A 16kB Tile-able SRAM Macro Prototype for an Operating Window of 4.8GHz at 1.12V VDD to 10 MHz at 0.5V in a 28-nm HKMG CMOS

Ming-Zhang Kuo¹, Henry Hsieh¹, Sang Dhong¹, Ping-Lin Yang¹, Cheng-Chung Lin¹, Ryan Tseng¹, Kevin Huang¹, Min-Jer Wang¹, Wei Hwang²

¹Taiwan Semiconductor Manufacturing Company, Hsinchu, Taiwan
²Department of Electronics Engineering, National Chiao-Tung University, Hsinchu, Taiwan

Abstract — This paper describes a tile-able 16-kByte 6-T SRAM macro in a High-K Metal-Gate (HKMG) 28-nm bulk technology with an operating window from 4.8 GHz at 1.12 V VDD down to 10 MHz at 0.5V, meeting almost all of the Dynamic Voltage Frequency Scaling (DVFS) requirements of Level-1 (L1) caches of a digital microprocessor SOC. It uses an unmodified technology-supported 0.156um² high-current (HC) SRAM cell. Innovative and carefully optimized circuit solutions provide the wide operating range measured in hardware.

We also discuss two circuit improvements, a cross-coupled PMOS-pair for each bitline-pair with mux readout and an independently-controlled precharge-and-write driver (ICPW), which gives a wider DVFS operating window with reduced sensitivities to Process-Voltage-Temperature (PVT) variations. Improved SRAM macros with new circuits have been designed and laid out and their performance and area verified in simulation.

Index Items — Cross-coupled PMOS-pair, DVFS, independently controlled precharge-and-write driver (ICPW), SRAM

I. INTRODUCTION

With the increasing demand for high performance and low power devices for mobile applications, a dynamic voltage and frequency scaling (DVFS) technique has become widely accepted as meeting the requirements for higher performance and lower power-consumption r in active and standby modes, respectively. A tile-able 16-kByte embedded SRAM macro with a wide operating voltage range has been designed to meet DVFS requirements of L1 memories. Fig. 1 shows high performance SRAMs published in recent years. Compared to other 32-nm designs, a 4% smaller VDD achieves roughly the same operating frequency, which should reduce the power roughly by 10–12% from a steeper-than quadratic dependence of power on VDD in a deep sub-micron technology.

The SRAM macro has hitherto been fabricated in a N28HP technology using the 0.156um² low-VDDM in high-current 6-T SRAM cell [1, 2]. A single-ended sensing (also known as ‘mux readout’) scheme with 16 cells per bitline has been selected to meet L1 speed requirements. There are 140 IOs, supporting a 137-128 single error-correction double error-detection (SECDED) ECC.

Fig. 1. Frequency vs. VDD plot of high performance SRAMs published in recent years. Compared to 32nm designs, there is an approxiamt 10-12 % energy reduction at 4GHz.

II. 16-kB SRAM MACRO

A. 16-kByte Macro Architecture

Fig. 2 shows the overall architecture and pipeline diagram of a 16-kB tile of the tile-able SRAM IP solution. It can be seen that the 16-kB tile is composed of 4 quadrant arrays of 4kB each with 2x8 sub-arrays in each 4kB. Each sub-array has 256 bytes of data arranged in 16 word lines and 140 bitline pairs, including ECC and column redundancy cells. Wordline drivers are located at the edge of each 4kB quadrant array. A decoded wordline-address latch drives the two selected
wordlines, one on each side of the quadrant arrays from the middle. The 16-kB tile has two pipelined cycles; cycles 2 and 3. The first cycle, cycle 2, decodes a word address and distributes the fully decoded wordline address to its WL latch. The decoded signal is latched at the sub-array WL latch, which is located next to the WL drivers of the left and right 4kB blocks. The second cycle, cycle 3 in the 16-kB tile is for an array access. Write data latched in the first cycle is written to a cell in the second cycle. Read data from the array is also latched at the cross-coupled NAND in the same cycle for a read.

B. 128kByte Macro Implementation

The tile-able SRAM solution is based on a 16-kB tile with optional blocks, such as e-fuse, MBIST, store queue, ECC, and thermometer code generator. A fully pipelined two-cycle access design with a 16-kB tile is expected to meet most L1 cache requirements up to 64kB with a minimal increase in the cycle time. A larger L1 memory such as 128kB is easily implementable by adding a cycle before cycle 2 and another after cycle 3 with minimal design effort.

C. Critical Path of 16-kByte SRAM Macro

Fig. 3 shows the critical path of the 16-kB tile. There are 16 cells per local bitline with a single-ended sensing scheme. In the read “0” case, the SRAM cell discharges the local bitline (BLT) and the signal propagates through the NAND2 (ND2) sense amplifier to a global bitline (GBL). Depending on the specific array, a column select and/or redundancy stage follows. The cross-coupled NAND stage converts the dynamic input into a static output. The data is captured and stored in the output read latch. In this scheme, the SRAM bit-cell performance directly impacts the read access time.

D. Cross Coupled PMOS in Local Bit Line Analysis

In a single-ended sensing scheme, PN ratio of NAND2 sense amplifier (ND2) affects the noise immunity and read out speed. A higher ND2 PN ratio gives a faster read out but at the cost of lower noise immunity. The two cross coupled PMOS shown in Fig. 3 are added as a bitline keeper. In read “1” operation, keepers hold BLT at high against the leakage current and noise coupling. In write operations, according to the write data (DLT/DLC), one of the bit lines (BLT or BLC) is pulled down by the write driver and PMOS keepers provide the pull high capability. Some designs [6, 7] improve the read speed by removing the cross coupled PMOS for a lighter bit line loading. Instead of removing keepers on local bitlines, we use cross coupled PMOS keepers with a high ND2 PN ratio, achieving better noise immunity and a faster read speed. Fig. 4 shows that keepers with a higher PN ratio have a faster read-0 speed than those without keepers.

Fig. 4. Local bit line with keeper scheme has better read “0” speed as the PN ratio of ND2 increases.

At a low operating frequency, VBLT is pulled down by bitline leakage current. Fig. 5 shows that VBLT degrades to 39% of VDD at 0.5V at 90°C. VBLT degradation will cause the pull down NMOS to turn on either partially or fully so as to induce leakage current or make the read “1” operation fail.

E. Keeper Size for Low Voltage Operation

The domino circuit in Fig. 3 used to meet the read-path speed requirements requires a keeper, which is called the GBL keeper. There are two keepers in the read path, the GBL keeper and another, which is a dynamic node keeper in the read latch. In general, any keeper size is a tradeoff between a read “0” speed and the glitch in VGBL level precharged VDD.
during a read “1”. A larger keeper size has smaller glitches at the cost of a slower read “0” speed.

Fig. 6 shows the Monte Carlo simulation of the minimum VGBl as a percentage of VDD arising from the glitches during a read “1” vs. keeper size at the technology nominal VDD of 0.85V and 0.5V, illustrating graphically the importance of choosing an optimal keeper size. For example, the minimum PMOS keeper allowed in technology design manuals can keep VGBl at “1” at 0.85V [1, 2]. When VDD is lowered to 0.5V, the strength of PMOS keeper degrades faster than that of NMOS, which makes VGBl discharge to 60.5% of VDD, thus leading to a read “1” fail.

Fig. 6. Monte Carlo simulation of GBL keeper size vs. VGBl level at the maximum glitch during a read “1”. A lower bit yield due to a larger VGBl variation happens with too small a keeper.

III. INDEPENDENTLY CONTROLLED PRECHARGE AND WRITE DRIVER SCHEME

A. Bitline behavior during write in a traditional SRAM

In a traditional write scheme, a write driver is activated after the completion of the precharge, eliminating any potential short current between the precharge PMOS and write driver NMOS. Let us assume the case of a cell storing “0” being written to “1” as shown in Fig. 7(a). The selected cell storing “0” makes the bitline true, the BLT previously precharged to VDD then droops toward the ground from its WL up to the activation of the write driver, which starts to pull BLT up toward VDD. The magnitude of this droop depends on the bitline loading as shown in Fig. 7(a), making a cell write time longer. The impact on the writing time of a cell increases drastically if the droop is large enough to activate the cross coupled PMOS pair. A new precharge and write scheme called IndeXedly Controlled Precharge and Write driver (ICPW) is therefore proposed to solve this issue.

B. Independently Controlled Precharge and Write Driver

Fig. 8 presents a schematic view of the ICPW. The four transistors (TP, CP, TN, and CN) are controlled independently during the precharge and write stage. TP and CP are on with TN and CN off in the precharge stage, precharging BLT and BLC to VDD. During a read, however, all four transistors are off, floating the bitline. In a deep sleep mode, bitlines float with all four transistors off as in a read. Writing a “0” or “1” is done by activating TN and CP with CN and TP off, and vice versa. Note that the precharge of any bitline is simultaneously terminated when a write happens, eliminating any timing window requirement between the precharge release and write. Additionally, bitlines can be driven by write drivers before the activation of WL, resulting in a fast write time. Detailed simulation based on the post-layout extracted netlist of the 16-kB tile with ICPW shows an approximate worst-case-write time improvement of 50 ps. At the time of submission of this paper, a 16-kB tile with ICPW is in the queue for the next available MPW run.

Fig. 7. Stylized bitline and WL waveforms. (a) Traditional write; (b) independently controlled Precharge and Write driver (ICPW).

Fig. 8. ICPW schematic and operating principles. The states of ICPW devices for various cases are shown. O represents on and X off.

IV. SILICON RESULTS

Fig. 9 shows a die photo of the 16-kB tile with no implementation of an ICPW and bitline cross-coupled pair. The silicon results in this paper are from the version without these components.
Data-in registers, control logic and decoder are located at the bottom of the floor plan. Read-out and write-in buffers are at the center of the macro area. A group of wordline drivers sits on the left side of each 4-kB mirrored-in-Y sub-array, two of which share the same WL latch. There are 16 word lines and 140 bit lines in each sub-array. Both sub-arrays will share the same single-ended NAND2 sense amplifier. The high-current 0.156um² bit-cell in 28-nm HKMG CMOS technology is used for implementation. The area of a 16-kB tile is 0.051mm², including redundancy and ECC cells. It is a tileable SRAM IP, primarily for L1 cache of embedded CPU and GPUs. The complete set of features is shown in Fig. 9.

The 16-kB IP is implemented in a universal high frequency testing structure [8]. In the test chip, we reserved four monitor pads to measure on-chip VDD and VSS as shown in Fig. 10. The measured mean and peak-to-peak of VDD at the monitor pads are used to calculate VDD at the circuit. Figs. 10 and 11 show the high frequency and low VDD shmoo, respectively. The 16-kB macro operates at 1.92GHz at 0.72V, 4.8GHz at 1.12V, and 0.5 V at low operating frequency.

V. CONCLUSION

We have presented a 16-kB SRAM IP fabricated in TSMC HKMG 28nm with a 6T high current low VDDMIN SRAM cell. The IP is fully pipelined with a two-cycle access and is designed mainly for a L1 cache. However, much larger density IP is implementable using 16-kB SRAM tiles. This supports a 137-128 ECC scheme and column redundancy and works up to 4.8GHz at 1.12V and 1.92GHz at 0.72V although it can also operate at a low frequency of 0.5V.

The paper is focused on the low-voltage design considerations and optimization needed for DVFS. Cross-coupled PMOS keepers in a local bit line, a higher P-to-N ratio NAND2 as the local readout circuit, optimizing the global bitline keeper size, and a new concept of ICPW have showed in simulation enhanced noise immunity and faster read-write speed.

REFERENCES