The use of embedded DRAM technology has become widespread, especially in higher-end system designs, because of its superior performance, silicon area savings, and low power compared to discrete memory solutions.

Traditionally, in cost-sensitive consumer applications, large memory arrays of 64 megabits and above were usually better suited to discrete commodity memory implementations. But as the supply of low-density DRAM wanes and prices rise, system designers are finding that embedding DRAM densities of 16 Mbits and below is more cost effective per chip than discrete alternatives. A highly integrated embedded DRAM approach also simplifies board design, thereby reducing overall system cost and time to market. Even more importantly, embedding DRAM enables higher bandwidth by allowing a wider on-chip bus, and saves power by eliminating DRAM I/O.

Today, designers can take advantage of these capabilities as various 0.25-micron embedded DRAM technologies enter production, with 0.18-micron to follow by the end of year 2000.

Embedded DRAM Technologies

The three commonly identified types of embedded DRAM are DRAM-based, blended (or hybrid), and logic-based. DRAM-based technology is practically the same as commodity DRAM--using DRAM periphery devices to build logic circuitry with perhaps the addition of one or two metal layers for logic routing. Blended technology uses additional front-end masks to enhance the performance of the DRAM periphery devices, to speed up logic performance. Logic-based embedded DRAM enables transistors with...
performance compatible with leading-edge logic processes, resulting in an improved DRAM-logic interface, and an on-chip logic performance path to implementing system-on-chip designs.

System designers are turning to embedded DRAM for several reasons. Unlike commodity DRAMs, which are only available in a standard range of densities--typically 4, 16 and 64 Mbits--the exact amount of memory required in a system can be specified in the embedded DRAM macro block, for example, 5, 9, or 17 Mbits. Thus, no memory is wasted, and area and cost are optimized. In addition, the exact configuration and memory interfaces can be specified in the macrocell, thus offering flexibility and optimum system performance.

Each of these three types combines the functions of both memory and logic on a single die. The elimination of the additional I/O bonding pads required for two separate chips saves about 5 to 10 percent of overall silicon area over discrete solutions. It can also help relieve the pad limitation problem of complex ICs by providing pad savings over discrete ICs, since DRAM driving pads are eliminated from both memory and logic parts. Depending on the particular design, an embedded array requires far fewer pads, thus saving space. This space saving is even more significant for smaller designs of 300K logic gates and below, because it alleviates the pad limitation problem common in these designs.

**DRAM-based Embedded DRAM**

DRAM-based embedded DRAM chips begin with a DRAM process architecture, usually one with two metal layers, on top of which one extra metal layer is added for logic routing. The philosophy behind this type of embedded DRAM is usually the same as that employed by discrete commodity DRAM manufacturers. This is to make the cell as small as possible, since a smaller cell means a smaller die, and thus a less expensive one. Typically, the DRAM cell size is 50 to 100 percent smaller than a cell of logic-based technology of the same generation. However, in this approach the peripheral circuitry used for logic design is the same as commodity-based DRAM circuitry. The high thermal cycles introduced in the DRAM-based process, just before the first metal level is processed, induce the diffusion of transistor dopants. This induced diffusion degrades
device performance.

The use in commodity DRAM of polycide in the polysilicon gate makes it impossible to introduce an advanced PMOS device. Polycide is necessary in order to make a self-aligned bitline contact in the DRAM cell, thus eliminating otherwise necessary design rule space between the transfer gate and the bitline contact, and reducing the cell size by at least 20 percent. In fact, because of this self-aligned contact, commodity DRAM can use only buried-channel PMOS, a technology that became extinct in logic processes after the 0.35-micron generation.

For these reasons, performance-wise, DRAM-based technology lags logic technology by at least two generations. For example, the performance of devices from 0.18-micron DRAM-based transistors is roughly equivalent to the performance of cutting-edge 0.35-micron logic process.

**Blended Embedded DRAM**

Blended, or hybrid, embedded DRAM is very similar to the DRAM-based type, but it is constructed with a couple of additional mask layers to enhance the DRAM periphery devices, which also serve as logic transistors. In essence, a blended process incorporates some additional steps lacking in a commodity DRAM process, in order to enhance the performance of peripheral circuitry. Normally, this involves slightly reducing the after-transistor thermal cycle and thereby reducing dopant diffusion; adding a source/drain silicide process outside of the DRAM array; and more aggressively reducing the channel lengths of peripheral transistors. But the blended embedded DRAM process architecture still looks much more like a DRAM-based device than a logic-based device because of features such as buried channel PMOS transistors, and possibly polycide instead of silicide gates.

Yet hybrid, like the DRAM-based process, is not library-compatible with logic. Therefore, designs with logic processes can't easily be ported to DRAM-based and hybrid processes, without re-designing the logic circuits. This is because system designers usually design a standalone logic chip first, and only later make the decision to create a second, embedded logic design for a more optimized, or higher-end, product offering.

The hybrid device speed/power figure of merit is closer to 1.5 generations behind
that of logic than the two generations behind of DRAM-based embedded DRAM. For example, the performance of 0.18-micron DRAM-based technology is roughly equivalent to 0.35-micron logic and 0.22-micron hybrid embedded DRAM. Note, however, that when comparing figures of merit, that several variables are involved, such as speed, power dissipation, design rule, and gate density. There are a wide range of hybrid types that have been introduced by manufacturers worldwide, and their performances vary depending on the manufacturing approach.

**Logic-Based Embedded DRAM**

Logic-based embedded DRAM derives from an existing logic process, so it has exactly the same design rules and SPICE models as the advanced standalone logic technology. Thus, there is no sacrifice of speed, as the speed/power figure of merit is exactly the same as the derivative logic process. Logic library compatibility also allows any design tested in a standalone logic technology to be easily ported into a logic-based embedded DRAM implementation without modification. In addition, logic-based embedded DRAM utilizes extensive libraries developed for standalone logic, thus making logic-based embedded DRAM designs more convenient for designers.

Knowing which approach is best is usually a simple task. For example, if the chip layout is dominated by logic, logic-based designs are more economical because logic design rules are denser than those of commodity DRAM periphery devices. But if the area balance shifts towards the DRAM array, DRAM-based or hybrid designs are more economical, even though they cannot offer performance as good as logic-based embedded DRAM designs.

It is possible to produce a small DRAM cell in a 0.18-micron logic-based embedded DRAM process. The approach shown in Figure 1 utilizes a self-aligned polysilicon bitline contact and polycided wordline. This allows a higher performance DRAM array, as well as a smaller cell. Yet, the DRAM structure utilizes metal as a bitline. This approach is good for reducing mask count and wafer cost. In fact, it allows the removal of at least two critical masks, compared to a commodity DRAM front-end process. Moreover, the resistance of a metal bitline is lower than that of a conventional polycide bitline typically used in commodity DRAM, thereby allowing higher speed and lower
power dissipation. Finally, the logic circuitry is similar to conventional logic technology, utilizing cobalt salicide, dual-gate poly (p + poly NMOS and n + poly PMOS), and abrupt p-n junctions for high performance.

Applications

Generally speaking, embedded DRAM is especially applicable to system-on-chip (SoC) designs because it integrates memory and logic on a single die; reduces total chip count in a system; reduces power consumption; and increases performance.

DRAM-based and blended embedded DRAM technologies are often used in applications that require high-density memory in a small area. Typically, these are systems with up to 128 Mbits of memory in an 0.18-micron process technology. These two technologies are also best for applications that are more cost-sensitive. A system design that requires considerably more memory than supporting on-chip logic, is an ideal candidate for DRAM-based or hybrid technology, especially from a cost standpoint as shown in Figure 2. Such applications include CD-ROM, DVD-ROM, disk drives, printers, lower-end graphics, 10/100-Mbits/sec switches, replacements for standalone SRAM, and custom-designed DRAMs.

A major benefit of logic-based embedded DRAM is higher performance. Thus, logic-based technology is often used in high-performance applications such as high-end consumer and networking. Applications that depend on video signal encoding, such as digital video cameras, laptop PC graphics, smart cellular phones and PDAs, also benefit from logic-based embedded DRAM. Lower power dissipation, another major embedded DRAM benefit, further advocates using this technology for portable applications.

In addition, some very fast custom memory designs are now possible using this technology. Because commodity memory standards do not apply to embedded DRAM, embedded DRAM is more flexible to use. Specialized designs can thus be created that are oriented toward speed, bandwidth and low power, rather than the emphasis on low price and efficiency that have historically been the aims of the DRAM macrocell. Architectural innovations made possible by this technology include higher-bandwidth DRAM with a very wide bus for handling a lot of parallel data, which is used in high-end graphics
applications and networking switches. Other designs emulate SRAMs with fast random access, rather than the typical DRAM page-mode access.

Design Tradeoffs

One of the advantages of embedded DRAM design is that JEDEC specifications do not necessarily apply to it. That implies that the tight refresh rate specs of standalone commodity DRAMs do not have to be met; instead, the cells can be refreshed more often. Normally, every bit in a commodity DRAM is refreshed every 64 milliseconds or longer, but embedded DRAM allows a much more frequent refresh spec as low as 2 ms, depending on the application and design. This contributes to a higher DRAM yield, because yield loss in mature DRAM processes is typically related to the designed-in refresh rate.

On the other hand, a high refresh rate spec may be desirable from the application point of view. If the embedded DRAM is used as a cache-like buffer, for example, a too-short refresh cycle increases the probability of cache misses. More important, if embedded DRAM is used for portable applications, it cannot be refreshed too frequently, because of standby power concerns. Indeed, the largest contribution to standby power dissipation comes from the DRAM refresh operation, which is essentially active circuit power dissipation. Therefore, it is crucial to reduce the refresh rate in portable systems, in order to minimize power dissipation and maximize battery life.

Several design approaches can be used to address the DRAM power dissipation problem. One approach is “smart refresh.” With this method, when the logic circuit is in the active mode--such as when a camcorder is recording or a computer game is being played--the chip is hot and DRAM is refreshed more frequently. When the logic circuit is off, the DRAM macro slowly goes into slow refresh mode, or sleep, thus reducing standby power dissipation. Even more, for certain applications DRAM does not need to be refreshed at all in standby mode, because data can be lost. Another approach is multi-bank design, or designing DRAM arrays in smaller banks, which allows shorter wordlines and bitlines. A line half the length will have a capacitance half as large. Since power dissipation is proportional to $CV^2$, power dissipation per bit read also decreases by roughly two times, not taking into account additional decoding. A multi-bank approach
allows faster DRAM access speeds for the same reason: bitline and wordline charging
time is proportional to CV/I. Here, I is the driver device current in the case of a wordline,
and the cell device current in the case of a bitline.

Testing Issues

DRAM ICs require extensive testing for functionality and reliability. The test flow
for embedded DRAM is even more complex because it involves testing and stressing
both the DRAM macro and the logic circuit. Figure 3 depicts two different test flows for
embedded DRAM: the flow used today, compared to the so-called “ideal” test flow.

The test flow commonly used today starts with the conventional Chip Probe 1
DRAM test. After failed bits are mapped out, they are replaced by redundant or repair
bits, which are always present for yield enhancement. The repair bits can be activated
using laser-programmable metal fuses. Laser repair is then verified in the CP 2 DRAM
test. If the DRAM has been repaired and is working at the wafer probe level, the logic
portion is then tested at the wafer probe level. After inking, bad dies are dropped, and the
good dies are assembled and tested for packaging yield (FT 1).

As can be seen in Figure 3, from the testing point of view it is less expensive to
replace the standard CP 2 memory testing with built-in self-test, or BIST, testing. BIST
translates the logic tester’s 1 and 0 signals into the timed signals needed for DRAM
testing. It allows the use of a logic tester to test DRAM, thereby combining logic and
DRAM testing simultaneously. Yet BIST circuitry can occupy 5% to 10% of the overall
area of the DRAM macro, depending on DRAM density and the amount of test patterns
to be covered. Although BIST is available in many DRAM IP offerings, most system
designers using embedded DRAM today still prefer separate DRAM and logic CP 2 tests.

Burn-in tests stress the DRAM part of the circuit at high temperature, in order to
screen out “soft” defects. These defects will not fail under normal test conditions, but will
slowly become worse and cause cell failures after years of failure-free operation.
Therefore, circuits with these defects must be screened out and marked as non-yielding,
in order to avoid field product failures. Package-level burn-in is more expensive for
embedded DRAM than for commodity DRAM for several reasons. First, the socket board
is custom-made and can be very expensive. More importantly, since embedded DRAM
ICs are usually large because they contain both DRAM and logic, each socket board fits few dies, making the costs of throughput, and therefore burn-in, variable and thus costly. Finally, package-level burn-in testing requires pins that allow special access to the DRAM, therefore reducing the pinout savings advantage of embedded DRAM.

The drawbacks of package-level burn-in can be addressed by wafer-level burn-in, in which the DRAM block is stressed on the wafer level at internal stress pads. This approach has a much faster throughput: wafer-level burn-in time is only a few minutes, compared to the hours required for package-level burn-in testing. At the same time, weak bits that have failed wafer-level burn-in can still be repaired, increasing overall product yield. However, some studies that correlate package and wafer burn-in results indicate that not all package-level failure modes can be screened out at the wafer level: only 80% or so of failures can be detected at the wafer level with the right burn-in conditions. Therefore, it appears that wafer-level burn-in cannot completely replace package-level burn-in; instead, a combination of the two can be a better solution. Wafer burn-in will screen out most failures at the beginning of the test flow so that they may be repaired, if possible; while package burn-in, either full or shortened types, will perform the ultimate reliability check of a packaged device.

In addition, there are other issues that make embedded DRAM testing different from commodity DRAM testing. Standalone DRAM is usually specified to work up to 85 degrees C, and all testing is done at this temperature. Yet logic is always more computation-intensive than DRAM, and it heats up the chip to more than 100 degrees C in a real operating conditions. Therefore, most semiconductor manufacturers benchmark their logic technology to up to 110 or even 125 degrees C, and the embedded DRAM block must be benchmarked accordingly. This restriction does not apply to every system, however, and appropriate refresh times and testing conditions should be determined for each design. For example, ICs used for mobile applications do not heat up as much as other designs, so DRAM refresh time can be increased and testing performed at lower temperatures.

Another issue is sensitivity to switching noise. Because of its logic content, embedded DRAM ICs are inherently noisier than standalone DRAM ICs. Noise issues can be addressed at the design level by higher sense amplitude margins and increasing the
charge coupling ratio in a DRAM array. Some designs will also utilize a shielding grounded metal plate above the DRAM block, to shield it from metal logic routing. The noise problem is even more acute if logic metal routing is done over the DRAM block. DRAM circuit robustness with respect to noise cannot be completely verified in every design by using macro verification test chips. Yet it can be partially addressed by using substrate noise generators near the on-chip DRAM well, and metal lines over the DRAM array that are swung up and down during testing.

**Conclusion**

In summary, embedded DRAM is becoming more common in consumer and communications applications because of its superior performance, silicon area savings, and low power compared to discrete memory solutions. Embedding DRAM enables higher bandwidth by allowing a wider on-chip bus. Therefore, an increasing number of system designers is using it particularly for high-performance or low-power applications requiring large memory buffers with fast access, such as networking, high-end digital consumer, and portable applications. As with other approaches, test issues still exist, but a combined test approach based on the requirements of the design can improve cost and throughput.
Figures

Figure 1: Cross-section of a 0.18-micron DRAM cell that utilizes a self-aligned polysilicon bitline contact and polished wordline to achieve higher performance and smaller size.

Figure 2: Optimal design spaces for DRAM-based vs. logic-based embedded DRAM technologies can be conceptually incorporated into a 2x2 matrix, with hybrid technology being somewhere in between, closer to DRAM-based.

Figure 3: The "ideal" test flow replaces the standard Chip Probe (CP) 2 memory testing of today's conventional DRAM test flow with BIST testing, so that logic and DRAM can be tested simultaneously.
Figure 1
Figure 2

Logic-based Embedded DRAM

Memory density

256Mb

64Mb

* Numbers best represent 0.18um technology

Logic density

Performance

Up to 1M logic 4T gates
50ns ring oscillator

Up to 4M logic 4T gates
28ns ring oscillator

DRAM-based Embedded DRAM

Hybrid Embedded DRAM
Figure 3

Today’s flow

CP1 : DRAM TEST

Laser Repair

CP2 : DRAM Test

CP2 : Logic Test

FT1 : DRAM & LOGIC

Package Burn-in

FT : DRAM & LOGIC

Ideal flow

DRAM Wafer Burn in

CP1 : DRAM TEST

CP2 : Logic Test + BIST on DRAM

Inking

Package

FT : DRAM & LOGIC