Ultra-High Density 3D SRAM Cell Designs for Monolithic 3D Integration

Chang Liu and Sung Kyu Lim
School of Electrical and Computer Engineering, Georgia Institute of Technology, USA

Abstract—This paper presents design options for 3D SRAM cells to enable ultra-high density 3D SRAM based on monolithic 3D integration. Our target technology offers one tier of NMOS devices, another for PMOS devices, and nano-scale inter-tier vias. Choosing the most compact 22nm 2D SRAM as our design baseline, we first achieve a 33% footprint area reduction by simply splitting the NMOS and PMOS devices in the 2D cell into two tiers. We then explore several alternative design options that fully take the advantage of monolithic 3D technology under area and reliability goals. Our options include (1) the conventional 2P4N cell with new sizing, (2) 3P3N cell with voltage enhancement, (3) 4P4N 8T cell. We achieve 44% and 45% area reduction with the first two and 40% with the last, all under high static noise margin, data retention voltage, and write margin under both nominal and statistical conditions.

I. INTRODUCTION

In today’s VLSI systems, embedded SRAM usually occupies large chip area (> 30%). Therefore, design of high density and reliable SRAM is becoming increasingly critical. 3D integration is an effective approach to reduce the chip footprint and increase the density. However, traditional TSV-based 3D technology is proved not suitable for 3D SRAM cell because of the prohibitively large TSV size. In comparison, monolithic 3D technology enables such tighter alignment precision of the strata and the nano-scale inter-tier vias, that it offers unparalleled opportunities for ultra-high density 3D SRAM compared with TSV-based approach.

Currently, several monolithic 3D integration processes have been proposed by CEA/LETI, Samsung and Rajendra[1][2][3]. And several works have also been proposed to investigate the monolithic 3D SRAM cell targeting two-tier monolithic 3D process that allow high density implementation under high stability operation.

The monolithic 3D technology used in this study is similar to the CEA/LETI process but targeted towards a 22nm technology [1]. Our 3D structure is shown in Figure 1. CEA/LETI showed [4] that regular 2D performance is achieved with their monolithic 3D process for a 3D structure is shown in Figure 1. CEA/LETI showed [4] that regular 2D performance is achieved with their monolithic 3D process for a

II. BASELINE 2D AND 3D SRAM CELL DESIGNS

We choose the recently reported SRAM design reported by IBM [6] in 22nm technology as our design baseline. We replicate the layout in [6] using the design rules provided in that paper, and achieve 0.1μm² area. The size of transistors used in this design (as well as others) is listed in Table I.

An obvious way to design the 3D versions of this 2P4N SRAM cell is to split the NMOS and PMOS transistors into separate device tiers. Figure 2(b) shows the layout of this 3D SRAM cell, where NMOS is located at the top tier and PMOS at the bottom. Each SRAM cell uses 4 internal inter-tier vias. The area saving of this 3D design compared with the 2D baseline is 33%. The reason why the area saving cannot reach the ideal 50% is twofold. First, the NMOS and PMOS counts are not balanced (4 vs 2). Second, the NMOS size is bigger than that of PMOS. The following sections present our solutions to these problems.

III. 2P4N 3D SRAM CELL DESIGN WITH NEW SIZING

Since simply slitting the NMOS and PMOS devices is not enough in terms of area reduction, we need to explore other design options that are suitable for monolithic 3D process. To facilitate comparison, we also simulate and present several key stability metrics on our designs, including static noise margin (SNM), write margin (WM), and data retention voltage (DRV) both under nominal and statistical conditions. Our statistical simulations consider Vth variations for each transistor using Monte-Carlo simulations.

To overcome the NMOS vs PMOS size skew, we propose a different sizing approach, where all NMOS sizes are decreased while the PMOS sizes are increased, as shown in Table I. Figure 2(c) shows the layout of this design option. With this new sizing option, we achieve an area reduction of 44% compared with the 2D SRAM.

One merit of this sizing approach is that the SNM can be maintained the same as the 2D design because the read voltage (Vread) and the inverter trip point (Vtrip) increase simultaneously. We see from Figure 5(b1,b2) that SNM is 224.4mV with 38mV standard deviation, which is basically the same as the 2D design. The major drawback...
of this new sizing approach is that the write stability is worse than the 2D design (220.7mV vs 190.9mV) because the PMOS becomes stronger while the pass transistor becomes weaker. To maintain the same writing stability, writing enhancement techniques will be required.

IV. 3P3N 3D SRAM CELL DESIGN

As discussed in Section II, the NMOS and PMOS count skew is another reason that prevents further 6T SRAM area reduction. A 5T SRAM structure can overcome this NP count skew problem. However, it is difficult to achieve a good write-ability because it only performs single-ended write. Therefore, we propose a new 3P3N SRAM cell structure, where one of the NMOS pass transistors is replaced with a PMOS transistor as shown in Figure 3. After this replacement, we need two word lines $W_L$ and $W_L'$ for NMOS and PMOS pass transistor separately. The layout of the two device layers becomes quite symmetric as shown in Figure 2(d). Table I shows the size of each transistor used in this design. The area reduction compared with 2D design is 45%, which results in a footprint reduction similar to the resized 2P4N case.

However, the replacement of NMOS pass transistor with PMOS pass transistor has significant impact on both read and write stability as well as the operation mode. For read operation, if we perform normal read as the original structure, two bit-lines are charged to Vdd and the two pass transistors are turned on by setting $W_L$ and $W_{L'}$ to Vdd and 0, respectively. As shown in Figure 3(a), $V_{gs}$ of PMOS pass-transistor M6 remains at Vdd constantly, and M6 is fighting against the pull-down NMOS M2, resulting in a significant rise at node N1, which will harm the read-stability. To solve this read-stability problem, the 3P3N needs single-ended read through the NMOS pass transistor M5 while turning off the PMOS pass transistor. In this case, the read operation is the same as a 5T SRAM cell, which has a very good SNM performance. Figure 5(c2) shows the SNM Monte-Carlo simulation results of our 3P3N design. This SNM is 265mV, which is better than the original 2D design (218.1mV) as expected.

In write operation, the PMOS pass transistor is turned on to help write in the value. Therefore, it has a better write-ability than 5T SRAM cell whose write operation is only through one side. However, compared with the original 2D 2P4N cell, the write-ability is degraded. Assuming the original state is $N_0=0$ and $N_1=1$. As shown in Figure 3(b), to flip the state, BL0 is set to 0 and BL1 is set to Vdd, and both NMOS and PMOS have difficulties in flipping the state because they are both in source-follower connection. Therefore, a write enhancement technique is needed in this case, including (1) word-line voltage boosting for NMOS pass-transistor, (2) negative word-line voltage for PMOS pass-transistor, and (3) lower cell Vdd. Figure 5(c3) shows the WM Monte-Carlo simulation results using these write enhancement techniques. We achieved a WM comparable with the original 2D design with several voltage configurations.

V. 4P4N 3D SRAM CELL DESIGN

As the nano-scale technology drives toward lower voltages, the 8T structure shown in Figure 4(a) becomes a viable competitor with the current 6T SRAM structure due to its excellent read-ability [7]. Table I shows the size of all the transistors in the original 2D 2P6N 8T SRAM cell. Figure 2(e) shows the layout. Compared with 2D 6T design, the SNM increases by 130%. The cost is that the area increases by 31% compared with 2D 6T structure. The area increases mainly because of the routing constraints. In the 8T structure, two global lines (RBL, RWL) are added along with another connection to Gnd, which causes routing congestion.

The original 8T SRAM cell has 6 NMOS and 2 PMOS transistors, which clearly will not result in an efficient use of space for 2-layer monolithic 3D. Therefore, we propose a new 8T structure where the
added 2 NMOS transistors are replaced with PMOS, as shown in Figure 4(b). After this replacement, the NMOS count and PMOS count are both 4. The major operation difference from the original 8T structure is that during read operation, RBL is set to 0 and RWL is set to 0 to turn on the pass transistor M7. Figure 2(f) shows the 3D layout of the new 4P4N 8T SRAM, the area reduction compared with the 2D 2P4N version is 40%. Since the transistor sizes except for the added transistors are the same as the original 8T structure, it has the same SNM, WM, DRV performance as the 2D 8T structure. Since we use a larger size for PMOS read-assist transistors, the access time is remained the same.

VI. CONCLUSIONS

Monolithic 3D processing provides new opportunities for transistor level 3D SRAM design. To further take the advantage of this technology, we proposed, optimized, and analyzed various SRAM design options. Our 3D designs based on 2P4N and 3P3N structure reduce the overall footprint by up to 44% and 45% compared with the conventional 2P4N 2D SRAM while maintaining comparable read/write stability. However, the 3P3N needs more read/write assistant techniques than the 2P4N structure to maintain read/write stability. Therefore, between these two 6T 3D structures, the 3D oriented 2P4N structure is preferred. In addition, our new 3D structure based on 8T (=4P4N) SRAM results in a 40% area reduction compared with the standard 2D 8T SRAM while maintaining the same read/write stability.

REFERENCES