reducing V_{DD} from 5.0V down to 1.0V. 99.2% of the total leakage power is saved.