Source: STANDARD HANDBOOK OF ELECTRONIC ENGINEERING

CHAPTER 6.3
MEMORY CIRCUITS

P. K. Vasudev, S. Tewksbury

DYNAMIC RAM

Dynamic RAM (DRAM) provides the highest density IC memory capacity of the various RAM/ROM technologies. With a market driven by the evolution of computers to higher performance products with increasingly powerful microprocessors (and correspondingly larger and faster main memory), the DRAM technologies are among the most sophisticated of the VLSI technologies. This section reviews the basics of DRAM circuits, drawing on the evolution from the earlier small DRAMs (1 kbit 3-transistor/cell memories to 64 kbit 1-transistor/cell memories) to more advanced DRAMs. Given the sophistication of contemporary DRAM ICs, it is not possible to describe in detail today's advanced technologies. However, it is important to recognize that advances other than the underlying transistor-level design and fabrication technology have played a substantial role in providing fast data transfers between the DRAMs and microprocessors.

DRAMs (like SRAMS and ROMS) are generally organized as an array of cells, each cell holding one bit of data. When an address of a data bit is applied, that address specifies which row (the row address part of the address word) of the array is to be selected and, given the row, which data element of the row is to be selected (according to the column address portion of the address word). Earlier DRAM technologies were designed to access a single data bit cell for a given input address, with a new address applied to access another data bit cell. The data bit is stored as a charge on a capacitor. To read that data bit, the row address is applied and the small output voltage from each cell of the row is connected to individual column lines. Each column line connects through a switch to the corresponding (e.g., *N*th) bit of each row and only the addressed row's switch is closed. The small signals from each of the selected row's cells are transmitted to a linear array of amplifiers, converting that small signal into a digital signal depending on whether charge was or was not stored on the cell's capacitor. In this sense, an entire row of the DRAM array is read at the same time and that entire row of data is available when the column address is applied. Earlier DRAM architectures selected one of these data bits as the desired output data, discarding the others. The inefficiency of this approach is clear and architectures have advanced to take advantage of the reading of an entire row of data. One approach was to retain the separate row and column addresses but to allow successive column addresses to be applied (without requiring a new read of the row) in order to select successive data bits already placed in the output stage (avoiding the delay time necessary to complete the reading of the row and amplification of the small signals from the cells of the row). However, the data transfer rates between the DRAM module and the microprocessor remained a substantial bottleneck to increased performance. Microprocessor architectures were advancing rapidly, including addition of cache memories to avoid, when possible, having to access the external DRAM to acquire instructions and data. With the advanced microprocessor architectures, a quite different approach became possible, greatly increasing the net data transfer rates between the DRAM module and the microprocessor. In particular, an entire row can be accessed by the row address and then, rather than reading individual bits of that row by applying a new column address each time, the entire row of data can be transferred to the microprocessor. Although somewhat simplified relative to the actual operation, the basic approach is as follows. A row

Downloaded from Digital Engineering Library @ McGraw-Hill (www.digitalengineeringlibrary.com)
Copyright © 2004 The McGraw-Hill Companies. All rights reserved.
Any use is subject to the Terms of Use as given at the website.

MEMORY CIRCUITS

address is applied and the small signals corresponding to data stored in the cells of that row are passed to the amplifier array, generating digital signals that are loaded into a parallel-load/serial-shift register. Once loaded, the data in the shift register can be clocked out at a high data rate without requiring application of a new address. It is through this process that the very high data rates can be achieved. If the address of the row that will be needed next is known when the current row of data is being shifted out of the DRAM, that address can be applied so that the small signals from that next row are being transferred to the amplifiers and converted to digital signals—ready to be loaded into the register for shifting out as soon as the present data in the register have been shifted out.

This provides a fast transfer of data from the DRAM to the microprocessor. A corresponding approach allows fast data transfer from the microprocessor into the DRAM. To write data, data corresponding to a row are shifted into an input data register. When the register has been filled, its contents are loaded into a parallel data register and are available for writing into the row specified by the write address. While the data in that parallel data register are being written into the DRAM, the next data to be written can be shifted into the serial shift register. The basic approach discussed above has actually been in use for some time, appearing in video RAM. In that case, an entire row of a screen is loaded from memory into a parallel-in/serial-out data register of the read operation to read the row and an entire row of a screen is loaded into the serial-in/parallel-out data register of the write operation. While a row of a data image is being shifted out, another row of the data image can be loaded.

The above example of architectural approaches to relax performance bottlenecks (in this case the data transfers between DRAM and the microprocessor) illustrates an important principle that will increasingly characterize VLSI components. In particular, system-level solutions will increasingly define the functional characteristics of mainstream technologies such as DRAM and microprocessors, providing performance enhancements unachievable through purely technology and circuit design approaches.

Space prohibits detailed discussion of these architectural schemes, or the rather sophisticated techniques that have been used to miniaturize the surface area of the capacitance. However, the underlying principles are similar to earlier memory designs, as summarized below.

DRAM is the lowest-cost, highest-density RAM available. Since the 4k generation DRAM (dynamic RAM) has held a 4 to 1 density advantage over static RAM, its primary competitor. Dynamic RAM also offers low power and a package configuration which easily permits using many devices together to expand memory sizes. Today's computers use DRAM for main memory storage with memory sizes ranging from 16 kbytes to hundreds of megabytes. With these large-size memories, very low failure rates are required. The metal oxide semiconductor (MOS) DRAM has proven itself to be more reliable than previous memory technologies (magnetic core) and capable of meeting the failure rates required to build huge memories. DRAM is the highest volume and highest revenue product in the industry.

Cell Construction

The 1k and early 4k memory devices used the three-transistor (3-T) cell shown in Fig. 6.3.1. The storage of a "1" or a "0" occurred on the parasitic capacitor formed between the gate and source of transistors. Each cell had amplification thus permitting storage on a very small capacitor. Because of junction and other leakage paths the charge on the capacitor had to be replenished at fixed intervals; hence the name dynamic RAM. Typically a *refresh* pulse of a maximum duration was specified.

The next evolution, the 1-T cell, was the breakthrough required to make MOS RAM a major product. The 1-T cell is shown in Fig. 6.3.2. This is the 16k RAM version of the cell. The 4k is similar except that only one level of polysilicon (poly 1) is used. The two-level poly process improves the cell density by about a factor of 2 at the expense of process complexity. The transistor acts as a switch between the capacitor and the bit line. The bit line carries data into and out of the cell. The transistor is enabled by the word line, which is a function of the row address. The row address inputs are decoded such that one out of **N** word lines is enabled. **N** is the number of rows which is a function of density and architecture. The 16k RAM has 128 rows and 128 columns. The storage cell transistor is situated such that its source is connected to the capacitor, its drain is connected to the bit line and the gate is connected to the word line. Higher-density DRAMs today have evolved more sophisticated structures, involving trenches and stacked capacitors to continue meeting the requirements of the 1-T cell. By simultaneously driving down the cell size, while maintaining the required storage capacitance of

Downloaded from Digital Engineering Library @ McGraw-Hill (www.digitalengineeringlibrary.com)
Copyright © 2004 The McGraw-Hill Companies. All rights reserved.
Any use is subject to the Terms of Use as given at the website.

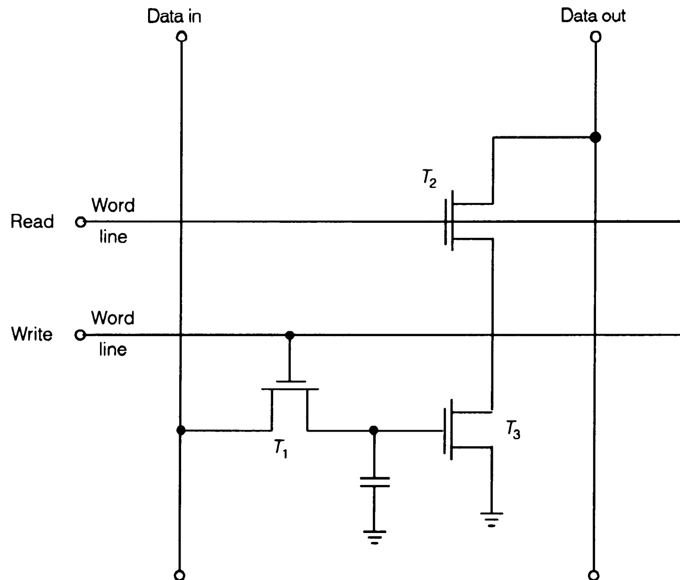

FIGURE 6.3.1 Dynamic RAM cell.

bit line, DRAMs have been able to sustain the steady increase in DRAM capacity, moving beyond the 1 GB level. This steady increase is expected to continue.

Cell Process

The principal process used in early generation DRAMs was the double-level polysilicon gate process. The process uses *n*-channel devices which rely on electron mobility. The 1k RAM (1103) used a *p*-channel process which relied on hole mobility. The *p*-channel process although simpler to manufacture (it is more forgiving to contamination) is by nature slower than the *n*-channel process (electrons are much faster than holes) and cannot operate a 5-V levels. Therefore once manufacturing technology advanced enough to permit *n*-channel to yield, it quickly displaced *p*-channel techniques for memory. Today the DRAM processes are extremely sophisticated, using trenches, stacked capacitors, and high-impedance dielectrics that allow densities in the 100-million gate level to be achieved. They typically use CMOS technology and power supplies as low as 1 V or so.

System Design Consideration Using Dynamic RAMs

Dynamic RAMs provide the advantages of high density, low power, and low cost. These advantages do not come for free; the dynamic RAM is considered to be more difficult to use than static RAMs. This is because dynamic RAMs require periodic refreshing in order to retain the stored data. Furthermore, although not generic to dynamic RAMs, most dynamic RAMs multiplex the address bits which requires more complex edge-activated multi-clock timing relationships. However, once the system techniques to handle refreshing, address multiplexing, and clock timing sequences have been mastered, it becomes obvious that the special care required for dynamic RAMs is only a minor inconvenience.

A synopsis of the primary topics of concern when using dynamic RAMs follows.

Address Multiplexing. The use of address multiplexing reduces the number of address lines and associated address drivers required for memory interfacing by a factor of 2. This scheme, however, requires that the RAM

Downloaded from Digital Engineering Library @ McGraw-Hill (www.digitalengineeringlibrary.com)
Copyright © 2004 The McGraw-Hill Companies. All rights reserved.
Any use is subject to the Terms of Use as given at the website.

MEMORY CIRCUITS

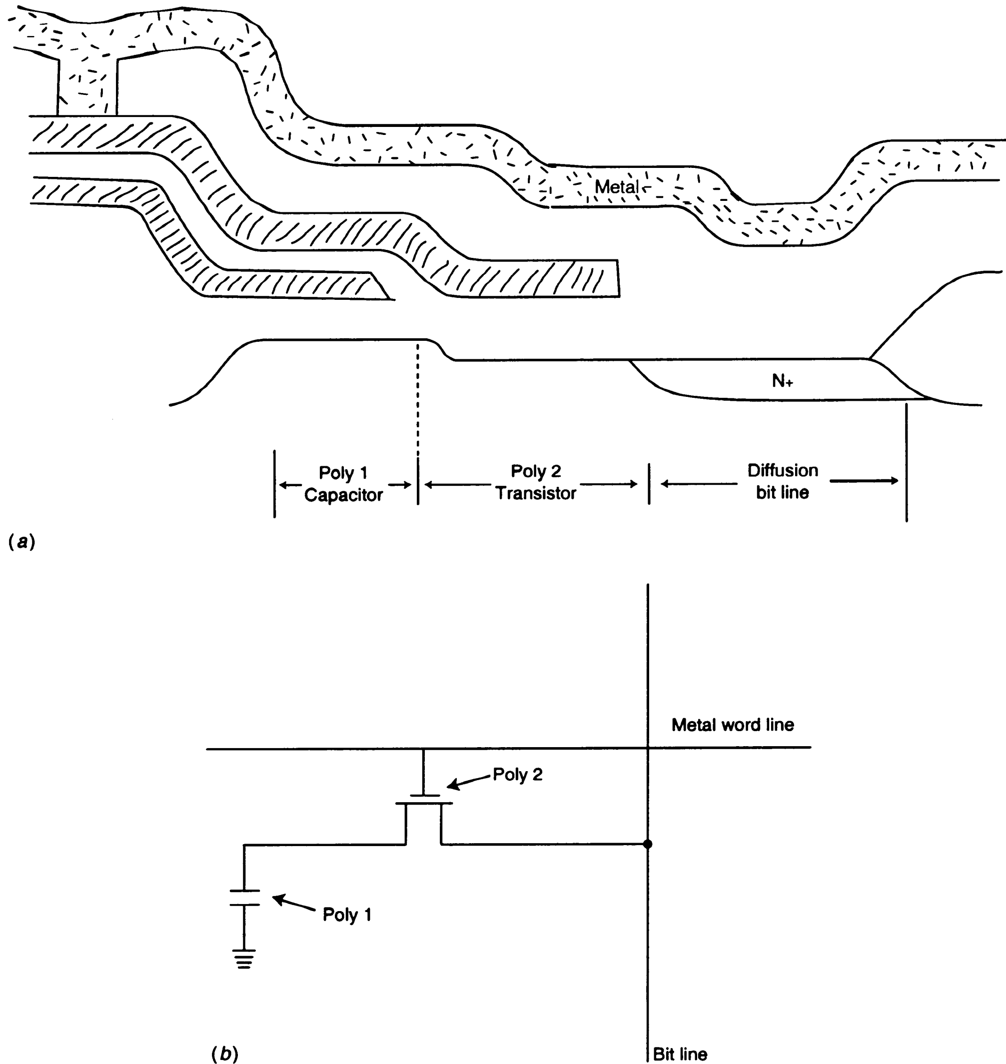

FIGURE 6.3.2 (*a*) DRAM cross section; (*b*) schematic of single-transistor cell DRAM.

address space be partitioned into an *X–Y* matrix with half the addresses selecting *X* or ROW address field and the other half of the address selecting the *Y* or column address field. The ROW addresses must be valid during the RAS clock and must remain in order to satisfy the hold requirements of the on-chip address latches. After the ROW address requirements have been met, the addresses are allowed to switch to the second address field, column addresses, a similar process is repeated using CAN (column address strobe) clock to latch the second (column) address field. Note, for RAS-only refreshing the column address field phase is not required.

Error Detection and Error Correction. In a large memory array there is a statistical probability that a soft error(s) or device failure(s) will occur resulting in erroneous data being accessed from the memory system. The larger the system, the greater the probability of a system error.

Downloaded from Digital Engineering Library @ McGraw-Hill (www.digitalengineeringlibrary.com)
Copyright © 2004 The McGraw-Hill Companies. All rights reserved.
Any use is subject to the Terms of Use as given at the website.

In memory systems using dynamic RAMs organized *N* by one configuration, the most common type of error-failure is single-bit oriented. Therefore, error detection and error correction schemes with limited detection/correction capabilities are ideally suited for these applications.

Single-bit detection does little more than give some level of confidence that an error has not occurred. Single-bit detection is accomplished by increasing the word width by 1 bit.

By adding a parity bit and selecting the value of this bit such that the 4-bit word has an even number of "1s" or an odd word number of "1s," complementing any single bit in the word no longer results in a valid binary combination. Notice that it is necessary to complement at least two bit locations across the word in order to achieve new or new valid data word with the proper number of "1s" or "0s" in the word. In order to extract the error information from the stored data word. This concept was introduced by Richard Hamming in 1950. The scheme of defining the number of positions between any two valid word combinations is called the "Hamming distance." The Hamming distance determines the detection/correction capability.

Correction/detection capability can be extended by additional parity bits in such a way that the Hemming distance is increased. Figure 6.3.2 shows how a 16-bit word with five parity bits (commonly referred to as "check bits") can provide single-bit detection and single-bit correction. In this case, a single-bit error in any location across the word, including the check bits, will result in a unique combination of the parity errors when parity is checked during an access.

As indicated above, the ability to tolerate system errors is brought at the expense of using additional memory devices (redundant bits). However, the inclusion of error correction at the system level has been greatly simplified by the availability of error correction chips. These chips provide the logic functions necessary to generate the check bits during a memory write cycle and perform the error detection and correction process during a memory ready cycle.

STATIC RAM

DRAM requires that the charge stored on the capacitance of the data cells be refreshed at a regular interval since the charge slowly leaks off through parasitic diodes. In addition, the signal provided by the cells when accessed is a small analog signal, requiring amplification and noise avoidance. These penalties are compensated by the requirement that only one transistor is needed for each memory cell. Static random access memory (SRAM) holds its data indefinitely, so long as power is applied to the memory. Each cell is essentially a flip-flop, set to logic "1" or "0" when written and providing a full-amplitude logic signal to the output stage. The number of transistors needed for this storage and to connect the flip-flop to the input and output data lines is typically six. With a regular digital signal provided when reading the SRAM, read delays are substantially smaller than in a DRAM, the fast access time justifying the smaller amount of data that can be stored on an SRAM IC (that lower storage capacity resulting from the larger number of cell transistors than needed in the DRAM). Several applications of this faster SRAM appear in a personal computer—including video memory, cache memory, and so forth. Since there is no need for special microfabrication steps to create the minimum area, large capacitance cells of the DRAM ICs, the SRAM can be fabricated on the same IC as standard digital logic. Although there have been demonstrations of the fabrication of microprocessors and DRAM on the same IC, the general need for more DRAM ICs than microprocessor ICs in a personal computer leads to no major advantage of this more complicated technology. On the other hand, the higher capacity of the DRAM memory technology may drive cointegration with logic circuitry for other applications requiring that higher-density RAM.

Just as in the case of the DRAMs discussed earlier, architectural principles are also a significant part of the SRAM IC design. With multiple levels of SRAM cache memory being used in contemporary microprocessor systems, the cache memories connected directly to the DRAM memory can draw upon some of the advantages of reading/writing entire rows of memory in parallel. Cache memories also generally require sophisticated algorithms to place memory from some real DRAM address location into a local RAM accessed with a different address. Translations and other system-level functions related to such needs can be embedded directly within the SRAM IC itself. As in the case of the DRAMs, space here does not permit a discussion of these advanced themes. However, the underlying principles of the basic SRAM cell are captured in the earlier technologies discussed next.

Downloaded from Digital Engineering Library @ McGraw-Hill (www.digitalengineeringlibrary.com)
Copyright © 2004 The McGraw-Hill Companies. All rights reserved.
Any use is subject to the Terms of Use as given at the website.

MEMORY CIRCUITS

Organization

Static RAMs in the 1k (k = 1024) and 4k generations are organized as a square array (number of rows of cells equals the number of columns of cells). For the 1k device there are 32 rows and 32 columns of memory cells. For a 4096 bit RAM there are 64 rows and 64 columns. The RAMS of the seventies and beyond include on-chip decoders to minimize external timing requirements and number of input/output (I/O) pins. Today, much higher-density SRAMs are being built using the same basic array design, but employing scaled silicon structures.

To select a location uniquely in a 4096-bit RAM, 12 address inputs are required. The lower-order 64 addresses decode to select one column. The intersection of the selected row and the selected column locates the desired memory cell.

Static RAMs of the late seventies and early eighties departed from the square array organization in favor of a rectangle. This occurred for primary reasons: performance and packaging considerations.

Static RAM has developed several application niches. These applications often require different organizations and/or performance ranges. Static RAM is currently available in 1-bit I/O, bout bit I/O, and 8-bit I/O configurations. The 1- and 4-bit configurations are available in a 0.3-in-wide package with densities of 1k, 4k, and 16k bits. The 8-bit configuration is available in a 0.6-in-wide package in densities of 8k and 16k bits. The 8-bit I/O device is pin compatible with ROM (read only memory) and EPROM (electrically programmable ROM) devices. All of the static RAM configurations are offered in two speed ranges. Ten nanoseconds typically are used as the dividing line for part numbering.

Construction

The static RAM uses a six-device storage cell. The cell consists of two cross-coupled transistors, two I/O transistors, and two load devices. There are three techniques that have been successfully employed for implementing the load devices: enhancement transistors, depletion transistors, and high-impedance polysilicon load resistors.

Enhancement transistors are normally off and require a positive gate voltage (relative to the source) to turn the device on. There is a voltage drop across the device equal to the transistors' threshold. Depletion transistors are normally on and require a negative gate voltage (relative to the source) to turn them off. There is no threshold drop across this device. Polysilicon load resistors are passive loads which are always on. Formed in the polylevel of the circuit they are normally undoped and have very high impedances (5000 MΩ typically). They have the advantage of very low current (nanoamperes) and occupy a relatively small area. All present large density, state of the art, static RAMs now use this technique for the cell loads. Further enhancements have been made by manufacturing a two-level polyprocess, thereby further decreasing cell area.

The operation of the polyload SRAM cell is shown in Fig. 6.3.3. The cross-coupled nature of the flip-flop is easily noticed.

Characteristics

Static RAM is typified by ease of use. This comes at the expense of a more complex circuit involving six or more transistors in a cross-coupled array. Thus, the total transistor count or density is four times higher than a DRAM.

The fundamental device is often referred to as a ripple through static RAM. Its design objective is ease of use. This device offers one power dissipation mode "on". Therefore its selected power (active power) is equal to its deselected power (standby power). One need simply present an address to the device and select it (via chip select) to access data. Note that the write signal must be inactive.

The address must be valid before and after read/write active to prevent writing into the wrong cell. Chip select gates the write pulse into the chip and must meet a minimum pulse width which is the same as the read/write minimum pulse width.

Read modify write cycles can be performed by this as well as other static RAMs. Read modify write merely means combining a read and write cycle into one operation. This is accomplished by first performing a read

Downloaded from Digital Engineering Library @ McGraw-Hill (www.digitalengineeringlibrary.com)
Copyright © 2004 The McGraw-Hill Companies. All rights reserved.
Any use is subject to the Terms of Use as given at the website.

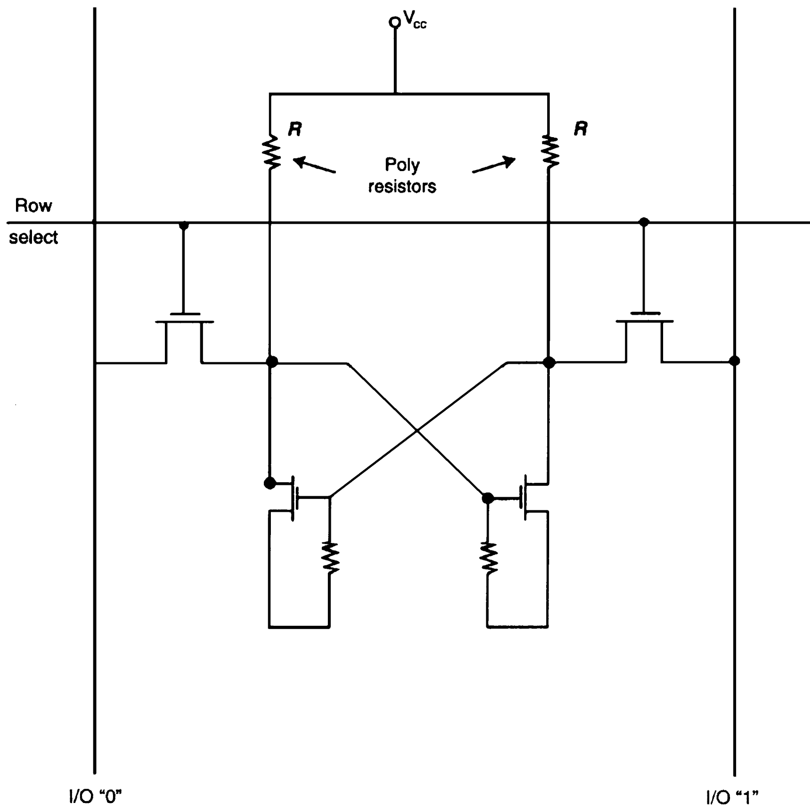

FIGURE 6.3.3 Polyresistor load static RAM cell.

then activating the read write line. The status of the data out pin during the write operation varies among vendors. The product's data sheet therefore be read carefully.

The device achieves its ease of use by using very simplistic internal circuitry. To permit a nonsynchronous interface (no timing sequence required) all circuitry within the part is active at all times. That static "NOR" row decoder is typical of the circuitry used. For a 4k RAM with 64 rows, this circuit is repeated 64 times with the address lines varying to uniquely identify each row.

In the static row decoder, a row is selected when all address inputs to the decoder are low making the output high. The need to keep the power consumption of this decoder at a minimum is in direct conflict with the desire to make it as fast as possible. The only way to make the decoder fast is to make the pull-up resistance small so that the capacitance of the word line can be changed quickly. However, only one row decoder's output is high; while the output of each of the other 63 decoders is low, causing the resistor current to be shunted to ground. This means that the pull-up resistance must be large in order to reduce power consumption.

Application Areas

Static RAM applications are segmented into two major areas based on performance. Slower devices, greater than 40 ns access time, are generally used with microprocessors. These memories are usually small (several thousands of bytes) with a general preference for wide word organization (by 8 bit). When by 1 organized memories are used, the applications tend to be "deeper" (larger memory) and desire lower power or parity.

Downloaded from Digital Engineering Library @ McGraw-Hill (www.digitalengineeringlibrary.com)
Copyright © 2004 The McGraw-Hill Companies. All rights reserved.
Any use is subject to the Terms of Use as given at the website.

MEMORY CIRCUITS

The by 4-bit product is an old design which is currently being replaced by 1k × 8 and 2k × 8 static RAM devices. Statics are used in these applications because of their ease of use and compatibility with other memory types (RAM, EPROM). The by 4 and by 8 devices use a common pin for data in and data out. In this configuration, one must be sure that the data out is turned off prior to enabling the data in circuitry to avoid a bus conflict (two active devices fighting each other drawing unnecessary current). Bus contention can also occur if two devices connected to the same bus are simultaneously on fighting each other for control. The by 8 devices include an additional function called output enable (OE) to avoid this problem. The output enable function controls the output buffer only and can therefore switch it on and off very quickly. The by 4 devices do not have a spare pin to incorporate this function. Therefore, the chip select (CS) function must be used to control the output buffer. Data out turns off a specified delay after CS turn off. One problem with this approach is that it is possible for bus contention to occur if a very fast device is accessed while a slow device is being turned off. During a read/modify/write cycle the leading edge of the read write line (WE) is normally used to turn off the data out buffer to free the bus for writing.

The second segment is high-performance applications. Speeds required here are typically 10–20 ns. This market is best serviced by static RAM because of their simple timing and faster circuit speeds. The requirement of synchronizing two "reads" as required in clocked part typically cost the system designer 10–20 ns. Therefore, fast devices are always designed with a ripple through interface. The by 1 devices have been available longer, and tend to be faster than available by 8 devices. As a result the by 1 device dominates the fast static RAM market today. Applications, such as cache and writable control store, tend to prefer wide word memories. The current device mix should shift over time as more suppliers manufacture faster devices.

Process

The pin-outs and functions tend to be the same for fast and slow static devices. The speed is achieved by added process complexity and the willingness to use more power, for a given process technology.

The key to speed enhancement is increasing the circuit transistor's gain, minimizing unwanted capacitance, and minimizing circuit interconnect impedance. Advanced process technologies have been developed to achieve these goals. In general, they all reduce geometries (scaling) while pushing the state of the art in manufacturing equipment and process complexity. Figure 6.3.4 shows the key device parameters of MOSFET scaling, one of which is device gain. The speed of the device is proportional to the gain. Because faster switching

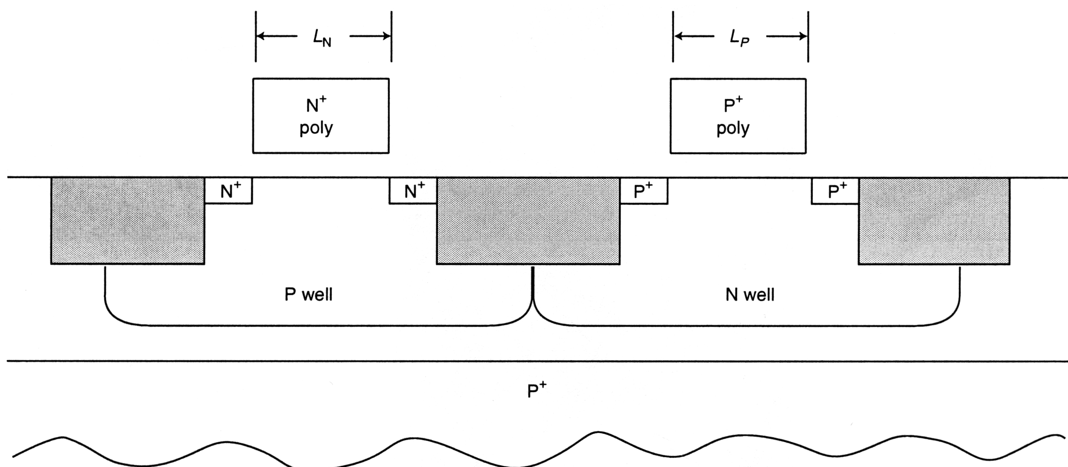

FIGURE 6.3.4 Cross section of submicron scaled CMOS gates for high-density SRAMs.

Downloaded from Digital Engineering Library @ McGraw-Hill (www.digitalengineeringlibrary.com)
Copyright © 2004 The McGraw-Hill Companies. All rights reserved.
Any use is subject to the Terms of Use as given at the website.

speeds occur with high gain, the gain is maximized for high speed. Device gain is inversely proportional to the gate oxide thickness and device length. Consequently, scaling these dimensions increases the gain of the device.

Another factor that influences performance is unwanted capacitance, which appears in two forms: diffusion and Miller capacitance. Diffusion capacitance is directly proportional to the overlap length of the gate and the source. Capacitance on the input shunts the high-frequency portion of the input signal so that the device can only respond to low frequencies. Second, capacitance from the drain to the gate forms a feedback path creating an integrator or low-pass filter which degrades the high-frequency performance.

One of the limits on device scaling is punch through voltage, which occurs when the field strength is too high, causing current to flow when the device is *turned off*. Punch through voltage is a function of channel length and doping concentration thus channel shortening can be compensated by increasing the doping concentration. This has the additional advantage of balancing the threshold voltage, which was decreased by scaling the oxide thickness for gain.

NONVOLATILE MEMORIES

DRAM and SRAM memories retain their data so long as power is applied but lose their data when the power is disrupted or turned off. In this sense, they are "volatile" memories. Nonvolatile memories retain their data even when the power is turned off. In the case of personal computers, the hard disk drive serves the essential "nonvolatile memory" function, in combination with a nonvolatile IC memory capable of starting the process of loading information from the hard drive into the computer system. The fundamental importance of nonvolatile IC memory can be seen by the rapid trend toward programmable components within many electronics applications not able to handle the inconvenience of an external hard drive. The term *embedded system* reflects this insertion of microprocessors into a vast range of applications. As one example, the microprocessor used to control a refrigerator requires storage of its *program*, a requirement that can be handled by a relatively simple programmable read only memory (PROM). Other applications of embedded systems require memories that can be written as well as read, with the additional requirement that the information written into the memory be preserved when power is removed. If this writing is very modest, then extensions of simple PROM ICs with slow write times is acceptable, as in the case of an electronic address book. But as the embedded systems have become more sophisticated and migrated closer to a personal computer in their general (not detailed) operation, the equivalent of a hard disk drive in IC form with not only fast reads but also fast writes becomes more important. To support such applications, there have been major advances in nonvolatile memory technologies recently, in several cases drawing on storage mechanisms quite different than those used in traditional LSI/VLSI components.

To illustrate the extent to which small embedded systems mimic larger personal computer systems, the example of the Motorola MC680xx 16/32-bit microprocessor can be used. This microprocessor (introduced in its initial version, the MC68000, in 1979 and requiring only about 68,000 transistors) was the processor that drove the early Apple computers. With today's VLSI able to provide several hundred millions of transistors in a single IC, placing the MC68000 on a VLSI circuit would require only about 0.001 percent of the IC area (assuming 68 million transistors). Equivalently, 1000 MC68000 microprocessors could be placed on a single IC. In this sense, adding a microprocessor to a digital circuit intended for embedded applications is not a major consumer of IC resources. As such embedded systems (particularly those that embed an earlier generation personal computer in its IC) continue to advance into a wider range of applications, the demand for high-performance nonvolatile memories with high-speed read and write capabilities (as well as low-voltage write conditions) will increase. As noted earlier, SRAM (with cells that are basically flip-flops) can be fabricated on the same IC as logic circuits, a major advantage over DRAM in many applications. Nonvolatile memories that can be fabricated on the same IC as logic will also be of major importance as the underlying microfabrication technologies continue to advance, seeking system-on-a-chip or system-on-a-(few) chip (set) capabilities.

The field of advanced technologies for nonvolatile memories is in a state of flux and a variety of quite different technical approaches are under study. Underlying these new technologies and approaches there remains the basic principles of nonvolatile memory seen in the traditional PROM, EPROM, and EEPROM technologies that continue to play a major role in systems designs.

Downloaded from Digital Engineering Library @ McGraw-Hill (www.digitalengineeringlibrary.com)
Copyright © 2004 The McGraw-Hill Companies. All rights reserved.
Any use is subject to the Terms of Use as given at the website.

Mask-Programmed ROM

In mask-programmed ROM, the memory bit pattern is produced during fabrication of the chip by the manufacturer using a mask operation. The memory matrix is defined by row (X) and column (Y) bit-selection lines that locate individual memory cell positions.

ROM Process

For many designs, fast manufacturing turnaround time on ROM patterns is essential for fast entry into system production. This is especially true for the consumers' "games" market. Several vendors now advertise turnaround times that vary from 2 to 6 weeks for prototype quantities (typically small quantities) after data verification. Data verification is the time when the user confirms that data have been transferred correctly into ROM in accordance with the input specifications.

Contact programming is one method that allows ROM programming to be accomplished in a shorter period of time than with gate mask programming. In mask programming, most ROMS are programmed with the required data bit pattern by vendors at the first (gate) mask level, which occurs very early in the manufacturing process. In contact programming, actual programming is not done until the fourth (contact) mask step, much later in the manufacturing process. That technique allows wafers to be processed through a significant portion of the manufacturing process, up to "contact mask" and then stored until required for a user pattern. Some vendors go one step further and program at fifth (metal) mask per process. The results in a significantly shorter lead time over the old gate-mask programmable time of 8 to 10 weeks; the net effect is time and cost savings for the end user.

ROM Applications. Typical ROM applications include code converters, look-up tables, character generators, and nonvolatile storage memories. In addition, ROMs are now playing an increasing role in microprocessor-based systems where a minimum parts configuration is the main design objective. The average amount of ROM in present microprocessor systems is in the 10 to 32 kbyte range. In this application, the ROM is often used to store the control program that directs CPU operation. It may also store data that will eventually be output to some peripheral circuitry through the CPU and the peripheral input/output device.

In a microprocessor system development cycle, several types of memory (RAM, ROM, and EPROM or PROM) are normally used to aid in the system design. After system definition, the designer will begin developing the software control program. At this point, RAM is usually used to store the program, because it allows for fast and easy editing of the data. As portions of the program are debugged, the designer may choose to transfer them to PROM or EPROM while continuing to edit in RAM. Thus, he avoids having to reload fixed portions of the program into RAM each time power is applied to the development system.

Electrically Erasable Programmable ROM

The ideal memory is one that can perform read/write cycles at speeds meeting the needs of microprocessors and store data in the absence of power. The EEPROM (electrically erasable PROM) is a device which meets these requirements. The EEPROM can be programmed in circuit and can selectively change a byte of memory instead of all bytes. The process technology to implement the EEPROM is quite complex and the industry is currently on the verge of mastering production.

EEPROM Theory. EEPROMs use a floating gate structure, much like the ultraviolet erasable PROM (EPROM), to achieve nonvolatile operation. To achieve the ability to electrically erase the PROM, a principle known as Fowler-Nordheim tunnelling was implemented. Fowler-Nordheim tunnelling predicts that under a field strength of 10 MV cm^{-1} a certain number of electrons can pass a short distance, from a negative electrode, through the forbidden gap of an insulator entering the conduction band and then flow freely toward a positive electrode. In practice the negative electrode is a polysilicon gate, the insulator is a silicon dioxide and the positive electrode is the silicon substrate.

Downloaded from Digital Engineering Library @ McGraw-Hill (www.digitalengineeringlibrary.com)
Copyright © 2004 The McGraw-Hill Companies. All rights reserved.
Any use is subject to the Terms of Use as given at the website.

MEMORY CIRCUITS

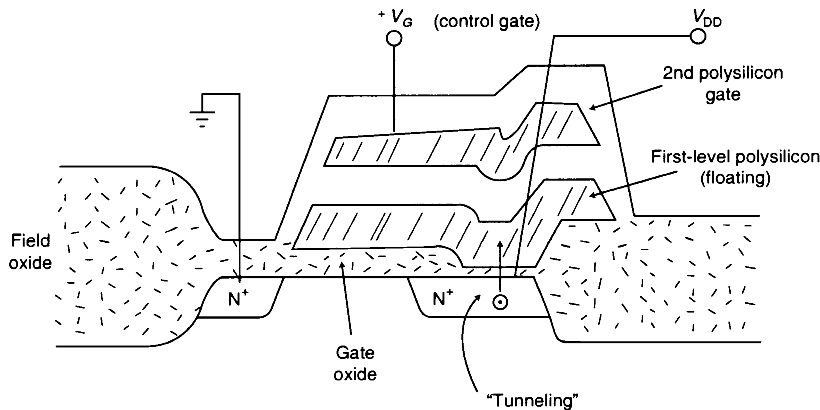

FIGURE 6.3.5 EEPROM cell using Fowler-Nordheim tunnellings.

Fowler-Nordheim tunnelling is bilateral, in nature, and can be used for charging the floating gate as well as discharging it. To permit the phenomenon to work at reasonable voltages (e.g., 20 V), the oxide insulator needs to be less than 200 Å thick. However, the tunnelling area can be made very small to aide the manufacturability aspects (20-nm oxides are typically one-half the previously used thickness).

Intel Corporation produces an EEPROM based on this principle. Intel named the cell structure FLOTOX. A cross section of the FLOTOX device is shown in Fig. 6.3.5.

The FLOTOX structure resembles the original structure used by Intel for EPROM devices. The primary difference is in the additional tunnel-oxide region over the drain. To charge the floating gate of the FLOTOX structure, a voltage V_G is applied to the top gate and with the drain voltage V_D at 0 V, the floating gate is capacitively coupled to a positive potential. Electrons will then flow to the floating gate. If a positive potential is applied to the drain and the gate is grounded, the process is reversed and the floating gate is discharged.

The EEPROM designs introduced today are configured externally to be compatible with ROM and EPROM standards which already exist. Devices typically use the same 24-pin pin-out as the generic 2716 (2k × 8 EPROM) device. A single 5 V supply is all that is needed for read operations. For the write and clear operations, an additional supply (V_{pp}) of 20 V is necessary. The device reads in the same manner as the EPROM it will eventually replace.

EEPROM Applications. The EEPROM has the nonvolatile storage characteristics of core, magnetic tape, floppy, and Winchester disks but is a rugged low-power solid-state device and occupies much less space. Solid-state nonvolatile devices, such as ROM and EPROM, have a significant disadvantage in that they cannot be deprogrammed (ROM) or reprogrammed in place (EPROM). The nonvolatile bipolar PROM blows fuses inside the device to program. Once set the program cannot be changed, greatly limiting their flexibility. The EEPROM, therefore, has the advantages of program flexibility, small size, and semiconductor memory ruggedness (low voltages and no mechanical parts).

The advantages of the EEPROM create many applications that were not feasible before. The low power supports field programming in portable devices for communication encoding, data formatting and conversion, and program storage. The EEPROM in circuit change capability permits computer systems whose programs can be altered remotely, possible by telephone. It can be changed in circuit to quickly provide branch points or alternate programs in interactive systems.

The EEPROMs nonvolatility permits a system to be immune to power interruptions. Simple fault tolerant multiprocessor systems also become feasible. Programs assigned to a processor that fails can be reassigned to the other processors with a minimum interruption of the system. Since a program can be backed up into EEPROM in a short period of time, key data can be transferred from volatile memory during power interruption and saved. The user will no longer need to either scrap parts or make service calls should a program bug be discovered in fixed memory. With EEPROM this could even be corrected remotely. The EEPROM's flexibility will create further applications as they become available in volume and people become familiar with their capabilities.

Downloaded from Digital Engineering Library @ McGraw-Hill (www.digitalengineeringlibrary.com)
Copyright © 2004 The McGraw-Hill Companies. All rights reserved.
Any use is subject to the Terms of Use as given at the website.

Erasable Programmable ROM

The EPROM like the EEPROM satisfies two of our three requirements. It is nonvolatile and can be read at speeds comparable with today's microprocessors. However, its write cycle is significantly slower, like the EEPROM. The EPROM has the additional disadvantage of having to be removed from the circuit to be programmed as contrasted to the EEPROM's ability to be programmed in circuit.

EPROM is electrically programmable, then erasable by ultraviolet (UV) light, and programmable again. Erasability is based on the floating gate structure of *n*- or *p*-channel MOSFIT. This gate, situated within the silicon dioxide layer, effectively controls the flow of current between the source and drain of the storage device. During programming, a high positive voltage (negative if *p*-channel) is applied to the source and gate of a selected MOSFET, causing the injection of electrons into the floating silicon gate. After voltage removal, the silicon gate retains its negative charge because it is electrically isolated (within the silicon dioxide layer) with no ground or discharge path. This gate then creates either the presence or absence of a conductive layer in the channel between the source and the drain directly under the gate region. In the case of an *n*-channel circuit, programming with a high positive voltage depletes the channel region of the cell; thus a higher turn-on voltage is required than on an unprogrammed device. The presence or absence of this conductive layer determines whether the binary 1 bit or the 0 bit is stored. The stored bit is erased by illuminating the chip's surface with UV light. The UV light sets up a photocurrent in the silicon dioxide layer which causes the charge on the floating gate to discharge into the substrate. A transparent window over the chip allows the user to perform erasing, after the chip has been packaged and programmed, in the field.

Programmable ROM

This PROM has a memory matrix in which each storage cell contains a transistor or diode with fusible link in series with one of the electrodes. After the programmer specifies which storage cell position should have a 1 bit or 0 bit, the PROM is placed in a programming toll which addresses the locations designated for a 1 bit. A high current is passed through the associated transistor or diode to destroy (open) the fusible link. A closed fusible link may represent a 0 bit, while an open link may represent a 1 bit (depending on the number of data inversions done in the circuit). A disadvantage of the fusible-link PROM is that its programming is permanent; that is, once the links are opened, the bit pattern produced cannot be changed.

Shadow RAM

A recently introduced RAM concept is called the shadow RAM. This approach to memory yields a nonvolatile RAM (data are retained even in the absence of power), by combining a static RAM cell and an electrically erasable cell into a single cell. The EEPROM shadows the status RAM on a bit-by-bit basis. Hence the name shadow RAM. This permits the device to have read/write cycle times comparable to a static RAM, yet offer nonvolatility.

The nonvolatility has a limit in that the number of write cycles is typically 1000 to 1 million maximum. The device can, however, be read indefinitely. Currently two product types are offered. One permits selectively recalling bits stored in the nonvolatile array, while the other product recalls all bits simultaneously.

Typical memory subsystems used in computers employ a complex combination of dynamic RAMs, static RAMs, and nonvolatile ROMs.

BIBLIOGRAPHY

Donnelly, W., "Memories—New Generations Push Technology Forward," *Electronic Industry*, October 1982.

Eaton, S. S., and D. Wooton, "Circuit Advances Propel 64 K RAM Across the 100 ns Barrier," *Electronics*, 24 March 1982.

Threewitt, B., "A VLSI Approach to Cache Memory," *Computer Design*, January 1982.

Whittier, R. J., "Semiconductor Memories," *Mini-Micro Systems*, December 1982.

Wilcock, J. D., "Semiconductor Memories," *New Electronics*, August 17, 1982.

Downloaded from Digital Engineering Library @ McGraw-Hill (www.digitalengineeringlibrary.com)
Copyright © 2004 The McGraw-Hill Companies. All rights reserved.
Any use is subject to the Terms of Use as given at the website.