Memories in Computers—Part 1

by

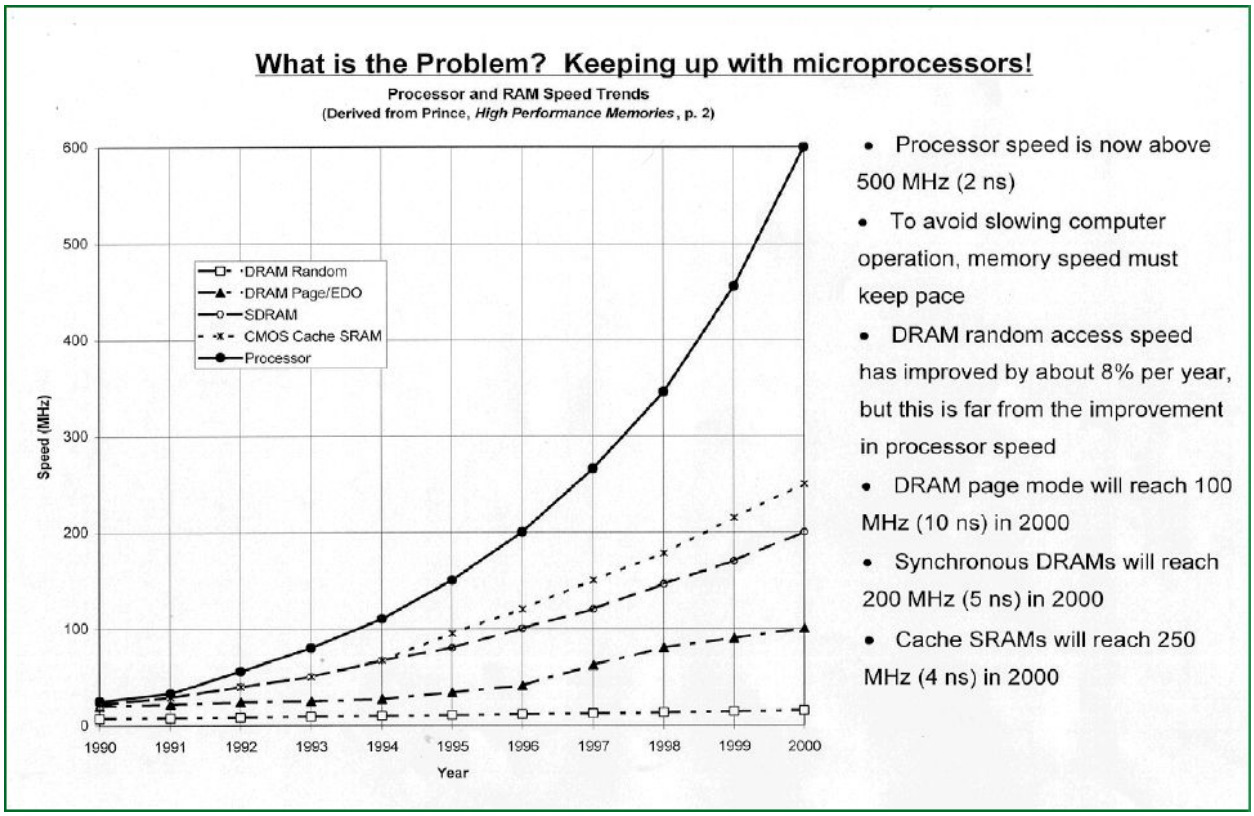
Dr. William R. Huber, P.E.

A. Memories in Computers

Memory devices are used to store information that is used by a central processing unit, or CPU, often in a personal computer. Almost 30 years ago, on August 12, 1981, IBM introduced the "Personal Computer", or "PC". The IBM machine used an Intel 8088 microprocessor with 29,000 transistors and had 16 KB[1] of Dynamic Random Access Memory (DRAM). The 8088 operated at 4.7 MHz (a cycle time of 213 ns). Today the "state-of-the-art" PC uses an Intel Core i5 microprocessor with 559 million transistors and has at least 4 GB of DRAM. The Core i5 operates at clock speeds up to 3.6 GHz (a cycle time of 0.28 ns).

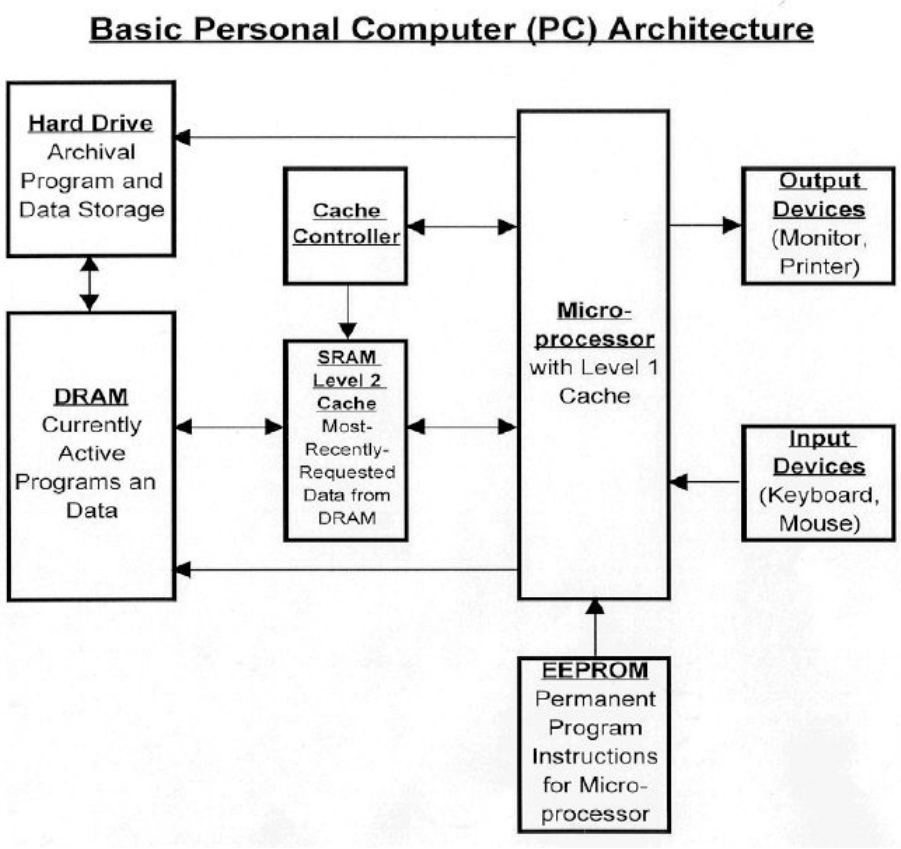
Because most computer cycles involve obtaining information from (or writing information to) memory, speed of data access is a critical factor in determining the speed of a computer as observed by the user. Therefore, as CPU speed has increased over 760-fold, memory designers and manufacturers have struggled to keep up. The graph on the next page illustrates how the gap between CPU speed and memory speed grew from 1990 to 2000.

[1] As the reader will see, the field of computers and memory devices is full of acronyms and abbreviations. KB stands for kilobytes, where a byte is 8 bits of information. MB stands for megabytes and GB for gigabytes. MHz stands for megahertz, or million per second; GHz stands for Gigahertz, or billion per second. ns stands for nanoseconds, or billionths of a second, a miniscule amount of time. Light travels about one foot in a nanosecond.



The two factors that have driven semiconductor memory development from its beginning in the early 1970s to the present are: (1) density (and therefore cost), i.e., increase the amount of information stored per square inch of silicon, thus reducing the cost per unit of information; and (2) access time, i.e., reduce the time required to obtain the information (reduce the access time) from the chip. As we shall see, improvements in density received far more attention until at least 1990 than did access time reductions.

To meet the dual and conflicting needs for massive amounts of data storage and increasingly rapid access times to that data, PCs use four major types of memory. The basic architecture of a PC, including the major memory types, is shown in the next diagram.



The slowest and cheapest memory, used for massive storage requirements, is magnetic storage in a “hard drive”. This type of storage has the added advantage that it is “non-volatile”; i.e., the information is retained even when power is turned off. Obviously, such non-volatility is essential for programs and user files. Recent developments in non-volatile semiconductor memory, in particular the “Flash” memory, could augment or supplant magnetic hard drives in the future if their cost comes down far enough.

The second level of memory, intermediate in access time and cost, is dynamic random access memory, or DRAM. DRAMs are semiconductor memories fabricated from silicon, and store their information on miniscule capacitors. A reasonable analogy for the capacitor is a balloon. An inflated balloon represents one state of information – say a “1”; and a deflated balloon represents the other state – say a “0”. A capacitor in a DRAM memory cell can have charge stored in it (a “1”) or no charge (a “0”). Just as a balloon loses air over time, the capacitor in a DRAM memory cell loses charge over time. If that charge is not replenished periodically, the information it represents is lost. The action of replenishing the charge is called “refreshing” the memory. Unlike

magnetic hard drives, DRAM storage is volatile; if the power is turned off, all of the information in the DRAM is lost.

The third level of memory is electrically erasable programmable read-only memory (EEPROM). The name itself is confusing, as how can a memory that is read-only be programmable? Rather than answer that unanswerable question, just call it Flash memory. This memory is for the very special purpose of storing the relatively small amount of information required to boot-up the computer. As such, this memory is written to very rarely; so erasing and writing speed are of little consequence. The essential feature is non-volatility—the boot-up information must be retained on a permanent basis—and Flash memory has that benefit.

The fourth level of memory in a PC is static random access memory, or SRAM. The access time of an SRAM is much faster than that of a DRAM, but its cost per unit of storage (each unit of storage is called a "bit") is much higher. As in a DRAM, SRAM storage is volatile; if the power is turned off, all of the information in the SRAM is lost.

The following chart summarizes important properties of the various memory types.

Overview of Memories for PC Applications

	Typical Storage Capacity/Unit	Random Read Access Time	Cost (Cents per MB)	Volatile (Loses data when power is removed)	Data Storage Mechanism
Hard Drive	1000 GB/Drive (8×10^{12} bits)	7,500,000 ns (7.5 ms)	0.0068	No	Magnetic
DRAM	128 MB/Chip (1×10^{9} bits)	30 ns	1.4	Yes	Charge on Capacitor
Flash	1 GB/SLC Flash Chip (8×10^{9} bits)	50 ns (Write times are much slower)	0.57	No	Charge on Ultra-Low-Leakage Capacitor
Flash in SSD (Solid State Drive for Hard Drive Replacement)	256 GB/Drive (2×10^{12} bits)		0.16		
SRAM	18 MB/Chip (1.4×10^{8} bits)	2.6 ns	115	Yes	Current Flow

B. DRAM History[2]

In early 1970, the newly-formed Intel Corp. announced a new type of memory device, a Dynamic Random Access Memory (DRAM). This memory was called the i1102[3]. Let's use this first DRAM to examine the characteristics and limitations of DRAMs in general, and see how those characteristics and limitations have evolved over a period of 40 years.

1. Storage Capacity

   The i1102 could store just 1,024 bits (~1 x 10^3) of information. Storage capacity has increased by a factor of 4 every 3 years, and now stands at 4 Gb (~4 x 10^9) per chip[4].

2. Organization

   The i1102 was organized as 512 words x 2 bits per word. 512 words is the number of uniquely addressable storage blocks on the chip. Each addressed word outputs a specific number of bits, in this case 2 bits per word. Modern DRAMs have up to 1 x 10^9 words, and word sizes ranging from 4 to 16 bits per word.

3. Cycle time

   It is a characteristic of DRAMs that cycle times for reading and writing are the same. For the i1102, that cycle time was 500 ns. In other words, 2 read or write cycles could be completed in 1 microsecond (1 us). Cycle times for modern DRAMs are in the range of 50 ns, allowing 20 full cycles in 1 microsecond.

4. Access Time

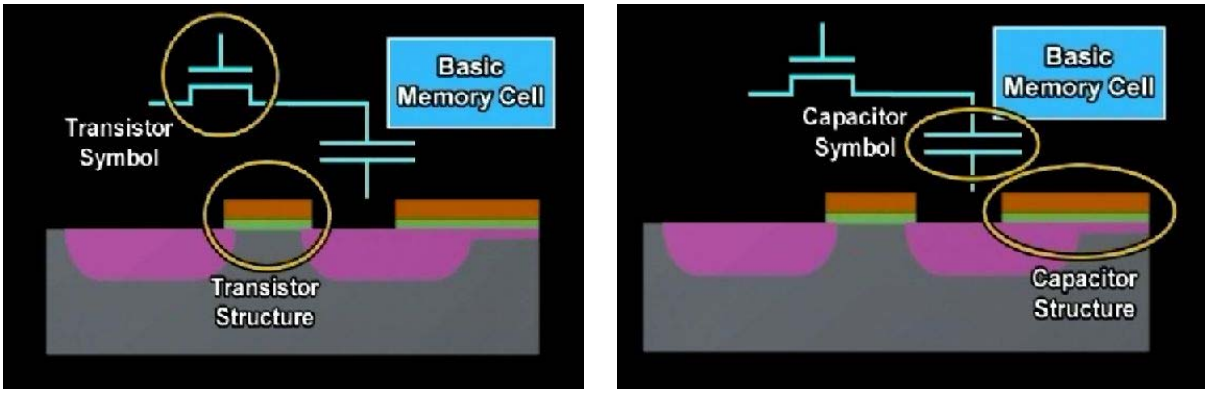
   Access time is the time from addressing the memory to obtaining valid data on the output pins. For the i1102, access time was 345 ns. The concept of access time for modern DRAMs is somewhat more complex, and we will discuss that complexity in detail later. But the comparable access time for modern DRAMs is 30 ns.

---

[2] A mini-glossary of DRAM terms is included as **Appendix A**.
[3] Regitz, W.M. and J. Karp; 1970 IEEE International Solid-State Circuits Conference; Digest of Technical Papers; "A Three-Transistor-Cell, 1024-Bit, 500 NS MOS RAM"; February 18, 1970; pp. 42-43
[4] Samsung K4B4G0446A/0846A Data Sheet, Rev. 1.0, July 2010

5. Memory Cell Structure

The i1102 memory cell was composed of 3 transistors and an inherent capacitor. Soon after the i1102, all DRAMs moved to a much smaller cell structure consisting of 1 transistor and 1 capacitor, abbreviated as the 1T-1C cell and shown below.



1T-1C Memory Cell Showing Structure and Symbol of the Transistor and Capacitor

6. Storage Mechanism

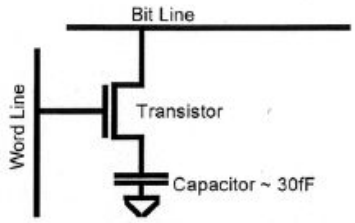
The i1102, and all DRAMs, store information as charge or lack of charge on a capacitor.

7. Storage Duration

The i1102 forgot everything if every bit wasn't reminded ("refreshed") every 54 milliseconds at room temperature. DRAMs have very short retention times, measured in milliseconds. If they are not refreshed, they forget the information that was stored in them. That need for periodic refreshing is still true today, but special circuitry provides options such as Auto-Refresh and Self-Refresh, so the need to manually cycle through each addressable row to accomplish the refreshing is eliminated.

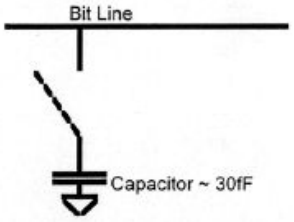
The mechanism of charge loss and the steps to perform a refresh operation are detailed on the next page.

Refreshing—Why Is It Needed?



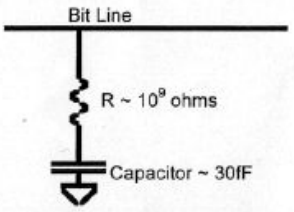
One Transistor-One Capacitor Memory Cell

- The Transistor is turned "on" or "off" by the voltage on the Word Line.
- The total capacitance of the Capacitor is about 30 x 10^{-12} Farads (30 femtofarads or 30 fF).



Ideal Equivalent Circuit--Deselected Word Line

- In the ideal case, when the Word Line is deselected, the Transistor is completely turned off. Then the charge on the Capacitor would remain forever.



Actual Equivalent Circuit--Deselected Word Line

- In the actual case, the Transistor remains very slightly on. Thus the charge on the Capacitor leaks off to the Bit Line.
- This lost charge must be replaced, or "refreshed" periodically. Otherwise the information will be lost.

Refreshing—How Is It Done?

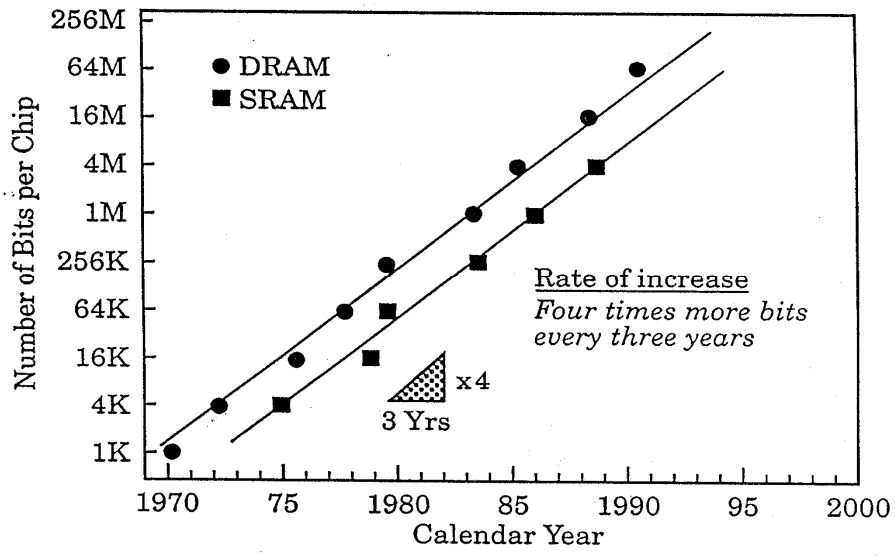
- A refresh cycle is performed at regular intervals to restore information in the memory cells
- Every row of the DRAM must be accessed for refreshing within the refresh interval (usually 4 to 64 milliseconds).
- To refresh a row:
  1. A row is accessed, and the state of every memory cell along the row is sensed by a sense amplifier.
  2. The sense amplifiers restore the information (replace the lost charge) in the memory cells of the row.
  3. Following the restore of the row, the lines of the DRAM are precharged to their standby condition.

8. Addressing

For the i1102, all addresses (both row and column) were delivered simultaneously on separate pins. To save package pins, both the row and column addresses were encoded in binary (e.g., Row 43 in decimal is 000101011 in binary). Thus the 512 word lines in the i1102 could be addressed by just 9 external row address lines (2^9 = 512). Binary coding of addresses is still used on all memories today.

Starting with the 16K DRAMs from Mostek in 1976[5], row and column addresses were delivered on the same set of pins, row addresses first followed by column addresses. This approach is called multiplexed addressing, and greatly reduces the number of device pins, and therefore package size, of DRAMs.

The i1102 DRAM contained 1024 bits and became available in 1970. Since then, the number of bits per chip has quadrupled every 3 years as shown in the following graph, a compound annual increase of 59%!



Bit Count vs. Year for DRAMs and SRAMs[6]

[5] Proebsting, R.J. and R.S. Green, "Dynamic Random Access Memory MISFET Integrated Circuit", U. S. Patent 3,969,706, July 13, 1976

[6] Elmasry, M.I. Editor, **BiCMOS Integrated Circuit Design**, IEEE Press, 1994

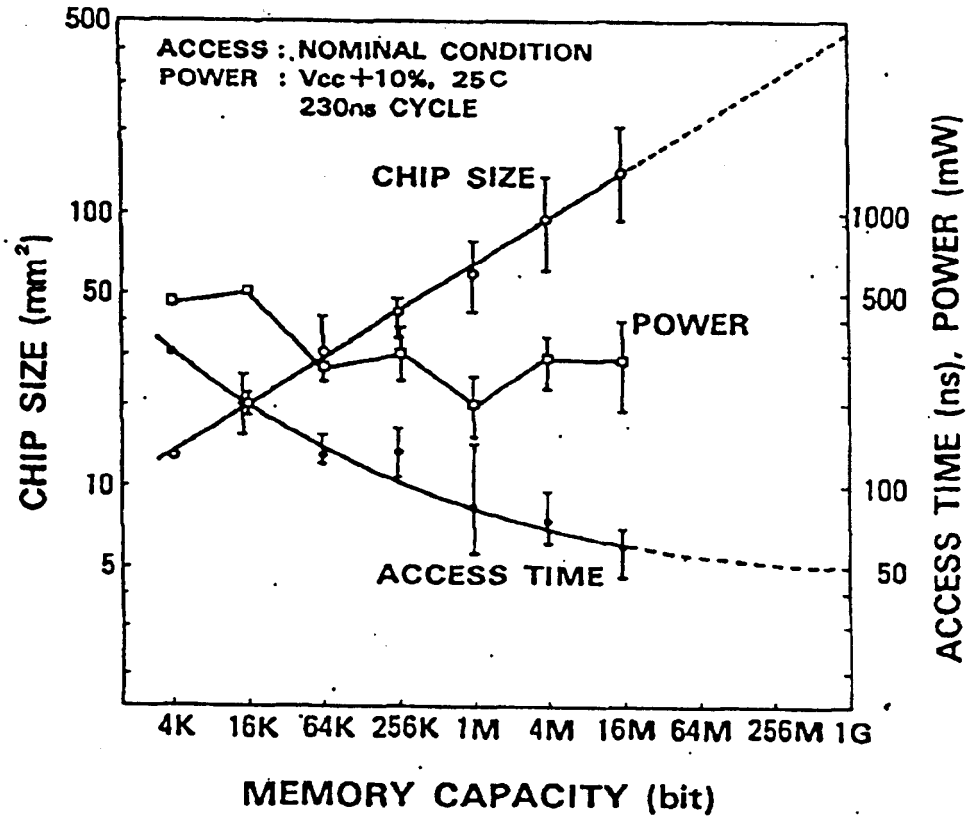
This increase in memory density is an illustration of "Moore's Law," a 1965 observation by Gordon Moore, the co-founder of Intel. Moore observed[7] that the number of transistors per square inch of silicon had doubled every year since the integrated circuit was invented, and predicted that such doubling would continue for the next 10 years. The density increase has slowed a bit, to a doubling every 18 months, but continues to this day. In fact future roadmaps of semiconductor innovation are based on the continued validity of Moore's Law. 1Gb[8] DRAMs are now the industry's mainstream product, and 4Gb parts are available.

Two factors have contributed to the dramatic increase in the number of bits per chip. First, the size of each memory cell has shrunken significantly, about 26% per year. Second, as shown in the next graph, overall DRAM chip size (area) has increased at the rate of about 12% per year. Over the same time period, both access[9] and cycle times have improved by about 8% per year. This graph also illustrates the access time changes accompanying the density increase from 4K to 1G bits per chip.

[7] Moore, G., *Electronics Magazine,* "Cramming more components onto integrated circuits", April 19, 1965

[8] In discussing memory capacity, K stands for 1024 (about one thousand), M stands for 1,048,576 (about one million) and G stands for 1,073,741,824 (about one billion).

[9] Access time here refers to random access time, also referred to as row access time, as opposed to page mode or column access times that will be discussed below.



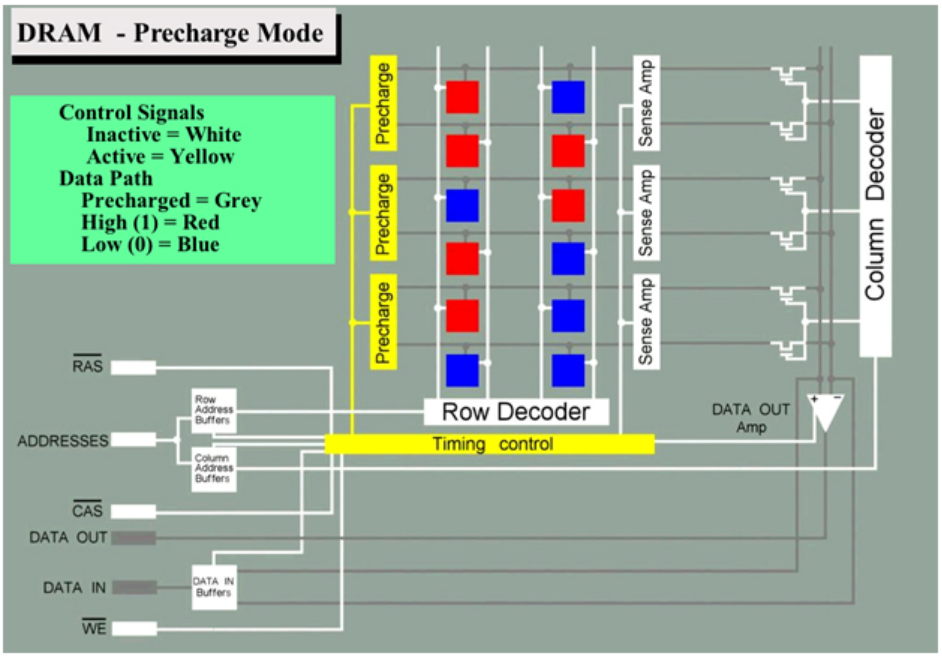
Trends in Standard DRAM Development[10]

C. DRAM Architecture and Basic Operation

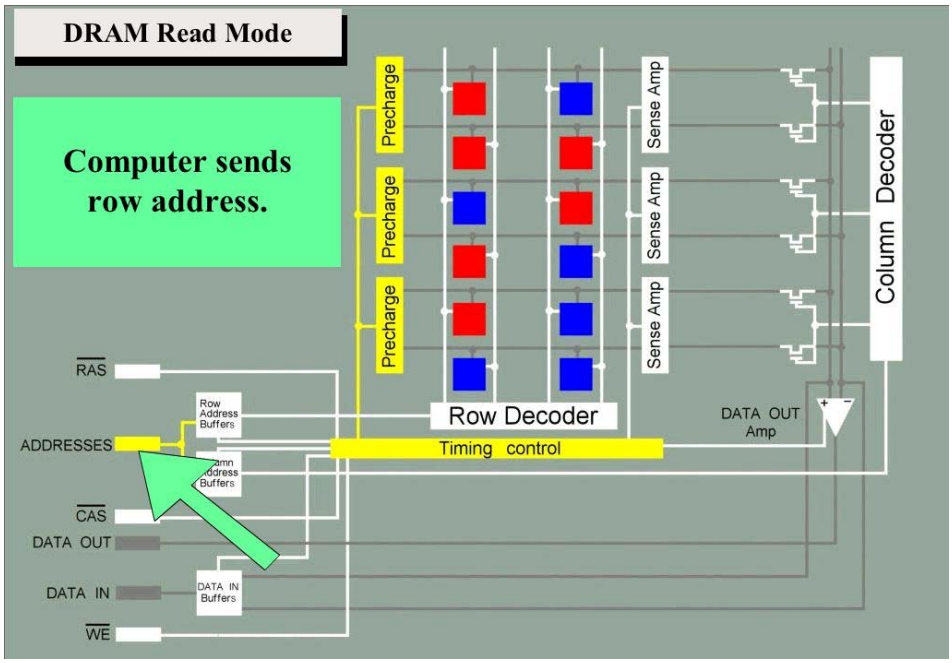
Now let us examine the basic architecture and operation of a DRAM. The figure on the next page illustrates one possible architecture for a simple DRAM.

Memory cells, represented by the red and blue rectangles, are arranged in a matrix pattern. Red indicates those cells storing a high level (1); and blue indicates those cells storing a low level (0). Of course, modern DRAMs contain about a billion times more memory cells (arranged in multiple sub-arrays, each with a plurality of sense amps, decoders, and other circuits) than are shown in the figure.

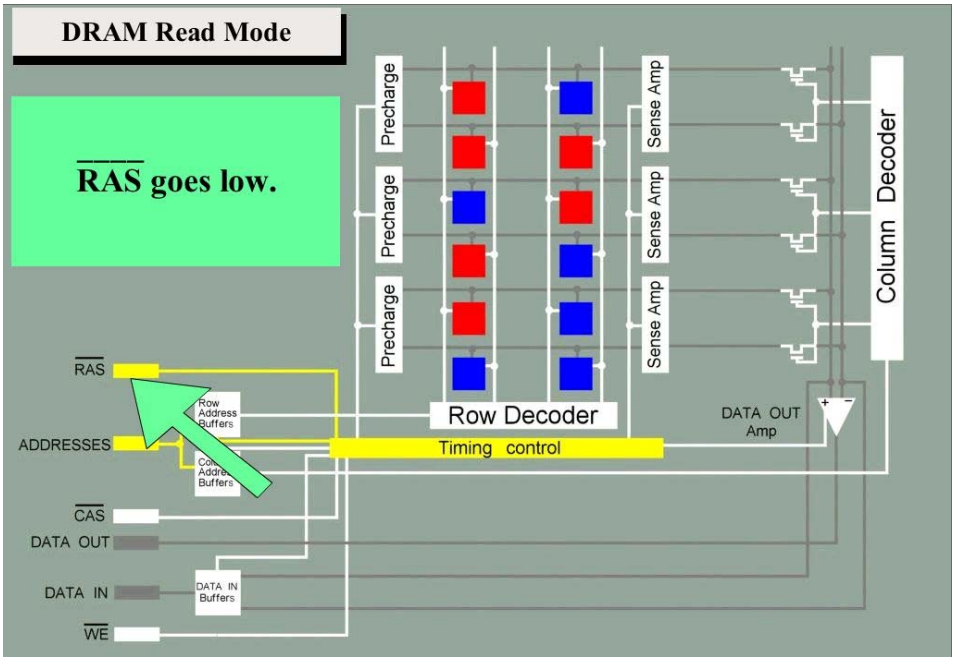
[10] Itoh, K., "Trends in Megabit DRAM Circuit Design", IEEE Journal of Solid-State Circuits, Vol. 25, No. 3, June 1990, p. 778



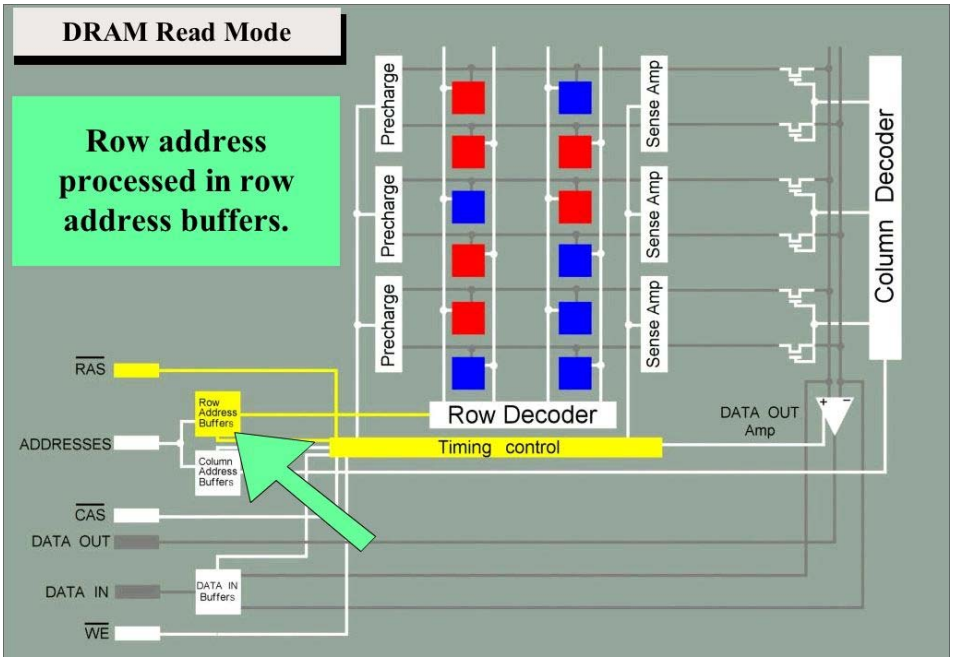
To illustrate memory operation, suppose that the memory is to be instructed to read the data in the top left cell and convey it to the output.



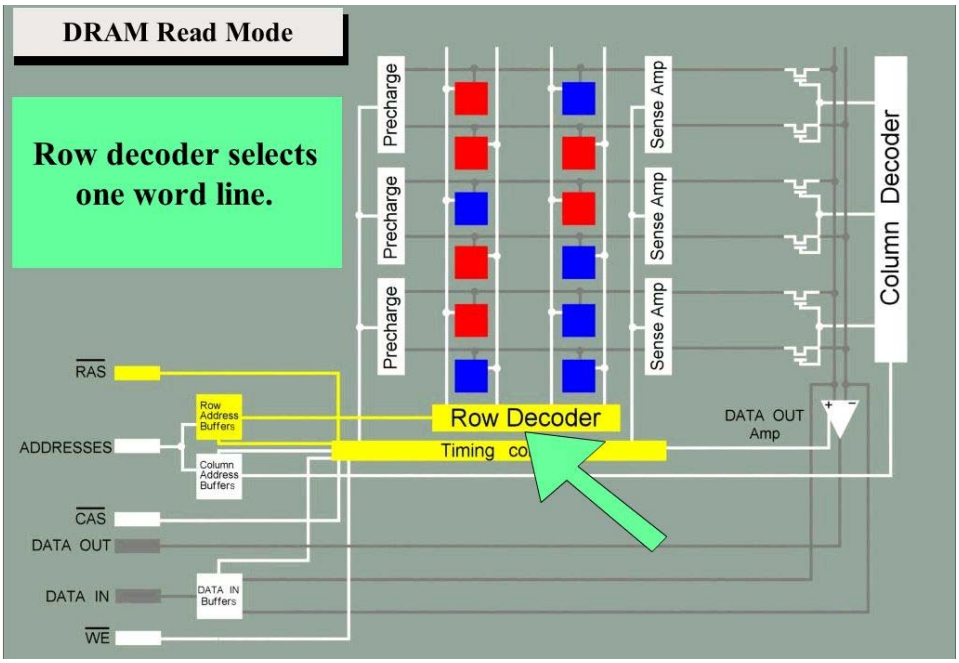
First, voltages representing the address of the row to be selected are applied to the Address Pins.



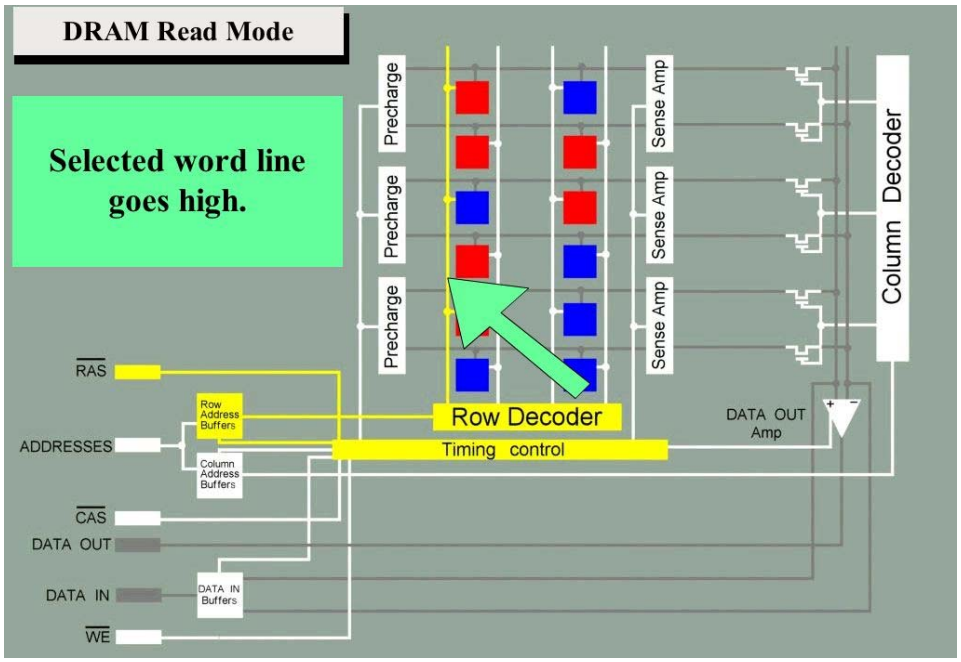
Then a signal called RAS/ (pronounced RAS-Bar) is applied to the memory, thus gating the voltages on the Address Pins into the Row Address Buffers.



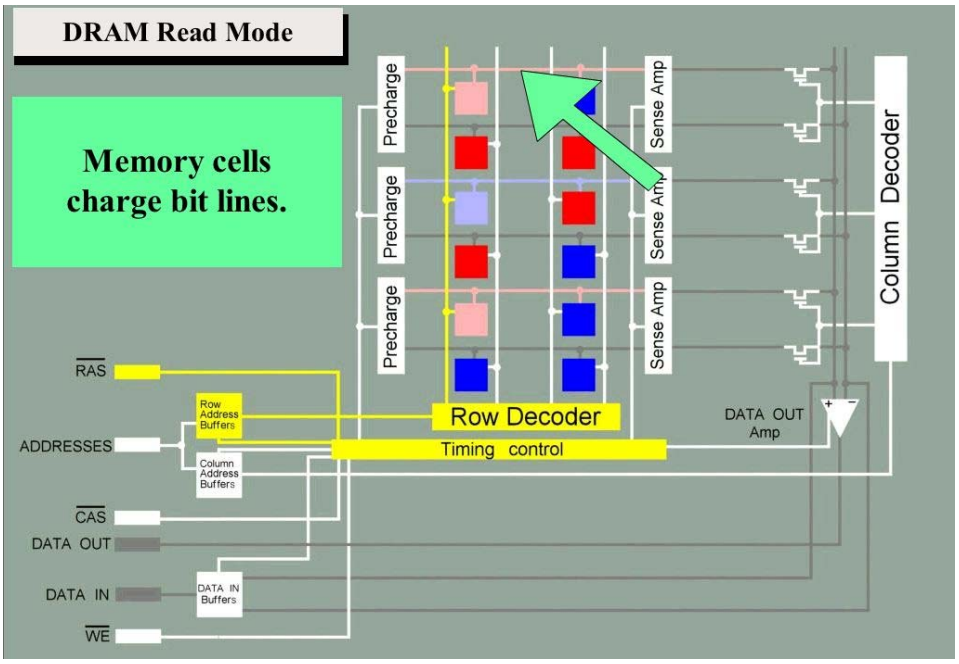
The Row Address Buffers sense the address input voltage levels and generate internal addresses that are sent on internal address lines to the Row Decoder.



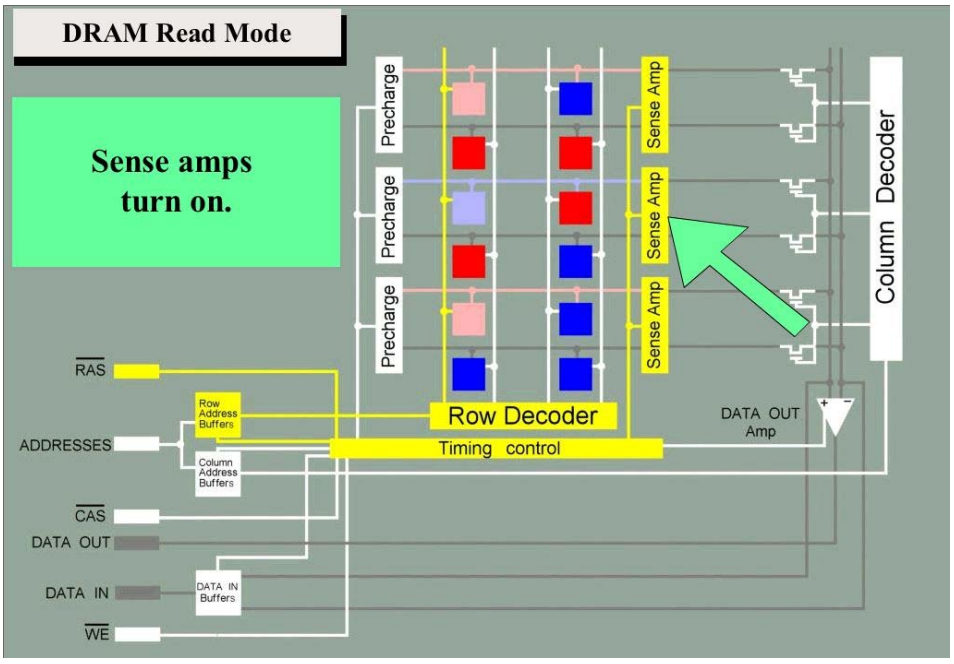
Based on the internal addresses applied to it, the Row Decoder identifies and activates the appropriate word line.



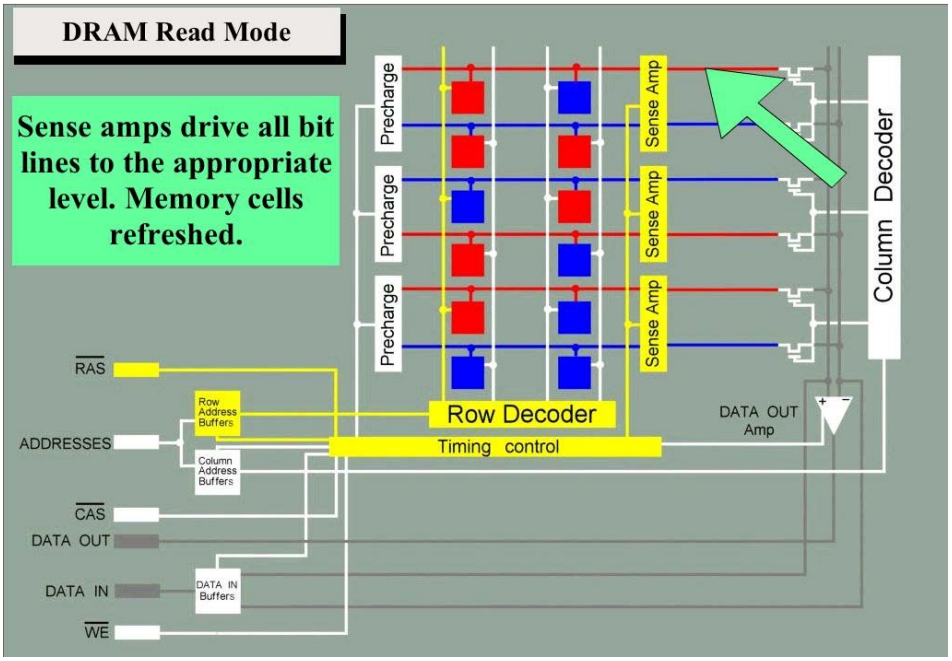
When the word line is activated, all of the memory cells attached to that word line are connected to their respective bit lines.



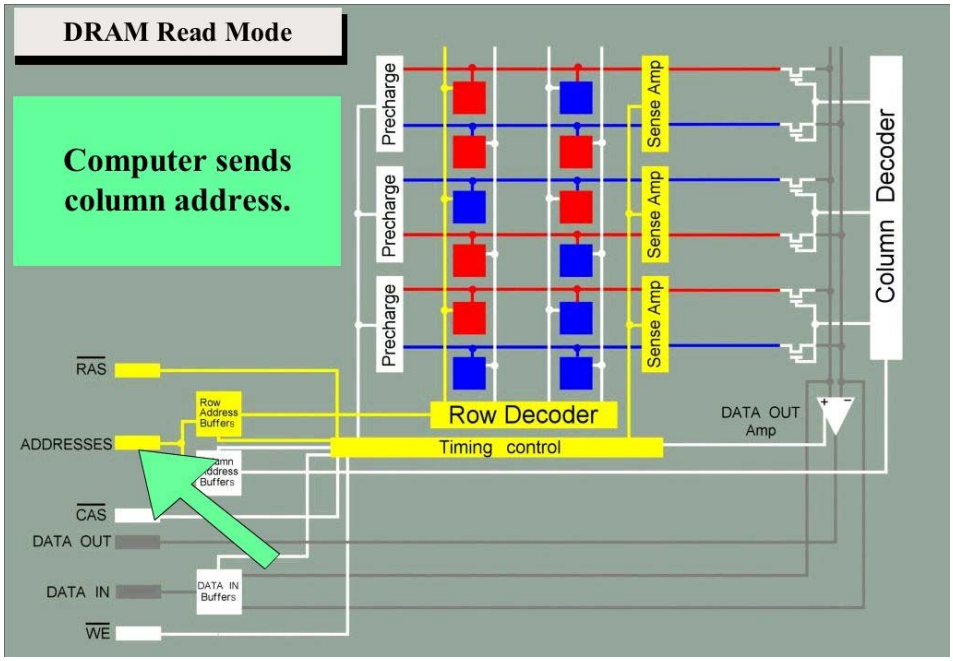
The information stored in those memory cells, represented as electrical charge or lack of charge, slightly changes the voltage on those bit lines. Therefore, there will be a voltage difference between each bit line and its corresponding complement bit line (the bit line pair), caused by the information in the memory cell now attached to that bit line.



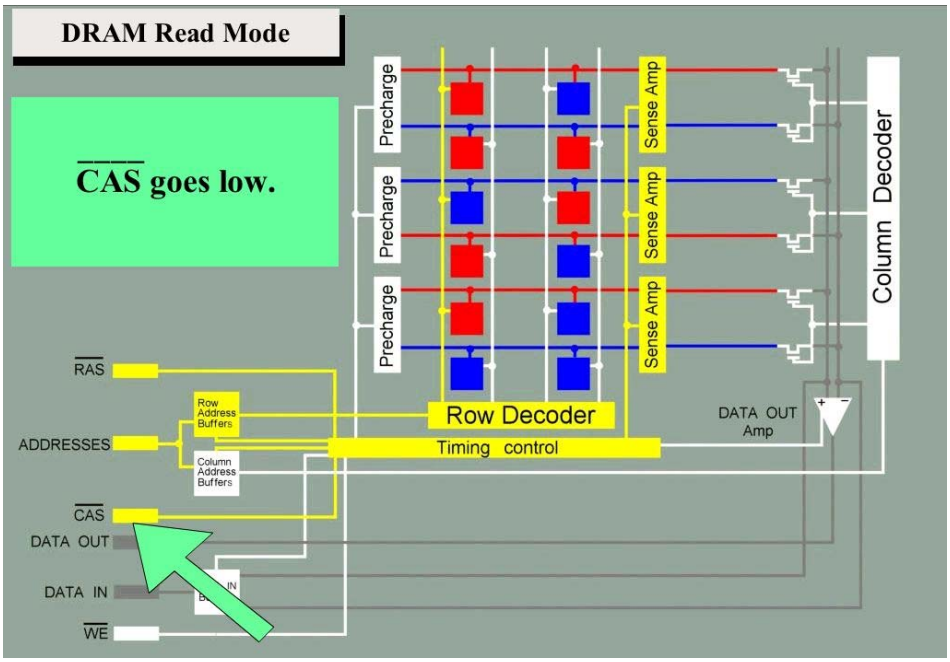
The bit line Sense Amplifiers amplify the voltage difference on each bit line pair,



and rewrite the information into the memory cells that initially provided it. As mentioned above, this rewriting is called “refreshing”, and must be done periodically to every memory cell on the chip.

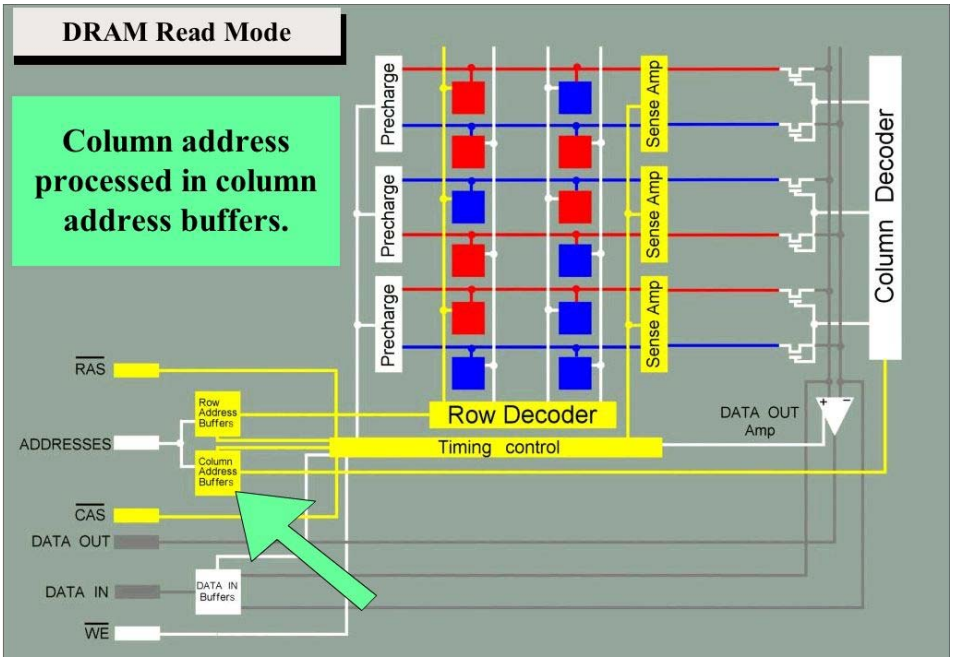


While the sense amplifier is operating, new voltages representing the address of the desired column are applied to the Address Pins.

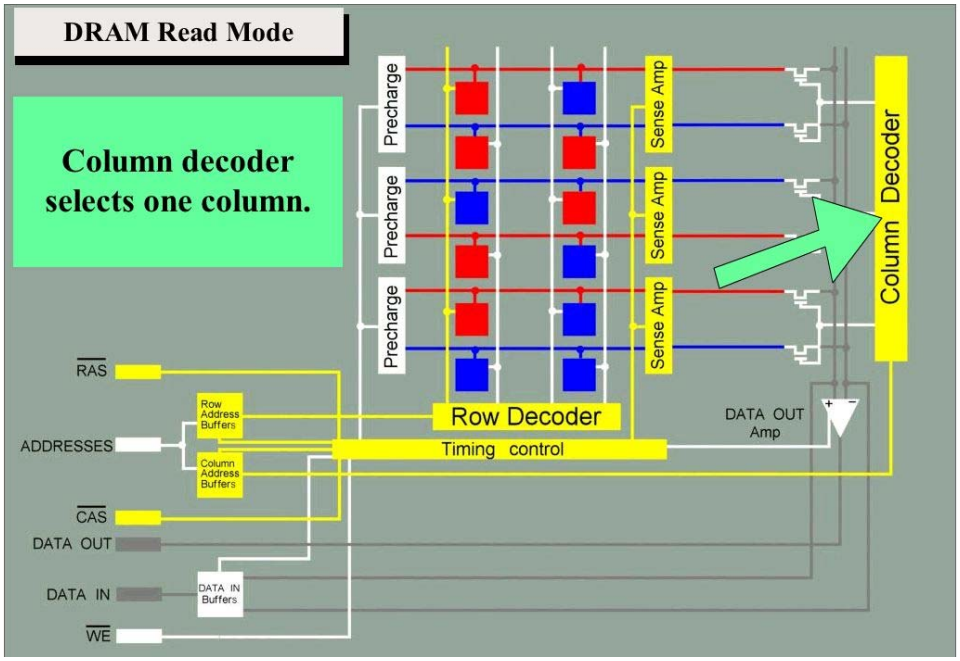


Then a signal called CAS/ (pronounced CAS-Bar) is applied to the memory[11], thus gating the voltages on the Address Pins into the Column Address Buffers.

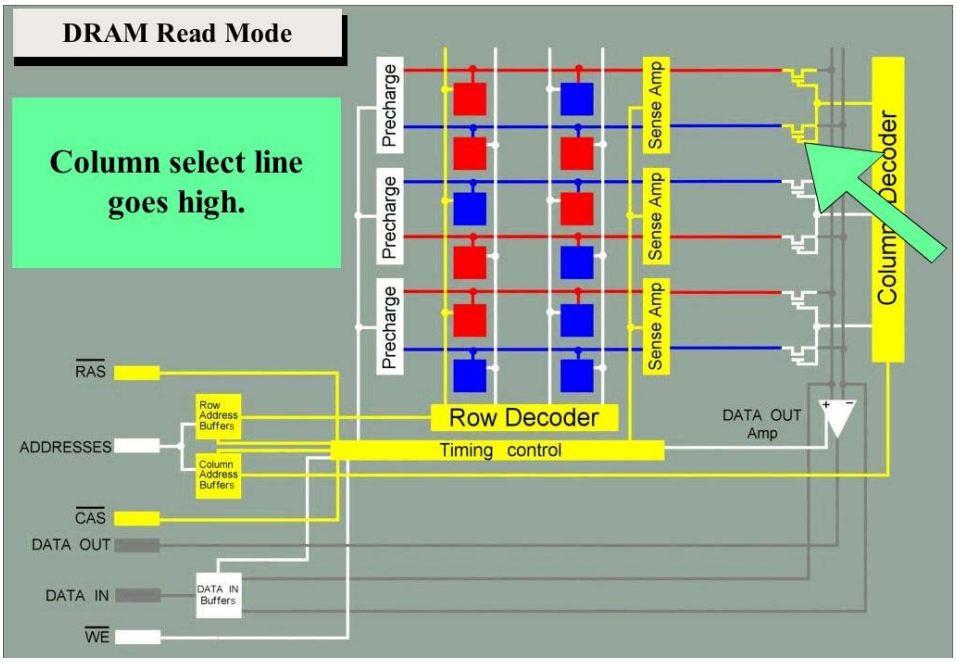
[11] As mentioned in the text earlier, the same address pins are used to receive both the row and the column addresses. The system controller identifies which set of addresses is present by sending either a RAS/ (for row addresses) or CAS/ (for column addresses) signal to the memory device. This sharing of address lines and pins is commonly called “address multiplexing”.



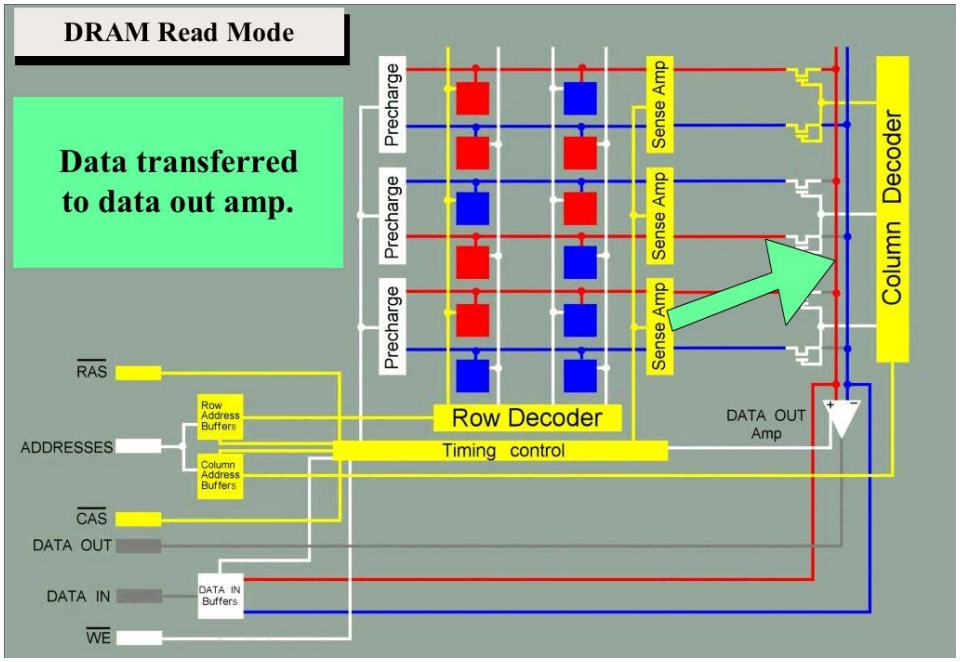
The Column Address Buffers sense the address input voltage levels and generate internal addresses that are sent on internal address lines to the Column Decoder.



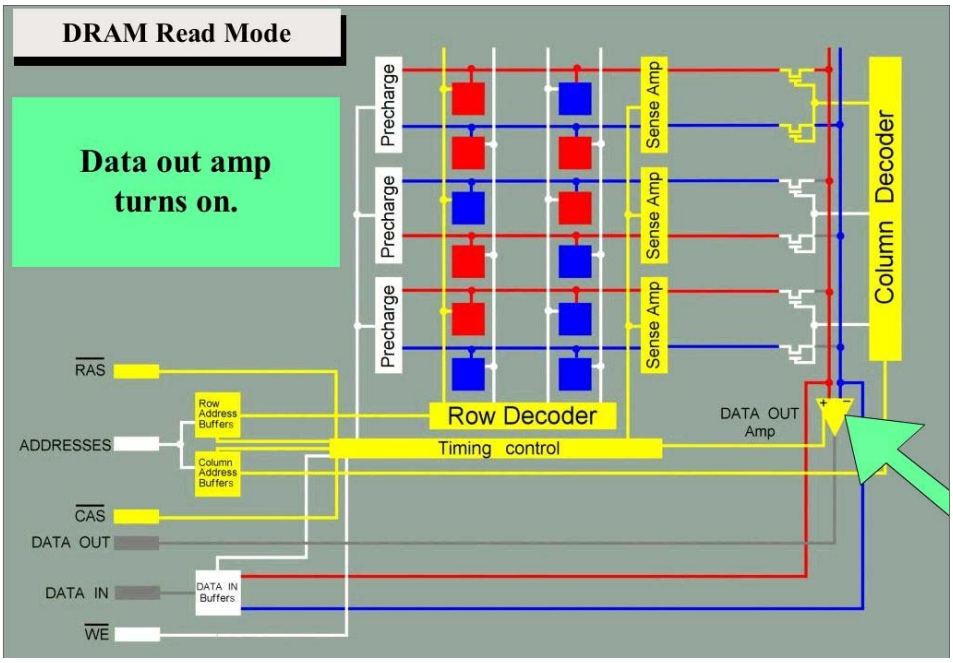
Based on the internal addresses applied to it, the Column Decoder is now activated and selects the desired column.



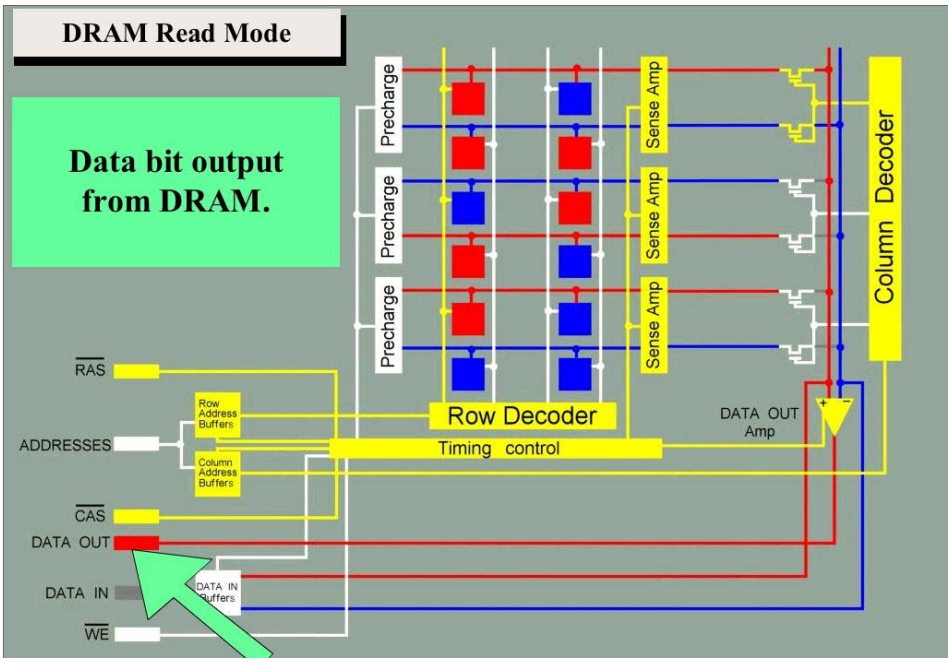
The Column Decoder selects and drives the Column Select line high.



The differential signal from the selected bit line pair is coupled to the internal data lines.

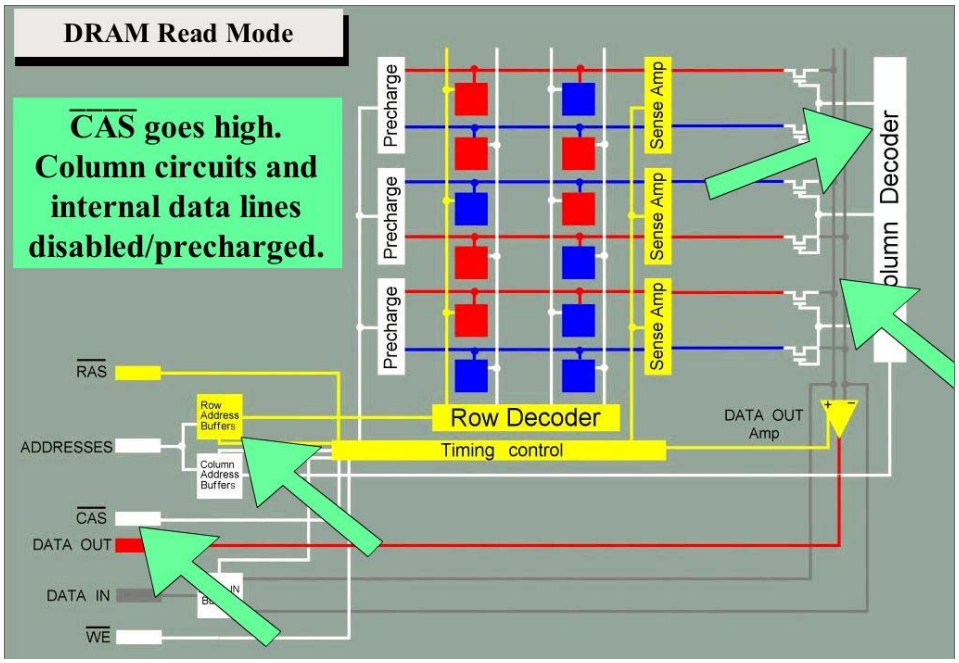


The signal on the internal data lines is amplified by the DATA OUT Amp.

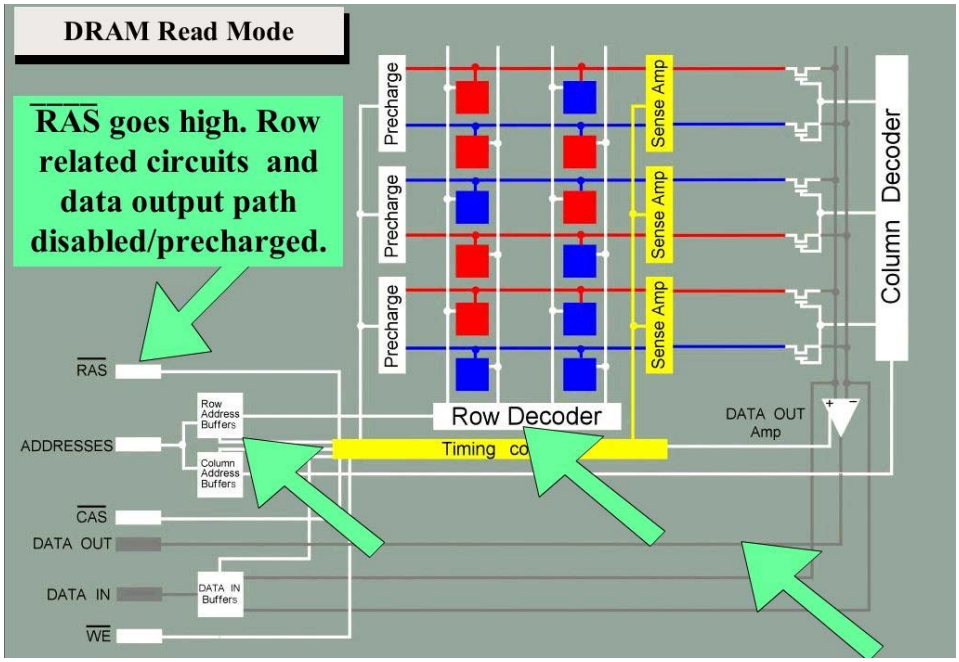


The output of the DATA Out Amp appears as valid data at the DATA OUT pin of the DRAM.

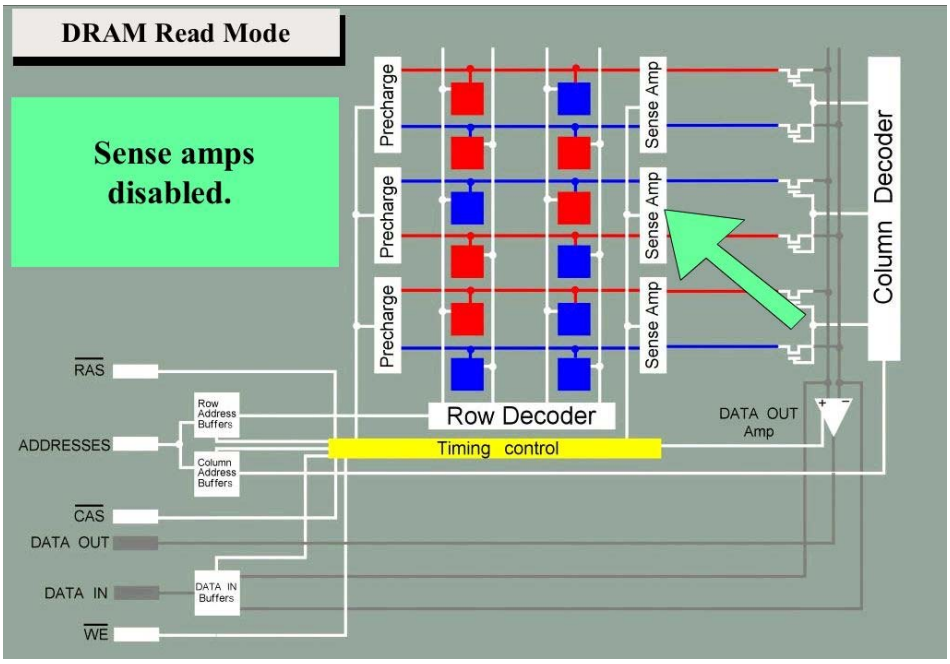
To complete the read cycle, the memory must be returned to its precharge state. The following steps accomplish this task.



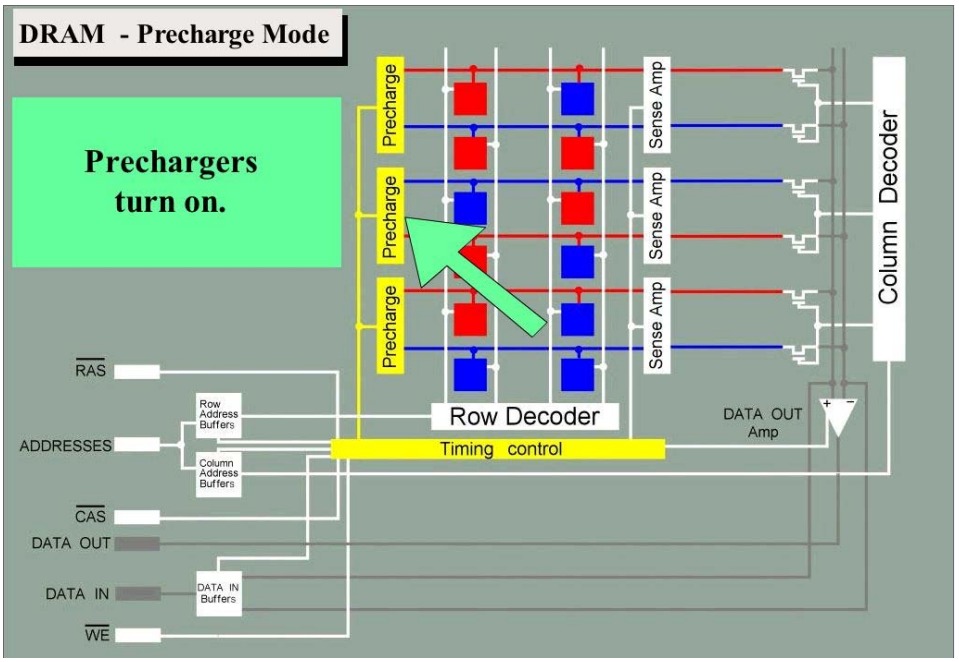
To start the recovery, CAS/ goes high, thus allowing column circuits and internal data lines to be recovered to their precharged levels.



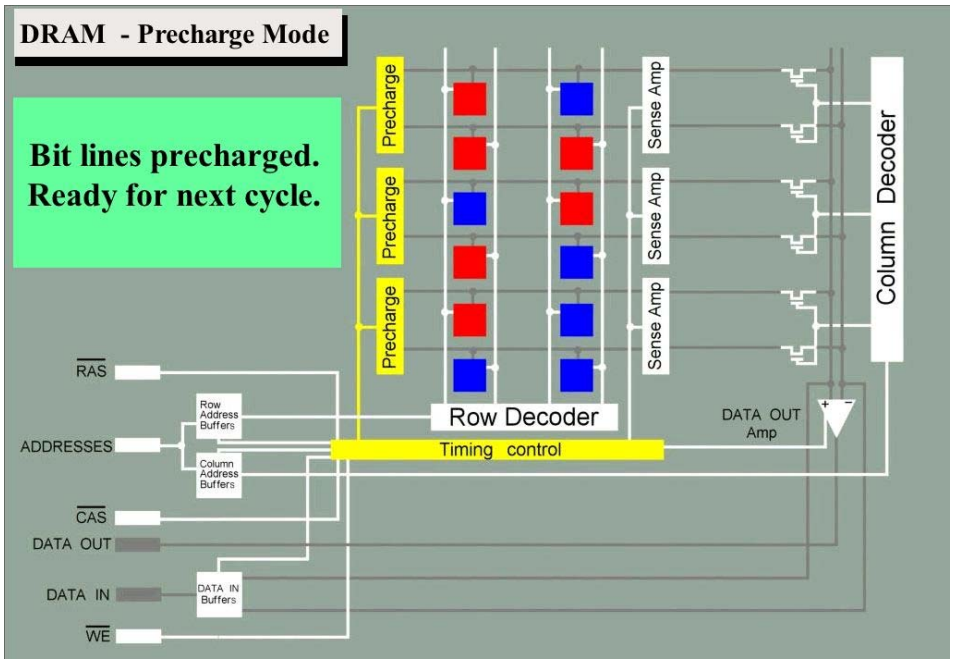
Then RAS/ goes high, first allowing the row circuits and data output path to be recovered to their precharged levels.



Next the bit line Sense Amps are disabled.



The bit line Precharge circuits are turned on.



Now the DRAM is ready for the next Read or Write cycle.

The operation just described is called a “random access read,” which means that any cell in the memory can be selected for the reading operation.

A similar operation, called a “random access write,” sends data from an external source, via the DATA IN pin and DATA IN Buffer, to any selected memory cell on the chip for storage there.

Notice that the DRAM has no external clock signal. All operations are initiated by either RAS/ or CAS/ changing state; and on a READ cycle, data appears at the DATA OUT pin as soon as the internal circuitry completes the necessary operations. Such a DRAM is now called “asynchronous,” as current DRAM architectures all have clock inputs to control all external and many internal operations.

D. Accelerated Access Modes

Over the years, access modes other than random read and random write have been developed, each with the objective of reducing the time required to read data from or write data to the memory. The names of some of these modes are page, static column and nibble. All of these accelerated access modes gain their speed advantage by restricting their access to a single row of the memory. In other words, row selection is done once at the beginning of the cycle and then data is read from or written to cells

in multiple columns along that row. Thus, the entire RAS/ portion of the memory cycle is performed once, and multiple data bits are provided very quickly from memory cells in columns along the selected row. Each of these accelerated access modes starts with row selection and sensing by the bit line sense amplifiers.

1. **Page Mode**

   In a page mode read cycle, after CAS/ initially goes active and gates in the first column address, CAS/ goes inactive. Then a new column address is applied to the Address Pins and CAS/ goes active again. The new column address selects a new column on the same row that was previously selected, and the data from the memory cell at the row-column intersection is provided to the DATA OUT Pin of the DRAM. As with all of the accelerated access modes, because the row selection has already been done, a page mode read cycle is much faster than a random read cycle.

2. **Static Column Mode**

   In a static column mode read cycle, after CAS/ initially goes active and gates in the first column address, it stays active for the balance of the accelerated access mode. A new column address can be provided by the external system at any time, and the change in the column address is detected on the memory chip by a circuit called an address transition detector (ATD). The ATD circuit causes the memory chip to respond to the new column address by selecting a new column on the same row that was previously selected, and the data from the memory cell at the row-column intersection is provided to the DATA OUT Pin of the DRAM.

3. **Nibble Mode**

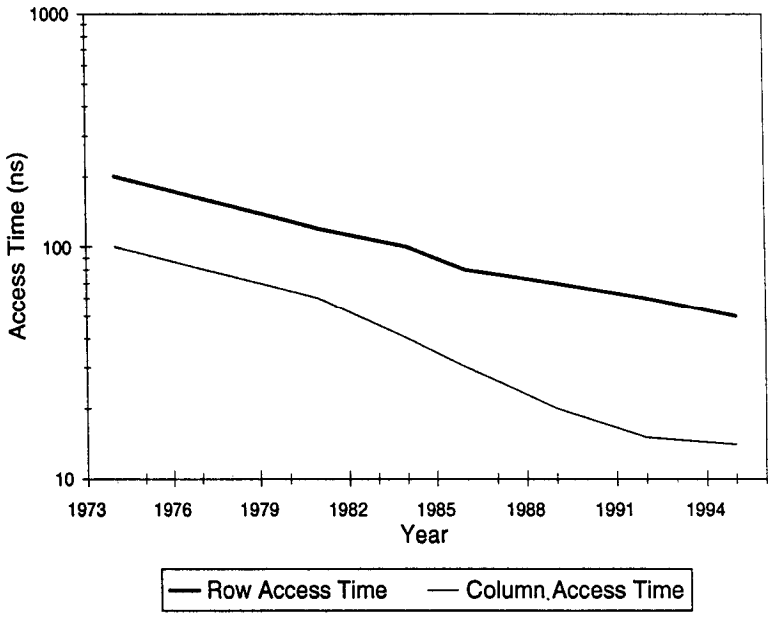
   In a nibble mode read cycle, CAS/ initially goes active and gates in the first column address. Similar to the DRAM described above, more than one column is accessed. Typically, four columns are accessed and their data amplified and stored in four separate I/O amplifiers. The data from the addressed column is provided to the DATA OUT Pin of the DRAM. When CAS/ goes inactive and then active again, the data from the next column in sequence is read from the next I/O amplifier. Two more CAS/ cycles result in data from the next two columns being read-out sequentially. Another CAS/ cycle would result in data from the first column being read-out again. Note that, in nibble mode, the initially applied column address determines which four bits will be available.

4. Summary of DRAM Accelerated Access Modes

The following table summarizes the significant characteristics of the various DRAM accelerated access modes. All of these modes start after row addresses are applied to the memory device and RAS/ becomes active (RAS/ must stay active until the end of the accelerated access mode).

Access Mode	CAS/ Activity	Column Address Within Selected Row	Address Transition Detector
Page	cycles	random; externally provided while CAS/ is inactive	not required
Static Column	stays active	random; externally provided while CAS/ is active, but must remain stable until data is obtained by system	required
Nibble	cycles	predetermined by initial column address; advances by one for each CAS/ cycle; typically, four bits are available	not required

While random or row access times have decreased by about 8% per year, the column access times important for the accelerated access modes have decreased by about 25% per year, as shown in the next graph.



Evolution of Row and Column Access Times[12]

E. Fill Frequency

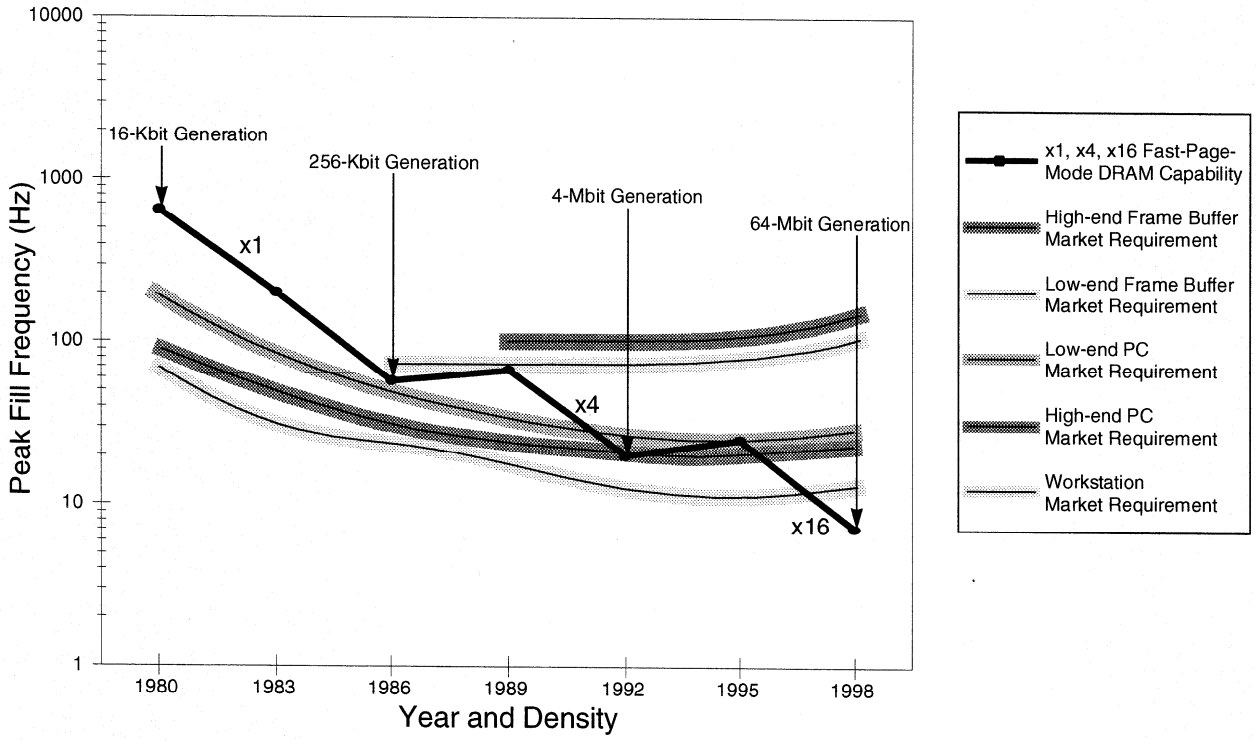
A detailed analysis of modern system requirements and the capability of various memory architectures to meet those requirements is provided by Przybylski[13]. To compare system requirements and memory capabilities, he introduces the concept of **fill frequency**. Very simply put, fill frequency describes how fast all of the bits in a memory can be read or written. Clearly, if two memories of different size operate at the same speed, the smaller memory can be read in its entirety more quickly than the larger memory. The smaller memory simply has fewer bits to read, and therefore has the higher fill frequency.

Likewise a memory with 4 input/output pins will have a fill frequency 4X higher than an otherwise identical memory with only 1 input/output pin. This fill frequency advantage arises because the 4 pins provide data simultaneously, so that 4X as much data flows from the memory in a given time, as compared to the memory with only 1 input/output pin. These concepts are captured in the following graph, which compares DRAM fill frequency capability to the requirements of various memory systems.

---

[12] from Przybylski, S., **New DRAM Technologies**, Second Edition; MicroDesign Resources, Sebastopol, CA; 1996; p. 14

[13] Op. cit.



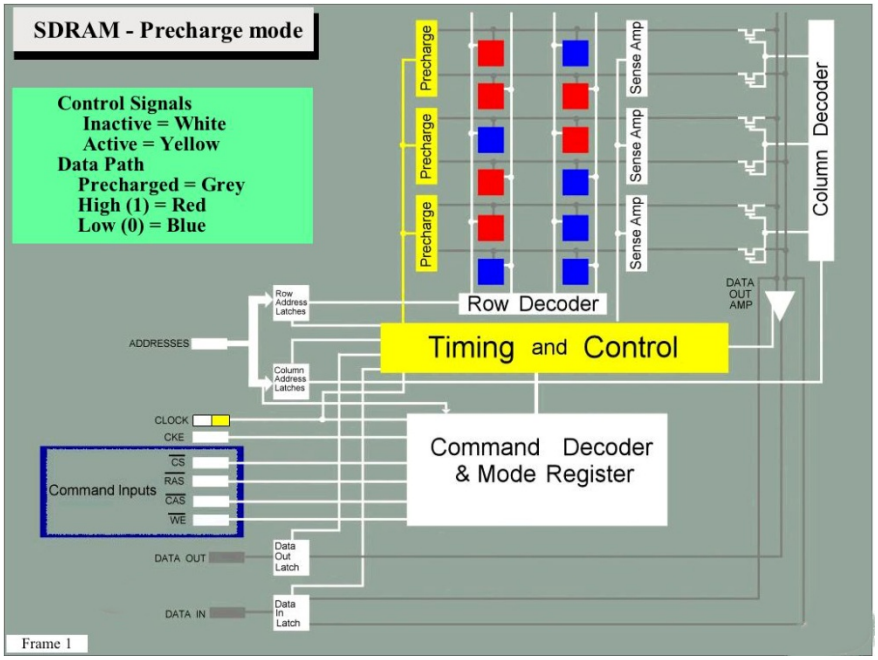
System and DRAM Fill Frequencies[14]

It is clear from the graph that DRAM fill frequency has decreased significantly as density has increased, even though wider data paths (x1→x4→x16) have been provided. Meanwhile, the advent of frame buffers for video applications has emphasized the need for higher fill frequencies, exceeding the capabilities of available DRAMs.

[14] op. cit.; p. 37

F. SDRAM Architecture and Operation

In the early 1990s, a new approach to providing ever-increasing amounts of data in ever-decreasing time was introduced. A memory architecture called synchronous DRAM ("SDRAM") retained the "core" of the DRAM chip, but significantly modified the earlier accelerated access modes. The following block diagram illustrates the major architectural features of an SDRAM.



Basic SDRAM Architecture

Note that the SDRAM architecture is very similar to a DRAM, with the addition of Command and Clock Inputs and a Command Decoder & Mode Register. Some of the significant differences in the operation of an SDRAM as compared to a DRAM are as follows:

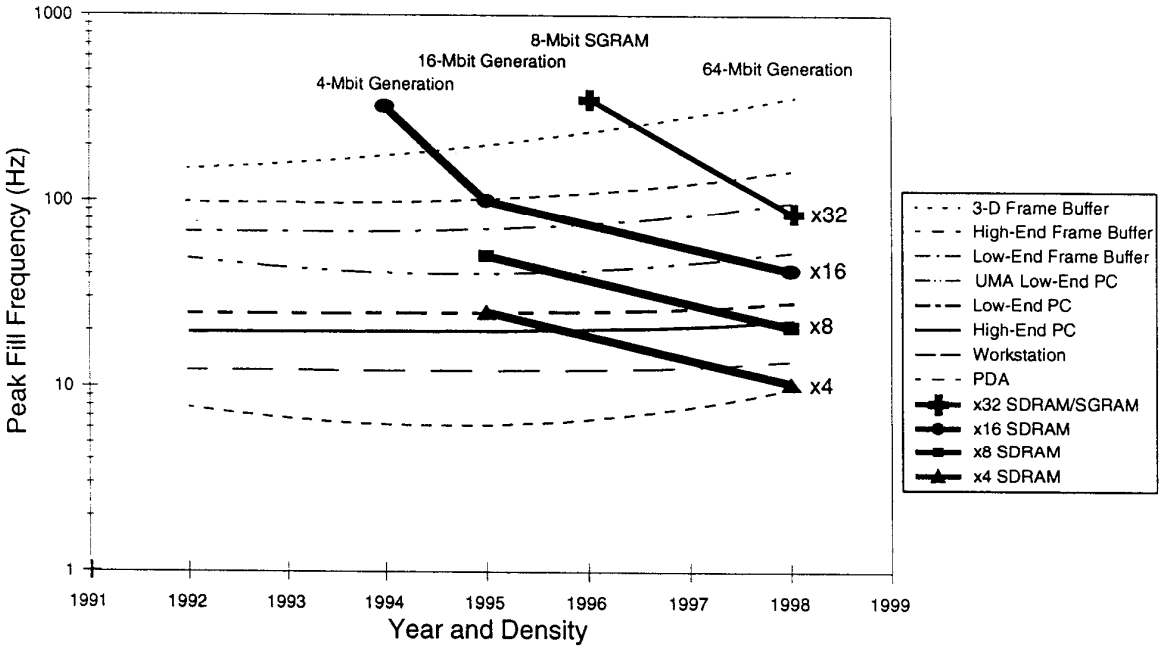
- SDRAM operation is initiated by the (low to high) transition of the external clock, and data bits are input or output in synchronism with the external clock;

- In an SDRAM, the operational mode (ACTIVE, READ, WRITE, REFRESH, etc.) is established by a command issued at the beginning of the cycle;

- The SDRAM command is defined by the logic state of several inputs (e.g.; RAS/, CAS/, WE/ and A10) when CS/ (Chip Select) is low, CKE (Clock Enable) is high, and CLK (external Clock) transitions from low to high; and

- Specific characteristics of the SDRAM operation, such as how many bits of data will be read out sequentially (burst length) and how many clock cycles will elapse between the READ command and the appearance of the first bit of data (latency), are specified by the system and stored in the SDRAM mode register prior to the READ command.

The impact of the SDRAM architecture on fill frequency is shown in the graph below.



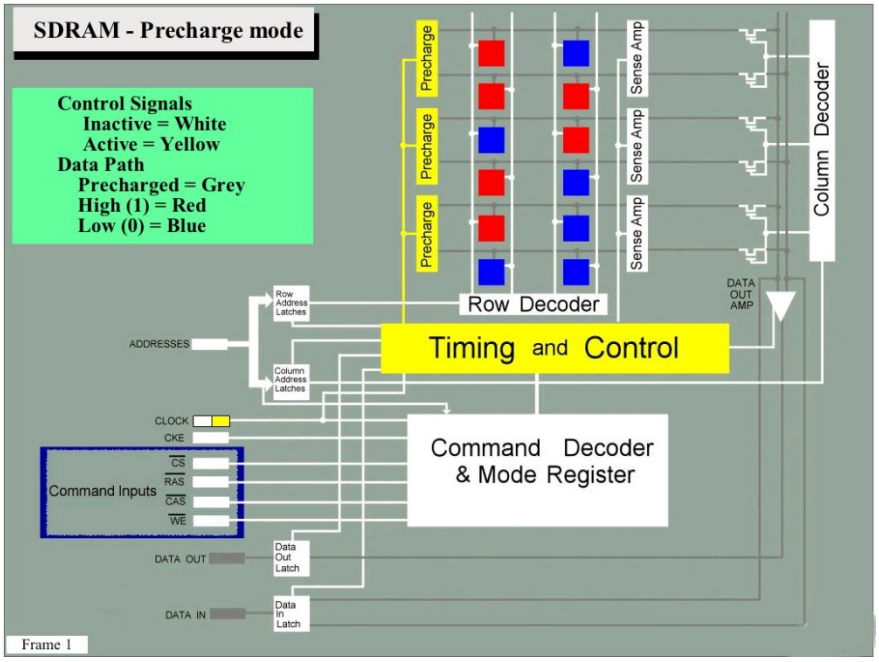
System and SDRAM Fill Frequencies[15]

Comparing the DRAM and SDRAM Fill Frequency graphs makes it clear that the SDRAM architecture greatly improves fill frequency performance.

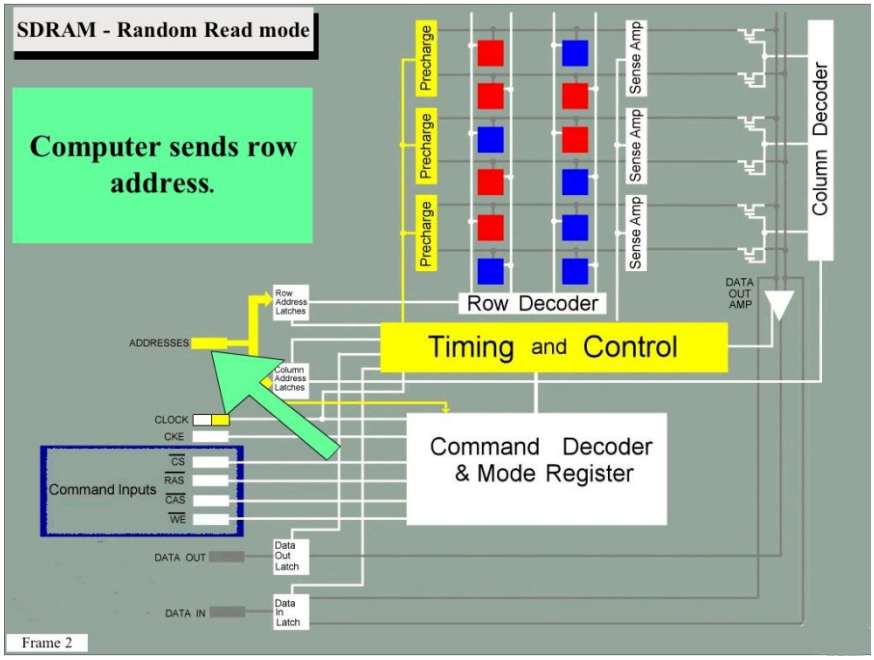
[15] op. cit.; p. 231

G. SDRAM Operation

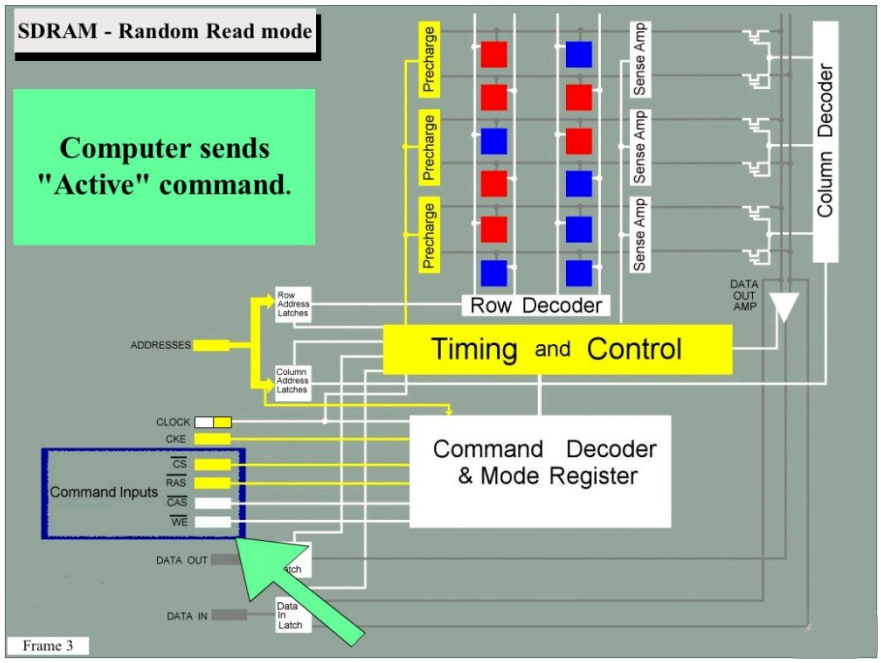
We will now examine the basic operation of an SDRAM, and highlight the important differences relative to a DRAM. All external and many internal SDRAM operations are initiated by the low to high transition of the external clock.



In the Precharge mode, the SDRAM is idle and awaiting inputs and a command to start an operation.



As the first step in starting an ACTIVE operation, the computer sends row addresses to the Address pins of the SDRAM.



Next the computer sends the ACTIVE command.

In a DRAM, RAS/ would be driven low to initiate the operation. In the SDRAM, a set of 4 signals (CS/, RAS/, CAS/ and WE/) is controlled to indicate what operation is to be performed. The table below[16] shows the possibilities:

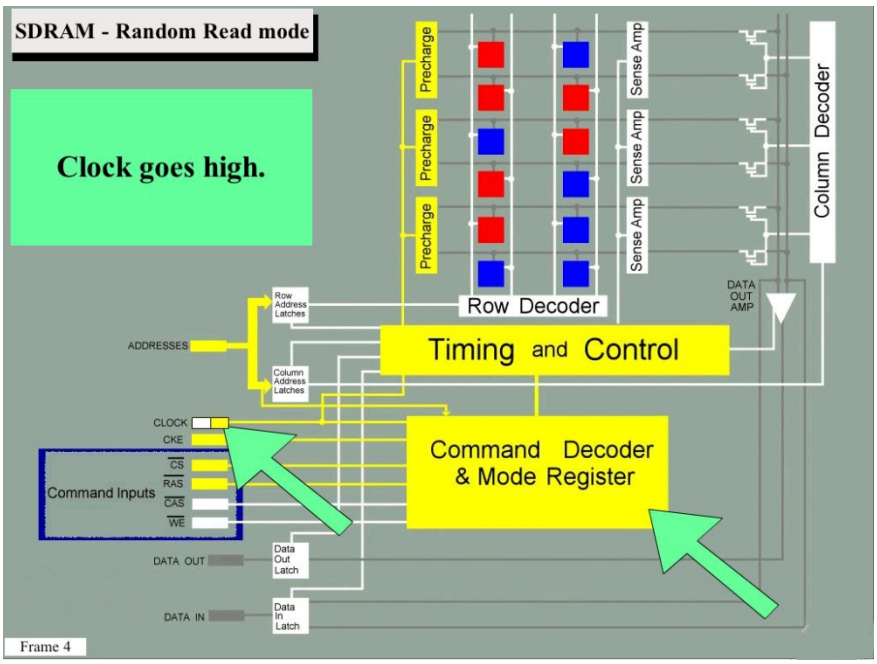
SDRAM COMMANDS

Name (Function)	CS#	RAS#	CAS#	WE#	DQM	ADDR	DQ
COMMAND INHIBIT (NOP)	H	X	X	X	X	X	X
NO OPERATION (NOP)	L	H	H	H	X	X	X
ACTIVE (select bank and activate row)	L	L	H	H	X	Bank/row	X
READ (select bank and column, and start READ burst)	L	H	L	H	L/H	Bank/col	X
WRITE (select bank and column, and start WRITE burst)	L	H	L	L	L/H	Bank/col	Valid
BURST TERMINATE	L	H	H	L	X	X	Active
PRECHARGE (Deactivate row in bank or banks)	L	L	H	L	X	Code	X
AUTO REFRESH or SELF REFRESH (enter self refresh mode)	L	L	L	H	X	X	X
LOAD MODE REGISTER	L	L	L	L	X	Op-code	X
Write enable/output enable	X	X	X	X	L	X	Active
Write inhibit/output High-Z	X	X	X	X	H	X	High-Z

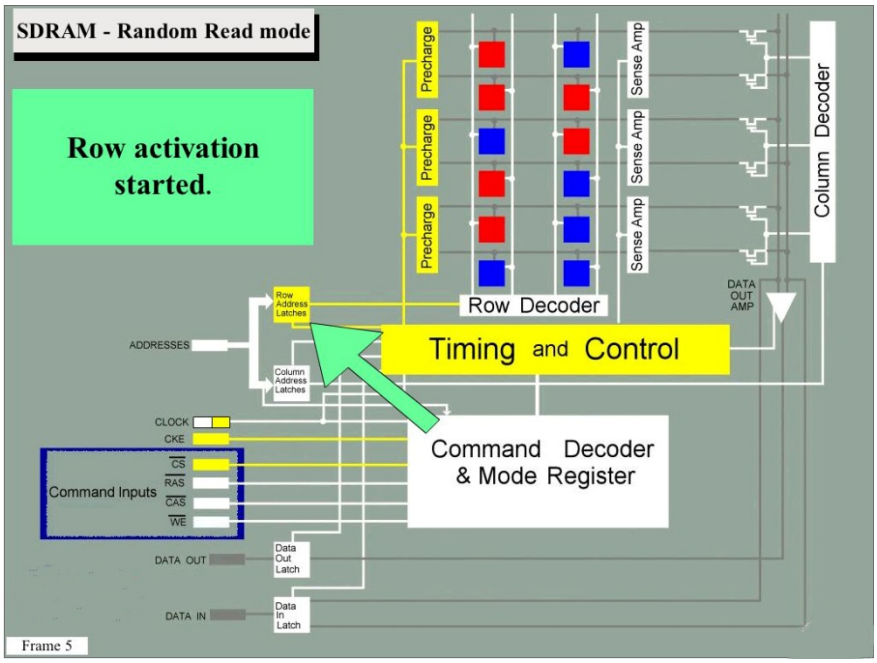
In the table, CS/ and other signals are shown as CS#, etc. Both the / and # indicate that the signal is active low.

For an ACTIVE operation, CS/ and RAS/ are low and CAS/ and WE/ are high.

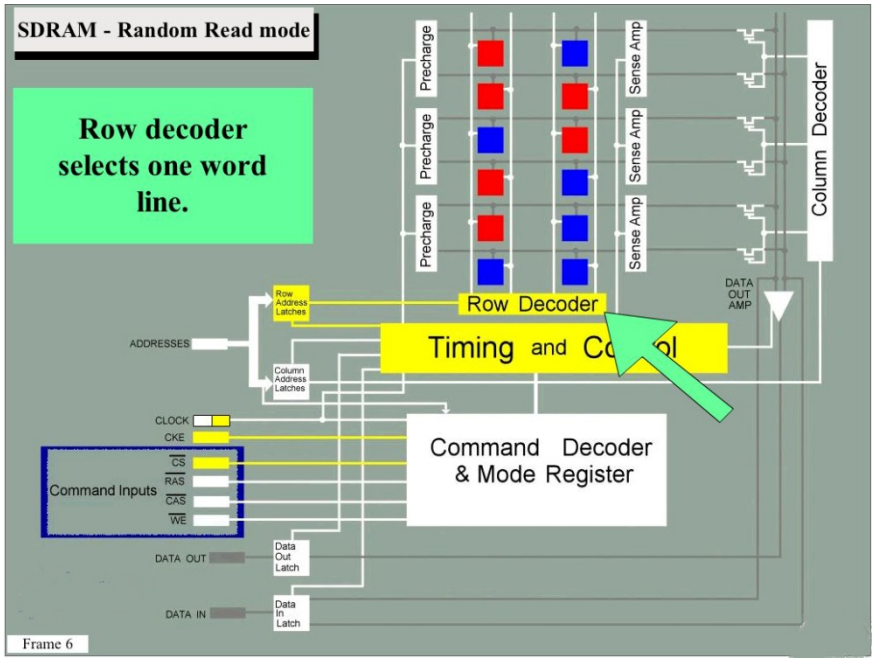
[16] Micron MT48LC128M4A2 512Mb SDRAM Data Sheet, Rev. M 6/10 EN, page 22



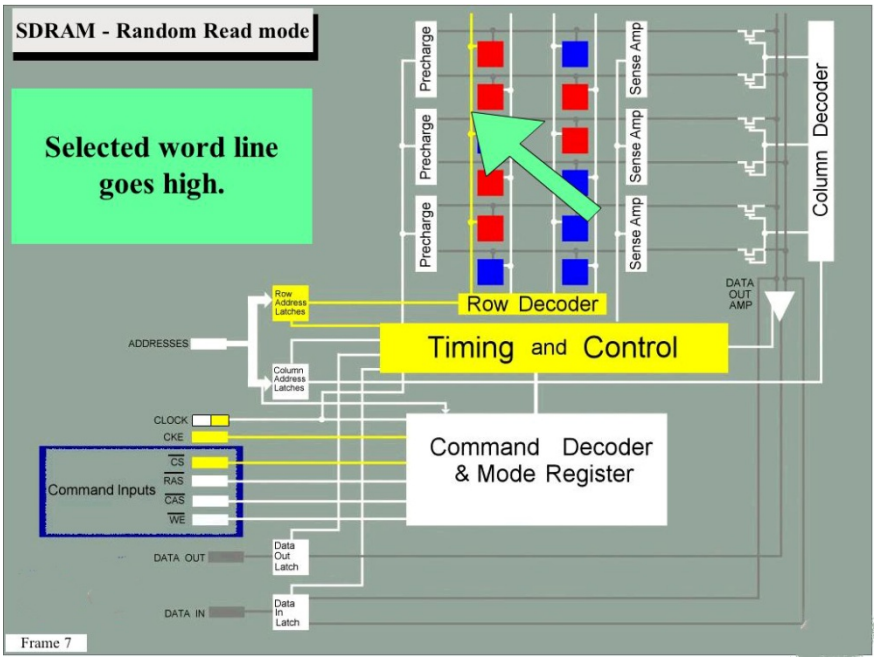
The logic states of the 4 command inputs, as well as the address inputs, are sampled and sent to the Command Decoder and Row Address Latches respectively when the Clock goes high.



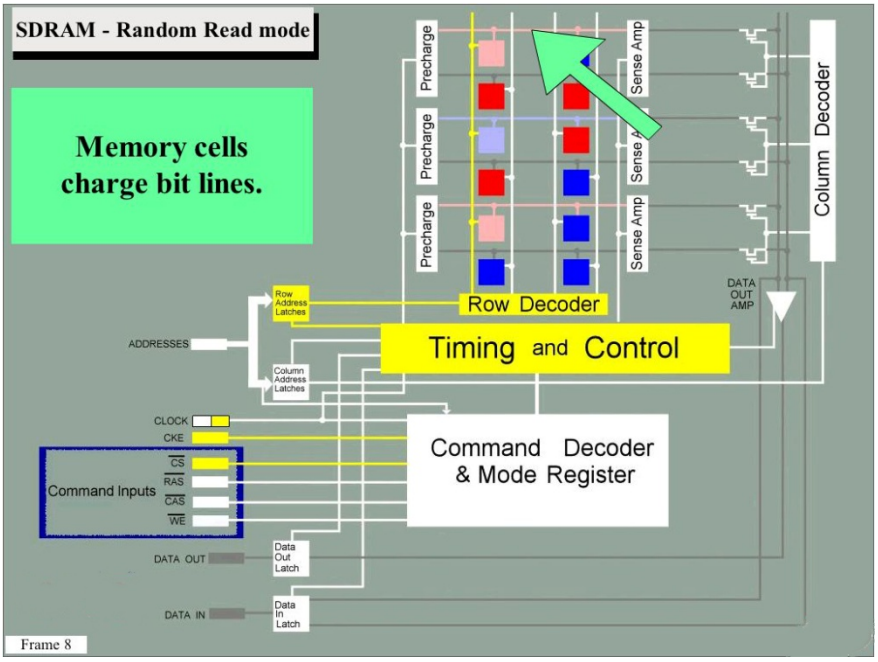
The Row Address Latches sense the address input voltage levels and generate and store internal addresses that are sent on internal address lines to the Row Decoder.



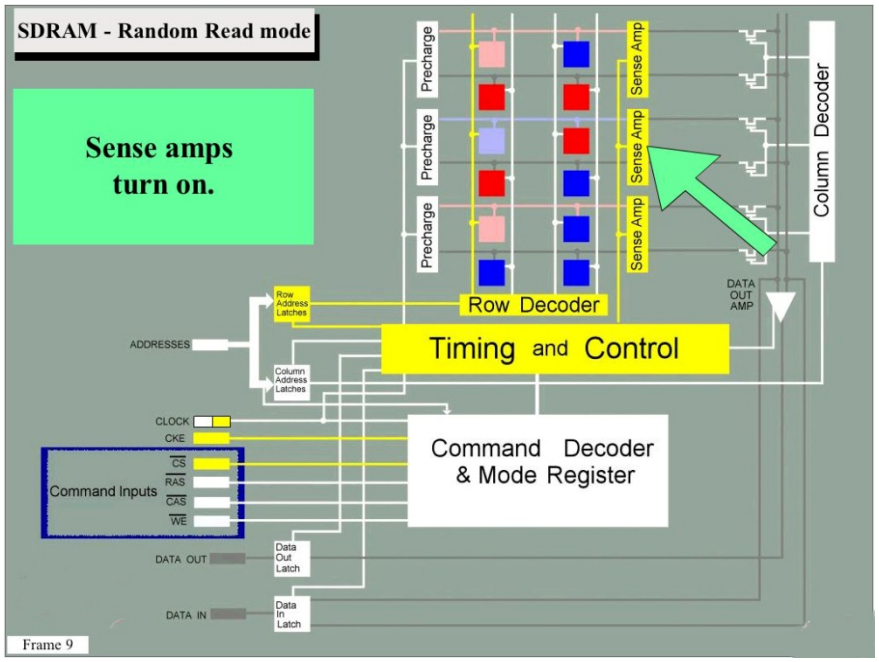
Based on the internal addresses applied to it, the Row Decoder identifies and activates the appropriate word line.



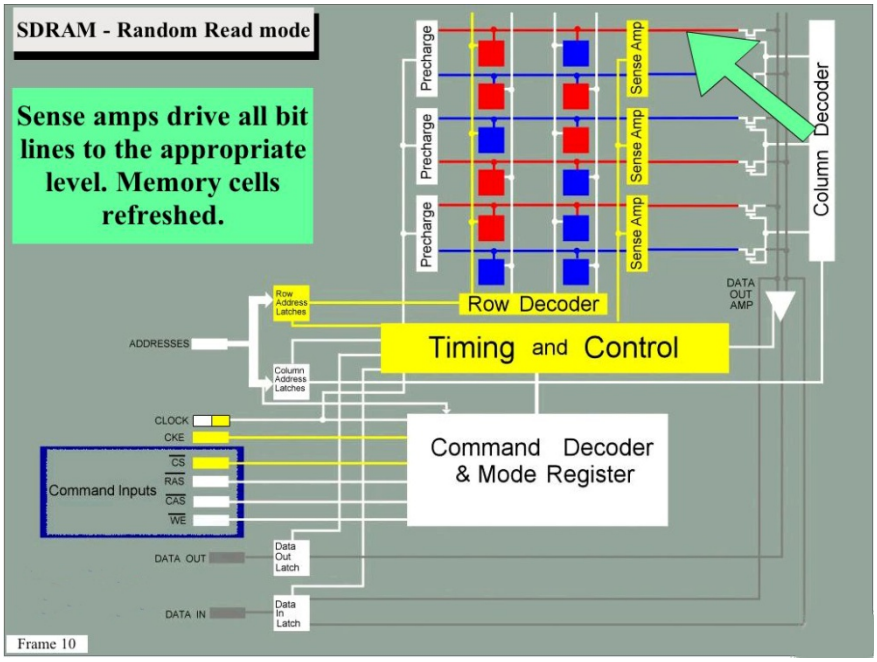
When the word line is activated, all of the memory cells attached to that word line are connected to their respective bit lines.



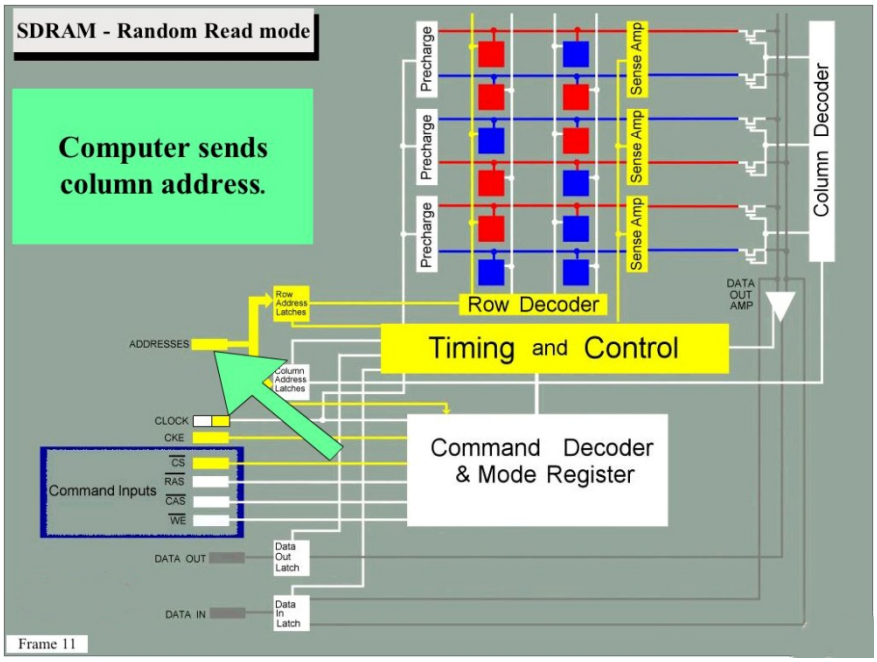
The information stored in those memory cells, represented as electrical charge or lack of charge, slightly changes the voltage on those bit lines. Therefore, there will be a voltage difference between each bit line and its corresponding complement bit line (the bit line pair), caused by the information in the memory cell now attached to that bit line.



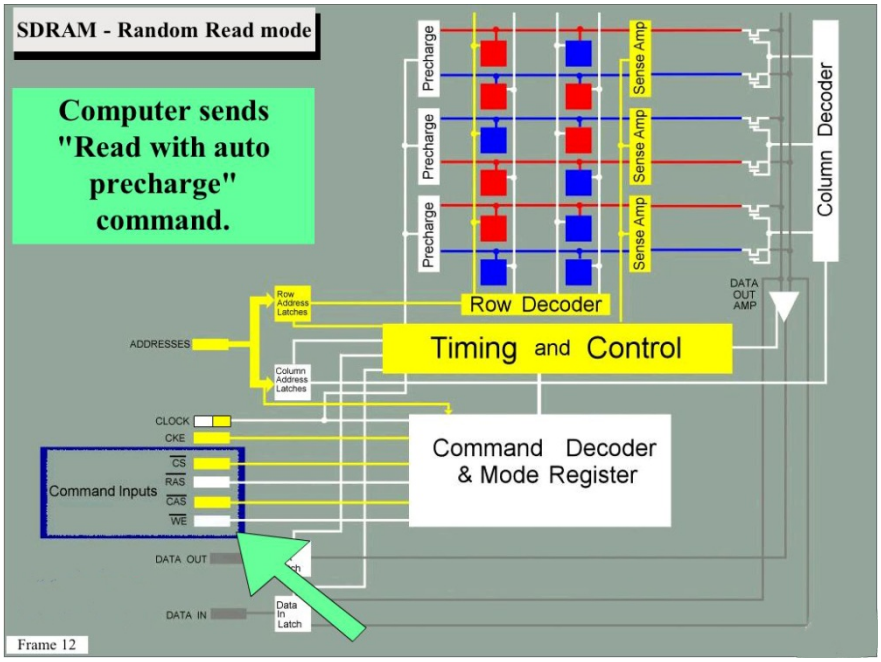
The bit line Sense Amplifiers amplify the voltage difference on each bit line pair,



and rewrite the information into the memory cells that initially provided it, thus refreshing the memory cells on the selected word line.

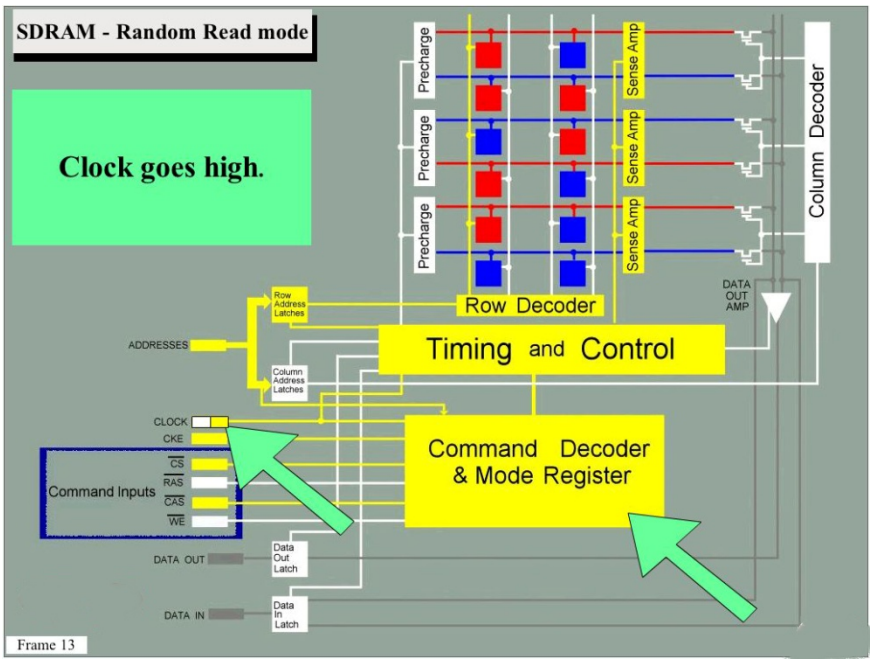


Another operation, as tabulated in the SDRAM COMMANDS table, can be initiated. Assuming we want to do a READ operation, the computer sends new voltages representing the address of the desired column to the Address Pins.

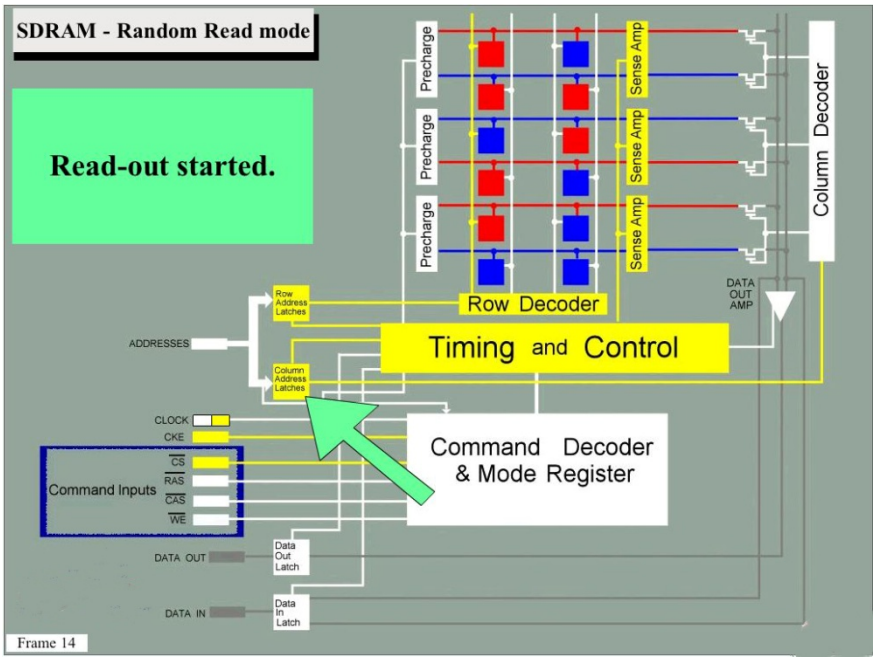


"Read with auto precharge" means that data will be read from the selected memory location, then the memory will be precharged so it is returned to the idle state.

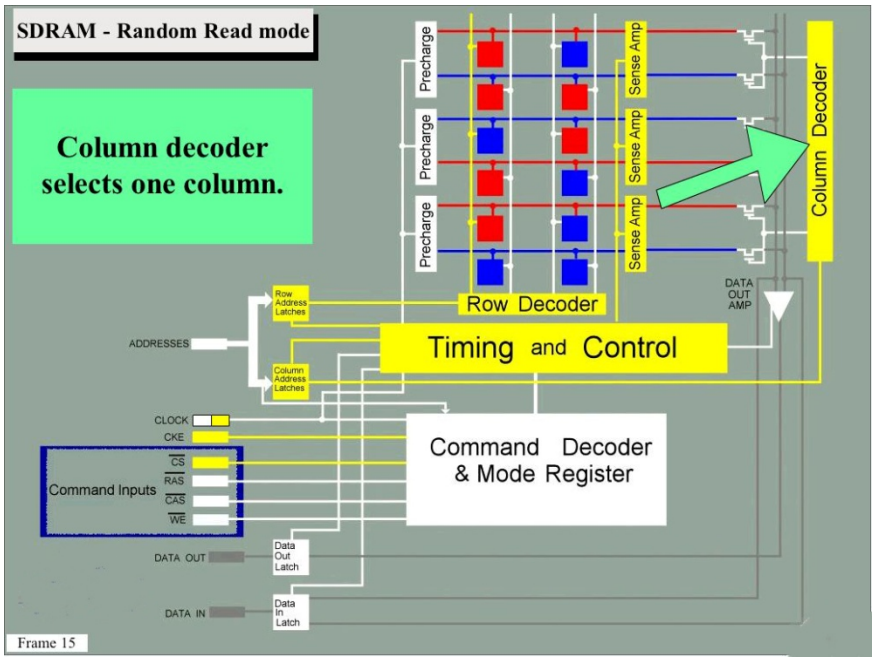
For a READ operation, CS/ and CAS/ are low and RAS/ and WE/ are high. Auto precharge is selected by address A10 high. Auto precharge is disabled if address A10 is low. Address input A10 is not used to select a column.



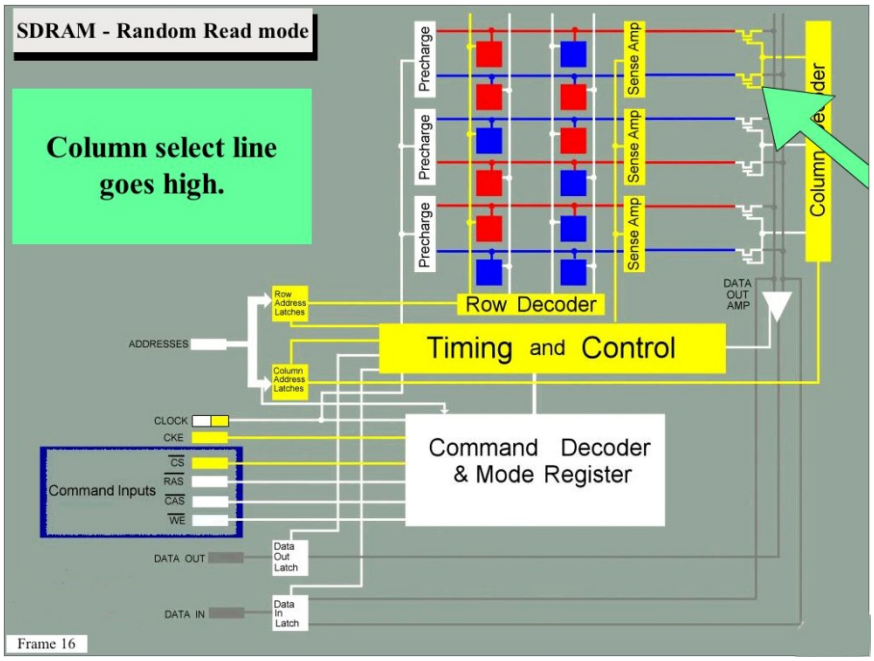
The logic states of the 4 command inputs, as well as the address inputs, are sampled and sent to the Command Decoder and Column Address Latches respectively when the Clock goes high.



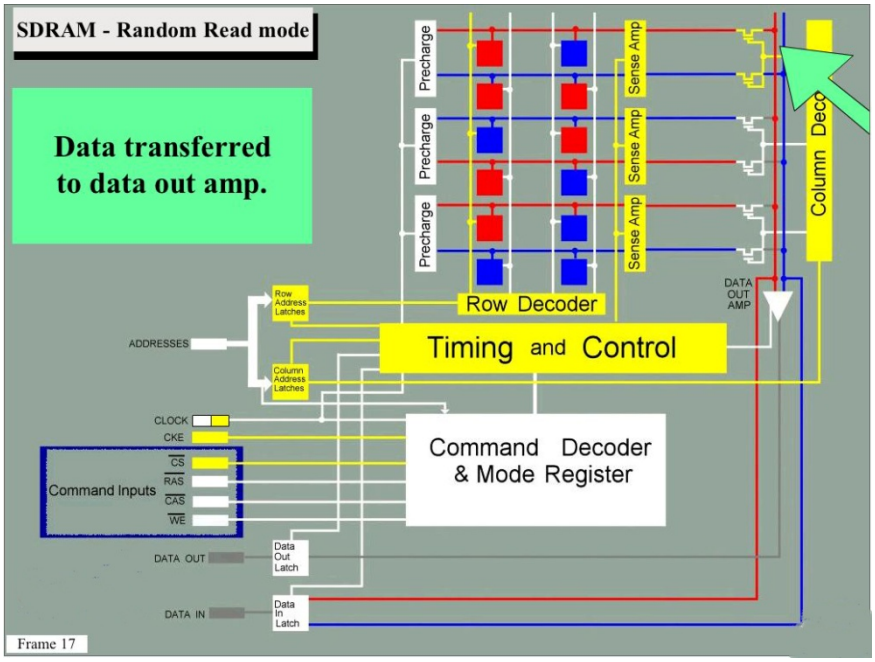
The Column Address Latches sense the address input voltage levels and generate and store internal addresses that are sent on internal address lines to the Column Decoder.



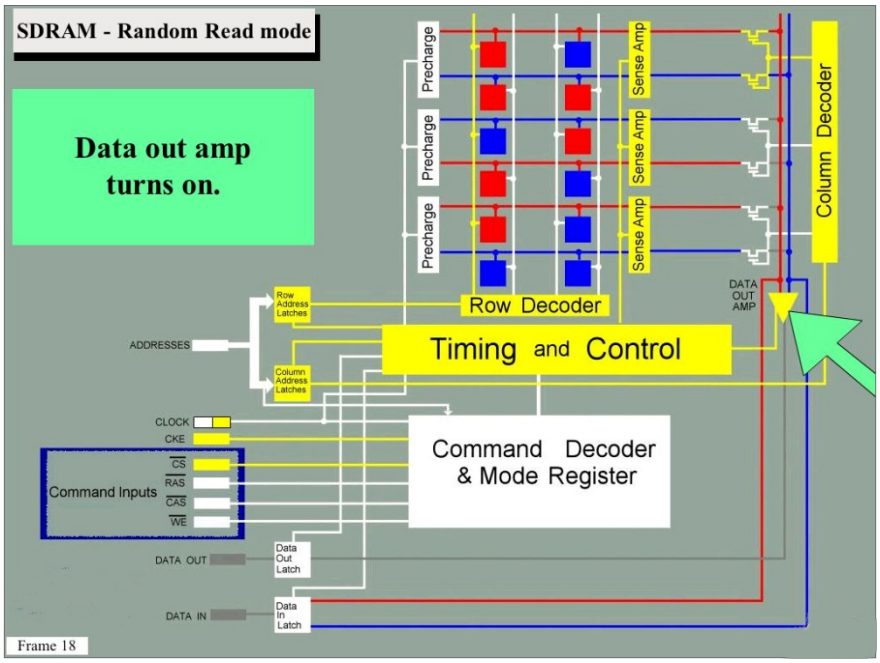
Based on the internal addresses applied to it, the Column Decoder is now activated and selects the desired column.



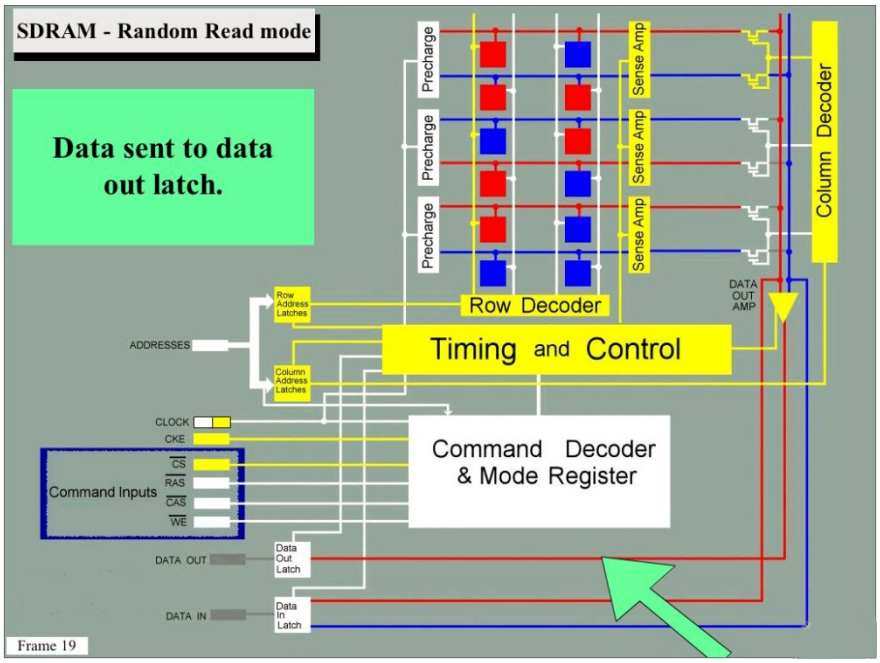
The Column Decoder selects and drives the Column Select line high.



The differential signal from the selected bit line pair is coupled to the internal data lines and sent to the DATA OUT AMP.



The signal on the internal data lines is amplified by the DATA OUT AMP.



The output of the DATA OUT AMP is sent to the Data Out Latch. Note that, unlike in the DRAM, valid data does not appear immediately at the DATA OUT pin of the SDRAM.

In an SDRAM, the time from the READ command to the appearance of valid data on the DATA OUT pin is measured in clock cycles, and is set by the user before the ACTIVE operation begins. The desired number of clock cycles (called the "CAS Latency") is stored as a code in the Mode Register. Typical values of CAS Latency are 2 and 3. From the Micron 512Mb SDRAM Data Sheet:[17]

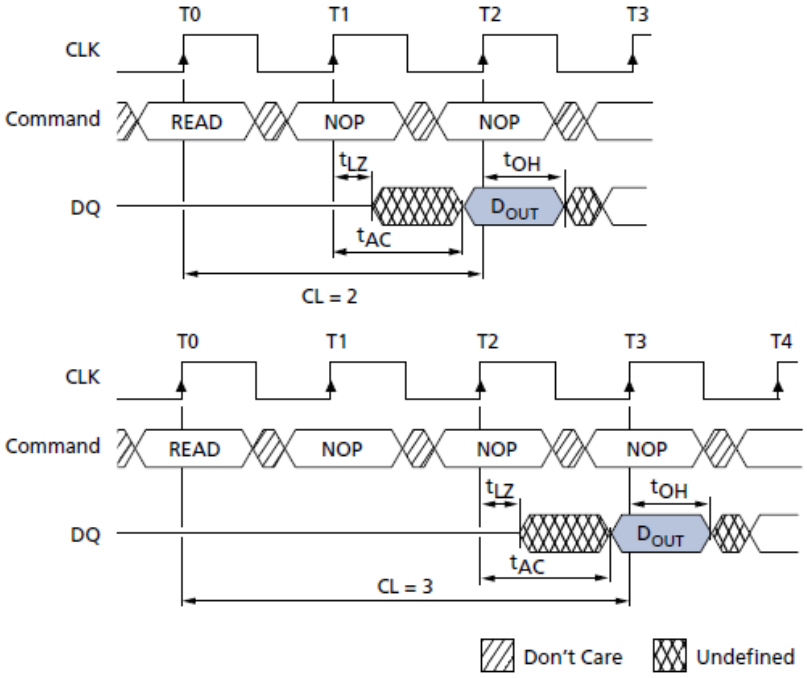
CAS Latency

The CAS latency (CL) is the delay, in clock cycles, between the registration of a READ command and the availability of the output data. The latency can be set to two or three clocks.

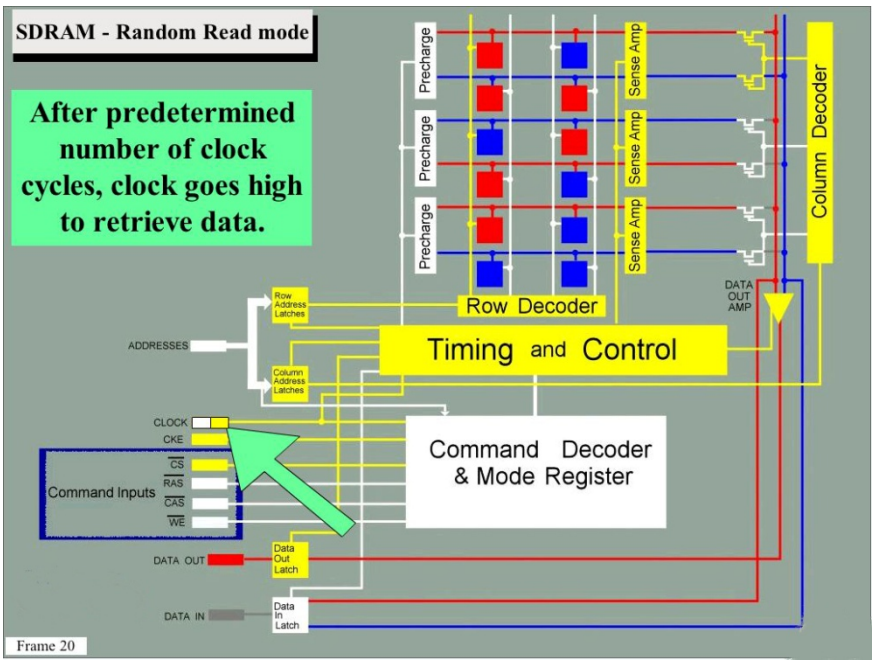
If a READ command is registered at clock edge *n*, and the latency is *m* clocks, the data will be available by clock edge *n* + *m*. The DQ start driving as a result of the clock edge one cycle earlier (*n* + *m* - 1), and provided that the relevant access times are met, the data is valid by clock edge *n* + *m*. For example, assuming that the clock cycle time is such that all relevant access times are met, if a READ command is registered at T0 and the latency is programmed to two clocks, the DQ start driving after T1 and the data is valid by T2.

Reserved states should not be used as unknown operation or incompatibility with future versions may result.

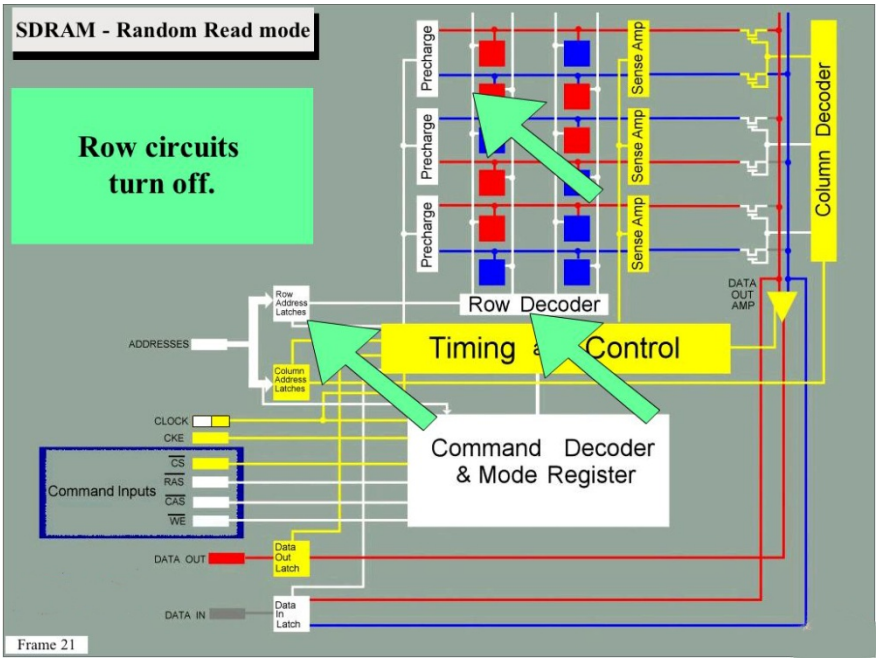
Figure 13: CAS Latency



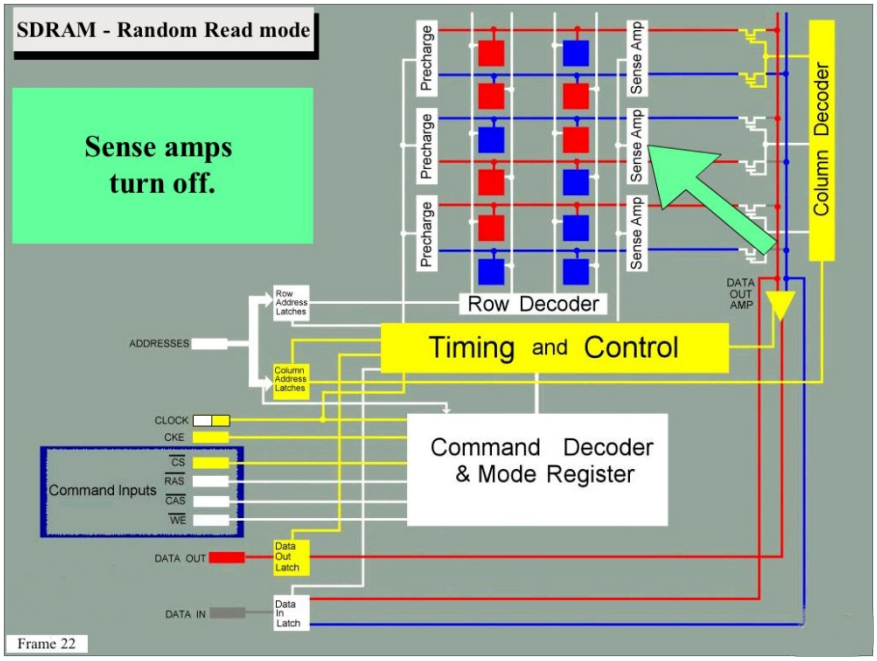
[17] Op. cit., p. 40



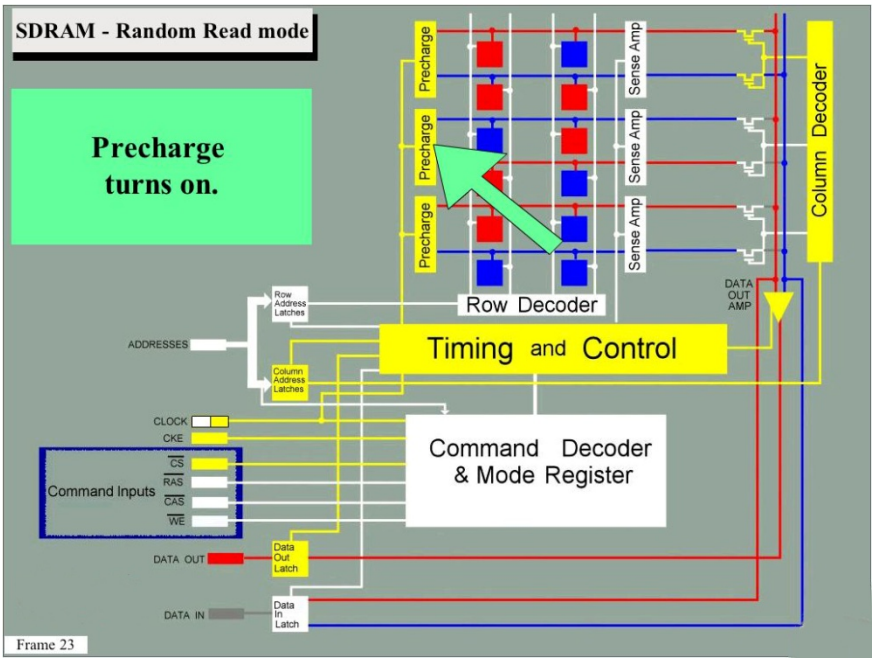
One cycle before the CAS Latency expires, the rising clock edge starts the read-out process from the Data Out Latch. Before the next rising clock edge (marking the end of the CAS Latency time), valid data appears at the DATA OUT pin.



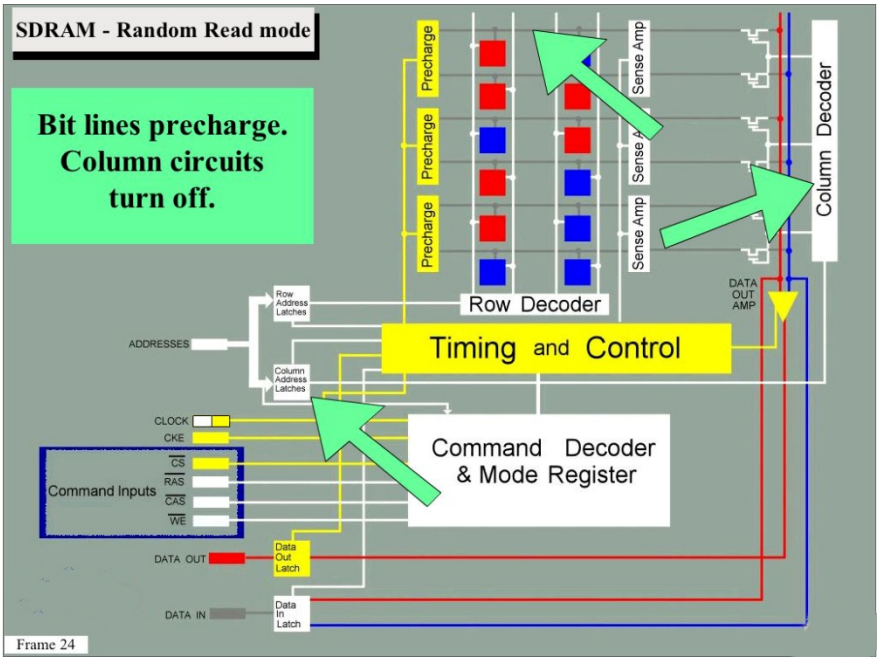
Because the Auto-precharge operation was chosen, the row-related circuits start to turn off immediately.



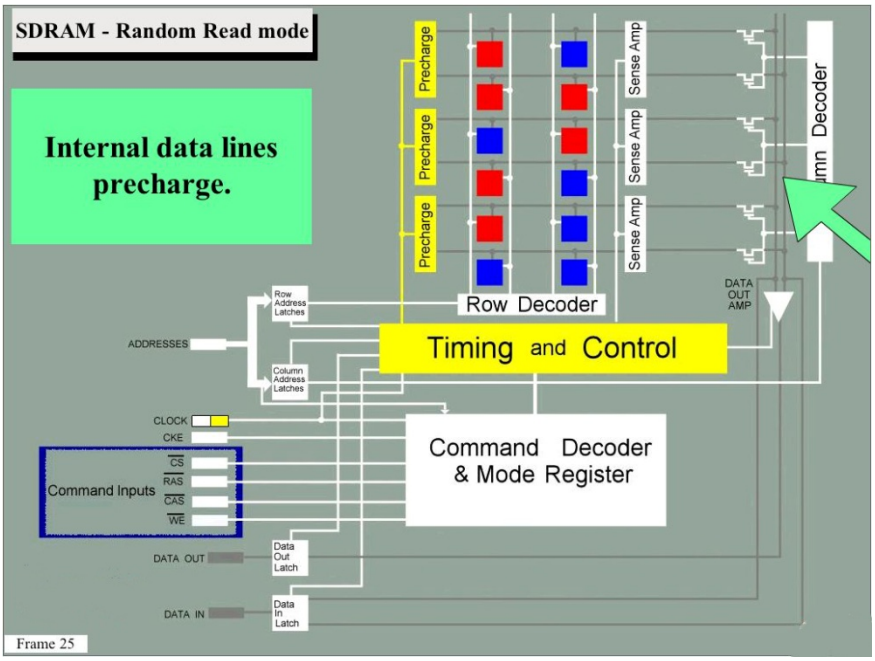
Next the bit line Sense Amps are disabled, and



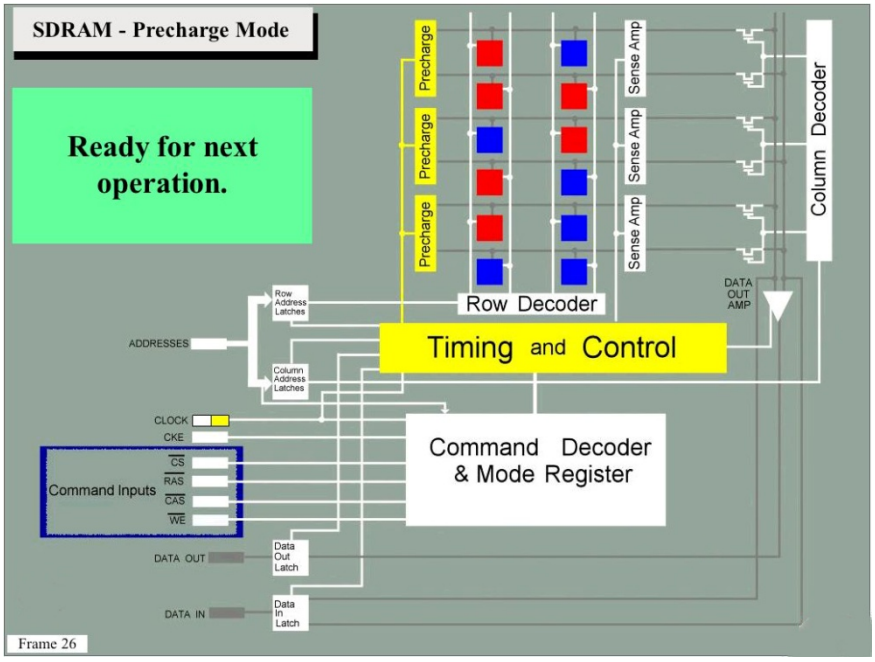
the bit line Precharge circuits are turned on.



The bit lines precharge and the column-related circuits turn off.



The internal data lines precharge.



The SDRAM is in the idle state and ready for the next operation.

H. Access Time in SDRAMs

In an SDRAM, reading data requires 2 operations, ACTIVE and READ. Therefore, 2 different access times can be defined: (1) a random access time, measured from the rising clock edge that initiates the ACTIVE operation; and (2) CAS access time, measured from the rising clock edge that initiates the READ operation.

Consider first the CAS access time. The sequence of steps above showed that the CAS access time is defined by the CAS Latency. From the Micron data sheet, "The CAS latency (CL) is the delay, in clock cycles, between the registration of a READ command and the availability of the output data."[18] Thus, the CAS access time is the CAS Latency multiplied by the clock cycle time. One big advantage of the SDRAM is that the system knows and controls the time when valid data appears on the system data bus. That time will be CL clock cycles after the READ command is initiated. With a DRAM, the time of valid data appearing is totally controlled by the speed of the specific DRAM, and can vary with the DRAM being accessed, with temperature and with power supply voltage.

Now consider the random access time, starting from the initiation of the ACTIVE operation. That is clearly the sum of (1) the time between the ACTIVE command and the READ command and (2) the CAS access time. The minimum time between ACTIVE and READ is specified by the memory vendor for each part type, as is the minimum clock cycle time.

[18] Loc. cit.

Let's look at the Micron data sheet[19] and calculate random and CAS access times.

Electrical Specifications – AC Operating Conditions

Table 11: Electrical Characteristics and Recommended AC Operating Conditions (-7E, -75)

Notes 1, 2, 4, 5, 7, and 20 apply to all parameters and conditions

Parameter		Symbol	-7E		-75		Unit	Notes
			Min	Max	Min	Max		
Access time from CLK (positive edge)	CL = 3	tAC(3)	–	5.4	–	5.4	ns	18
	CL = 2	tAC(2)	–	5.4	–	6		
Address hold time		tAH	0.8	–	0.8	–	ns	
Address setup time		tAS	1.5	–	1.5	–	ns	
CLK high-level width		tCH	2.5	–	2.5	–	ns	
CLK low-level width		tCL	2.5	–	2.5	–	ns	
Clock cycle time	CL = 3	tCK(3)	7	–	7.5	–	ns	14
	CL = 2	tCK(2)	7.5	–	10	–		
CKE hold time		tCKH	0.8	–	0.8	–	ns	
CKE setup time		tCKS	1.5	–	1.5	–	ns	21
CS#, RAS#, CAS#, WE#, DQM hold time		tCMH	0.8	–	0.8	–	ns	
CS#, RAS#, CAS#, WE#, DQM setup time		tCMS	1.5	–	1.5	–	ns	
Data-in hold time		tDH	0.8	–	0.8	–	ns	
Data-in setup time		tDS	1.5	–	1.5	–	ns	
Data-out High-Z time	CL = 3	tHZ(3)	–	5.4	–	5.4	ns	6
	CL = 2	tHZ(2)	–	5.4	–	6	ns	
Data-out Low-Z time		tLZ	1	–	1	–	ns	
Data-out hold time (load)		tOH	2.7	–	2.7	–	ns	
Data-out hold time (no load)		tOHn	1.8	–	1.8	–	ns	19
ACTIVE-to-PRECHARGE command		tRAS	37	120,000	44	120,000	ns	
ACTIVE-to-ACTIVE command period		tRC	60	–	66	–	ns	
ACTIVE-to-READ or WRITE delay		tRCD	15	–	20	–	ns	
Refresh period (8192 rows)		tREF	–	64	–	64	ms	
AUTO REFRESH period		tRFC	66	–	66	–	ns	
PRECHARGE command period		tRP	15	–	20	–	ns	
ACTIVE bank *a* to ACTIVE bank *b* command		tRRD	14	–	15	–	tCK	
Transition time		tT	0.3	1.2	0.3	1.2	ns	3
WRITE recovery time		tWR	1 CLK + 7ns	–	1 CLK + 7.5ns	–	ns	15
			14	–	15	–		16
Exit SELF REFRESH-to-ACTIVE command		tXSR	67	–	75	–	ns	12

[19] Op. cit., p. 18

First calculate CAS access time. It is simply the CAS Latency (CL) multiplied by the clock cycle time. Thus, for a CL = 2, the CAS access time is 2 x 7.5ns = 15ns for the -7E device or 2 x 10ns = 20ns for the -75 device. For CL = 3, the corresponding CAS access times are 21ns (-7E) and 22.5ns (-75). The -7E and -75 designations indicate different speed grade devices, which are sorted during final test and typically have different selling prices.

Now consider random access time. The minimum interval between ACTIVE and READ is specified as tRCD, and is 15ns for the -7E and 20ns for the -75. Adding these numbers to the CAS access times gives the random access times. The following table summarizes the results:

	-7E		-75	
	CL = 2	CL = 3	CL = 2	CL = 3
CAS Access Time	15ns	21ns	20ns	22.5ns
Random Access Time	30ns	36ns	40ns	42.5ns

I. Data Burst Operation

So far we have seen that SDRAMs provide predictable times for valid data appearing on the system data bus as compared to the unpredictability of DRAMs. But in fact, because the SDRAM has to wait for the CAS Latency to elapse, the SDRAM data appears later than that for an equivalent DRAM. That obviously does little to solve the critical problem of providing more data in less time. The answer to that problem is in the feature called Data Burst, or simply Burst.

In Burst operation, a predefined number of data bits from one row of the memory are output from the DATA OUT pin in rapid sequence. Burst operation is somewhat similar to Nibble Mode in a DRAM; except that instead of CAS/ cycling to advance the address of the next data bit in a DRAM, each low-to-high clock transition advances the address in the SDRAM. Burst lengths of 1, 2, 4, 8, or continuous are typically available for both READ and WRITE operations. Here is what the Micron data sheet says about Burst Length:[20]

[20] Op. cit., p. 38

Burst Length

Read and write accesses to the device are burst oriented, and the burst length (BL) is programmable. The burst length determines the maximum number of column locations that can be accessed for a given READ or WRITE command. Burst lengths of 1, 2, 4, 8, or continuous locations are available for both the sequential and the interleaved burst types, and a continuous page burst is available for the sequential type. The continuous page burst is used in conjunction with the BURST TERMINATE command to generate arbitrary burst lengths.

Reserved states should not be used, as unknown operation or incompatibility with future versions may result.

When a READ or WRITE command is issued, a block of columns equal to the burst length is effectively selected. All accesses for that burst take place within this block, meaning that the burst wraps within the block when a boundary is reached. The block is uniquely selected by A[8:1] when BL = 2, A[8:2] when BL = 4, and A[8:3] when BL = 8. The remaining (least significant) address bit(s) is (are) used to select the starting location within the block. Continuous page bursts wrap within the page when the boundary is reached.

J. Multiple Data Pins

Thus far we have looked primarily at memory devices with a single data pin. In practice, both DRAMs and SDRAMs have multiple data pins, and usually accommodate both Data In and Data Out on the same pin (DQ pin). In an SDRAM, each of these multiple DQ pins includes burst operation. The following table summarizes common pinouts and burst lengths:

Data Bits Handled per READ or WRITE Operation

DQ Pins	Burst Length			
	1	2	4	8
1	1	2	4	8
4	4	8	16	32
8	8	16	32	64
16	16	32	64	128

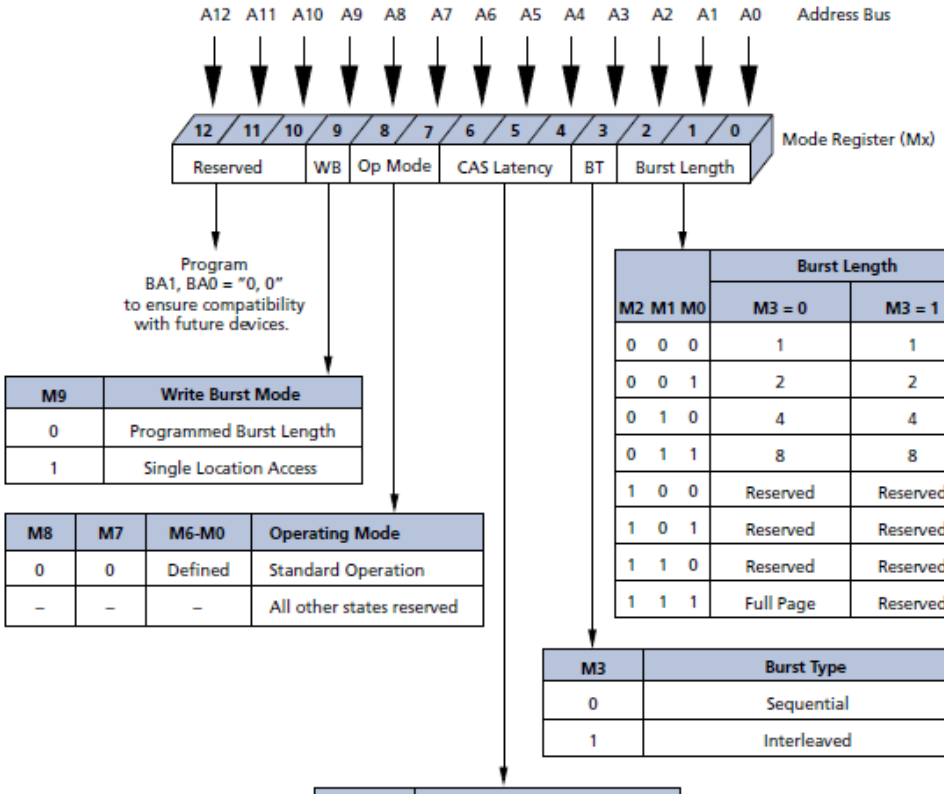
K. Multi-Bank Architecture

SDRAMs typically have 4 or more banks of memory that can be operated separately. For example, one bank can be refreshed or precharged while data is being written to or read from another bank. Thus, the penalty for having to do refresh and precharge operations is virtually eliminated. Banks can be refreshed or precharged separately or all at once.

L. Mode Register

Operating parameters of the SDRAM, such as CAS Latency and Burst Length, are provided by the user and stored on the SDRAM in a special purpose memory called the Mode Register. In the Micron 512Mb SDRAM,[21] the mode register has 12 bits, each with a specific function as defined below. Mode Register bits are loaded during a special LOAD MODE REGISTER operation when the system boots up. The information in the Mode Register can be modified later if necessary, and is retained until power is turned off.

Figure 12: Mode Register Definition



M9	Write Burst Mode
0	Programmed Burst Length
1	Single Location Access

M8	M7	M6-M0	Operating Mode
0	0	Defined	Standard Operation
–	–	–	All other states reserved

M2 M1 M0	Burst Length M3 = 0	Burst Length M3 = 1
0 0 0	1	1
0 0 1	2	2
0 1 0	4	4
0 1 1	8	8
1 0 0	Reserved	Reserved
1 0 1	Reserved	Reserved
1 1 0	Reserved	Reserved
1 1 1	Full Page	Reserved

M3	Burst Type
0	Sequential
1	Interleaved

M6 M5 M4	CAS Latency
0 0 0	Reserved
0 0 1	Reserved
0 1 0	2
0 1 1	3
1 0 0	Reserved
1 0 1	Reserved
1 1 0	Reserved
1 1 1	Reserved

[21] Op. cit., p. 37

M. Memory Standardization

DRAMs as a memory category (mostly SDRAMs of all types) were a $38B (USD) business in 2010. Five vendors supply about 90% of the DRAM market. DRAMs are indeed a commodity, and must be interchangeable in most system sockets. How did this interchangeability occur?

1958 was a pivotal year in the semiconductor industry. Jack Kilby at Texas Instruments and Bob Noyce at Fairchild separately invented the integrated circuit. And in New Jersey, the Electronic Industries Association formed an organization to develop standards in the semiconductor industry. That organization was called the Joint Electron Device Engineering Council, and is today known simply by the acronym, JEDEC. In the mid-1970's, JEDEC formed a committee specifically chartered to focus on integrated circuit memories, the JC-42 Committee.

Made up of technical personnel from memory suppliers and users, JC-42 develops pinout, package, and function standards so that products from various manufacturers can be used interchangeably in systems. As products become more complex, the standards must follow suit. For example, the most recent DRAM standard contains over 170 pages and defines specific package types and sizes, pinout, required functionality, DC parameters and AC timing requirements.

Appendix A

Mini-Glossary of DRAM Terms	
/A	Logical complement of A. If A = 1, then A/ = 0. If A = 0, then A/ = 1.
A_n	Address inputs; "Those inputs that select (address) a particular cell or set of cells within a memory array for presentation on the device outputs. The integer (n) serves to differentiate the address inputs, one from another."[22]
asynchronous	capable of receiving signals and data, and sending data, at arbitrary times independent of a system clock
bit	single unit of information
bit line (or column line or digit line)	conducting line in a memory array that carries information from a selected cell to a sense amplifier
byte	8 bits of information
capacitor	physical structure capable of storing electrical charge
CAS/	Column Address Strobe; / denotes active low; "An enable signal that on some dynamic RAMs actuates only the column oriented internal circuits and the data input/output circuits."[22]
CE/	Chip Enable; / denotes active low; "The input that, when true, permits active operation including the input and/or output of data, and when false, prevents active operation and causes the memory to be in a reduced power standby mode with the outputs floating."[22]
CK	Clock
CKE	Clock Enable; The signal that, when true, gates the clock signal into the memory chip.
column select line	The wire from the column decoder to the control terminal(s) of the transistor(s) connecting the bit lines to the I/O lines
CS/	Chip Select; / denotes active low; "The input(s) that, when any one is false, causes the device to be disabled without any significant change in the power consumption."[22]
differential	(1) responding to the difference (in voltage or current) between two signals, as opposed to the absolute level of either signal; as in "differential amplifier"
	(2) a pair of lines carrying true data and complement data; especially related to I/O lines

DQ	"The pins that serve as data output(s) when in the read mode and as data input(s) when in the write mode. When the device is not selected or enabled, the output(s) are in a floating state."[22]
equalize	cause two or more nodes or wires to have the same voltage
gate	(1) control terminal of an MOS transistor
	(2) logic circuit; e.g., Inverter, NOR, NAND
	(3) action of allowing a signal or voltage to pass from one node to another
I/O lines	Internal wires that carry (usually differential) input and output data to and from the bit lines; usually separated from the bit lines by column decoder circuitry
nibble	4 bits of information
node	a connection point in a circuit; in practice, the connection point and all wires connected to it
OE/	Output Enable; / denotes active low; "The input that, when false, disables the outputs and causes them to go to an inactive state, but that does not affect the writing function."[22]
precharge	establish a predetermined voltage on nodes and wires in a circuit
RAS/	Row Address Strobe; / denotes active low; "A chip enable signal that, on certain RAMs, actuates only row oriented internal circuitry"[22]
refresh	process of sensing, amplifying and rewriting information stored in a DRAM cell. All cells must be refreshed within a specified time interval, called the refresh interval.
synchronous	A synchronous circuit is a circuit in which events, or the performance of operations, start as a result of a signal generated by an external clock[23].
T-gate	short for transmission gate; a circuit that passes (when enabled) or blocks (when disabled) a signal from one node to another
V_{BB}	(Negative) Substrate bias voltage
V_{CC}	Positive power supply voltage

[22] from JEDEC Standard No. 21-B, December 1988

[23] from The New IEEE Standard Dictionary of Electrical and Electronics Terms, Fifth Edition; Christopher J. Booth, Editor; The Institute of Electrical and Electronics Engineers, Inc.; 345 East 47th Street; New York, NY 10017-2394; January 15, 1993; derived from definition of synchronous computer on p. 1326

VDD	Positive power supply voltage
VPP	Boosted voltage (usually above VCC)
VSS	Zero volt potential, ground
WE/	Write Enable; / denotes active low; "The input that, when true, causes the data present on the D or DQ pin(s) to be written into the address[ed] cell(s) of the device."[22]
word	16 bits of information
word line (or row line)	conducting line in a memory array that causes connection of multiple memory cells to their corresponding bit lines

