A 32-KB HIGH–SPEED 8T SRAM WITH FINE-GRAINED BITLINE STACKING FOR LEAKAGE REDUCTION IN 7NM TECHNOLOGY

Fei Zhou[1], *Xiaoli Hu*[2], *Guoxing Wang*[1]
[1]Department of Micro-Nano Electronics, Shanghai Jiao Tong University, Shanghai, China
[2] GLOBALFOUNDRIES, 7F, Century 333, Shanghai, China
Email: zhoufei2012@sjtu.edu.cn

ABSTRACT

A conventional dual port 8-transistors (8T) SRAM circuit allows an asynchronous reading and writing feature, but its leakage power is significantly higher compared to the conventional 6T cell in the same technology. We proposed transistors stacking technique to reduce leakage power consumption in the fine-grained 8T cell's read bitline. Furthermore, power gating and floating write driver techniques are proposed to further reduce the leakage power consumption in the bank-level wordline (WL) driver. Post-layout simulations show that the leakage power consumption of the 32-Kb SRAM is reduced by 75% in stand-by mode with only 13% read access time increase with 1.5% area penalty.

Keywords—8T SRAM, leakage power reduction, power gating, read bitline stacking.

INTRODUCTION

Growing demand for energy and area efficient Application Specific Integrated Circuit (ASIC) products drives the need for low power embedded Static Random Access Memory (SRAM) [1]. Leakage power consumption has exponentially increased in planar transistor device technology as threshold voltage decreases. While FinFET transistor device technology brings considerable leakage power reduction in SRAM bit-cell and also in peripheral memory circuits [2], it is still a challenge to reduce leakage power [3].

A conventional 6T bit-cell design has read disturb issue, also called read stability, when operating at low supply voltage condition. 8T bit-cell was proposed to overcome the read disturb issue [4], with an independent read port to eliminate read disturbance. 8T bit-cell also provides a feature of reading and writing asynchronously [5]. However, an 8T bit-cell increases the total area by 30% and increases the leakage power by five times compared to the 6T bit-cell in 7nm technology.

In this paper, a transistors stacking technique is proposed to reduce leakage power in the fine-grained 8T cell's read bitline. Power gating and floating write driver techniques are proposed to further reduce the leakage power consumption in the bank-level wordline (WL) driver and write driver respectively.

Our work is based on a 32-Kb SRAM using 8T bit-cell, composing a bit-cell array, sense amplifiers (SA), WL drivers, DATAIO, SA drivers. Additionally, DATAIO includes write driver, address decoder, and output latches.

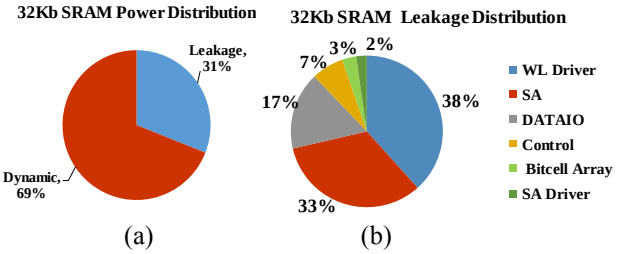

Fig. 1: (a) Total power consumption in a 32Kb SRAM, (b) Leakage power consumption in a 32Kb SRAM.

Fig. 1 shows that this SRAM's leakage power occupies about 31% of the total power. SA, WL driver and DATAIO occupy 88% of the total leakage power. Therefore, we proposed three techniques to reduce leakage power of these three blocks respectively.

PROPOSED TECHNIQUES

Fine-grained Bitline Stacking Technique

The read port of 8T bit-cell causes the main source of leakage power in SA. SA is used to amplify the voltage difference between two bitlines RBL and RBLB in a read operation. NT is the storage node in 8T bit-cell. RPD1 and RPD2 are the read port transistors of 8T bit-cell, as shown in Fig.2.

There is a total of 32 bit-cells on each bitline. Leakage current path at the read port is composed of transistor RPD1 and RPD2. In standby mode, if the node NT stores "1", then the read port transistor RPD2 will be turned on to pass GND to the source of transistor RPD1. Moreover, the current leaks from node RBL to GND terminal with only one switched-off transistor RPD1 as shown in Fig. 2.

32 bit-cells is selected for each bitline considering the tradeoff between area and performance. If using 8 or 16 bit-cells on each bitline, more SAs need to be added in SRAM, occupying more area. If using 64 or 128 bit-cells on each bitline, there will be large capacitance from bit-cells, degrading the read performance.

As shown in Fig. 2, an additional n-type transistor, *Ntail* is proposed to provide a common connection to all read ports of bit-cells on one bitline, creating a virtual ground to all bit-cells. The *Ntail* forms a stack structure

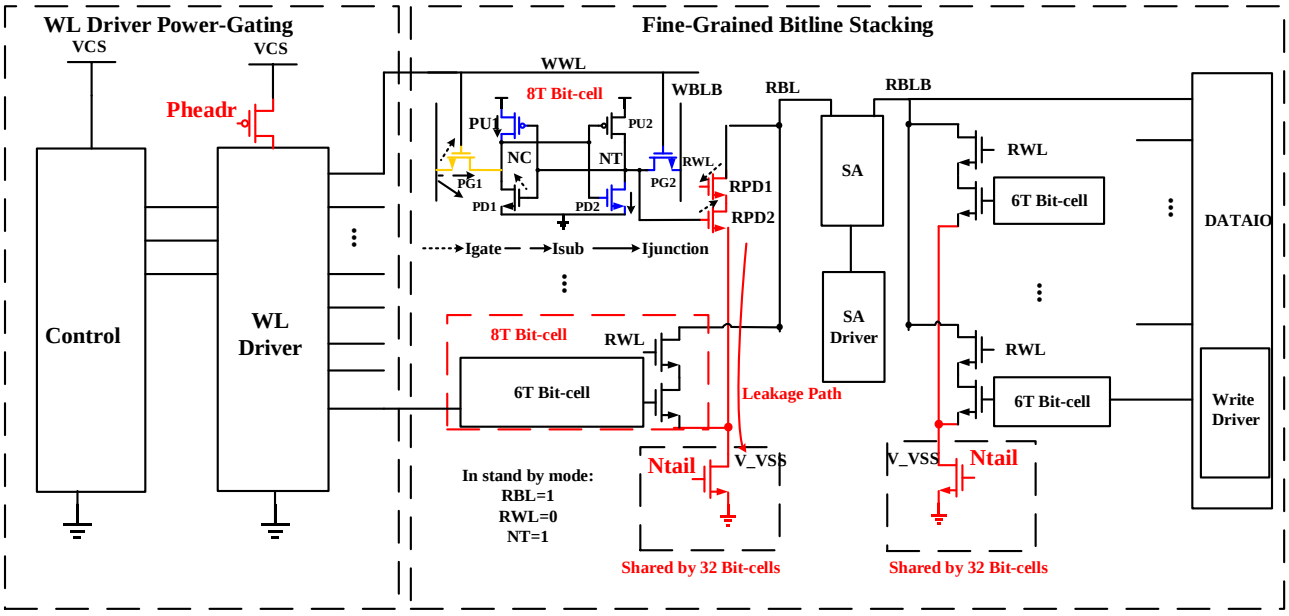

Fig. 2: 32-Kb SRAM architecture, fine-grained bitline stacking circuit, and WL driver power-gating circuit.

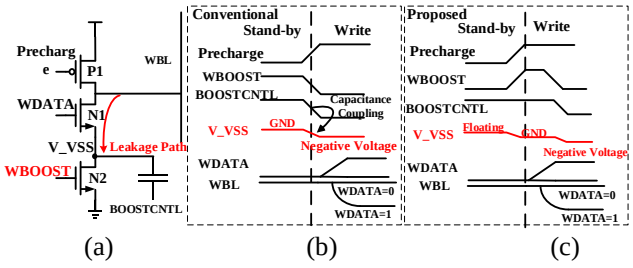

Fig. 3: (a) Write driver Circuit Structure, (b) Conventional negative bitline signal control, (c) Proposed negative bitline signal control.

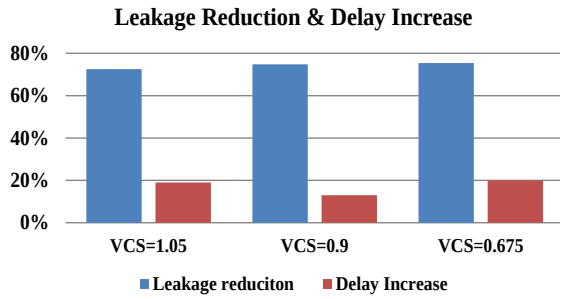

Fig. 4: Leakage reduction and access time increase at high, normal and low voltage.

together with the RPD1 in every bit-cell on the same bitline. During standby state, the *Ntail* is turned off (Vgs=0), so the virtual ground becomes floating. If the leakage current increases from the read port, it will also increase the Vds of the *Ntail*, thus the node *V_VSS* gets higher. This will lead to a negative Vgs in the RPD1 to reduce leakage current.

As there are a total of 32 bit-cells for each bitline, only one bit-cell's RWL is required to be on in one read cycle. One *Ntail* is sufficient to supply one bit-cell's read current and thus, *Ntail* is shared by 32 bit-cells to minimize area penalty.

Ntail also functions as a read mask to determine whether specific bit can be reached. When the SRAM is operating in reading mode, *Ntails* are turned off for masked bits to cut off the read current.

Bank-Level WL Driver Power-Gating

In our 32Kb SRAM design, we have a total of 512 WL and WL drivers. Each driver requires a significant amount of drive current to control the 64 bit-cells in each WL. Hence, the WL drivers are one of the main contributor of the total leakage power consumption.

Power gating technique [6] is adopted in our SRAM to reduce the leakage power consumption of the inactive WL drivers. P-type transistor, *Pheader*, is added between WL driver and supply voltage, and turned off to reduce leakage power when WL is not in use. Significant leakage power reduction is achieved with slight performance degradation.

Floating Write Driver Technique

DATAIO is composed of write drivers, address decoders and output latches. Write driver is a significant leakage power component in all of them.

To operating at low supply voltage condition, write driver uses a technique [7], but it increases the leakage current in the driver.

Node *V_VSS* generates a negative voltage coupled from falling edge of BOOSTCNTL in write cycle as shown in Fig. 3. N2 is switched-on to pull node *V_VSS* to GND in stand-by mode. It causes that the current to leak from WBL=1 to GND with only one switched-off n-type transistor *N1*.

Floating write driver technique is proposed to reduce leakage power as shown in Fig. 3. N-type transistor *N2* is switched-off to generate floating *V_VSS* in stand-by mode. Therefore, *N1* and *N2* form a stacked structure to decrease leakage current [8].

Since the negative bitline assist technique require *V_VSS* to be GND at the beginning of write mode [9], the control signal WBOOST needs to be pulled up before data signal WDATA comes. Additional signal control logic circuits are added to ensure this time margin.

RESULTS AND ANALYSIS

The design was implemented in 7 nm technology. IDDQ MOSFET model is used to measure leakage current accurately.

Using the "fine-grained bitline stacking technique", the leakage power consumption of SA reduces by 89.6% at the cost of 16% read access time. 32 bit-cells share *Ntail* for each bitline, and the incremental area of each bit-cell is only about 1.6%.

Using the "bank-level WL driver power-gating technique", 96.8% leakage power reduction of WL driver is achieved. 4 sub-arrays composes this SRAM, and each one shares one *Pheader*. So this SRAM needs only four *Pheaders*. There are 512 WLs in each sub-array. The incremental *Pheaders* are placed in the edge when designing layout, so there is no area penalty.

There is a 78.7% reduction in the leakage power of write driver using the "floating write driver technique", is achieved. The area of additional signal control logic circuit is negligible.

Fig. 4 summarizes the trade-off between leakage power reduction and maximum operating speed of 32-Kb SRAM using our proposed techniques. When the SRAM is operating at different supply voltage conditions (1.05-V, 0.9-V and 0.675-V), we can achieve a total leakage power reductions about 75% with the increase of access time by 19%, 13%, 20%, respectively.

For area penalty, besides bit-cell array, SA, WL driver, and DATAIO, controller logic circuits also occupy a large percent area of the whole SRAM. The incremental area of whole SRAM by the proposed techniques is only about 1.5%.

In all, the leakage power consumption contributes 9% of the total SRAM power after optimizing the SRAM circuit.

CONCLUSIONS

A 32-Kb SRAM in 7 nm technology is designed with several leakage power reduction techniques. Fine-grained bitline stacking technique is proposed to decrease the large leakage current from the read port of 8T bit-cell. Bank-level WL driver power-gating is proposed to get rid of the significant leakage power from both read and write WL drivers. Floating write drive technique is proposed to get effective leakage power reduction. With the use of our proposed techniques, SRAM achieves 75% leakage power reduction with only 13% performance degradation and 1% area penalty. Leakage power consumption contributes only 9% of the total SRAM power after optimizing the circuit.

REFERENCES

[1] Z. Guo, et al., "A 23.6 Mb/mm 2 SRAM in 10nm FinFET technology with pulsed PMOS TVC and stepped-WL for low-voltage applications," in *Proc. IEEE Int. Solid-State Circuits Conf.,* Feb. 2018, pp. 196-198.

[2] C. Hou, et al., "A smart design paradigm for smart chips," in *Proc. IEEE Int. Solid-State Circuits Conf.,* Feb. 2017, pp. 8-13.

[3] K. Mahmood, et al., "Average-8T differential-sensing subthreshold SRAM with bit interleaving and 1k bits per bitline," *IEEE Transactions on VLSI Systems*, vol. 22, no. 5, pp. 971-982, May. 2014.

[4] L. Chang, et al., "An 8T-SRAM for Variability Tolerance and Low-Voltage Operation in High-Performance Caches," *IEEE J. Solid-State Circuits*, vol. 43, no. 4, pp. 956-963, Apr. 2008.

[5] H. Fujiwara, et al., "A 64kb 16nm asynchronous disturb current free 2-port SRAM with PMOS pass-gates for FinFET technologies," in *Proc. IEEE Int. Solid-State Circuits Conf,* Feb. 2015, pp. 1-3.

[6] Y. Wang, et al., "A 1.1 GHz 12/Mb-leakage SRAM design in 65 nm ultra-low-power CMOS technology with integrated leakage reduction for mobile applications", *IEEE J. Solid-State Circuits*, vol. 43, no. 1, pp. 172-179, Jan. 2008.

[7] T. Song, et al., "A 10 nm FinFET 128 Mb SRAM with assist adjustment system for power, performance, and area optimization," *IEEE J. Solid-State Circuits*, vol. 52, no. 1, pp. 240-249, Jan. 2017.

[8] H. Pilo, et al., " A 64Mb SRAM in 22nm SOI technology featuring fine-granularity power gating and low-energy power-supply-partition techniques for 37% leakage reduction," in *Proc. IEEE Int. Solid-State Circuits Conf.,* Feb. 2013, pp. 322-323.

[9] E. Karl, et al., " A 0.6 V 1.5 GHz 84Mb SRAM design in 14nm FinFET CMOS technology," in *Proc. IEEE Int. Solid-State Circuits Conf.,* Feb. 2015, pp.1-3.

32-Kb High–Speed 8T SRAM With Fine-Grained Bitline Stacking For Leakage Reduction in 7nm Technology

Fei Zhou[1], Xiaoli Hu[2], Guoxing Wang[1]
[1]Shanghai Jiao Tong University, Shanghai, China
[2] GLOBALFOUNDRIES, Shanghai, China

Introduction

Low-power Static Random Access Memory (SRAM)

- Applications: mobile devices, IoT, DL processors.
- Advantage: FinFET transistor device technology brings leakage power ↓
- SRAM leakage power ➔ dominant with the advanced process nodes. .

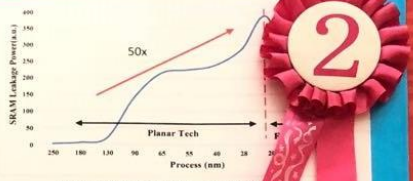

Fig 1. The trend of leakage power in

System Architecture

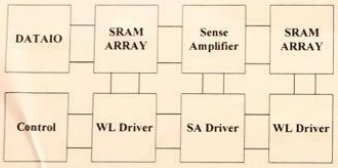

Fig 2. The system block diagram of the SRAM.

Technique #1: Fine-grained bitline stacking

- ✓ N-type transistor *Ntail* is proposed to provide a common connection to all read ports of 32 bit-cells on one bitline, creating a virtual ground to all bit-cells.
- ✓ Function: Determine specific bit can be reached.

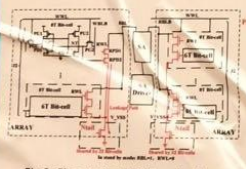

Fig 3. Fine-grained bitline stacking technique.

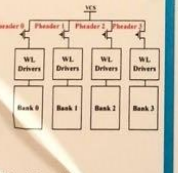

Fig 4. Bank-level WL driver power-gating.

Fig 5. Floating write driver.

Technique #2: Bank-level WL driver power-gating

- ✓ P-type transistor *Pheader* is added between WL driver and supply voltage.
- ✓ It is turned off to reduce leakage power when WL is not in use.

Technique #3: Floating write driver

- ✓ N-type transistor *N2* is proposed to generate floating *V_VSS* in stand-by mode.
- ✓ Pulse generating circuit is used to control floating write driver.

Experimental Result

➢ Power and leakage power distribution

- ✓ SRAM's leakage power occupies about 31% of the total power. Leakage power of SA, WL driver and DATAIO occupy 88% of the total leakage power.

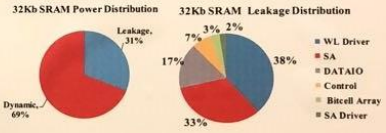

Fig 6. Power and leakage power distribution in 32-Kb SRAM.

➢ Leakage power reduction and read access time

- ✓ With the three proposed techniques: Leakage power of SRAM ⬇ 75%, Area ⬆ 1.5%, performance ⬇ 13%.

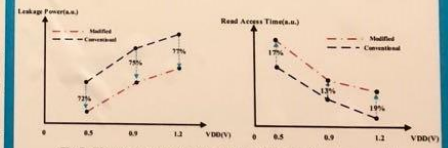

Fig 7. The comparison of leakage power and read access time between conventional and modified SRAM at different voltage.

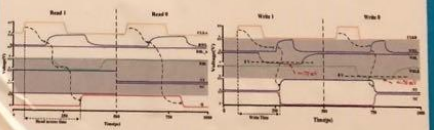

Fig 8. The simulation waveform of 32-kb SRAM in read and write operation.

Conclusion

- ✓ A low leakage 32-Kb SRAM in 7 nm technology is designed with several leakage power reduction techniques.
- ✓ With the use of our proposed techniques, SRAM achieves 75% leakage power reduction with only 13% performance degradation and 1% area penalty.
- ✓ It works well at supply voltage ranging from 0.5V to 1.2V, and, specifically, it is able to work on a frequency as high as 2GHz at 0.9V.