Flash Memory Heads Toward Mainstream

NOR Architectures Dominate, But Challengers Prepare New Entries

by Curtis P. Feigel

Flash memory, the only new solid-state memory technology to appear in a decade, is on the rise. Its combination of low cost, nonvolatility, and in-system programmability is unique, creating a large market opportunity in computers and consumer electronics. Flash memories are already replacing ROMs in some applications and disk drives in others. Flash even has the potential to reach a lower cost-per-bit than DRAM.

Although flash memories are built from relatively simple IC structures, they are difficult to manufacture, leading vendors to form development partnerships and spawning a variety of clever designs. This article examines current and potential applications for flash chips and describes the technology used to build them, including the competing NOR, NAND, and AND architectures.

Starting with EPROM designs, Toshiba began in the late 1970s to develop what is now called flash memory. The company described the new technology in 1985 but failed to develop it into a commercial product until recently. Intel, however, followed through with the concept and introduced its flash memories in 1988 (see MPR 4/88, p. 2). Since then, Intel has been the market leader, although the company has inevitably lost market share as competitors entered the arena.

The latest flash devices have read-access times as short as 60 ns, rivaling those of DRAM. Most are segmented into blocks that may each be electrically erased and programmed. Single flash chips are available today in capacities as large as 16 Mbits, and Intel is sampling an IC containing two 16M dice.

Problems Lead to Partnerships

Neither Intel nor any other vendor was prepared for the dramatic rise in demand for flash memory that began in the second half of 1992. Previously, the market had totaled less than $200 million per year. The total for 1992, according to Dataquest, was more than $300 million and nearly doubled in 1993. Projections vary, but it's likely that 1996 will see a flash market of $1.5–$3.0 billion. This market explosion has caught the attention of literally all major semiconductor makers.

Even companies with experience making DRAM have found that building working flash cells is a challenge. Intel, which developed its 0.8-micron flash technology at a facility in Santa Clara (Calif.), planned for production units to be manufactured at NMB (now NPNX) in Tateyama, Japan *(see **0615MSB.PDF**)*. But after months of technology-transfer problems, Intel sent a team of engineers to facilitate the learning process and transfer the art of making flash oxides that function properly; the company has had two decades of practice building stacked-gate transistors since the start of its EPROM business in 1971. Sharp, Intel's partner for the 0.6-micron process, did not have problems, so its production was ramped up to help meet demand while the NMB fab got started.

While Intel had the requisite experience, others found the vagaries of flash memory to be a black art. As shown in Figure 1, most have sought partnerships with other vendors, pooling resources and experience to ease the development struggle. Toshiba, despite its problems, will continue to pursue the NAND approach while concurrently producing NOR-based flash memories with help from National. Toshiba is also working with IBM to develop NAND-based PCMCIA cards for file storage.

Mitsubishi's 16M part is scheduled for announcement in 1Q95, but partner Hitachi's plans for AND devices won't be revealed until 2Q95. NEC is not in volume production but is currently aiming to produce a 16M flash device sometime in mid-1994. The company presented a paper describing a 64M, 3.3-V flash design at the recent ISSCC. NEC may try to produce this high-capacity device before any other vendor, leapfrogging Intel in the process. Last but not least, AMD is strongly committed to its partner Fujitsu and is building a $700-million flash-memory fab in Japan that is scheduled to come on-line in late 1994.

Leaders Take Different Positions

To maintain market share in the face of increasing competition, Intel and AMD are both investing heavily in flash fab capacity and dropping prices on 4M and smaller

Cell Type	Company or Partnership
NOR	NEC Atmel Catalyst AMD, Fujitsu NKK, Macronix Intel, NPNX, Sharp Matsushita, SunDisk
NAND	Toshiba, National, Samsung
AND/DINOR	Hitachi, Mitsubishi

Figure 1. Most major semiconductor makers have leapt into the billion-dollar flash-memory market, but the difficulty of the new technology has forced many to form partnerships.

Intel Promises 5-V Only

Intel is planning to introduce flash memories with a twist later this year. Called SmartVoltage parts, the devices will work with either 5 V or 12 V as the programming supply (V_{PP}) and allow either 5 V or 3.3 V on the V_{CC} pin. Thus, the chips have 5-V-only capability using on-board charge pumps but exhibit improved write and erase times with the external 12-V supply. Read operations, which do not require high voltage, can be performed with just a 3.3-V supply.

The SmartVoltage approach gives designers flexibility to trade complex power-supply circuitry for slower access times. Equipment that is normally battery powered can supply the higher voltages when connected to AC and benefit from faster flash-memory speeds.

The new devices will be available in 2M, 4M, and 16M capacities. Although prices have not been released, Intel says that it will not charge a premium for SmartVoltage parts.

devices. Flash is a difficult technology with a long learning curve; both Intel and AMD are well up the slope, but most Japanese and other contenders have barely gotten off the ground. To successfully challenge the market leaders, they must invest time and gain experience.

In an attempt to maintain position or even overtake Intel, AMD and other vendors of single-supply flash memories make the obvious point: dual-supply devices require a more complex power system or an external DC-DC converter. Initially, Intel argued that single-supply devices force the customer to pay for a charge pump with every chip and that the external circuitry is easily amortized across multiple flash memories. Thus, the two camps had positioned themselves for different markets: single-supply devices fit in applications that need the capacity of only one or a few chips, whereas dual-supply devices made more sense where large memory arrays were needed.

But AMD's pricing put single-supply chips on a par with Intel's dual-supply devices, eliminating any cost advantage. Countering that move, Intel has revealed plans to introduce single-supply flash memories this year (see "Intel" sidebar).

Flash Enters Memory System

Flash memory's nonvolatility makes it attractive as a replacement for both ROMs and hard-disk drives. Ultimately, it may fundamentally alter the hierarchy of memory in a computer system.

The ROM that holds a PC's BIOS is a popular target for flash memory. In the past, computer makers have judged it impractical to release updates, bug fixes, and enhancements for their firmware, not trusting users to disassemble systems and properly extract and install chips. Because flash memory can be erased and reprogrammed while installed, this is no longer an issue. A new concern is the potential for problems caused by virus software, so provisions are made in the hardware to write-protect the flash memory.

Another obvious home for flash memory is as a mass-storage medium emulating hard-disk drives. Although it is much more expensive per megabyte than rotating media, its low power requirements and ruggedness (due to lack of moving parts) make flash memory a logical choice for portable applications. Flash memory also has a read-performance advantage of two orders of magnitude over disks. Over time, the cost of flash may drop more rapidly than the cost of rotating media and perhaps even cross over, but this is a point of much dispute in the industry *(see **071003.PDF**)*.

SunDisk (Santa Clara, Calif.) markets a flash product designed from the start as a hard-disk replacement. Using patented flash technology from AT&T and custom flash-memory chips built by fab partner Masushita, SunDisk makes 40-Mbyte PCMCIA cards that duplicate the 512-byte sectors of a hard drive. It uses a custom controller designed with Motorola. The device controls access to the flash memory array, performs ECC and wear leveling, and manages the host interface. Because the flash card looks just like a hard disk, standard software runs without modification. Seagate, a major manufacturer of hard disks, now holds a 25% interest in SunDisk.

An extension of the disk-replacement concept is called XIP (execute in place). Rather than simply emulating a disk drive, the flash memory acts as a file storage area that the system can access randomly. Instead of copying an application program from mass storage into main memory and executing from there, an XIP system directly runs code stored in the flash memory, using a small amount of standard RAM to store constantly changing data and variables.

Removing one whole layer from today's common memory hierarchy and eliminating the lengthy disk-access time represents a major performance benefit, but operating systems and application software must be altered to take advantage of it. Application programs typically must request the operating system to allocate memory space, so the OS must know when and where to provide flash space versus DRAM space.

For applications demanding low power, an emerging trend in file storage is to program the flash memory in a desktop system, where power is not a limitation. Once it has been loaded, the memory can be installed in a portable system that uses a 3.3-V supply for reading.

FETs Are Basic Building Block

Flash-memory design encompasses two basic considerations: the structure of the memory's cells and the arrangement of those cells in a usable device. In com-

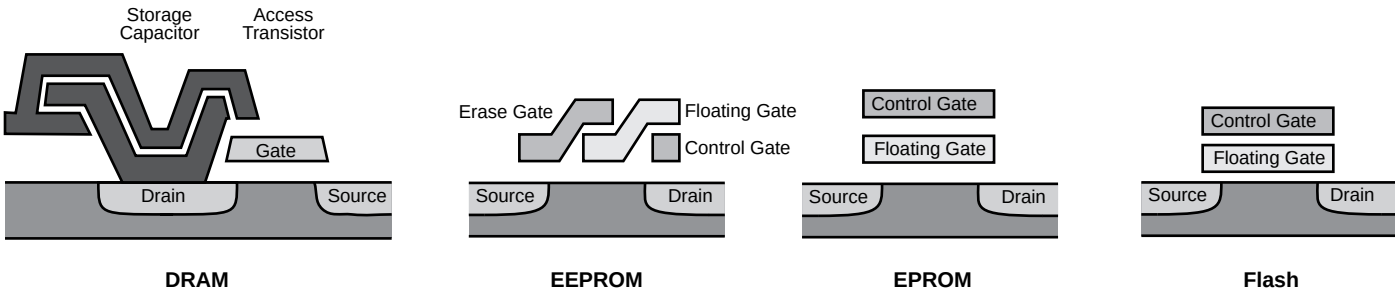

Figure 2. The flash cell is among the simplest of the semiconductor memories, combining a field-effect transistor with a floating gate. The flash cell is derived from the EPROM cell but uses thinner oxide and dielectric layers.

mercial flash memories, these two aspects are inextricably linked, but the field may be divided into three broad categories: NOR, NAND, and AND architectures. The NOR architecture is by far the most prevalent, supported by market-leader Intel, second place AMD, and most of the other players. The NAND architecture is under development by Toshiba and its partners National and Samsung but is not yet on the market. Hitachi is the proponent of the AND architecture, with its partner Mitsubishi pursuing its own NOR variant called DINOR.

A flash memory cell is simply a CMOS field-effect transistor (FET) with a floating (electrically isolated) gate stacked between the transistor's gate and its channel. It is this floating gate that is the device's actual storage element. Figure 2 compares this simple structure to the cell structures of other memory types.

DRAM cells combine a transistor with a capacitor that is the storage element. Although the transistor's size scales with the fabrication process, the capacitor is constrained: it must remain large enough to retain data with a reasonably long refresh period and to resist soft errors caused by alpha-particle strikes. As fabrication geometries get smaller, the capacitor dominates DRAM cell size, despite complex structures like the one shown in Figure 2. Other limitations are that the cell requires constant refresh due to capacitor leakage and that only a small voltage potential is available for reading the value stored in the cell. But the DRAM-fabrication process is well understood, and many companies have years of experience producing it in high volume.

Because of their two-transistor cell, EEPROMs are larger, more complex, and more expensive than these other memory types. The second transistor selects individual bytes of the memory for erase and reprogramming. The extra transistors and selection circuitry give EEPROM the highest per-bit cost of all nonvolatile memories. Its high internal voltages achieve fast write and erase times but break down oxides and limit the chip's life.

EPROMs use a simple FET with stacked gates but a relatively thick oxide layer. The complete memory array is bulk-erased when energetic UV (ultraviolet) photons discharge its floating gates. EPROM is another technology that has difficulty as cell geometries become smaller: metal layers do not scale proportionately and thus block the UV light, so it is harder to erase the cells. In addition, UV light requires a special window that precludes inexpensive plastic packaging. (None of which is a problem with one-time-programmable EPROMs—one of the most cost-effective memory types.)

The flash-memory cell is derived from the EPROM design but uses thinner oxides that allow electrical erase (although both use the same programming mechanism). The main difference between the various flash devices is the mechanism each uses to charge and discharge the floating gate to write and erase the cell.

Flash Cell Uses Electron Tunneling

Figure 3(a) shows how a typical flash cell, in this case Intel's ETOX (EPROM with tunnel oxide), is written. Circuitry in the chip links the cell's control gate to the V_{PP} supply voltage (12 V) and the cell's source to ground. Capacitive coupling biases the floating gate, inverting the p-type channel in the substrate (that is, it

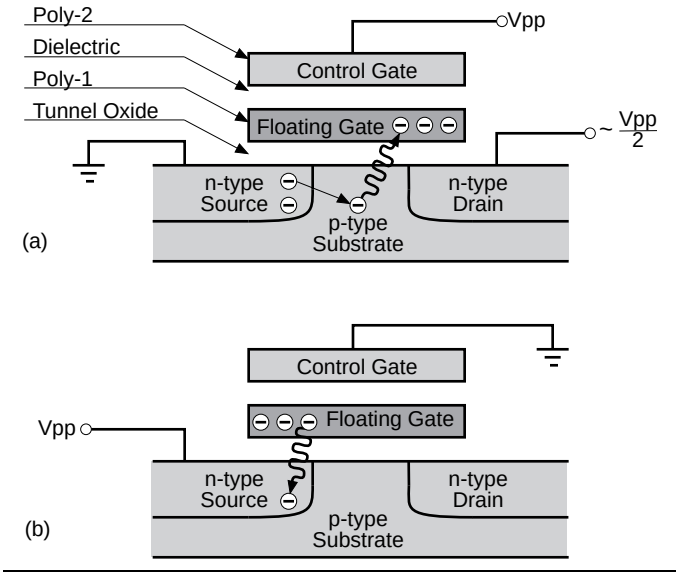

Figure 3. The floating gate of a field-effect transistor serves as the storage element of Intel's ETOX flash-memory cell. The basic mechanisms for (a) programming and (b) erasing require charging or discharging the floating gate.

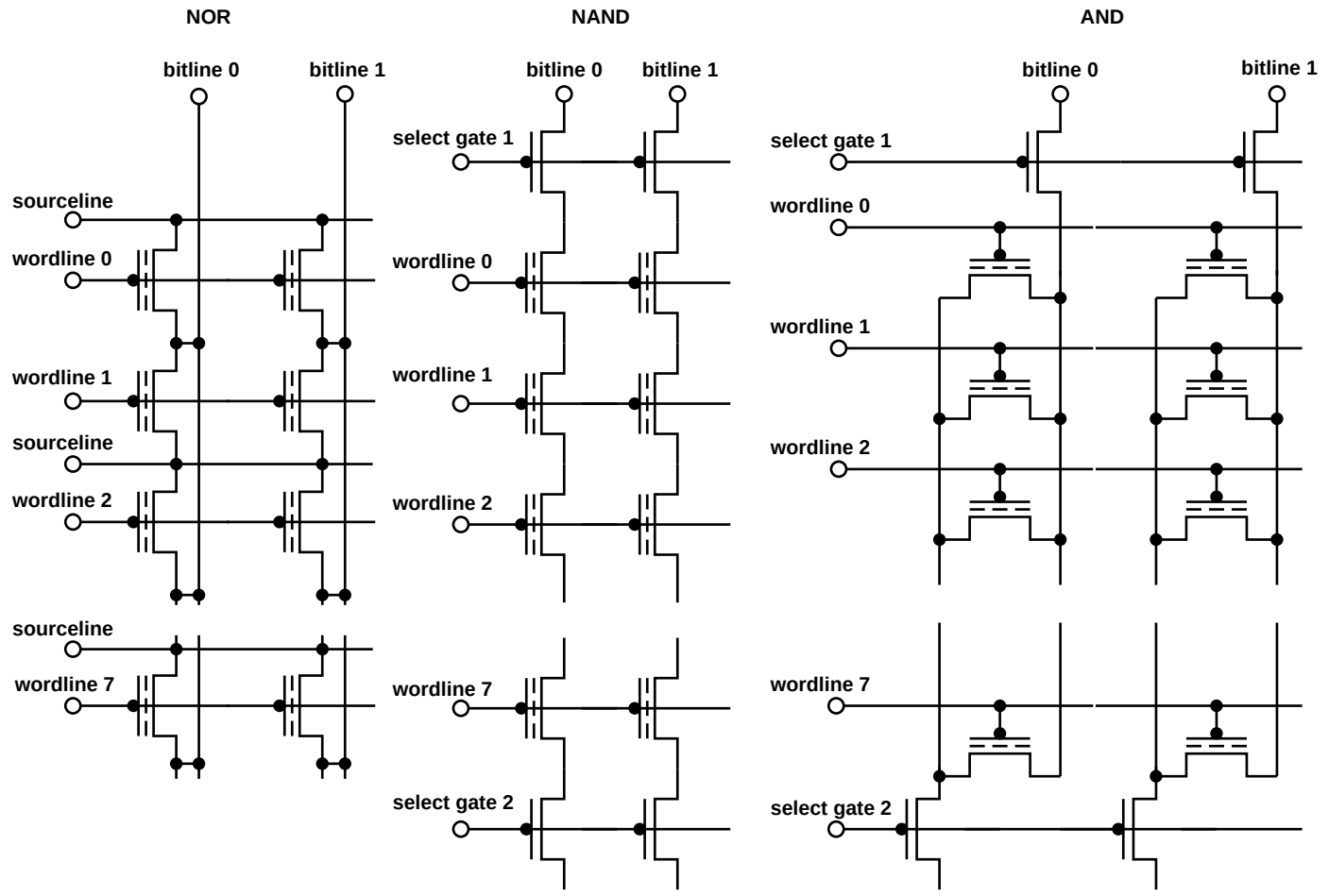

Figure 4. The NOR architecture is the most prevalent arrangement of cells in today's flash-memory devices. The NAND, with its fewer connections, has the potential to reach higher density and lower cost, but its serial arrangement means slow access times. The AND architecture blends some characteristics of the other two by minimizing interconnects but allowing random parallel access.

takes on the characteristics of n-type material, with electrons becoming the primary charge carriers). Then the chip's circuitry connects the cell's drain to an intermediate voltage about one-half of the V_{PP} supply.

With the drain at a higher voltage than the source, electrons flow from source to drain, colliding with and energizing atoms in the substrate. Some electrons reach a high enough energy level—they become "hot" enough—that they can overcome the tunnel-oxide barrier and accumulate on the floating gate. This effect, called hot-electron injection, raises the threshold voltage required to turn on the cell's transistor.

In the flash design, there is no difference between an erased cell and a cell with a 1 stored in it—the user must recognize and keep track of memory areas that have not been programmed. To read a cell, the chip connects the V_{CC} supply voltage (5 V or 3.3 V) to the control gate. If the cell has been programmed, the transistor will conduct less current than if it has not been programmed. Thus, the chip's comparator logic can generate a 0 if the floating gate has been charged or a 1 if not.

The erase procedure is essentially the reverse of programming: removing the charge on the floating gate. To do this, the chip connects the cell's source to V_{PP} and grounds the control gate while the drain is left unconnected. As Figure 3(b) shows, electrons on the floating gate of a programmed cell are attracted to the higher potential of the source and pierce the oxide layer by a process called Fowler-Nordheim tunneling.

NOR Logic Forms Common Architecture

NOR-based memories get their name from the layout of the cells in the memory matrix, shown in Figure 4. Multiple cells are connected to the same bitline; groups of bitlines are connected together to one of the chip's sense amps to drive its data I/O lines. This is commonly called a wire-OR connection but, since the stored values are low-true, it is referred to as NOR.

Because flash-memory chips are typically byte wide, cells are read in groups of eight, so the control gates of eight parallel cells are connected to a common wordline. During a read, address-decoding logic within

the chip selects a single row and column of the memory matrix to activate a single wordline. Each cell on a wordline generates its output on a separate bitline, thus producing a parallel byte output. Cells on the same bitline that are not selected allow their outputs to float.

Unprogrammed or erased parts of a flash chip may be programmed byte-by-byte, but unlike random-access memories (RAM), it is not possible to erase a byte in flash memory simply by overwriting it. First-generation flash memories had to be completely erased, much like EPROMs, but modern chips are divided into blocks of memory cells that may be individually write protected and erased. Compared with erasing cells individually, blocking simplifies the selection circuitry and saves precious die area. The drawback is that the whole block must be erased to alter a single byte.

Some flash chips with very small blocks use two-transistor cells, as Table 1 shows. This 2T cell increases cost, however, and most new devices use larger blocks with the single-transistor cell described earlier. The size and number of the blocks is one way manufacturers differentiate their flash-memory products.

Other Architectures Hold Promise

Although the NOR architecture is the most prevalent at this time, others are possible, and each has its advantages. Toshiba has been exploring a substantially different architecture called NAND. The NAND scheme still uses FETs, but they are constructed to use Fowler-Nordheim tunneling for both writing and erasing. The cell's oxide layer must, at times, allow tunneling in one direction or the other—but not at all times and never in both directions.

Rather than being wire-ORed, the NAND-based memory cells are connected in series and linked to the bitline through a single switching transistor. To read the first memory location, the switching transistor must be enabled. To read the second memory location, the first must be programmed to pass the second transistor's signal. To read the third location, the first two must be programmed, and so on. Cells within a block cannot be accessed randomly, and information read from a block must be stored in another to be retained.

Compared with the NOR architecture, the NAND design results in slower read- and write-access times. Its advantage is that it requires only a single metal-layer contact for every string of 8 or 16 cells, versus the NOR design's contact per two cells. Metal-layer contacts don't shrink in the same proportion as other IC fabrication geometries, so they are a limiting factor in the size of flash-memory cells. Fewer contacts means that the NAND cells themselves can be smaller. The NAND architecture does require more complex address decode and support logic on the chip, however, so the cost per bit may be about the same as for NOR-based devices.

	AMD 29F016	Atmel 29C040	Atmel 29LV040	Intel 28F032SA	Intel 28F400
Architecture	NOR	NOR	NOR	NOR	NOR
Capacity (Mbits)	16	4	4	32	4
Block Size (bytes)	64K	256	256	64K	8K–128K
Read Time (ns)	90	120	200	120, 70*	150, 65*
Write TIme (µs/byte)	8	30	30	6	6
Erase Time (seconds)	1.0	n/a	n/a	0.4	2.0
Erase Cycles (1000s)	100	10	10	1,000	100
Transistors/Cell	1	2	2	1	1
Voltages	5	5	3.3	3.3, 5/12	3.3, 5/12
Price (10,000)	$40	n/a	n/a	$190	$21
Samples	3Q94	Now	Now	Now	Now
Production	4Q94	Now	Now	2Q94	Now

Table 1. A representative sample of currently available flash memories shows that, while all use the NOR architecture, there is a range of capacities and speeds to fit a variety of applications. *Faster times at 5 V, slower times at 3.3 V. (Source: vendor data)

The necessary characteristics of the NAND's oxide layer are difficult to produce, and Toshiba has run into problems in both operation and yield. Earlier this year, the company was forced to redesign the memory cell's structure because the high programming voltages reduced its durability.

The structure of Hitachi's AND cell is similar to the NAND cell, depending on tunneling for both write and erase processes. But the sense of the floating gate is inverted, with a programmed cell indicated by a discharged floating gate. This scheme combines some of the best features of the other two designs: only one metal-layer contact is required for about a hundred cells, and the programming voltage and the field it produces are reduced, resulting in a small cell that is durable and can be accessed randomly.

Mitsubishi has come up with a variation on the NOR architecture that it calls DINOR. It also aims to reduce the number of metal contacts. In the DINOR design, 8–64 memory cells connect to a polycide sub-bitline that is connected to a main bitline through a switching transistor.

Mitsubishi is the only developer of a non-NOR architecture that hasn't publicly admitted to fab problems (the company indicates that its first commercial device will be a 16-Mbit part). So far, claims of speed, power, or cost advantages for NAND and AND are difficult to prove—only NOR designs are available.

Automation Hides Slow Write, Erase

The read-access time of flash memory is in the same ballpark as DRAM or slow SRAM: currently available devices can be read in as little as 60 ns. As Table 1 shows, however, erasing or writing flash memory cells is much slower. The fastest flash devices require 4 µs per byte written, and erasing a block or the whole chip can take more than one-half second. Even though, in a typical file-storage application, reads are five times more

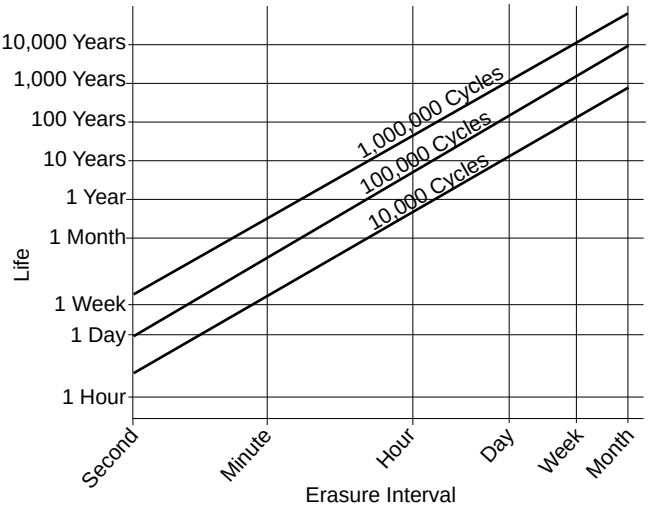

Figure 5. Depending on the interval between erase/write cycles, flash memories can have life times of many years.

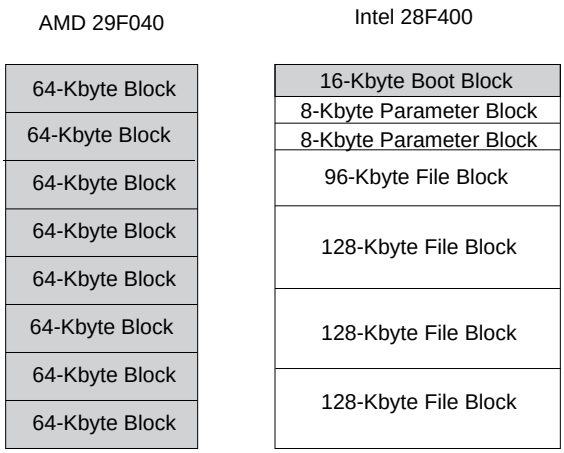

Figure 6. The arrangement of erasable blocks within a flash memory is determined its target application. The AMD device, with equally sized blocks, is suited to file storage, while the Intel device is intended to store BIOS information for a personal computer. In this diagram, shading represents blocks that can be write-protected by programming in a PROM-type device programmer.

common than writes, flash memory's slow erase and write times can be a serious bottleneck. In the 16M arena, Intel's flash memories contain input buffers that hide the slow writing speed from the host. But modern flash-memory chips also incorporate other hardware to hide the slow erase and write processes and to balance cell wear to maximize the device's life.

Because the erase process can take so long, most flash memories are designed to perform the function autonomously and provide an interrupt signal to the host processor to indicate completion. Before erasing a block, a chip's internal control logic typically preconditions the block by programming all its cells. With all floating gates at the same potential, the whole block can then be erased in parallel. Logic within the chip verifies that the charges on all the floating gates are reduced to a uniform level. If a cell produces current above a certain threshold, the block of cells is erased again.

Flash's tunneling process gradually wears out the oxide layer, so memories have a lifetime determined by the number of erase and write cycles they can endure: typically from 10,000 to 1,000,000. Figure 5 shows that this can translate into years of useful life; for example, a cell with a 10,000-cycle lifetime that is written three times per day will last for 10 years. Even with a million-cycle endurance, however, flash cannot be used to store frequently altered program data, as DRAM is.

The problem is that the cells are not equally receptive to being erased—in fact, they follow a Gaussian distribution with fast and slow outliers causing problems: cells that are slow to erase must be erased again, while cells that erase too quickly may be driven into depletion mode and fail to function properly.

System software, such as Microsoft's Flash File System, spreads cycling wear across all of a device's blocks. By periodically moving stable data, a formerly "quiet" block is made available for erase and write cycles. The software also logs the number of erase cycles each block has suffered and can identify blocks that are approaching the end of their useful life.

Flash memory's write and erase processes both depend on a difference of potential and so are affected by the supply voltages. If the programming voltage is low, these processes will take longer. Low temperatures can also slow the erase process, but they enhance the write process. As the temperature declines, the breakdown voltage at a cell's source drops, clamping (limiting) the erase potential and slowing the tunneling process. Lower temperatures enhance electron mobility through the p-type channel of the substrate, however, so during a write, electron collisions are more energetic and the floating gate is charged faster.

NOR Spawns Variants

Following in Intel's footsteps, AMD has the second-largest share of the market and is currently the leading supplier of 5-V-only (single-supply) flash memories. AMD's NOR devices use FET structures very similar to Intel's ETOX devices (but with different oxides), and, like most other vendors, AMD makes some chips that match Intel's pinouts. The write and erase mechanisms are similar too, but AMD chooses different internal voltages to reduce current requirements. The cells of AMD's devices still require higher potentials than are provided by the single supply, so all its chips include charge pumps to boost the programming voltage.

To write a cell, AMD also uses hot-electron injection but, because only a single supply is available, must use a charge pump to drive the cell's drain to about 6.7 V and its control gate to 10.5 V. Again, the potential difference

between the cell's grounded source and its drain causes a current flow, and hot electrons, attracted by the control gate, traverse the tunnel oxide and form a charge on the cell's floating gate.

The company has developed a technique called negative-gate erase to remove the charge and thus erase the cell. As in the Intel erase process, AMD devices create a difference of potential that causes electrons to tunnel from the cell's floating gate to the source. But Intel devices use the external V_{PP} to create this potential difference and supply a peak current of as much as 10–20 mA. For a single-supply device to provide this much energy would necessitate a large on-chip voltage converter.

Instead of grounding the cell's control gate and applying a high positive voltage to the cell's source, AMD connects the source to V_{CC} and drives its control gate to –10.5 V. The company claims that the strength of the electric field generated is the same as for an ETOX device, but that the current required is less than 10 µA—small enough that it can be generated economically by an on-chip charge pump.

As Figure 6 shows, AMD's devices use uniform-sized blocks, each of which may be write-protected by hardware. Intel and others also make devices with uniform blocks, all aimed at the file-storage market. Flash memories with asymmetrical block divisions are designed to replace the ROMs that hold PC BIOS code. The memory in these "boot block" devices is divided into sensible block sizes: large blocks for code that will not be changed often, small blocks for storing more frequently changed system parameters, and a block segregating and protecting the system code that controls the loading of the other blocks.

Flash Cheaper Than DRAM, Eventually

Replacing hard-disk drives is the holy grail of flash-memory vendors. It's a difficult quest: the price-per-megabyte of hard-disk storage has dropped dramatically in the past nine months. As Figure 7 shows, moderately sized drives are currently priced at about $1 per megabyte, and large drives of a gigabyte or more approach one-half that. Even if hard-drive prices stabilize without going lower, flash memory's current $30-per-megabyte price puts it out of reach for most file-storage applications unless they demand low power consumption or ruggedness.

But flash memory can fill another file-storage niche. At present, there is a "floor" of about $200 below which disk drives cannot go—for less than 100M or so, the mechanism costs too much. In this realm, flash memory solves a real problem: bringing a low-power, rugged file-storage medium to portable devices such as subnotebook computers and PDAs. It also represents an important advance for storing changeable configuration information in consumer devices. Until the advent of flash memory, storing data in a portable device had to be done with battery-backed SRAM; flash doesn't drain the battery when storing data and costs less than SRAM.

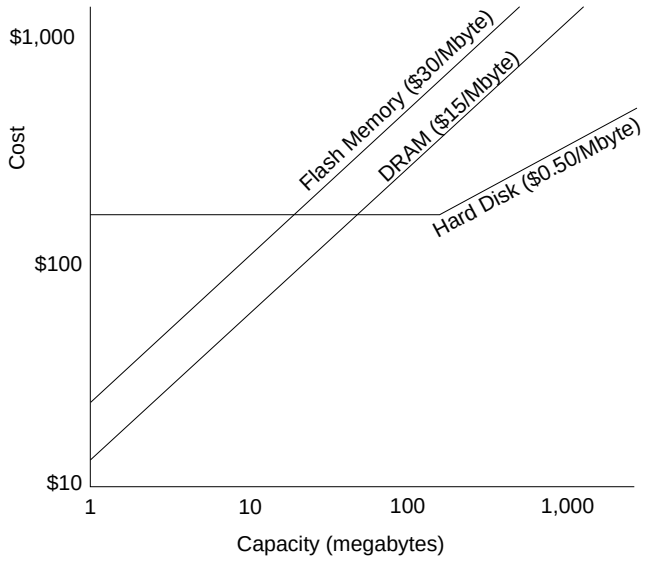

Figure 7. For more than a few megabytes of storage, hard disks deliver the lowest cost, but because of their complex mechanisms they have a "floor" price of about $200. Although not priced competitively with DRAM, for applications requiring a few megabytes of nonvolatile storage, flash memory can be cost-effective.

Choosing flash over other semiconductor memory types is another battle entirely. According to recent figures, the rate of increase in DRAM density has slowed. Intel's Gordon Moore claims flash memory is on a steeper discount curve than DRAM. Understanding the DRAM cell's construction, it's easy to see why: with its large capacitive-storage element, its cell size cannot be reduced as much when fabrication geometries decrease. Without the capacitor, flash memory's single-transistor cells eventually will cost less than DRAM's. Even now, flash memories require less die area than DRAMs of the same capacity built to the same design rules.

Depending on which pundit you believe, flash memory will be cheaper than DRAM sometime between 1995 and 2000. But flash memory still can't compete in applications where bytes or bits must be individually altered or frequently written. Its slow write speed is a serious bottleneck for most DRAM-like applications. And while the flash memory of the future looks good in many ways when compared with today's DRAM, the latter technology is still growing more dense and less expensive.

Flash has made a rapid impact on the design of computer and embedded systems. With the growing interest in portable computer devices, flash's popularity will continue to increase. Its biggest impact on system design will be as a potential supplement to DRAM, allowing fast, low-cost storage of information that is often read but infrequently changed. ♦

