nonvolatile memory technology a promising candidate for low-voltage, high-density, low-cost, future-generation products.

See also: Conductivity, Electrical; Electrons and Holes; Integrated Circuits; Memory Devices, Volatile; Semiconductor Devices; Transistors.

PACS: 85.25.Hv; 84.30.−r; 85.30.Tv; 84.30.Bv; 84.30.Sk

Further Reading

Brown WD and Brewer JE (1997) *Nonvolatile Semiconductor Memory Technology – A Comprehensive Guide to Understanding and Using NVSM Devices*. New York: Wiley IEEE Press.

Cappelletti P, Golla C, Olivo P, and Zanoni E (1999) *Flash Memories*. London: Kluwer Academic Publishers.

Compardo G and Micheloni R (eds.) (2003) *Special Issue on: Flash Memory Technology*. Proceedings of the IEEE, vol. 91, no. 4, April.

Prince B (2002) *Emerging Memories – Technologies and Trends*. Norwell: Kluwer Academic Publishers.

Memory Devices, Volatile

G Baccarani and E Gnani, University of Bologna, Bologna, Italy

Introduction

Modern computer architectures rely on a hierarchy of memories, which differ in access times and density. The higher speed is typically achieved at the expense of a larger cell area, smaller memory capacity and, thus, higher cost per bit. At the upper levels of the hierarchy one finds registers, which are deeply rooted within the logic circuits of the processor, and three levels of cache memory, with the first two levels usually embedded within the processor chip. Caches typically employ static random access memories (SRAMs) based on either six or four transistors per cell. At the next level, the computer central memory uses dynamic random-access memories (DRAMs) based on the one-transistor memory cell. At the lower levels of the hierarchy, one finds several mass-memory devices, namely, magnetic disks optical disks, such as CDs and DVDs, and flash memory cards. Magnetic tapes are used to archive large amount of data to be seldom retrieved. As opposed to SRAMs and DRAMs, which are volatile memories, mass-memory devices are nonvolatile, that is, they retain the stored information upon power loss or shutdown. This is an essential requisite, which allows information to be safely stored and retrieved whenever necessary.

To date (2004), memory devices have found widespread use not just within computers, but also in a variety of professional systems, such as telecommunication networks and phone exchanges, robotic and control systems, automotive and medical electronics systems, as well as commodities and consumer products, such as cell phones, personal digital assistants (PDAs), digital video recorders and still cameras, digital TV (DTV), set-top boxes, GPS-based navigation systems, toys, and smart cards. Despite this wide variety of applications, which require a different throughput, latency and power consumption, and the related diversification of the memory market, the emphasis, in what follows, is on unifying factors, such as basic physical principles, and general architectural features, which allow one to treat this subject within the constraints of a tutorial presentation.

This treatment specifically addresses semiconductor memories (mass memory devices based on magnetic and optical storage are discussed elsewhere in the encyclopedia) and is split into two parts: one part is devoted to volatile memories, which are by far the most widespread devices in view of their excellent performance and density; the other part is devoted instead to nonvolatile memories, which are being used only when the nonvolatility is an essential requisite of the application.

The classification of semiconductor memories is presented in the next section. The section "Volatile memories" discusses the general features of such devices, including access time, packing density, and reliability issues. The section "Static random access memory (SRAM)" is devoted to the description of the basic cell topologies, cell-array organization, and sense-amplifier operation. The section "Dynamic random access memory (DRAM)" addresses the structural and functional properties of the one-transistor memory cell, the organization of the cell array, the structure and operation of the sense amplifier and the memory architecture. Finally, the section "Content-addressable memory (CAM)" contains a short description of the structure and operation of associative memories, as well as their basic application as a tag memory of set-associative caches.

Memory Classification

Semiconductor memories can be classified according to both structural and functional criteria. Following

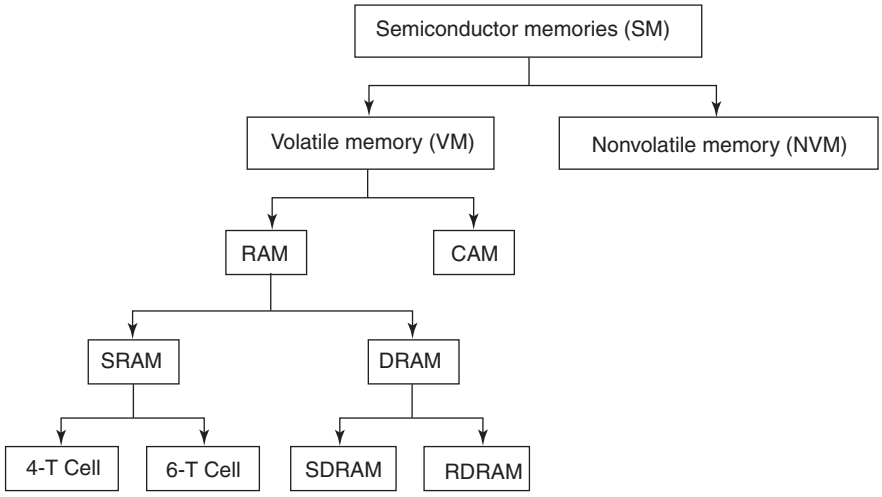

Figure 1 Semiconductor volatile memory classification according to functional criteria.

the latter approach, they are split first into volatile and nonvolatile memories, as indicated in **Figure 1**. As pointed out in the section, "Introduction," volatile memories do not retain the stored information if the power supply is turned off, while nonvolatile memories do retain the stored information in the above conditions. Also, volatile memories can be written and read on field in very short times. Nonvolatile memories, instead, can only be read on field in very short times, whereas repeated programming and erasing operations may not be possible or, if possible, typically require much longer times. Different physical approaches are used to store information in the above memories.

Most volatile memories are randomly accessed, and are thus referred to as random access memories (RAMs). Sequential memories, where bits are written and read sequentially, are rarely used, being outperformed by RAMs both in terms of speed and power consumption.

Content-addressable memories (CAM) are functionally and structurally different from RAMs. In the latter memories, every single bit, byte, or word can be accessed from the knowledge of its physical address. CAMs are, instead, addressed by an associative search using a data word as input, and provide an address as a result of this search.

Volatile memories store the information either statically (SRAM) or dynamically (DRAM). In the former case, the storing mechanism is based on a regenerative circuit characterized by two stable states. In the latter, information is stored as a charge on a capacitor.

SRAMs use a basic cell containing either six or four transistors per cell. The former type of cell is obviously larger, but is compatible with the standard CMOS technology, and is typically used for small, fast, embedded SRAMs employed in first- and second-level caches. The latter type requires a thin-film transistor (TFT) technology not compatible with the standard CMOS process. On the other hand, it provides a higher density and is usually preferred for large third-level caches outside the processor chip.

Dynamic RAMs may be characterized by different architectures. In the past, DRAMs used to be asynchronous and could start a memory cycle whenever two external commands referred to as row address strobe (RAS) and column address strobe (CAS) were successively asserted. More recently, the need to improve the cycle time of DRAMs has suggested more sophisticated architectures, which operate synchronously under the control of the system clock. To date, the synchronous DRAM (SDRAM) is the most widely used memory architecture. An improved performance is made possible by the Rambus DRAM (RDRAM), by which a careful design of the memory board allows for very large clock frequencies.

Volatile Memories

The most important parameter characterizing the memory performance is the access time, that is, the time elapsed between a read request and the availability of the requested word. The cycle time is, instead, the minimum time between successive read requests; its inverse is the memory throughput. The access time of SRAMs may vary from one to a few nanoseconds, depending on the memory size and the resulting capacitances of the word and bit lines. Therefore, these memories are used at the highest levels of the hierarchy within advanced computers and PCs.

DRAMs feature the smallest bit size and are typically used for the computer central memory. To date,

a central memory of 1 GB or more is common practice for high-end desktop and even portable PCs. Servers or mainframes typically employ several tens or hundreds of giga bytes of central memory; hence, the one-transistor DRAM cell is by far the most largely produced semiconductor device in modern industry. Under the assumption of an average main memory of 1 GB and a worldwide production of 100 million computers per year, it turns out that $\approx 10^{18}$ DRAM cells are manufactured every year.

The access time of DRAMs is $\sim$40–60 ns. Therefore, accessing data stored within the central memory of a computer requires several tens of clock cycles for a modern processor.

Volatile memories can suffer data loss due to soft errors, which are caused by the interaction of energetic nuclear particles, such as α, β, and cosmic rays, with the silicon substrate. These events may occur due to the radioactive decay of some impurity atoms contained within the package material and/or the metal layers of the memory chip. Alternatively, as for cosmic rays, these energetic particles may come from the outside space and occasionally hit the memory device. A charged energetic particle penetrating through a semiconductor substrate generates a large number of electron–hole pairs, which are then separated by the electric field within the device-active region. The generated carriers can possibly compensate the charge stored within a cell capacitor, and modify its state. Even static memories are not immune to soft errors, for an energetic particle impinging onto a bi-stable circuit can upset it and change its state. In both cases, a soft error would occur.

In order to ensure the full reliability of the system, several strategies are pursued. At the device level, the geometry of the space charge region is designed in such a way as to drain excess carriers generated by the energetic particle by reverse-biased *p–n* junctions, so that the charge stored within a cell capacitor can hardly suffer a substantial change; also, the bits of a word are not stored at consecutive physical locations, so that, at most, one bit of a word can be changed due to the occurrence of a soft error. At the system level, a few more parity bits are stored together with a 32-bit word. When a word is retrieved from the memory, the examination of the parity bits makes it possible to detect the occurrence of one or more soft errors and also correct the wrong bit.

Static Random Access Memory

The simplest bi-stable circuit is obtained by cross-connecting two CMOS inverters M_1–M_2 and M_3–M_4, as shown in **Figure 2**. Clearly, such a circuit can be biased in either of the two states, characterized by

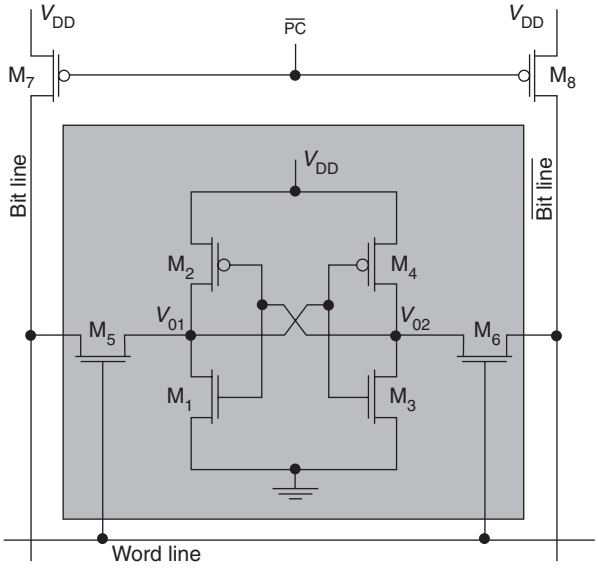

Figure 2 The six-transistor CMOS SRAM cell.

$V_{01} = 0$, $V_{02} = V_{DD}$ or $V_{01} = V_{DD}$, $V_{02} = 0$, respectively. By associating the logical value "0" to the former, and the logical value "1" to the latter bias condition, one stores one bit of information within such a circuit. So long as the power supply is on, this cell will indefinitely preserve the stored information, hence the definition of static memory.

In order to access the information, however, two more devices M_5 and M_6 are needed within the cell, as shown in the same figure. These devices act as pass transistors, which isolate the cell if the word line is low, or connect the cell to the bit lines if the word line is high. Hence, the total number of transistors per cell is six. The bit lines, which are shared by all cells within the same column, provide the readout information to the sense amplifier, that is, the circuit which generates a logic output data from a slight unbalance of the bit lines.

In four-transistor memory cells, the pullup metal-oxide-semiconductor field effect transistor (MOSFETs) M_2 and M_4 are replaced by either passive resistors, as shown in **Figure 3**, or TFTs made by polycrystalline silicon.

In the former case, their resistance R must be very large, since the standby current consumption of a cell is given by V_{DD}/R. The resistance value is thus defined according to the need of supplying the minimum current to preserve the high voltage of the node. In practice, $R \approx 1$–$10\ G\Omega$, which can be achieved using lightly doped polycrystalline silicon films. The use of TFTs as pullup devices reduces the leakage current by nearly two orders of magnitude with respect to the passive resistor cell, and improves the current-drive capability of the pullup devices in the on state.

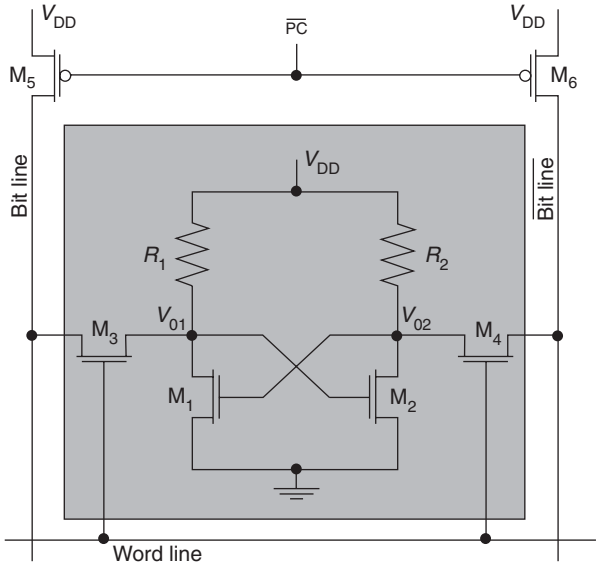

Figure 3 The four-transistor CMOS SRAM cell.

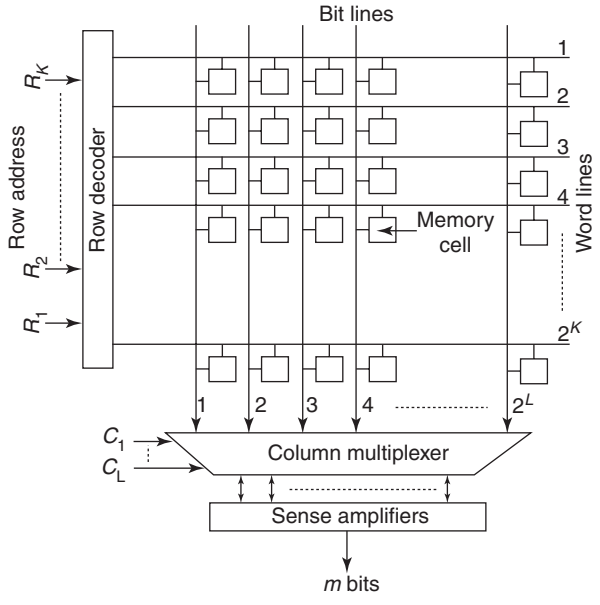

Figure 4 SRAM cell array organization.

This improvement can be achieved at the expense of a more complex technology, involving an additional masking step, and the deposition of an undoped polycrystalline silicon film.

Cell Array Organization

The cell array of an SRAM is organized as indicated in **Figure 4**. A row decoder receives the row address of the cell at the input and raises the corresponding word line high. Hence, all the cells pertaining to the above row become connected to the respective bit lines.

The output multiplexer acts as a column decoder, and selects the appropriate bit, byte, or word. The corresponding bit lines are then connected to the sense amplifier(s), which detect the stored information within the selected cells. If the depth of the column decoder is too large, it may be worth interposing the sense amplifiers between a first partial column decoder, acting on the bit lines which carry analog signals, and a second column decoder, acting on the digital outputs of the sense amplifiers. Clearly, reducing the number of sense amplifiers is an advantage from the standpoint of power consumption, but, on the other hand, interposing too many pass transistors between the bit lines and the inputs of the sense amplifier could be detrimental to the accuracy of the readings.

It should be noticed that, if the row decoder receives a K-bit address and the column decoder gets an L-bit address, then the array contains 2^K rows and 2^L columns. Thus, a total number of 2^N cells is present within the array, with $N = K + L$. The number of selected outputs m depends on the memory organization: if the SRAM is fully decoded, $m = 1$. In this case, one single bit is read from, or written into, the array. Alternatively, m can be either 8 (byte decoding), 16 (half-word decoding), or 32 (word decoding).

Cell Reading and Writing

A reading cycle proceeds as follows: (1) the bit lines, precharged at V_{DD}, are turned to a high-impedance state, (2) the word line is raised, thus connecting the cells within the addressed row to the bit lines, (3) the pull-down transistors of the addressed cells discharge one of the two bit lines, while leaving the other at V_{DD}, (4) the selected bit lines are connected to the sense amplifiers via the column decoder, (5) the sense amplifiers read the content of the addressed cells and make the bit information available to the output circuitry, (6) the row and column decoders are reset, and (7) the bit lines are reset to V_{DD}.

Likewise, a writing cycle proceeds as follows: (1) the row and column decoders are activated, thus connecting the addressed cell to the respective bit lines and the latter to the output circuitry, (2) either one of the two bit lines connected to the cell via the column decoder is discharged to zero, (3) the bit line connected to ground either confirms the status of the cell or upsets the cell in order to make it biased according to the desired input, (4) the row and column decoders are reset, and (5) the bit lines are precharged to V_{DD}.

The timing of the above operation sequence is controlled by internally generated voltages.

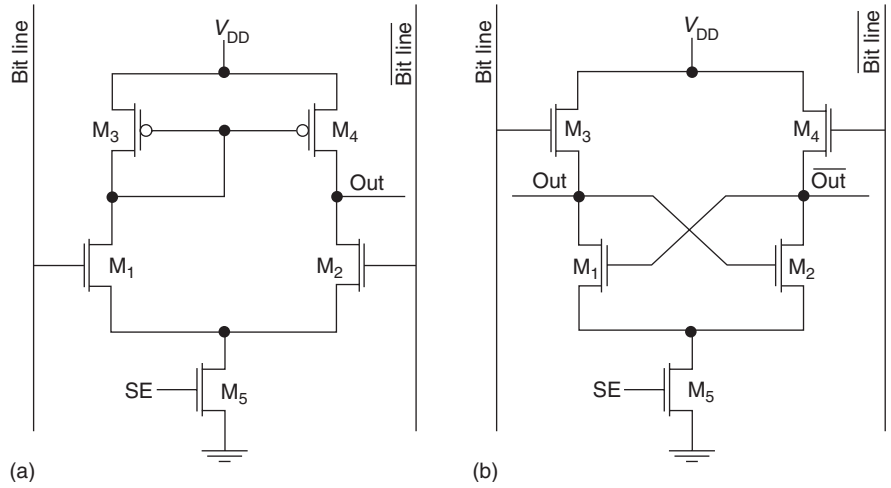

Figure 5 SRAM sense amplifier. (a) voltage amplifier, (b) level shifter.

Sense Amplifier

For static memories, the sense amplifier is often the standard one-stage CMOS differential amplifier with single-ended output, as shown in **Figure 5a**.

The function of the sense amplifier is, in this case, that of allowing for a quick reading of the differential signal on the bit lines, so that the discharging process can be stopped as soon as possible, to the advantage of speed and power consumption. One major drawback of this simple scheme is that, the bit lines being pre-charged to V_{DD}, the differential-pair M_1, M_2 operate in the triode region, with a severe degradation of the voltage gain.

A better solution is a two-stage scheme, with interposition of a differential-level shifter, shown in **Figure 5b** between the bit lines and the voltage amplifier. This circuit features a positive feedback intended to increase the differential voltage gain of the level shifter above 1. Most important, the appropriate DC bias of the differential pair optimizes the amplifier gain, resulting in an excellent performance of the sense amplifier.

Dynamic Random Access Memory

Information can be dynamically stored as a charge on a capacitor, as indicated in **Figure 6**. If the capacitor is charged, its voltage $V_c = V_{DD}$; otherwise, $V_c = 0$. Here again the MOSFET acts as a pass transistor, which can either isolate the cell or connect it to the bit line. In order to charge the cell capacitor to V_{DD}, the word line must be driven to a higher supply voltage V_{PP}. A disadvantage of this storage scheme is that, due to the transistor leakage current, the charge is not indefinitely conserved within the capacitor but, rather, leaks out in a time approximately a few milliseconds. This requires the information to be refreshed some thousand times per second; hence, the definition of dynamic memory.

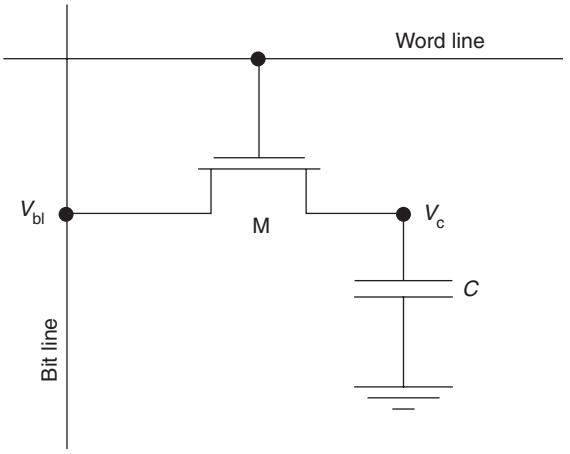

Figure 6 The one-transistor DRAM cell.

The cell capacitance must be as large as possible to ensure a long retention time as well as a safer reading of the stored information. At the same time, it must occupy a small silicon area, in order to allow for a large density and a small cost per bit. These conflicting requirements are pursued with either a trench or a stacked cell. In the former case, schematically shown in **Figure 7a**, the capacitor is placed in a deep trench engraved within the silicon substrate by reactive-ion etching. The trench surface is thermally oxidized and filled with polycrystalline silicon, which acts as the upper electrode of the capacitor and is electrically connected to the MOSFET source. The plate electrode is formed instead by the silicon substrate. In a recent implementation, even the pass transistor of the cell is fabricated along the trench sidewall and is therefore vertically oriented.

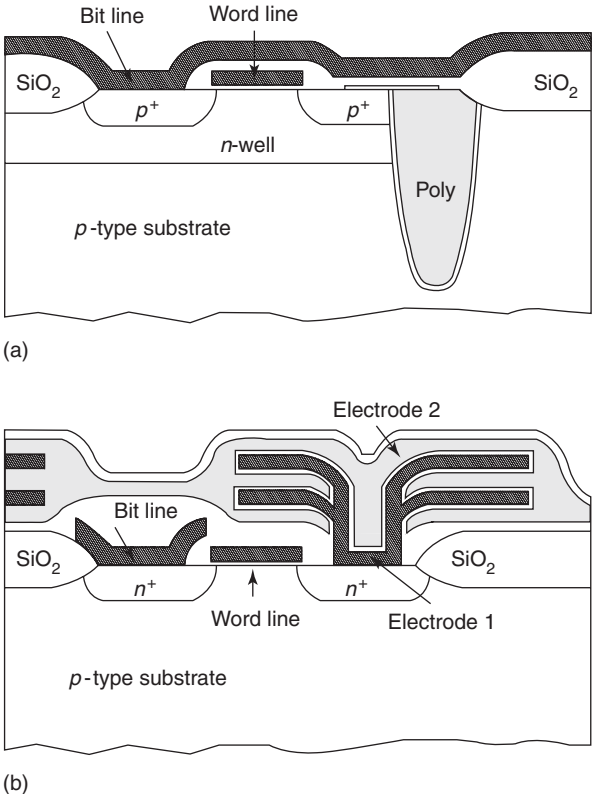

Figure 7 Schematic view of DRAM-cells. (a) DRAM-trench cell and (b) DRAM-stacked cell.

The stacked cell, shown in **Figure 7b**, aims to deploy a large capacitance by extending the surface of the upper electrode in a complex elevated structure. This is achieved by a double-fin structure, generally made by polycrystalline silicon, which is electrically insulated and surrounded by the second electrode, representing the capacitor plate.

An electron micrograph, showing a realistic image of a trench cell, is represented in **Figure 8a**. The aspect ratio of the trench cell is remarkably elongated, stressing the need for a large capacitance, typically larger than 35 fF, within a small silicon area. The polycrystalline silicon finger is electrically connected to the source of the *p*-channel transistor, whose drain is shorted to the bit line via a tungsten plug.

Figure 8b shows instead an electron micrograph of a realistic stacked cell. The double-fin structure represents the upper electrode of the cell capacitor and appears to be light as the silicide layer overlapping the transistor gate. The latter is connected to the word line by vertical W plugs shown in the upper part of the micrograph. The cell connection to the bit line is not shown in this cross section.

DRAM Cell Array Organization

The DRAM cell array may be organized as indicated in **Figure 9**. The bit line is split into two identical

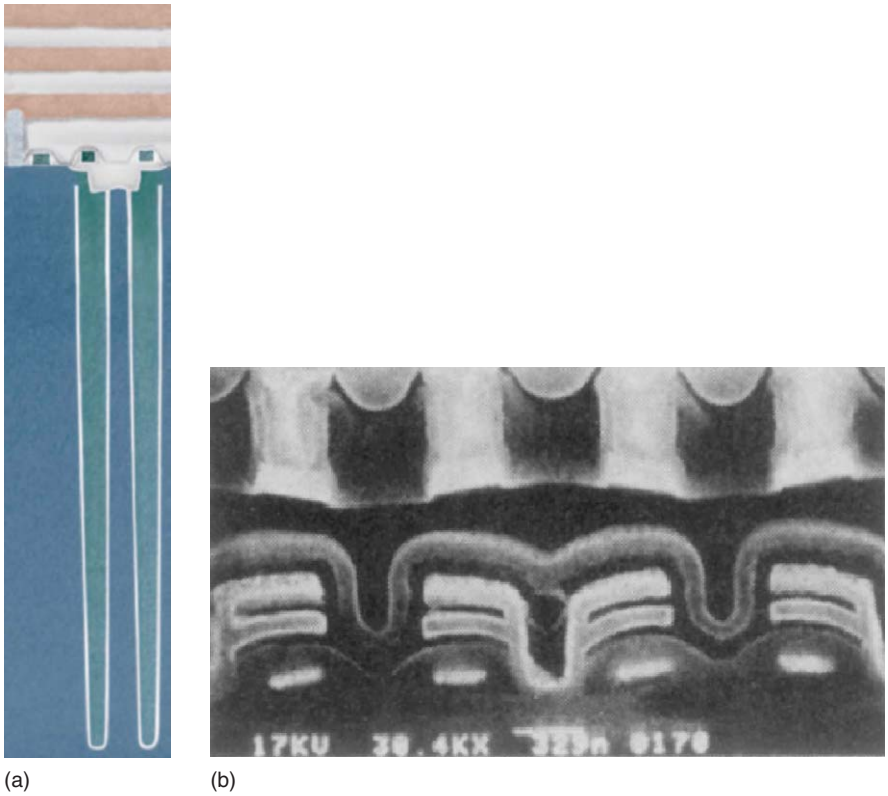

Figure 8 Electron micrographs of DRAM cells. (a) Trench cell (courtesy of IBM Corporation), and (b) stacked cell. (Courtesy of Fujitsu Ltd.)

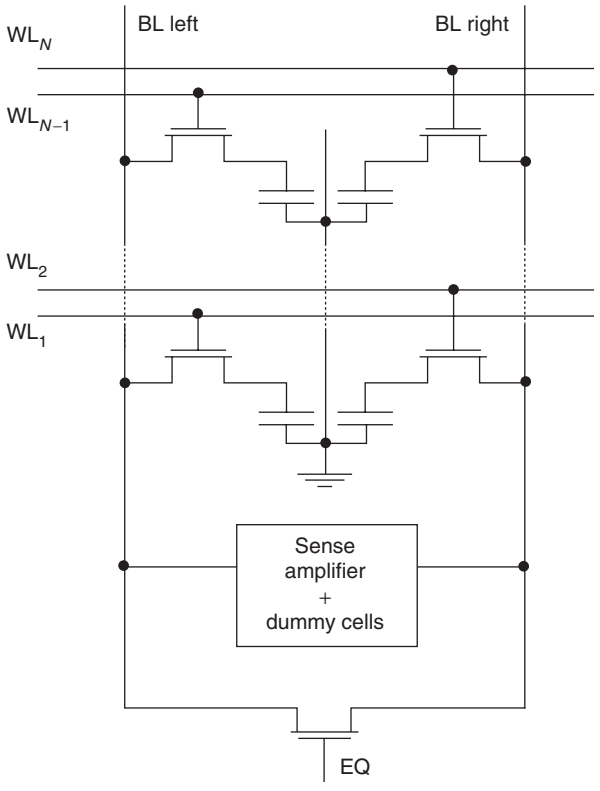

Figure 9 Folded bit-line organization of the DRAM cell array.

segments connected to the sense amplifier, which allows for a differential reading of the stored information and halves the bit-line capacitance for a given column size. In the folded bit-line architecture, shown in the same figure, both segments of the bit line are laid out on the same side of the sense amplifier. The bit-line folding makes it easier to match the layout of the sense amplifier onto the wider pitch of the two half-columns, and allows for an easy placement of the column decoders.

The bit lines are initially precharged at $V_{DD}/2$ by turning on transistor EQ, and are then kept in a high impedance state. As for SRAMs, the row decoder addresses a row of cells and turns the corresponding pass transistors on, thus connecting the cell capacitors to their respective bit lines. Due to the large capacitance ratio between the parasitic bit line and the cell capacitances C_{bl} and C_c, respectively, the bit-line voltage is only slightly affected. More specifically, the voltage change ΔV_{bl} of the bit line connected to the cell equals

$$\Delta V_{bl} = [C_c/(C_{bl} + C_c)](V_c - V_{DD}/2)$$

where V_c is the cell voltage, which may be either 0 or V_{DD}. If the ratio $C_o/C_{bl} \approx 0.1$, the voltage change ΔV_{bl} is only a few hundred millivolts. Due to an unavoidable offset of the sense amplifier, the cell capacitance must be as large as possible, while being dimensionally small to allow for a high bit density.

It is clear from the previous expression that the bit-line voltage increases with respect to $V_{DD}/2$ if the cell voltage $V_c = V_{DD}$, and decreases if $V_c = 0$. The second input of the sense amplifier is connected to a reference voltage generated by a dummy cell precharged at $V_{DD}/2$, as the bit lines. Therefore, the properties of the sense amplifier must be the following: (1) It must detect and amplify a small voltage difference between the two bit lines, eventually generating logic levels and (2) upon reading, it must refresh the content of the cell by reestablishing the original value of the charge stored within the capacitor.

DRAM Sense Amplifier

The schematic of the DRAM sense amplifier is reported in **Figure 10**. Here M_1–M_4 form a regenerative, bi-stable circuit connected to ground via M_5 and to V_{DD} via M_6. An additional pass transistor M_7 may either connect or isolate the two bit lines, thus enabling their precharge to $V_{DD}/2$. Two dummy cells are located on the opposite sides of the sense amplifier and their capacitor is precharged to $V_{DD}/2$ by a suitable circuit: their function is that of balancing the two bit lines by letting them undergo the same transient conditions before readout.

At the start of a reading, SE is low, keeping M_5 and M_6 off; the bit lines are precharged to $V_{DD}/2$ and left in a high-impedance state. Thus, the cross-coupled transistors M_1–M_4 are self-biased on their respective thresholds, and the voltages V_3 and V_4 set at $V_{DD}/2 - V_{Tn}$ and $V_{DD}/2 + |V_{Tp}|$, respectively, V_{Tn} and V_{Tp} being the threshold voltages of the *n*- and *p*-channel MOSFETs, respectively.

The reading sequence is the following: (1) the word line of the addressed cell is activated, thus connecting its capacitor to the corresponding bit line, and so is the word line connected to the dummy cell on the opposite side of the addressed cell, (2) SE is gradually raised high, allowing the sense amplifier to unbalance and reach its final state, (3) the sense amplifiers are readout via a column multiplexer, (4) the word lines are reset to their low value, (5) the signal SE is switched low, thus turning off M_5 and M_6, (6) the bit lines are shorted by raising the signal EQ, which drives them and the dummy-cell capacitors to $V_{DD}/2$, and (7) M_7 is turned off by lowering EQ and leaving the bit lines in a high-impedance state. So doing the initial reading conditions of the sense amplifier are reset.

It may be worth pointing out that, following the gradual raise of SE in step 2, M_5 and M_6 turn on lowering V_3 and raising V_4. The cross-coupled

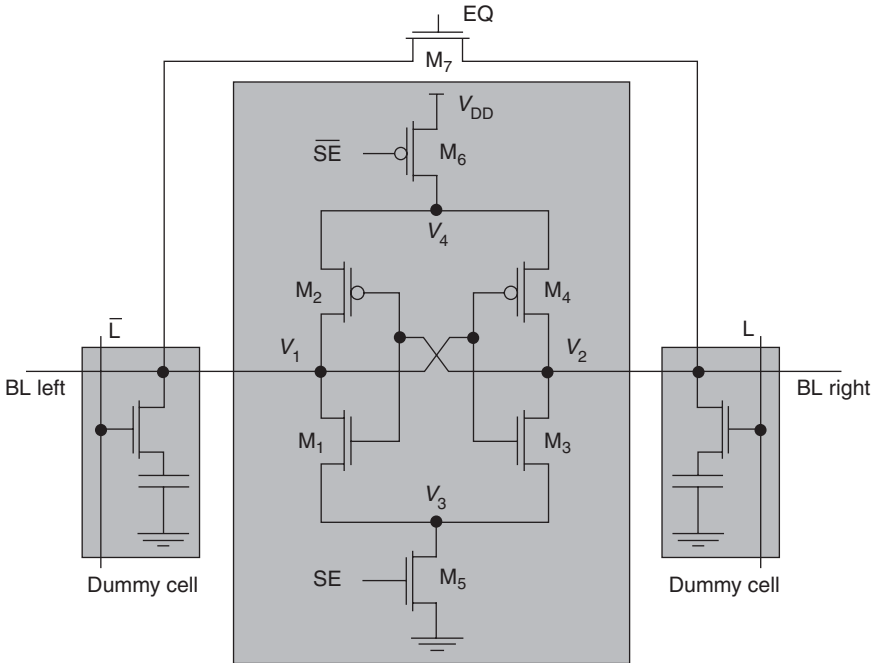

Figure 10 Schematic depiction of the DRAM sense amplifier.

transistors M_1–M_4 turn on and unbalance the sense amplifier according to the initial voltage mismatch. If the initial bit-line voltage $V_{bl}^{(left)} > V_{DD}/2$, which reflects the content of a logic "1" in the addressed cell, the transient drives it to V_{DD} while the right bit-line voltage $V_{bl}^{(right)}$ goes to zero. The opposite situation occurs if $V_{bl}^{(left)} < V_{DD}/2$, which reflects the content of a logical "0" within the same cell. Thus, reading the addressed cell restores the original voltage within the cell capacitor, which makes the reading nondestructive. Also, if the two bit-line capacitances are equal, precharge to $V_{DD}/2$ is simply achieved with no additional power consumption by connecting them via M_7.

DRAM Architecture

The block diagram of a DRAM bank is shown in **Figure 11**. The row addresses are delivered to the row decoders, which activate one word line. All the cells associated with that word line are thus connected to the bit lines, and their content is readout by the sense amplifiers. The column addresses set the column decoder (actually a simple multiplexer) and select the requested bytes or words, which are delivered to the output registers. Multiple column readings can be performed in sequence for any given row address. This improves the data rate of the DRAM. The above structure may be mirrored and replicated as many times as needed. The size of a DRAM bank is defined according to conflicting constraints: increasing the size of the cell array makes the layout more compact for a given DRAM capacity, but deteriorates the access time and, ultimately, the reading integrity, due to the larger word- and bit-line capacitances.

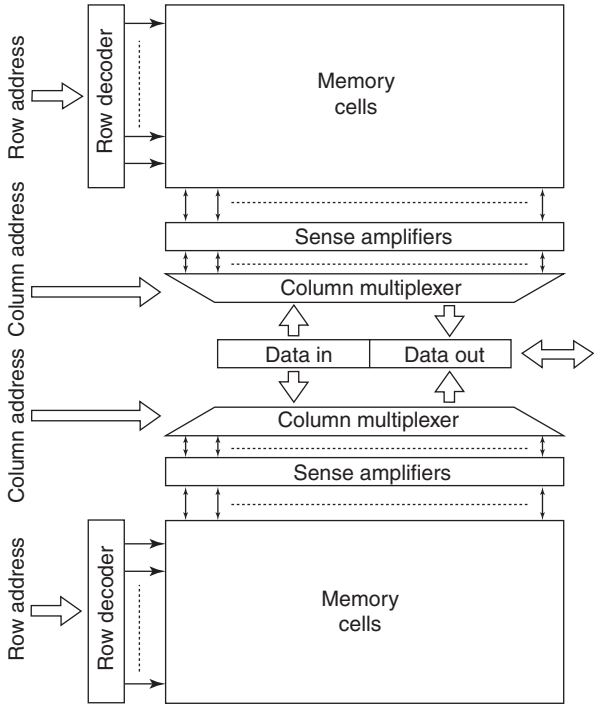

Figure 11 DRAM bank architecture.

In order to save package pins, a multiplexed addressing scheme, where the row and column addresses are presented in sequence to the same bus, is typically used. In this addressing scheme, two main control

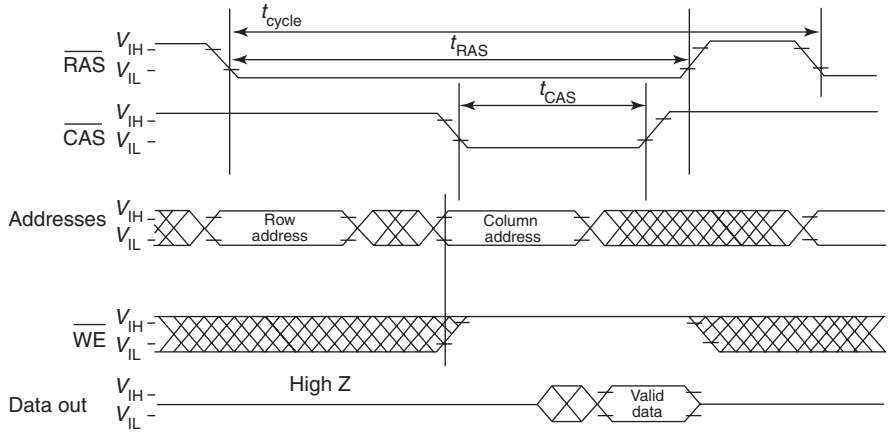

Figure 12 Timing diagram of the DRAM input signals.

signals must be provided to the DRAM, namely, RAS and CAS. Another control signal, write enable WE, indicates if the intended operation is a read or a write.

The timing diagram of a read operation is shown in **Figure 12**, where the control signals RAS′, CAS′, and WE′ are shown in complementary form. The RAS′ signal is switched low only after valid row address bits are present at the DRAM input buffer. In turn, the CAS′ signal goes low only when valid column address bits and a valid WE signal are set to their respective values. The time t_{CAS} is the minimum time that CAS′ must be kept low to generate valid output data; the time t_{RAS} is the minimum time that RAS′ must be kept low to generate valid output data; finally, the cycle time t_{cycle} is the minimum time needed between two successive RAS′ commands.

Synchronous DRAM

Early DRAM architectures used to operate asynchronously under the control of the external commands outlined in the previous section. To date, most DRAM chips are synchronous devices driven by the system clock, and are thus referred to as SDRAMs. Synchronous operation has several advantages: most notably, the possibility of a pipelined DRAM architecture with concurrent row and column addressing, and high-speed I/O operations. Taking advantage from the inherent parallelism of the memory core, this architecture improves the DRAM data rate, which is the most important performance parameter, while leaving its latency scarcely affected. Due to the hierarchical memory organization in modern computers, entire data blocks are, in fact, retrieved from the central memory when a block miss occurs at a higher level of the hierarchy. To date, SDRAM data rate may be as large as $1.6\,\mathrm{GB\,s^{-1}}$. This comes at the penalty of extra latches and buffers, as well as high-speed circuitry to support the I/O interface.

SDRAM commands, addresses, and data are latched to the rising edge of the system clock. Basic SDRAM commands are chip select (CS), RAS, CAS, WE, data mask (DM), and data strobe (DQS). The CS signal is used to let the chip know that the commands coming in over the bus are intended for it. RAS, CAS, and WE retain the usual meanings of row and column address strobe and write enable, respectively. DM masks the input data during a write; DQS is instead a bi-directional edge aligned data strobe which toggles at the same time as the output data. This is made possible by a delay locked loop (DLL), which shifts the output data in order to align DQ and DQS. The bus width is most often 64 bit.

The basic SDRAM operations are: activate (ACT), read (RD), or write (WR) followed by a precharge. SDRAM operation can be configured for CAS latency and burst length by setting the 12 bits of the load mode register (LMR). The burst length determines the amount of data transferred in consecutive cycles between the memory controller and the memory after applying one start address.

The typical delay between the RAS and CAS signals is two clock cycles, and the CAS latency is again two clock cycles; thus, the access time is four clock cycles. The SDRAM cycle time t_{cycle} depends, in this case, on the bus width and the burst length.

Rambus DRAM

The direct RDRAM architecture is based on a system approach, which maximizes the data rate on the memory board by carefully synchronizing data, addresses, and commands with the clock signal. The key elements of this technology are the following: (1) a 16-bit wide, 800 MHz channel, (2) an on-board interface that allows the memory controller talk to the RDRAM and, (3) an in-line memory module called RIMM.

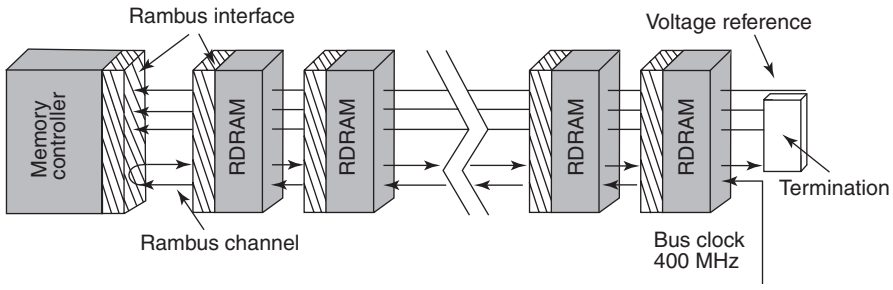

Figure 13 Direct Rambus DRAM architecture.

A Rambus channel includes a controller and a variable number of RDRAM chips connected together via a linear common bus. As shown in **Figure 13**, the controller is located at one end, and the RDRAMs are distributed along the bus, which is terminated at the opposite end with the characteristic impedance of the lines connected to the high-voltage level. Therefore, the bus driver operates in the open-drain configuration. The 400 MHz clock signal is generated at this same end; it propagates along a bus line connecting the clock generator to the controller, and turns back to the terminating end of the bus line. All commands, addresses, and data moving from the controller to the RDRAM chips and back are synchronized on the edges of the clock which propagates in the same direction. In doing so, the clock skew is minimized.

The channel uses 18 data pins, two of which for error correction code (ECC), cycling at $800\,\mathrm{MB\,s^{-1}}$ per pin to provide a bandwidth of $1600\,\mathrm{MB\,s^{-1}}$. This is achieved by latching data on both the rising and the falling edges of the clock signal.

The RDRAM has a pipelined microarchitecture, which fully supports concurrent RAS and CAS operation, as well as read/write buffering. Also, it provides 16 bytes every 10 ns on the internal bus. The Rambus interface transforms the 10 ns internal bus into an external, two-byte wide, 1.25 ns external bus.

All signal wires have an equal loading and fan-out and are routed parallel to each other on the top trace of a PCB with a ground plane located underneath. The addition of more RDRAM chips linearly increases the load and the delay, but leaves the phase relationships of the clock, data, addresses, and commands unaltered. Also, the memory granularity is simply given by the memory capacity of one chip.

Content-Addressable Memory

As already stated in the introduction, modern computers rely on a hierarchy of memories comprising three levels of cache in addition to the central memory. This complex organization is due to the design consideration that smaller memories are faster, and to the locality principle, according to which programs tend to reuse either data which have already been used in previous cycles (temporal locality), or data sets stored next to the used data (spatial locality).

Cache memories are classified according to the criteria by which the central memory is mapped onto it. In a fully associative cache, data blocks can be stored anywhere within the cache, as opposed to direct-mapped memories, where the block address uniquely defines its location. In most cases, cache memories are *n*-way set associative; this means that the cache memory is split into several sets, each containing *n* blocks: the block address uniquely defines the set where it must be stored, but the location of the block inside the set (way) is free.

In any case, a tag memory containing a list of the block addresses stored within the cache is needed. When a new instruction or a new data is requested by the processor, an associative search must be performed on the tag memory in order to find out if the block containing that instruction or data is stored within the cache and, in case it is, in what location. As opposed to standard memories, the input of the tag memory is a data and its output is the local block address within the cache. The tag memory is, therefore, a CAM. The basic cell of a CAM, shown in **Figure 14**, contains ten transistors.

The information bit is stored within a static latch made by the two cross-coupled inverters I_1 and I_2; two pass transistors M_1 and M_2 driven by the write line (WL) connect the latch outputs to the input data D and its complement D′. Two couples of pass transistors M_3–M_4, M_5–M_6, driven by the input data D and D′ and by the latch outputs as shown in the figure, connect to ground a precharged match line (ML). If the stored bit is opposite to the data D, either of the two couples of pass transistors is on and discharges the match line. Otherwise, the match line remains biased at V_{DD}. The former case indicates that no match occurs between the input and the stored data; the latter indicates the opposite condition.

The above cells are organized within an array, and all the cells of a column, sharing the same write line

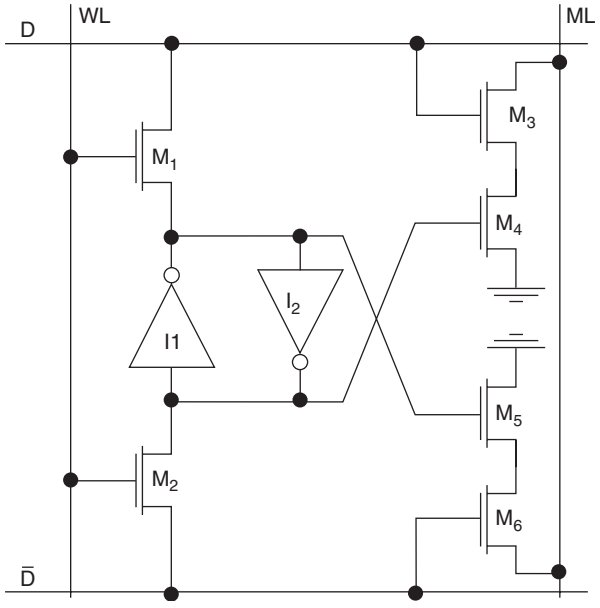

Figure 14 The ten-transistor CAM cell.

and the same match line, contain the address of one memory block. If any one of the above cells does not match its input bit *D*, it drives the match line to ground, indicating that no match occurs between the input word and the stored address. If none of the column matches the input word, this means that a block miss is occurring; all the match lines are at ground potential and an NOR gate raises the block-miss signal high.

If, on the other hand, all the input bits of a column match their respective bits stored within the cells, the match line remains high. This is an indication that the requested block is stored within the cache, and the high match line indicates the internal address of the block within the cache memory.

See also: Conductivity, Electrical; Electrons and Holes; Integrated Circuits; Memory Devices, Nonvolatile; Semiconductor Devices; Transistors.

PACS: 85.25.Hv; 84.30.−r; 85.30.Tv; 84.30.Bv; 84.30.Sk

Further Reading

Haraszti TP (2001) *CMOS Memory Circuits*. Dordrecht: Kluwer.
Prince B (1999) *High Performance Memories: New Architecture DRAMs and SRAMs – Evolution and Function*, Rev. edition. Chichester: Wiley.
Sharma AK (2002) *Advanced Semiconductor Memories: Architectures, Designs, and Applications*. Hoboken: Wiley-IEEE Press.

Meso- and Nanostructures

A Horsfield, University College London, London, UK

Introduction

Meso- and nanostructures are "small" structures. It is possible to assign a rough length scale to them (micrometers and below), but it is more instructive to define smallness in terms of the physical phenomena that it makes possible, as these are what make small structures important. In the following section, selected phenomena are briefly described, together with the principal characteristics that structures need in order to exhibit them. In the section "Contemporary structures", some contemporary structures in which these phenomena can be observed are listed. These structures are organized into categories according to the number of their dimensions which are not small. New structures are being invented constantly, so this list should be seen as illustrating possibilities rather than being comprehensive. However, some structures – such as carbon nanotubes – will remain important for the foreseeable future.

Phenomena

The following phenomena are all a consequence of smallness in one or more dimensions, though they vary in the way smallness is exploited. Many make use of the wave nature of the electron, while others (notably Coulomb blockade) have a classical origin.

Continuum Electronic States Become Discrete

In an infinite solid, the electronic states form continuous bands of states (Bloch states). When the system is large but finite, the states are discrete, but have very small energy separations, so that a continuum description remains accurate. As the size of the system is reduced, the energy spacing between levels increases roughly as $1/L^2$, where L is the size of the system (quantum confinement). When the spacing is large enough (the system size is small enough) that it exceeds any broadening (e.g., thermal), then this discreteness can be exploited. In order to obtain discrete states, the structures need to be small in three dimensions with weak coupling of the discrete states to the environment (e.g., quantum dots). Thus, nonlinear