

Embedded memory evolves

IN THE SEARCH FOR ON-CHIP RAM, SOI TECHNOLOGY OFFERS A NEW ANSWER.

Ever since its invention in the 1970s, DRAM (dynamic RAM) has been the most popular form of semiconductor memory. Indeed, modern computing and communication systems could not exist without DRAM. The first volume-production device had a capacity of only 4 kbits, but, within 30 years, the capacity has increased a millionfold. At about the same time that DRAM emerged, the microprocessor appeared, evolving from early calculator chips. This development quickly led to the development of the PC. When people started trying to use these new marvels, two limitations became apparent: The processor chip needed to be faster, and more memory was necessary. The same pressure for processor speed and memory size came from the communications sector, with the result that logic and memory have followed two divergent evolutionary paths.

LOGIC AND MEMORY DIVERGE

The drive for speed in processors led to the development of logic processes to produce transistors with fast-switching characteristics. Power consumption and, to some extent, cost were not primary considerations. Process scaling enabled vast numbers of transistors to find use in designs, and the lower voltages these designs employed mitigated the rise in power consumption. But, with fast transistor switching came leakage, which, together with low-voltage operation, made the inclusion of DRAM in ASICs effectively impractical. In recent years, the upward trend in switching speed has reached a saturation point, and processor designers have employed architectural developments such as multiprocessor configurations to achieve greater performance. However, one recent technological development—the move from bulk silicon to SOI (silicon-on-insulator) structures—has provided a step-function increase in processor performance. Following this trend, AMD (Advanced Micro Devices, www.amd.com) has announced the move to SOI for its new processor families (Reference 1).

While processors were chasing speed, DRAM continued to emphasize size. Every couple of years, manufacturers were introducing a new process reduction. They developed the trench and stacked capacitors to cram more memory cells into a unit area. But the manufacturers emphasized size and cost, rather than performance. So, memory processes became more complex for each generation, diverging from logic processes. As time passed, memory and logic processes became more and more incompatible.

Fast-forward to the 1990s with the introduction of digital

phones, digital TV, and the Internet, spawning a massive demand for sophisticated personal electronics. The use of complex ASICs, almost always containing a processor and memory, made these products economically viable. The high-performance processor communicated with on-chip SRAM (static RAM) because SRAM macros were both fast and compatible with logic processes.

Now, you need to know a little more history. Because DRAM development had gone down the large and cheap but slow path, DRAM couldn't match the speed demands of processors. Hence, the concept of cache memory evolved. Small and fast, cache memory stores a segment of the information that the bulk DRAM contains, with clever logic determining which data the processor is most likely to require next. Cache copies the selected "snapshot" of data from bulk memory before the processor makes the next memory access.

So, designers of cache memories employed the "forget-expense; make-it-as-fast-as-possible" principle. Technology limitations at the time meant that cache memories were not only expensive, but also small—just a few kilobytes. It was impossible to make small DRAMs because of the overhead

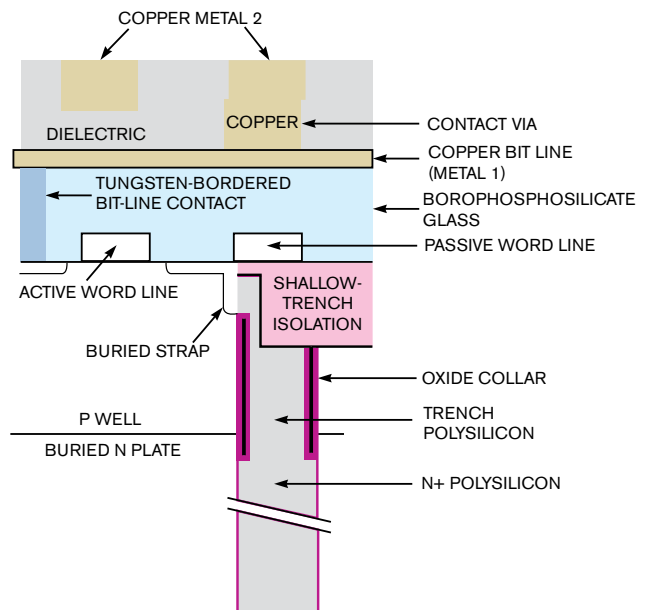


Figure 1 IBM's embedded DRAM stores charge on the walls of a deep-trench structure.

of the refresh logic necessary to restore the data that would otherwise fade away after a few milliseconds. Therefore, manufacturers made cache memories using SRAMs, which required no refresh. The application in cache memories pushed SRAM development along the same evolutionary path as microprocessors. They had to be fast, so they thrived on the same logic-process improvements as microprocessors; this co-evolution meant that they remained process-compatible.

JOINING DRAM AND LOGIC

In the 1990s, designers were building ASICs using embedded processor cores and embedded SRAM. They used SRAM because they had always done so and because they believed that DRAM was much slower and incompatible with ASIC processes. However, researchers at IBM (www.ibm.com) suspected that these beliefs were not necessarily true. They postulated that designers could build DRAM to be almost as fast as SRAM and to occupy less chip area (Reference 2).

In parallel, researchers at the IMEC (Interuniversity Micro-Electronics Center, www.imec.be) in Leuven, Belgium, were studying a parasitic phenomenon that occurred with SOI devices. Because manufacturers fabricated the MOS transistors in electrically isolated islands, they could accumulate charge. This charge, or “floating-body” effect, influenced the current that would flow when you apply voltage to the gate. The researchers concluded that they could harness this effect to make a memory device—DRAM—and filed a patent. They encountered problems, however. The effect was difficult to control, and, when they fabricated a memory array, activity in one cell tended to cause adjacent cells to switch. The technical hurdles at the time were too formidable, and the researchers let the patent lapse.

Now, fast-forward to 2000 and beyond. IBM had figured out how to make fast DRAM and make it compatible with logic processing. The company started to offer embedded DRAM as part of its ASIC portfolio. NEC (www.necel.com) has taken the same tack. Meanwhile, Serguei Okhonin, PhD, and Pierre C Fazan, PhD, had cracked the problems of making floating-body-effect DRAM work. The two men formed a company, ISI (Innovative Silicon Inc, www.innovative-silicon.com), to develop and exploit the invention, which they patented as Z-RAM (zero-capacitor RAM). In early 2006, AMD licensed the ISI floating-body-effect-memory IP (intellectual property).

Now, two approaches are available for building high-performance processors and ASICs: Either build in bulk silicon CMOS with compatible embedded DRAM, or build in SOI using the floating-body-effect DRAM. Many engineers’ immediate reaction to this choice is to point out that SOI wafers are 10 to 15% more expensive than bulk silicon, making any use of SOI appear more expensive. However, several facts about SOI design make this observation incorrect in practice.

First, because of the nature of SOI circuits, a given design can occupy less die area and operate as much as 35% faster than an equivalent design in bulk CMOS. Second, die yield

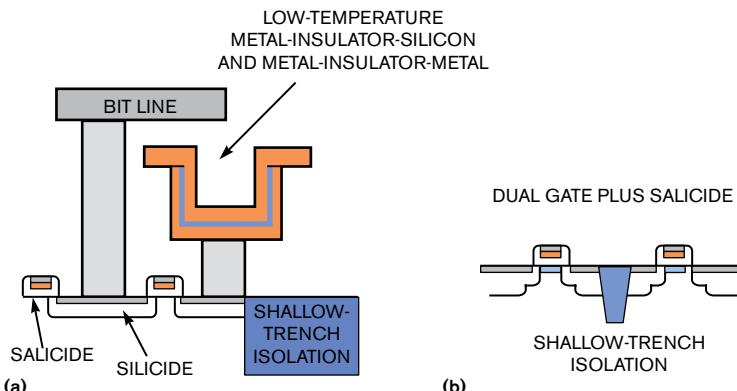


Figure 2 NEC’s stacked-capacitor embedded DRAM (a) uses a metal-insulator-metal capacitor above the silicon in the interconnect stack (b).

falls by approximately one-fourth the power of the die size, so any die-size reduction immediately translates to a major saving in cost. Clearly, these two factors work together on cost for good dice. ISI estimates that, if embedded memory occupies more than 18% of the die area, the smaller die resulting from using Z-RAM generates a cost saving that counterbalances the extra cost of SOI wafers and processing. The bigger the memory, the greater the cost advantage—along with the speed and power improvements of SOI.

And the comparison to bulk CMOS is itself not quite accurate. Although manufacturers can make embedded DRAM process-compatible with logic, the DRAM still requires a trench or stacked capacitor to provide sufficient charge storage for reliable operation. Fabricating this capacitor requires additional process steps, some of which are critical. But floating-body-effect DRAM on SOI requires no additional capacitor and no additional process steps. All this approach requires is standard logic processing.

KEY MEMORY PARAMETERS

Memory has become the largest single component in modern high-performance processor-chip designs. Hence, choosing the optimum design compatible with high-performance logic is critical to a new design’s success. Designers must consider many parameters, including cell size, standby power, SER (soft-error rate), and performance.

Cell size: Dynamic cells tend to be five to seven times smaller than six-transistor static cells. However, using DRAM entails the overhead of voltage-level shifters and refresh circuitry. This overhead causes the area of dynamic circuitry to be only three to four times smaller than that of six-transistor cells.

Standby power: With SRAMs, background leakage current was once negligible. Unfortunately, with each process generation, the “device-off” or subthreshold leakage doubled to the extent that designers now must contend with static-memory cells that exhibit approximately 1000 times more leakage current on a cell-per-cell basis than dynamic cells. However, dynamic cells require level shifters and consume a refresh current. When you take these factors into account, embedded DRAMs tend to have a six- to eight-times power advantage over embedded SRAM.

SER: An SRAM would once hold its data without error

indefinitely as long as the system maintained power. This situation is no longer always the case. A soft error occurs when a single bit flips to the opposite state, resulting in data corruption. High-energy particles, such as cosmic rays, which are constantly in the background in our environment, cause these errors. The susceptibility of a memory to soft errors is a function of the stored charge maintaining a given logic state in each cell. With embedded DRAMs in submicron technology, each cell holds approximately 20 times more charge per bit than SRAMs. Hence, embedded DRAMs tend to have as much as 1000 times fewer SER-induced failures than those in SRAMs, according to NEC (Reference 3). With SOI technology, no conduction can occur through the substrate; hence, embedded DRAM on SOI exhibits significantly better SER performance than bulk CMOS DRAM. You can mitigate the SER problem by including error-correction circuits within the memory array. However, these circuits cost silicon area, and the occurrence of multiple simultaneous “hits” can overwhelm them.

Performance: DRAM is not inherently slower than SRAM, but DRAM is unavailable during the refreshed cycle. So, when constructing a small memory with a simple interface, it would be natural to use SRAM. Several vendors now offer DRAM IP with built-in refresh and SRAM interfaces, so, for ASIC designers, it is no more difficult to use a DRAM block than an SRAM of the same size. When choosing between SRAM and DRAM, you must consider many density and performance trade-offs; however, the most important parameter is latency. “Latency” refers to the time interval between presenting a memory macro with a random data access and when that information becomes available. For a large ASIC containing a multiple megabyte RAM, the dominant element in latency is not the time to read the cells, but the time for the data request and its response to travel through the interconnect. Using a smaller die reduces this latency. In this case, embedded DRAM wins because, for a given memory size, DRAMs are three to four times smaller than SRAM. As an example, substituting DRAM for a 9-Mbyte embedded SRAM results in an area reduction of about 70% and a worst-case latency—that is, the longest signal path—reduction of more than 30% (Reference 4).

The field of bulk-CMOS embedded DRAM has two technology camps: the trench capacitor, which IBM champions, and the stacked capacitor, which NEC and TSMC (Taiwan Semiconductor Manufacturing Co, www.tsmc.com) advocate. ISI is currently the sole supplier of commercial SOI floating-body-effect-memory IP. Other suppliers have R&D efforts under way. For example, Renesas (www.renesas.com) is developing an SOI floating-body-effect device, and Toshiba (www.toshiba.com) is working

on both floating-body-effect memory and FERAM (ferroelectric RAM).

TRENCH- VERSUS STACKED-CAPACITOR DRAM

IBM early on discovered the difficulties of combining DRAM with ASICs and labored through the 1990s with a one- to two-generation penalty in ASIC-logic performance when it combined ASICs with DRAM. IBM claims that this problem no longer occurs with its 130-nm and smaller process geometries (Reference 4). The company developed its Blue Gene/L supercomputer chip in 130-nm technology using embedded DRAM for the Level 3 cache. The company achieved 3.3-nsec, 300-MHz operation for the memory array, well within the design requirement of 250 MHz. Designers estimated that if they had fabricated the Level 3 in SRAM, it would have occupied 66% of the die. The area saving they achieved by using DRAM equated to a cost saving of 40%, without any compromise in performance.

The production of the trench capacitor is the first stage in wafer processing (Figure 1). The embedded DRAM follows the layout rules of the technology, and, after the deep trench processing, a planar wafer surface presents itself for designers to build the logic stages. Hence, the presence of the DRAM has minimal impact on the logic yield. The 130-nsec process requires four additional masks to implement the DRAM. IBM anticipates that future generations of embedded DRAM can achieve cycle times of less than 2.5 nsec and data rates of greater than 1.5 GHz and still maintain three- to four-times better density than SRAM.

NEC uses a stacked-memory architecture for its ML embedded DRAM with a zirconium-dioxide, high-K dielectric (Reference 5). This process includes an MIM (metal-insulator-metal) stacked capacitor to avoid the high-temperature process steps that conventional SIS (silicon-insulator-silicon) stacked DRAMs require (Figure 2). The low-temperature processing minimizes degradation of the logic performance and enables the DRAM to be compatible with standard logic CMOS, although this approach requires eight or nine additional masks.

The company is currently using a 90-nm process with a DRAM-cell size of 0.22 microns square and is moving toward 65- and 45-nm processes.

The difficulty with producing a stacked capacitor is that you must fabricate it in the middle of the process between the active devices and the wiring. Also, you must use novel materials to build the complex 3-D structure. Furthermore, additional planarizing steps are necessary to avoid the yield problems that would arise if metal tracks had to negotiate over tall capacitor structures. At a 90-nm process, NEC can incorporate as much as 256 Mbits of embedded DRAM on a 15×15-mm die; assuming that the memory occupies half the silicon area,

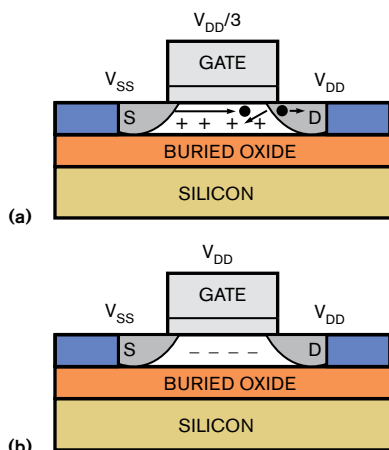


Figure 3 A floating-body-effect DRAM (a) stores information in the charge that becomes trapped in the body of an MOS SOI transistor (b).

NEC claims this method achieves speeds greater than 250 MHz. Nintendo (www.nintendo.com) selected NEC as the supplier for a complex system LSI chip for the GameCube product, which also incorporates memory IP from MoSys (www.mosys.com). Manufacturers have built millions of devices for this application.

MORE AT EDN.COM

Go to www.edn.com/ms4226 and click on Feedback Loop to post a comment on this article.

ANOTHER ALTERNATIVE

The issue confronting both the trench- and stacked-capacitor makers is that these capacitors are the final steps in an evolutionary path of 33 years that has taken them from planar capacitors in 8-micron technology to complex 3-D structures in 90-nm processes. The aspect ratio of these capacitors now exceeds 30-to-1, and the practical limit of fabricating capacitors in this way is looming. Floating-body DRAM, on the other hand, eliminates the need for a separate capacitor and provides a technology path that researchers have demonstrated experimentally at the 22-nm level (**Reference 6**).

Floating-body DRAM uses a MOS transistor instead of a capacitor to store each bit of data. The key point is that the memory employs an SOI wafer. Designers construct each transistor on its own silicon island and electrically isolate the transistors from one another, enabling each one to accumulate a charge. The quantity of this charge affects the current that flows through the transistor when you apply a voltage between the source and the drain (**Figure 3**). You can sense this difference in current and use it to indicate the storage of a one or a zero. To store a one in a cell, you apply the charge to the body of the transistor using impact ionization. This process is similar to the technique that flash memories employ, except that impact ionization is faster and uses less energy. Also, flash memories retain stored data for a long time, whereas Z-RAM loses its data after a few milliseconds if you do not refresh it. The much lower capacitance of the floating-body cell has additional implications. Because there is no bulk capacitance, drive and decoding circuitry can be faster. And the smaller capacitance means that refresh requires significantly less energy than with a capacitor-based DRAM.

Whereas each storage node in a conventional DRAM requires a capacitor and an access transistor, each node of a floating-body DRAM requires only one NMOS transistor. This simplicity enables the devices to achieve double the density of standard embedded DRAM at a 90-nm process, and this advantage will become greater at each succeeding process generation.

In January 2006, AMD endorsed floating-body-DRAM technology; the company subsequently licensed the IP from ISI to develop larger cache memories for its new generation of processors. "The agreement is part of our ongoing research into higher density and more energy-efficient on-chip memory," says an AMD spokesman. "We're looking at this technology for potential use in future AMD processors." Because AMD manufactures processors using SOI technology, floating-body DRAM as on-chip cache or working storage could be a natural fit, enabling the company to increase performance and reduce cost.

It is clear that embedded DRAM has a firm place in all

application sectors. In consumer, communications, and automotive applications, a complex system ASIC with onboard DRAM may require no external memory at all. In computing applications, especially high-end processors, embedded DRAM enables the use of a large Level 3 cache, greatly increasing the memory bandwidth. IBM has followed this path with its Blue Gene/L supercomputer chip, and AMD will likely follow, bringing L3 cache into the PC market. The final determinant is technology. SOI processing is quickly moving into the mainstream. It is a case of simple arithmetic to demonstrate that the combination of SOI and embedded floating-body DRAM produces the most economic high-performance ASIC technology. **EDN**

AUTHOR'S BIOGRAPHY

Raymond Ambrose operates Randa Creative (www.randacreative.com), which he founded in 2005 after 33 years in electronics. The first 10 of those years, he worked in telecom, and the last 23, he worked in microelectronics for STMicroelectronics, where he managed the development of high-end graphics accelerators and contributed to the development of digital-satellite television. Ambrose also writes technical articles for the electronics industry and is a photographer, specializing in microphotography of electronic components. You can reach him at ray.ambrose@randacreative.com.

REFERENCES

- 1 "Advanced Micro Devices licenses embedded memory from Innovative Silicon," Innovative Silicon Inc, Jan 23, 2006, www.innovativesilicon.com/en/news_isi.php.
- 2 Matick, Richard E, and Schuster, Stanley E, "Logic-based eDRAM: Origins and rationale for use," *IBM Journal of Research and Development, Volume 49, No. 1*, January 2005, pg 145, www.research.ibm.com/journal/rd/491/matick.pdf.
- 3 "Advantages Over Embedded SRAM," NEC Electronics, Advanced Process Technology, www.necel.com/process/en/index.html.
- 4 Gara, A, MA Blumrich, D Chen, GL-T Chiu, P Coteus, ME Giampapa, RA Haring, P Heidelberger, D Hoenicke, GV Kopcsay, TA Liebsch, M Ohmacht, BD Steinmacher-Burrow, T Takken, and P Vranas, "Overview of the Blue Gene/L system architecture," *IBM Journal of Research and Development, Volume 49, No. 2/3*, 2005, pg 195, <http://research.web.watson.ibm.com/journal/rd/492/iyer.html>.
- 5 "Future Memories are Made of This: Embedded DRAM Completes the System on the Chip," NEC Electronics, Volume 39, April 21, 2005, www.necel.com/en/channel/vol_0039/vol_0039_1.html.
- 6 Fazan, Pierre, PhD, "Z-RAM zero capacitor Embedded Memory Technology addresses dual requirements of die size and scalability," Innovative Silicon Inc, 2005 www.innovative-silicon.com.
- 7 "Renesas Technology develops capacitorless twin-transistor RAM, enabling faster, more power-efficient embedded memory for SOC devices," Renesas, Sept 26, 2005, <http://eu.renesas.com>.