Embedded DRAM Finds Growing Niche

Applications Abound in Portable Devices, But Merging Processes Adds Cost

by Peter Song

As packing millions of logic transistors or tens of millions of bits of DRAM on a chip becomes cost effective, many vendors are looking at merging DRAM and logic transistors onto a single die. At the moment, their customers have mixed views on the benefits of this technology. While most ASIC designers see huge market opportunities for integrated chips, many CPU designers see little advantage to merged technology that results in logic transistors and interconnects that are bigger and slower.

Integrating memory and logic onto one chip increases bandwidth, reduces latency, and results in more flexible memory size and organization than in discrete DRAM chips. By eliminating pin-to-pin connection to external DRAM chips, embedded-DRAM products also reduce power consumption, board space, and electromagnetic interference while improving overall reliability. These chips can command premium prices in a growing number of portable applications—such as notebook PCs, handheld computers, and cellular telephones—that have a compelling need for one or more of the intrinsic advantages of this technology.

After years of being optimized for either DRAM- or logic-intensive processes, however, most IC fabs are ill suited to produce high-speed or cost-effective embedded-DRAM chips. Typical DRAM processes yield slower logic transistors, primarily due to their use of higher threshold voltages and longer gate lengths than used in typical logic processes. DRAM processes also require more area for interconnections, due to fewer metal layers and more-resistive contacts. In contrast, logic fabs are not well suited to building the high-capacitance and low-leakage capacitors used in single-transistor DRAM cells. Substantial investments are needed in either type of fab to produce embedded-DRAM chips that are competitive with combinations of discrete devices.

To solve these problems, many ASIC and DRAM vendors are aggressively advancing their merged technology, seeing an opportunity to add value to their products and services. They are committed to building faster and smaller logic transistors without unduly compromising DRAM density. Over the next few years, they aim to minimize the added cost of merging the two divergent processes.

Embedded DRAMs Offer Huge Bandwidth

Embedded DRAMs provide much flexibility in designing efficient memory systems. They easily support wide and fast interfaces as well as memory sizes and organizations that are tailored to an application's needs—a feat that is becoming more difficult with commodity DRAMs, as each new generation packs four times as much memory into a package with at most twice as many data pins. Embedded DRAMs offer attractive options to those applications that find the bandwidth or flexibility of commodity DRAMs inadequate.

Conventional DRAMs actually have much more internal bandwidth than is made available externally. Their two-dimensional row and column organizations make all bits in a row accessible to the sense amplifiers. For instance, a row access in a 1M×16 DRAM reads 4,096 bits from the memory array to the sense amps, of which only 16 bits are selected in a column access.

Embedded-DRAM designs can take advantage of the memory-array organization to provide bandwidth beyond what any discrete DRAM can offer. A good example is the Accelerix AX256-1M, which uses 4,096 pixel processors designed into the columns of the 13-Mbit DRAM array. This arrangement provides enormous processing and memory bandwidth but also requires each processor to be small enough to fit into a few columns—thus, each pixel processor performs only simple operations like add, subtract, invert and shift for 8-bit pixels.

Unconstrained by a slow system bus, embedded DRAMs can also offer faster interfaces than discrete DRAMs. Due to the many electrical constraints imposed by IC packages and printed-circuit boards, buses between chips operate at lower frequencies than those on a single die. Although system architectures are moving toward segregating peripheral devices and main memory into separate buses—to allow for both faster and wider memory interfaces as well as inexpensive I/O interfaces—external memory buses generally operate at lower frequencies than internal buses. Furthermore, internal buses can be as wide as desired, since there are no pin limitations on the chip. These wider, faster buses can deliver much better bandwidth than external memory buses.

Embedded-DRAM designs also reduce memory latency by eliminating the time needed to push signals through pin interfaces and to synchronize with slower system buses. The Mitsubishi M32R/D, for example, uses a 128-bit interface, which matches the width of the instruction queue and unified cache, to its 16M embedded DRAM (see MPR 5/27/96, p. 10). It needs only 5 cycles to access its embedded DRAM, saving 3–9 cycles over accessing external memory.

Embedded DRAMs Fill Growing Portable Market

Due to their cost premiums, embedded DRAMs have not been widely accepted, but their key advantages—high bandwidth, flexible memory size, low power, and small footprint—are directly applicable to one specific application: the graphics subsystem of a portable computer. Today's DRAM

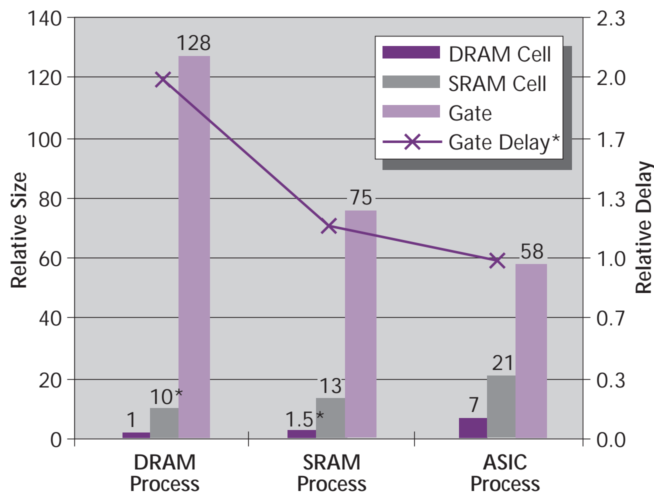


Figure 1. The relative sizes of a DRAM cell, an SRAM cell, and a gate (in a gate array) vary dramatically when they are designed following the design rules of 0.5µm DRAM, SRAM, and ASIC processes, as do the relative speeds of logic gates in three processes. (Source: Richard Foss, ISSCC '96, p. 260, except *MDR estimates)

processes can pack a 1M or 2M frame buffer onto half a die, leaving the other half for graphics-acceleration logic. This combination provides higher bandwidth than a discrete solution while reducing power and physical size—critical factors in a portable system.

The first vendor to implement this concept was NeoMagic (see MPR 3/6/95, p. 20), working with Mitsubishi's embedded-DRAM process. The success of NeoMagic's 2D graphics chips has spawned a slew of imitators, including Silicon Magic (see MPR 2/17/97, p. 5) and Trident (see MPR 6/23/97, p. 5), the former using Oki as a foundry and the latter going with Samsung.

For desktop graphics chips, however, the higher cost of the embedded-DRAM design becomes a burden, as does the inability to easily expand the size of the frame buffer. Where high-performance 3D logic is needed, the slower embedded-DRAM transistors may be a problem. Thus, while we expect other graphics vendors to offer embedded-DRAM graphics chips for portable systems in the future, we don't think the technology will take over on the desktop.

Other markets that are well suited to embedded-DRAM chips include smaller portable devices such as PDAs, handheld (Windows CE) PCs, and cellular telephones. In these cases, the CPU could be combined with enough DRAM to handle the needs of the entire system, reducing power and the size of the device. Mitsubishi's M32R/D is the first such processor aimed at these applications, but so far it has had limited market success. Others, however, are likely to follow.

Some Vendors Offer Quick, Temporary Solutions

To meet these emerging market demands, many ASIC vendors are offering quick but temporary embedded-DRAM solutions. Some supply three-transistor (3T) or single-transistor (1T) DRAM designs that can be fabricated in existing ASIC processes. Those with DRAM experience will integrate existing DRAM designs with limited numbers of logic transistors built in what are essentially DRAM processes. Few vendors currently offer truly merged processes that combine high-density DRAMs with high-speed logic transistors and interconnections.

Embedded-DRAM chips can be implemented in existing ASIC processes. For instance, LSI Logic currently offers up to 8 Mbits of 3T DRAM cells in its 0.25-micron G11 process. This process can pack up to 64 million transistors onto a die. LSI will merge the G11 process with Micron's DRAM process to deliver up to 128 Mbits of DRAM on a single chip. Mosaid Technologies offers 1T DRAM cells using TSMC's 0.35-micron ASIC process, enabling a die to have up to 16 Mbits of DRAM; in comparison, using a similar 0.35-micron process, Toshiba builds its second-generation 64-Mbit DRAM. Thus, while ASIC processes have the advantage of high-speed transistors, they offer significantly fewer bits of memory than do DRAM processes.

Figure 1 shows the relative sizes of a DRAM cell, an SRAM cell, and a gate (in a gate array) designed using various 0.5-micron design rules for DRAM, SRAM, and ASIC processes. The actual numbers are less significant than their magnitudes, since many variations are unaccounted for in this comparison. The figure shows that a DRAM cell built in an ASIC process is seven times larger than one built in a standard DRAM process, but it is still only one-third of the size of an SRAM cell built in the same ASIC process.

Building embedded-DRAM chips in a DRAM process yields the most bits for a given die size, but it also results in slower and larger logic designs than in an ASIC process. As Figure 1 shows, the gate delay increases by 2× and the size inflates by 2.2× when the gate is designed for a DRAM process. It is worth noting how small a DRAM cell is compared to a gate when optimized in the DRAM process—this data shows a gate is 128 times bigger than a DRAM cell. Similarly, Toshiba's 0.25-micron merged process shows a gate to be 234 times bigger than a DRAM cell.

By offering DRAMs as hard macros—design blocks that cannot be modified and generally cannot have routing over them—and porting logic gates to what are essentially DRAM processes, DRAM vendors can quickly offer embedded-DRAM ASIC libraries. As expected, these libraries typically contain slow and large gates and offer limited choices in DRAM megacells. Most vendors are also making design rules for their DRAM-intensive processes available as COT/foundry (customer-owned tools/foundry) services for customers willing to design chips to these specifications. This alternative offers customers an option of designing faster and smaller logic for their merged chips at the risk of producing single-sourced products.

The vendors that have both ASIC and DRAM processes have an advantage in establishing merged processes that promise the best of both worlds. Among ASIC vendors, Mitsubishi, NEC, Samsung, Toshiba, and TSMC are already

	Logic	DRAM
Capacitor Type	P-channel gate	Trench, stacked
Transistor Features		
Threshold Voltage	0.5 V	0.75 V
Gate Oxide	70 Å	90 Å
Drawn Gate Length	0.35 µm	>0.6 µm
Polysilicon Gate Types	P+, N+	N+ only
Salicide Source/Drain	Yes	No
Interconnect Features		
Polysilicon Layers	1–2	3–4
Metal Layers	4–5	2
Contact, Via Materials	Tungsten	Polysilicon

Table 1. Basic differences between the processes show that logic processes yield fast transistors, whereas DRAM processes yield low-leakage-current transistors. (Source: Mark Horowitz; MDR)

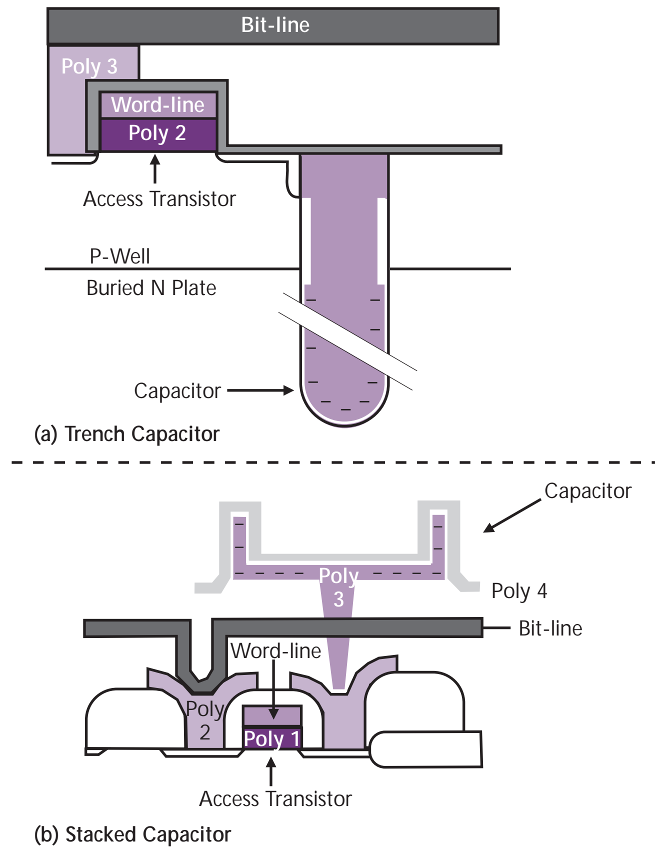


Figure 2. In trench-capacitor DRAMs, the capacitor is formed before the access transistor and its word-line and bit-line connections are made. In stacked-capacitor DRAMs, the capacitor is formed after the access transistor and its connections are made.

offering merged 0.35-micron processes that support up to 24 Mbits of DRAM and 140,000 gates on one chip. Toshiba will be the first to offer a 0.25-micron process that supports up to 128 Mbits or 32 Mbits with 410,000 gates. Other vendors that can spend billions of dollars, such as Hitachi, LSI/Micron, and UMC, will soon join them.

As more vendors offer 0.35-micron and 0.25-micron merged processes, the current DRAM-intensive and logic-intensive processes are likely to become less attractive for embedded-DRAM products. As more of today's high-volume embedded-DRAM chips demand the best of both worlds—high-density DRAMs and high-speed logic transistors—from merged processes, their cost premium will fall. These new processes will offer plenty of DRAM cells and transistors with reasonable die sizes.

DRAMs Use High-Quality Capacitors

Although DRAM and logic processes differ in many aspects, as Table 1 shows, a few key differences highlight the tasks of merging DRAM and logic processes. The major difference is that, in DRAM processes, elaborate processing steps are used to build capacitors that are physically small but electrically large, whereas design rules in logic processes are intended to minimize capacitance. Capacitors built in logic processes are many times larger and discharge at a faster rate than capacitors built in DRAM processes. For example, Mosaid's 1T DRAM cells, built in a standard logic process (using a P-channel gate isolated in an n-well), must be refreshed 100 times more often than typical DRAM cells.

High-performance logic processes use extra processing steps to make faster transistors. Key features that enhance transistor performance include low threshold voltage, thin gate oxide, salicided source and drain, N-type and P-type polysilicon gates, and short gate lengths. As some of these features are easier and less expensive to implement than others, semiconductor makers are likely to add different combinations of speed-enhancing features to their merged-DRAM processes.

Both DRAM and logic processes use multiple layers of interconnections, but their uses are different. In logic processes, these layers are made of metals, such as aluminum or copper, that have very low resistance. The contacts and vias that connect the different layers are typically made of tungsten, which also has relatively low resistance. In DRAM processes, however, some of the interconnect layers are made of polysilicon, for forming large surfaces of capacitors, with contacts and vias also made of polysilicon. Only two interconnect layers are typically made of metal.

Fast Transistors Have High Leakage Currents

As a first step toward a merged process, many vendors are lowering the threshold voltage to make transistors that switch faster and have a higher saturation current. To produce high electric fields under the gates, the gate oxides are made as thin as possible. By adjusting the amount of dopants, the doping profiles under the gates are changed to produce the desired threshold voltage. These enhancements also increase leakage current and, as a consequence, greatly reduce the DRAM cells' charge-retention time.

Reducing the charge-retention time is undesirable, since it forces more frequent refresh cycles and thus higher power consumption. Refreshing—reading the cell and writing back the same value—is required because reading the

binary state of a DRAM cell involves sensing the amount of charge stored in the cell's capacitor. Increased leakage current causes the capacitor to discharge at a faster rate. An error occurs if the cell is read after its capacitor has discharged too much. To prevent such errors, all DRAM cells are periodically refreshed to restore the charge in their capacitors.

Another challenge is that the gates of the DRAM access transistors are driven to 1 V higher than the supply voltage during write cycles, a technique known as word-line boosting. This higher voltage is needed to account for the voltage drop across the access transistors and to fully charge the capacitors. To withstand this higher voltage, thicker gate oxides are needed for the access transistors in DRAM cells.

To satisfy the conflicting requirements of thick gate oxides for access transistors and thin gate oxides for fast logic transistors, some vendors are using dual-thickness gate-oxide processes. For example, Toshiba's 0.25-micron merged-DRAM process uses 80-Å gate oxides for DRAM cells and 55-Å oxides elsewhere. Although an additional mask step is needed to grow the oxides to different thicknesses, many logic processes that support dual supply voltages—for example, 2.5 V for the core and 3.3 V for the I/O—already use dual-thickness oxides. Since the use of word-line boosting forces gate oxides in DRAM cells to scale at a slower rate than in logic transistors, dual-thickness gate oxides will become a common feature in future generations of merged-DRAM processes.

To improve performance of PFETs, logic processes use P-type polysilicon gates for these transistors. Since PFETs are built in N-type substrates, P-type polysilicon gates form surface channels—that is, the conducting regions lie close to the surface of the silicon. In contrast, DRAM processes use N-type polysilicon gates for both NFETs and PFETs to reduce cost, since high-performance PFETs are not needed. This method results in buried channels for PFETs, increasing the channel length and reducing the speed of the PFETs.

Reducing gate length is the most effective way to make transistors switch faster and to have higher saturation current. It is, however, the most expensive feature to implement, because it requires precise photolithography equipment for gate patterning. Although DRAM processes also use precise photolithography equipment, it is used for building capacitors in many layers of polysilicon or for building single-crystal silicon and insulators between the word-lines, bitlines, and capacitors. In fact, insulator spacings require finer critical dimensions than access transistors or capacitors, due to the need to pack as many DRAM cells as possible on a die while maintaining physically large capacitors and access transistors for high yields.

Stacked Capacitors Disrupt Logic-Processing Flow

Starting with the 4M generation, DRAM processes have evolved in two different directions: trench-capacitor cells that have the capacitor under the access transistor, and stacked-capacitor cells that have the capacitor above the access transistor, as Figure 2 shows. Although IBM, Toshiba, and Siemens are the minority among the DRAM vendors in using trench-capacitor processes, they seem to have an easier task of merging DRAM and logic processes.

Although both processes have advantages and disadvantages, a trench-capacitor process has two major advantages when merging DRAM and logic processes. The first is that burying the capacitors beneath the silicon surface makes planarization easier than when the capacitors are sandwiched between the transistors and the first metal layer. Planar surfaces (see MPR 4/18/94, p. 16) provide smaller deviations during photolithography for patterning metal lines and vias. Smaller deviations, in turn, allow for closer spacing between adjacent metal lines as well as more accurate placement of metal lines and vias across multiple metal layers. Establishing a planar surface across the die is essential for adding four or more metal layers with tight pitches.

The second advantage is that, in a stacked-capacitor process, forming capacitors after the transistors are built disturbs the dopant profiles already established in the silicon. That is, the electrical adjustments made to the wells, diffusions, and channel regions through carefully controlled doping steps are disturbed by the high temperatures needed for depositing insulation and polysilicon materials during the formation of the capacitors, causing dopants to diffuse further into the silicon. These diffused dopants yield slower transistors than do more abrupt dopant profiles.

According to Samsung, which uses a stacked-capacitor process, these processing difficulties are solvable. To reduce changes in dopant profiles during capacitor formation, lower-temperature deposition steps can be used. Chip-wide planarization can occur after deposition of a thick oxide layer to cover the protruding stacked capacitors, followed by the formation of tall contacts between the first metal layer and the transistors.

Etching this oxide layer for the contact openings is cheaper than etching the single-crystal silicon for the trench capacitors, simply because oxide etching is a common semiconductor process. Furthermore, etching trenches damages the single-crystal structure of the silicon and results in more leakage current, higher defect density, and generally lower yield. Although IBM admits to these problems, it claims to have solved them in its 16-Mbit generation.

Adding Logic Lowers Yield, Raises Cost

DRAMs' long product life and huge volume allow manufacturers to continually reduce their cost. In addition, DRAMs' simple and regular array structures make redundancy an effective way to tolerate processing defects. Unfortunately, adding more logic to DRAM chips takes away these benefits and increases the per-bit cost of memory.

DRAMs enjoy at least two years of product life before they are displaced by the next generation of chips with higher density. Many logic products, in contrast, are displaced in less than a year by competitors' designs with better

performance. Since they inherit the demands for best performance from logic products, merged products have a short product life. Their short life, coupled with much lower production volume, makes them less suitable for ongoing cost reduction. Merged products also make failure analysis more difficult than for DRAMs. Thus, economies of scale allow the commodity DRAM to be optimized for high yields and low manufacturing costs.

A typical memory array is organized into rows and columns of memory cells, with a few redundant rows and columns. When defective rows (or columns) are detected during tests, the row decoders are reprogrammed (by vaporizing appropriate polysilicon fuses) to select the redundant rows instead of the defective rows. A few redundant rows and columns can improve yields by many times.

In logic areas where redundancy is not practical, relaxed design rules keep the failure rate lower than in the memory array. As a result, many commercial DRAMs have yields in excess of 90%. The dense logic of a typical embedded DRAM, however, will have much lower yields, significantly reducing the overall yield of the merged product and thus increasing its manufacturing cost.

DRAM chips take longer to test than logic chips, due to their lower speed and the need to wait for cells to lose much of their charge for most tests. Logic chips, on the other hand, need more expensive test equipment due to their higher operating frequency and larger number of pins. embedded-DRAM chips need expensive logic-test equipment and longer test times—the worst of both worlds.

To minimize test costs, the merged DRAM and logic in the Mitsubishi M32R/D are tested separately on DRAM- and logic-test equipment. Another option is to use BIST (built-in self-test) for the memory arrays to reduce test times. Some test experts believe that, with BIST, embedded-DRAM chips can be tested quickly using only logic-test equipment.

Merged DRAMs also suffer from two new problems that neither DRAM nor ASIC chips have. The first is that most embedded-DRAM chips are single-sourced products, since DRAM processes differ vastly among DRAM vendors. The second is that embedded-DRAM chips require a longer production time than ASIC products, because they use the same production flow as a standard DRAM. While most ASIC chips need 3–4 weeks of production time, embedded-DRAM chips need 12–14 weeks. Embedded DRAMs are unlikely to succeed in markets that do not tolerate single-sourced products or long turnaround times.

Embedded DRAMs Open New Opportunities

Embedded DRAMs offer opportunities to many vendors, from semiconductor giants to entrepreneurial startup companies. embedded-DRAM technology gives DRAM-only vendors ways to add value to their products and services. It gives them a chance to enter the profitable ASIC and COT/foundry business while leveraging their DRAM expertise. It also gives them products that are not commodities and can still provide profits in times of DRAM surplus.

For More Information

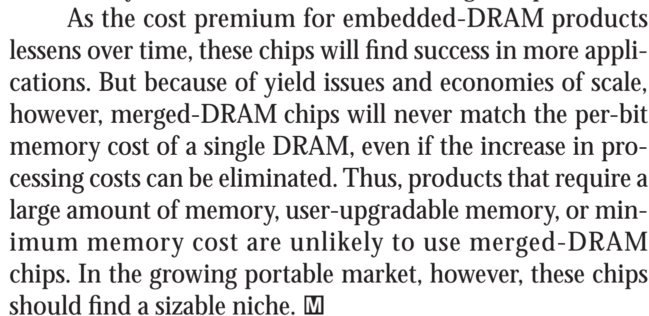
For more information on embedded-DRAM products, you can contact the vendors mentioned in this article via the Web at the following addresses: Accelerix, *www.accelerix.com;* Hitachi, *www.hitachi.com;* LSI Logic, *www.lsilogic.com/products/unit5_2.html;* Mitsubishi, *www.mitsubishi.com/TechShowcase/tsItem07.html*; Mosaid, *www.mosaid.com*; NEC, *www.nec.com;* Samsung, *www.sec.samsung.com;* Toshiba, *www.toshiba.com/taec/components/family/family.html.*

Steven Przybylski of the Verdande Group (San Jose, Calif.) is a consultant specializing in DRAM and embedded-DRAM chips. He can be reached at 408.984.2719 and *www.verdande.com.*

For immediate but perhaps temporary solutions, DRAM vendors are offering DRAM-intensive ASICs while ASIC vendors are offering logic-intensive ASICs. Due to their lower yields and premium pricing, both processes offer only limited opportunities. As merged processes become available, chips built in these one-sided processes will be less attractive than those built in merged processes.

Embedded DRAMs can provide higher bandwidth from fewer bits of memory than commodity DRAMs. Most memory systems have conflicting needs to provide the highest bandwidth, imposed by the applications, from the smallest amount of memory, imposed by low-cost competition. Because commodity DRAMs are offered in only a few combinations of size, data width, and speed, many systems end up sacrificing one to meet the other—they generally use more bits than required to meet their bandwidth requirements. Embedded DRAMs make it easier for memory systems to meet their conflicting requirements.

But for the moment, embedded-DRAM chips are more expensive to build and consequently can succeed only in markets that can pay a premium price for their improved performance, reduced power consumption, and decreased board space. These markets are primarily in portable systems that can fit an entire memory array (either for a frame buffer or for the system software itself) onto a single chip.



As the cost premium for embedded-DRAM products lessens over time, these chips will find success in more applications. But because of yield issues and economies of scale, however, merged-DRAM chips will never match the per-bit memory cost of a single DRAM, even if the increase in processing costs can be eliminated. Thus, products that require a large amount of memory, user-upgradable memory, or minimum memory cost are unlikely to use merged-DRAM chips. In the growing portable market, however, these chips should find a sizable niche. M