Dynamic Random Access Memory (DRAM)

Overview
Dynamic RAM (DRAM) is a computer's main volatile memory. It resides in modules inserted into the motherboard, and is most commonly referred to as RAM. See diagram to the right. It is called Dynamic because it must be continually refreshed or data will be lost.

Dynamic RAM can be contrasted with Static RAM (SRAM) which is used for cache inside or near the CPU. Both are volatile which means that they lose data when power is removed. However, SRAM is faster (and more expensive) because it doesn't need to be refreshed.

Memory Hierarchy
DRAM fits into the computer's memory hierarchy as shown in the diagram below. The top of the pyramid represents memory that is closest to the CPU and the bottom represents memory
that is further away. Registers, SRAM, and DRAM are all forms of volatile memory. Non-volatile memory, known as storage, sits at the bottom.

**Two Different Definitions of the Term 'Bank'**

There appears to be two different meanings of the word bank in DRAM literature. They are defined below.

**Definition one for bank** - The logical subdivision of a DRAM chip into (typically) four or eight addressable zones. This type of bank is divided into columns that are usually 4-, 8-, or 16 bits wide.

**Definition two for bank** - The amount of data that is read from a memory channel at one time. This type of bank is 64-bits wide.

The first type of bank is contained within a DRAM chip. The second type of bank is the parallel output of four to sixteen DRAM chips "ganged" together. This document uses bank to express both meanings. Hopefully, its definition will be clear from the immediate context.

**Organization of DRAM**

Modern DRAM comes in the form of modules or "sticks" which consists of a green circuit board containing multiple separate DRAM chips with a row of gold contacts along one edge (see diagram below). The contact edge is inserted into special slots on the motherboard. There are special clips on either side of the slot for securing the DRAM stick.
A DRAM chip is a large array of microscopic capacitors. Each capacitor (paired with a transistor) is capable of storing one bit of information (i.e., 0 or 1). A DRAM chip is logically subdivided into **banks, columns, and rows** for addressing. There are typically four or eight equally sized banks per DRAM chip (see diagram to the left). Each bank is numbered (starting at zero), and is itself divided into numbered columns and rows.

Columns in a bank are typically four-bits (x4), eight-bits (8x), or 16-bits (16x) wide. However, each row is only one bit high (see diagram below). When columns are four bits wide, it is known as x4 DRAM; when columns are eight-bits wide, it is known as x8 DRAM, etc.

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Segment of a Bank Showing Eight-Bit (x8) Addressable Cells

**Pipelining**

Pipelining is the parallel addressing of "ganged" chips on a DRAM module and the corresponding parallel output of data from each chip to fill a bank. This is possible because each DRAM chip uses the same bank, row, and column addressing scheme. With an eight chip gang, each chip contributes eight bits to the bank. With a four chip gang, each chip contributes 16 bits to the bank. With a 16 bit gang, each chip contributes 4 bits. See diagram below.
Multiplexing

Multiplexing is the performing of row addressing and column addressing in two phases to minimize the number of DRAM pins required. This allows one set of pins to perform both functions, thereby decreasing cost. However, this technique also increases latency. Here are the steps taken in DRAM addressing when multiplexing is used:

1. The DRAM bank is addressed using a dedicated set of wires for that purpose (typically three).
2. The bank row is addresses and latched.
3. The bank column is addressed and latched.

The multiplexing process was invented by the Mostek DRAM manufacturing company in 1973.

Following is an example of how a 1 GB DRAM module might be organized and multiplexed:

The 1GB module contains eight separate 1 Gbit DRAM chips - each providing 128 MB of storage. Each of the 128 MB chips is divided internally into eight banks of $2^{27} = 128$ Mbits. Each bank contains $2^{14} = 16384$ rows of $2^{13} = 8192$ bits each. The 8192 bits per row are divided into 128 8-bit columns. One byte of memory (from each chip; 64 bits total from the whole DIMM) is accessed by supplying a 3-bit bank number, a 14-bit row address, and a 10-bit column address to each chip.

Legacy DRAM

The table below provides a summary of legacy DRAM. Each entry is explained in further detail below.
Original Dram

The first successful integrated-circuit DRAM was the Intel 1103 chip. It was introduced in 1970 and could store 1 Kbit of data. The 1103 was packaged in an 18 pin dual-inline pin (DIP) chip that was manually inserted into parent circuit boards.

Fast Page Module (FPM)

In addition to pipelining and multiplexing, FPM added the following features to DRAM:

- **Use of Single Inline Memory Modules (SIMMs)** - DRAM chips are packaged in 32 or 72 SIMMs to make it easier to install. FPM DRAM used SIMMS with 30 or 72 contacts to accommodate 16 or 32 bit data busses.
- **Burst Mode** - In burst mode, four consecutive reads are performed to take advantage of the fact that most memory reads are done to sequential locations. The setup overhead (latency) is paid for on the first read, so the follow-on reads can be done in fewer cycles. With FPM DRAM, for example, the first read take 5 cycles, but the remaining three reads take only take 3 cycles each. This is known as a 5-3-3-3 burst mode, which saves a total of 6 cycles over non-burst mode DRAM.
- **Interleaving** - Two separate banks of memory are used together - alternating access from one to the other as even and odd bytes. That way, when one is accesses the other is being pre-charged. This meant that two identical banks of memory must be installed at a time.

Most 386, 486, and Pentium PCs sold between 1987 and 1995 used FPM memory.

Extended Data Out (EDO)

EDO RAM made the following improvements over FPM DRAM:

- Data output circuitry is not turned off between boost mode cycles which saves 10 ns per cycle. This allows burst mode cycling of 5-2-2-2 for EDO RAM compared to the 5-3-3-3 of FPM.
- Interleaving is eliminated which means that EDO RAM modules do not have to be installed in pairs.

EDO RAM was limited to a 66 MHZ front-side bus speed, which led to its fading from the marketplace starting in 1998 when 100 MHz bus speeds were introduced.

Modern DRAM

The table below provides a summary of modern DRAM. Each entry is explained in further detail below.

<table>
<thead>
<tr>
<th>Gen</th>
<th>Era</th>
<th>Type</th>
<th>Package</th>
<th>Pins</th>
<th>Synch</th>
<th>Data Rate</th>
<th>Channels</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>1996-2000-</td>
<td>Synchronous DRAM (SDRAM)</td>
<td>DIMM</td>
<td>168</td>
<td>Yes</td>
<td>single</td>
<td>1</td>
</tr>
</tbody>
</table>
Synchronous DRAM (SDRAM)

Synchronous DRAM gets its name from the fact that it is synchronized with the system clock. It made its debut in 1996 with the Pentium processor which had a 64 bit address bus. It is often called Single Data Rate (SDR) SDRAM to distinguish it from Double Data Rate SDRAM that was released in 2002. Three features made SDRAM a significant improvement over previous DRAM:

- 64 bits of data could be processed at a time versus 32 bits. This means that one stick would fill a complete 64 bit bank. Two sticks of older asynchronous 32 bit RAM were required to fill a bank on a 64 bit address bus.
- It was tied to the system clock which allowed instructions from the memory controller to be pipelined. Pipelining means that the chip could accept a new instruction before it has finished processing the previous one.
- It was able to interleave operations to multiple banks of memory, thereby increasing effective bandwidth. However, it would still work if only one bank was filled.

SDRAM is packaged in dual inline memory modules (DIMMs). The main difference between SIMMs and DIMMs is that DIMMs have separate electrical contacts on each side of the module, while the contacts on SIMMs on both sides are redundant. Another difference is that standard SIMMs have a 32-bit data path, while standard DIMMs have a 64-bit data path. Since Intel's Pentium has (as do several other processors) a 64-bit bus width, it requires SIMMs installed in matched pairs in order to complete the data bus. The processor would then access the two SIMMs simultaneously. DIMMs were introduced to eliminate this practice.

SDRAM DIMMs came in a variety of pin sizes as shown in the table below:

<table>
<thead>
<tr>
<th>Pin Size</th>
<th>Type</th>
<th>Application</th>
</tr>
</thead>
<tbody>
<tr>
<td>168 pins</td>
<td>Standard</td>
<td>Desktops</td>
</tr>
<tr>
<td>68, 144, or 172 pins</td>
<td>micro-DIMM</td>
<td>Laptops</td>
</tr>
<tr>
<td>72, 144, 200 pins</td>
<td>Small Outline DIMM (SO-DIMM)</td>
<td>Laptop and small form factor desktops</td>
</tr>
</tbody>
</table>

Since DIMM sticks are 64 bits wide, each slot is called a bank. DIMM sticks do not need to be installed in pairs; however, performance is improved when they are.

The most popular forms of SDR SDRAM came in 66, 100, and 133 MHz speeds, and were labeled PC66, PC100, and PC133, respectively. The table below provides a summary of these three types of SDRAM.
**Rambus DRAM (RDRAM)**

RDRAM was created by the Rambus Corporation and came in sticks called RIMMS. It was the first SDRAM to incorporate double pumping technology that was later used in Double Data Rate (DDR) SDRAM. Like Fast Page Module DRAM, RDRAM uses interleaving. This means that modules must be installed in pairs. All RIMM slots on a motherboard must be filled with a RIMM or a CRIMM. A CRIMM, or continuity RIMM, is simply a stick with no memory chips on it.

Due to cost, licensing issues, and technology setbacks, RDRAM saw limited market penetration. It was somewhat popular between 1999 and 2003 and can still be purchased today. The table below summarizes RDRAM module speeds and transfer rates:

<table>
<thead>
<tr>
<th>Marketing Name</th>
<th>Chip Type</th>
<th>Clock Speed (MHz)</th>
<th>Cycles per Clock</th>
<th>Bus Speed (MTps)</th>
<th>Bus Width (Bytes)</th>
<th>Transfer Rate (MBps)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PC66</td>
<td>10ns</td>
<td>66</td>
<td>1</td>
<td>66</td>
<td>8</td>
<td>533</td>
</tr>
<tr>
<td>PC100</td>
<td>8ns</td>
<td>100</td>
<td>1</td>
<td>100</td>
<td>8</td>
<td>800</td>
</tr>
<tr>
<td>PC133</td>
<td>7ns</td>
<td>133</td>
<td>1</td>
<td>133</td>
<td>8</td>
<td>1,066</td>
</tr>
</tbody>
</table>

**Double Data Rate (DDR) SDRAM**

DDR SDRAM made its debut in high-end graphics cards around 2000 and in memory modules around 2002. DDR SDRAM is inherently twice as fast as regular SDRAM because, like RDRAM, it is "double pumped." Double pumped means that DDR SDRAM outputs data twice a clock cycle whereas SDR SDRAM outputs data once a clock cycle - as depicted in the diagram below.

![Diagram of SDR Output - Once a Cycle](image1)

![Diagram of DDR Output - Twice a Cycle](image2)

The DDR modules use uses 184 pin DIMMs for desktops, and 200 or 172 pin SO-DIMMs for laptops. The DDR module has a unique notch configuration that distinguishes it from other memory types.
DDR DRAM has a "DDR" nomenclature (e.g., DDR200) and a "PC" nomenclature (e.g., PC1600). DDR200, in fact, is another name for PC1600. The "DDR" nomenclature reflects the effective memory data frequency (in MHz), and is given by multiplying the system clock speed (MHz) by 2. A multiplier of 2 is used because DDR memory moves data twice a cycle. A DDR module with a 200 MHz system clock, for example, is DDR400. Following is formula for calculating the DDR nomenclature:

\[
\text{DDR nomenclature} = \text{system clock speed (MHz)} \times 2 \text{ cycles per clock}
\]

The "PC" nomenclature or rating reflects the overall data throughput (in MBps), and is given by multiplying the data frequency (in MHz) by 8 bytes. The 8 byte multiplier is used because all DDR modules transmit 64 bits at a time. A DDR module with a 200 MHz system clock, for example, is PC3200. Following is a formula for calculating the PC nomenclature:

\[
\text{PC rating} = \text{system clock speed (MHz)} \times 2 \text{ cycles per clock} \times 8 \text{ bytes}
\]

These formulas for calculating DDR nomenclature and PC rating are the same for all DDR memory, including DDR2 and DDR3 memory. The table below summarizes DDR module specifications.

<table>
<thead>
<tr>
<th>Marketing Name</th>
<th>Chip Type</th>
<th>Clock Speed (MHz)</th>
<th>Cycles per Clock</th>
<th>Bus Speed (MTps)</th>
<th>Bus Width (Bytes)</th>
<th>Transfer Rate (MBps)</th>
<th>Dual-Channel Rate (MBps)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PC1600</td>
<td>DDR200</td>
<td>100</td>
<td>2</td>
<td>200</td>
<td>8</td>
<td>1,600</td>
<td>3,200</td>
</tr>
<tr>
<td>PC2100</td>
<td>DDR226</td>
<td>133</td>
<td>2</td>
<td>266</td>
<td>8</td>
<td>2,133</td>
<td>4,266</td>
</tr>
<tr>
<td>PC2700</td>
<td>DDR333</td>
<td>166</td>
<td>2</td>
<td>333</td>
<td>8</td>
<td>2,667</td>
<td>5,333</td>
</tr>
<tr>
<td>PC3200</td>
<td>DDR400</td>
<td>200</td>
<td>2</td>
<td>400</td>
<td>8</td>
<td>3,200</td>
<td>6,400</td>
</tr>
</tbody>
</table>

**Dual Channel Technology**

Dual channel technology uses a dual channel memory bus which theoretically doubles the memory speed. When double pumping and dual channels are used together, memory speed is quadrupled. This combination is often used with quad pumped front side busses. DDR2 uses an improved tri-channel configuration. With either dual- or tri-channel technology, each channel must be connected to a separate stick of memory for best performance.

**DDR2 SDRAM**

Systems implementing the follow-on technology to DDR, called DDR2 began to appear in mid-2004. DDR2 achieves speeds beyond that of DDR, delivering bandwidth of up to 8.5 GB per second. Frequently, DDR2 based systems can use memory installed in pairs to run in "dual channel mode" to increase memory throughput even further. The table below summarizes speeds and transfer rates for standard DDR2 modules:
### DDR3 SDRAM

The latest generation of memory technology, DDR3, began to appear in systems in late 2007. DDR3 is an evolutionary step beyond DDR2 and operates at lower voltages, thereby consuming less power. Typically, DDR3 based systems can address memory modules in banks of 1, 2 or 3. If a system supports it, installing memory in matched sets of 2 or 3 modules (dual channel or triple channel modes) will deliver greatly increased memory performance over running a single memory module by itself. The table below summarizes speeds and transfer rates for standard DDR3 modules:

<table>
<thead>
<tr>
<th>Marketing Name</th>
<th>Chip Type</th>
<th>Clock Speed (MHz)</th>
<th>Cycles per Clock</th>
<th>Bus Speed (MTps)</th>
<th>Bus Width (Bytes)</th>
<th>Transfer Rate (MBps)</th>
<th>Dual/Channel Rate (MBps)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PC3-6200</td>
<td>DDR3-800</td>
<td>400</td>
<td>2</td>
<td>800</td>
<td>8</td>
<td>6,400</td>
<td>12,800/19,200</td>
</tr>
<tr>
<td>PC3-8500</td>
<td>DDR3-1066</td>
<td>533</td>
<td>2</td>
<td>1,066</td>
<td>8</td>
<td>8,533</td>
<td>17,066/25,600</td>
</tr>
<tr>
<td>PC3-10600</td>
<td>DDR3-1333</td>
<td>667</td>
<td>2</td>
<td>1,333</td>
<td>8</td>
<td>10,667</td>
<td>21,333/32,000</td>
</tr>
<tr>
<td>PC3-12800</td>
<td>DDR3-1600</td>
<td>800</td>
<td>2</td>
<td>1,600</td>
<td>8</td>
<td>12,800</td>
<td>25,600/38,400</td>
</tr>
</tbody>
</table>

As shown in the diagram to the left, DDR, DDR2, and DDR3 DIMMS are keyed differently because the three standards are incompatible.
DRAM Ranks

A rank is an independent set of DRAM chips on a module that is addressed independently as if it were another DIMM. Dram ranking is most often used with servers as a means of increasing DRAM capacity without needing more DIMM socket.

A rank is an independent set of dram chips that is 64 bits wide - except on memory that supports Error Conrrection Code (ECC). When ECC is supported, a rank is 72 bits wide. A stick (module) of memory that has DRAM chips on one side can only have one rank. A module that has chips on both sides can have one, two, or four ranks, as outlined in the table below.

<table>
<thead>
<tr>
<th>Module Type</th>
<th>Ranks</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single Sided</td>
<td>One 64-bit wide data area</td>
</tr>
<tr>
<td>Double-Sided</td>
<td>One 64-bit wide data area (1/2 on each side)</td>
</tr>
<tr>
<td>Double-Sided</td>
<td>Two 64-bit wide data areas (One on each side)</td>
</tr>
<tr>
<td>Double-Sided</td>
<td>Four 64-bit wide data areas (Two on each side)</td>
</tr>
</tbody>
</table>

Latency

Latency is the delay between when a request is made and when it is answered - it plays an important role in memory timing. Following are the general step taken when memory is access and the most important latencies that occur.

1. The CPU sends memory request to the memory controller over an address bus.
2. The memory controller receives the address and divides it up into high-order bits and low order bits. The high-order bits provide the column address, and the low order bits provide the row address.
3. The memory controller accesses the appropriate row by activating a Row Access Stobe (RAS) at the address given by the low order bits. A few cycles of delay (latency), known as RAS Precharge (tRP), occurs while the RAS stabilizes.
4. The memory controller accesses the appropriate column by activating the Column Access Strobe (CAS) high-order bits. The delay of a few clock cycles occurs between RAS access and CAS access - known as RAS-to-CAS delay (tRCD).
5. The data is then output back to the memory controller. The delay of a few clock cycles between when the CAS is activated and data is output is called CAS latency (tCL).
6. A delay known as Cycle Time (tRAS) occurs if the RAS has to be moved to a new bank.

The following table provides a summary of the most important types of memory latency.

<table>
<thead>
<tr>
<th>Latency Type</th>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>CAS Latency</td>
<td>CL, tCL, or tCAS</td>
<td>CAS Latency is the delay between the CAS signal and the availability of data on the data pins.</td>
</tr>
<tr>
<td>RAS to CAS Delay</td>
<td>tRCD</td>
<td>The time required between RAS and CAS access</td>
</tr>
<tr>
<td>RAS Precharge</td>
<td>tRP</td>
<td>The time required to switch from one row to the next row in the same bank.</td>
</tr>
<tr>
<td>Cycle Time</td>
<td>tRAS</td>
<td>The time required to switch from one row to the next row in another bank.</td>
</tr>
</tbody>
</table>
Memory Errors

There are two types of errors that occur with memory: **hard fails and soft errors**. A hard fail occurs when there is some kind of physical defect in the memory. These are errors that are consistent and are normally fixed by replacing the faulty module. A soft error is one that occurs due to a transient condition such as:

- Cosmic Rays
- Power fluctuations or noise
- Incorrect memory type
- Incorrect timing settings
- Electromagnetic interference (EMI)/ Radio frequency interference (RFI)
- Static discharges
- Heat buildup

Soft errors are intermittently and normally do not cause permanent damage. There are two basic techniques for mitigating the effects of soft errors: **Parity checking**, and **error correcting code (ECC)**. These techniques must be enabled in the motherboard chipset before they can be used - they and are normally only found on high-end motherboards.

**Parity Checking**

Parity checking requires that an extra bit of memory be used with each eight bits of data to track whether the sum of the bits is odd or even. It often requires that an extra chip be used on the module. The way that parity works is by summing the ones in the byte of data. If the sum is even, then the parity bit is set to one. If the sum is odd, then the parity bit is set to zero. This describes odd parity. (With even parity, the parity bit is set to zero if the sum is even.) Before the byte of data is used, the parity bit is recalculated and compared to the existing value. If they do not match, then a parity error is thrown by the BIOS. Parity checking only detects a single-bit error. Studies have shown that 98% of memory errors are single bit.

When adding or replacing memory, it is important to determine whether it requires error detection or correction capability. This can be determined by checking the motherboard manual or BIOS settings. Memory that uses parity checking normally cost about 10 - 15% more than regular memory.

**Error Correction Code (ECC)**

ECC not only detects errors, but can correct single-bit errors. ECC is a good choice for servers and mission critical workstations, but is always slower than non-ECC memory because of the overhead.

**Buffering**

A technique used primarily with high-end motherboards to compensate for the interference caused by having more than four DIMMs.