

Departamento de Arquitectura Instituto de Computación
Universidad de la República
Montevideo - Uruguay

Notas de Teórico

Memorias RAM

Arquitectura de Computadoras
(Versión 4.3b - 2016)

9 MEMORIAS RAM

9.1 Introducción

Una característica distintiva de los flip-flops es su capacidad de recordar el valor anterior de una entrada, o lo que es lo mismo: tienen memoria.

Esta capacidad de memoria es fundamental no solo para la construcción de circuitos secuenciales, sino que también es la base para el almacenamiento de la información a ser procesada en las computadoras en la forma en que éstas están organizadas en la actualidad (arquitecturas de programa almacenado), como se verá más adelante.

La base de la memoria es el flip-flop que es capaz, como vimos, de almacenar un bit. Lo normal es utilizar conjuntos de bits, organizados en **palabras** de **n** bits. Cuando la memoria es de una sola palabra recibe el nombre de **Registro** (también puede recibir el nombre de **Latch**) cuando tiene múltiples recibe la denominación genérica de **RAM**.

En general una memoria RAM es un array de **2^m palabras** de **n** bits, de forma similar a lo que ya analizamos para las ROMs, salvo que en esta oportunidad se comportan como un arreglo (array) tal como los conocemos en las materias de programación (si escribo un valor en una posición y leo esa posición el valor que leo es el que escribí).

El nombre RAM es una sigla que corresponde a Random Access Memory. Este nombre no es del todo feliz y seguramente fue elegido históricamente por tres causas:

- para diferenciar estos dispositivos de las primeras memorias que funcionaban de modo "secuencial" (para leer una posición había que leer previamente todas las anteriores). Lo de Random estaría apuntando al tipo de acceso "directo" a la posición requerida.
- para que se pareciera a la sigla de la "memoria" ROM, lo que seguramente llevó a Random (por la R) en vez de Direct, que sería lo correcto. Pensemos que nadie querría almacenar información en un dispositivo que la devolvería en forma aleatoria, por lo que definitivamente las memorias RAM **no** son Random.
- por esa manía existente en la industria informática de nombrar a las cosas utilizando tres letras.

Nota: en francés a las memorias RAM se las llama "memoire vivre" (memoria viva) y a las ROM se las llama "memoire morte" (memoria muerta), términos que parecen mucho más apropiados a las características de estos dispositivos.

9.2 Tipos de Memoria

Hay distintos tipos de organizaciones de memorias del tipo RAM. Los elementos que determinan estas variantes son:

- si los datos de entrada (lo que se va a escribir) y de salida (lo que se lee) comparten los mismos caminos físicos o están separados.
- si tienen más de una única palabra de n bits.
- si requieren de circuitos de "refresco" (veremos qué significa esto más adelante).

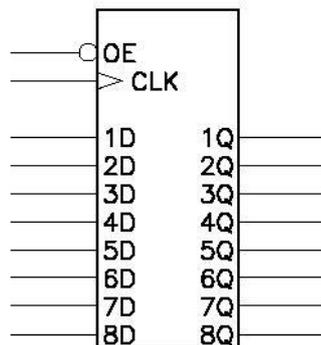
Cuando se tiene una memoria de una única palabra, con entradas y salidas diferenciadas se habla de **Latch** o **Registro**, cuando se tienen múltiples palabras y no se requiere de circuitos de "refresco", se habla en general de **SRAM**, por **Static RAM** (otro nombre para la polémica, ¿cómo algo que puede ser modificado recibe el nombre de "estático"?). Lo normal son las SRAM que tienen la entrada y salida coincidentes en las conexiones, mientras que las de múltiples palabras con entradas y salidas diferenciadas, si

bien existen, están reservadas a aplicaciones específicas (ej: memorias para tarjetas de video). Finalmente las memorias de multiples palabras que necesitan circuito de refresco, reciben el nombre de **DRAM** por **Dynamic RAM** (no existen memorias tipo Registro que sean a su vez Dinámicas).

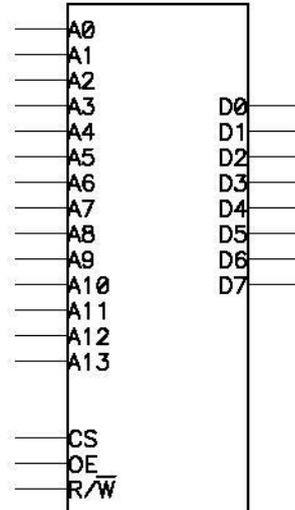
Nota: la denominación "Static" se comenzó a utilizar para diferenciar a las memorias clásicas a partir del momento en que aparecieron las memorias "Dynamic", término que hace referencia a su necesidad de contar con un circuito de "refresh" (refresco).

Los símbolos de los distintos tipos de memoria son:

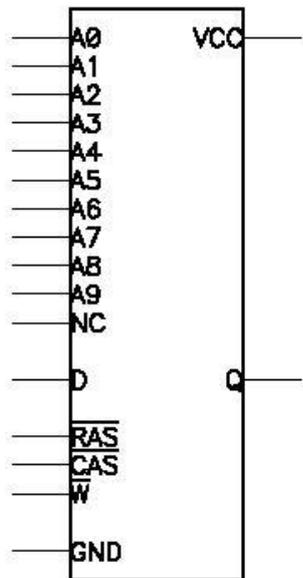
Registro (ó Latch):



Static RAM:



Dynamic RAM:



9.3 Circuito Interno

Los distintos tipos de memoria están implementados por circuitos característicos.

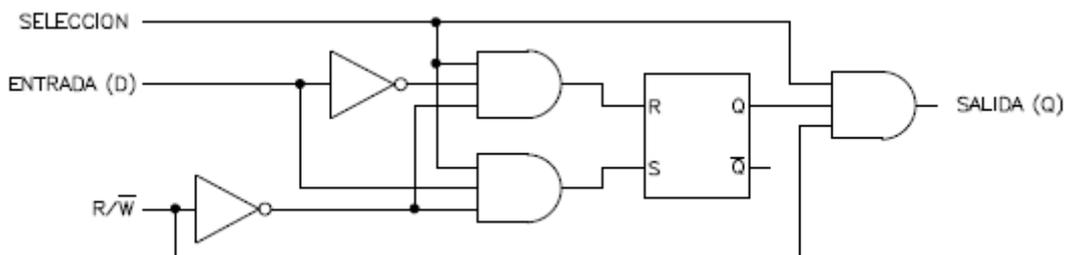
Registro ó Latch:

Tienen el circuito mas simple. Son un conjunto de flip-flops tipo D, sincrónicos (flanco ascendente) conectados con su entrada de reloj en común y un control de tercer estado, también común, en sus salidas.

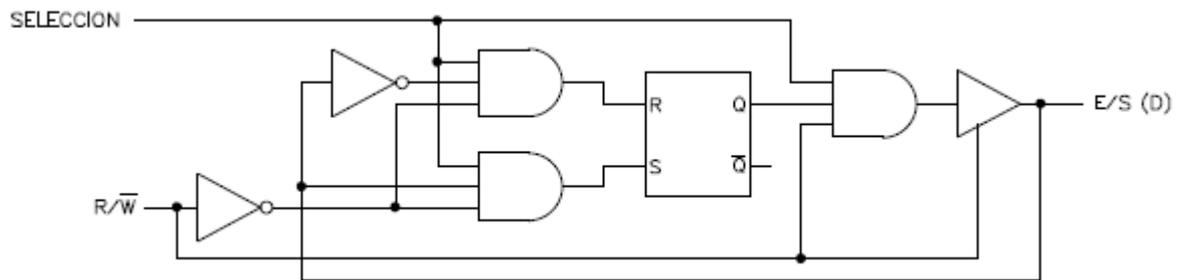
Static RAM (SRAM):

Están construidas en base a flip-flops tipo D, los cuales tienen sus entradas de habilitación controladas por un circuito de selección en base a la decodificación de las entradas de dirección (Am).

Veamos como es el circuito asociado al bit D_{ij} (bit j -ésimo de los n de la palabra i -ésima de las 2^m palabras de la memoria), para una memoria SRAM con entrada y salida independiente:

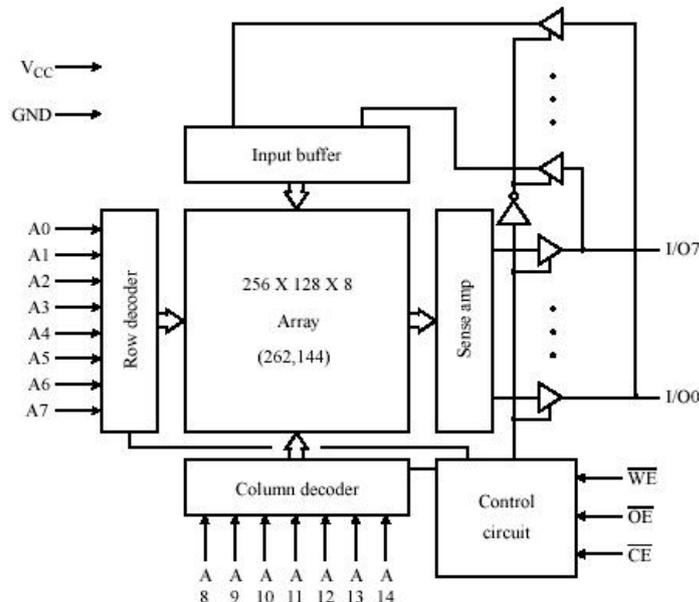


En el caso que fuera una memoria SRAM con entrada y salida combinadas, el circuito del elemento de memoria sería de la forma:



El circuito de selección está formado por un decodificador y lógica de control asociada (para hacer intervenir el CS). A su vez a la salida se requiere de lógica (compuertas OR) y de buffers de tercer estado para implementar el OE general de la memoria.

El siguiente es un diagrama de bloques de un circuito SRAM real (el chip AS7C3256A de 32 K x 8 de Alliance Semiconductor):



Dynamic RAM (DRAM):

Aunque en la práctica se realiza una optimización a nivel del circuito con transistores que implementa el circuito lógico mostrado para el elemento de memoria, se requieren al menos 6 transistores para construir cada bit de una memoria del tipo SRAM. Dado que en cada momento de la evolución de la tecnología la cantidad de transistores que se pueden colocar en un mismo chip está limitada por la tecnología disponible en ese momento, es un dato muy importante el número de transistores que se requieren para cada bit, ya que una disminución en ese número tiene un inmediato impacto en la capacidad del chip de memoria obtenido.

A este punto se dirige el diseño de las DRAM. Estas memorias utilizan una propiedad de los transistores: la existencia de la junta entre la base y el emisor (o entre el gate y el source) produce como efecto lateral la formación de un condensador (capacidad parásita).

Normalmente la existencia de este condensador es un efecto no deseado, ya que la acumulación de carga eléctrica que se produce en él perjudica la velocidad de conmutación del transistor (ya que para cambiar su estado de conducción hay que desalojar la carga acumulada en este condensador parásito y eso lleva su tiempo). Sin embargo los inventores de las DRAM utilizaron esa propiedad para "guardar" un bit en forma de carga acumulada en ese capacitor "parásito".

Esto lleva a que en este tipo de memorias se requiera solamente un transistor por cada bit a almacenar. Este hecho ha resultado determinante para el éxito de este diseño, ya que permite hasta 6 veces mayores capacidades de almacenamiento para cada estadio tecnológico de la fabricación de chips y ha disimulado las notorias complicaciones de la circuitería auxiliar requerida por estas memorias para funcionar.

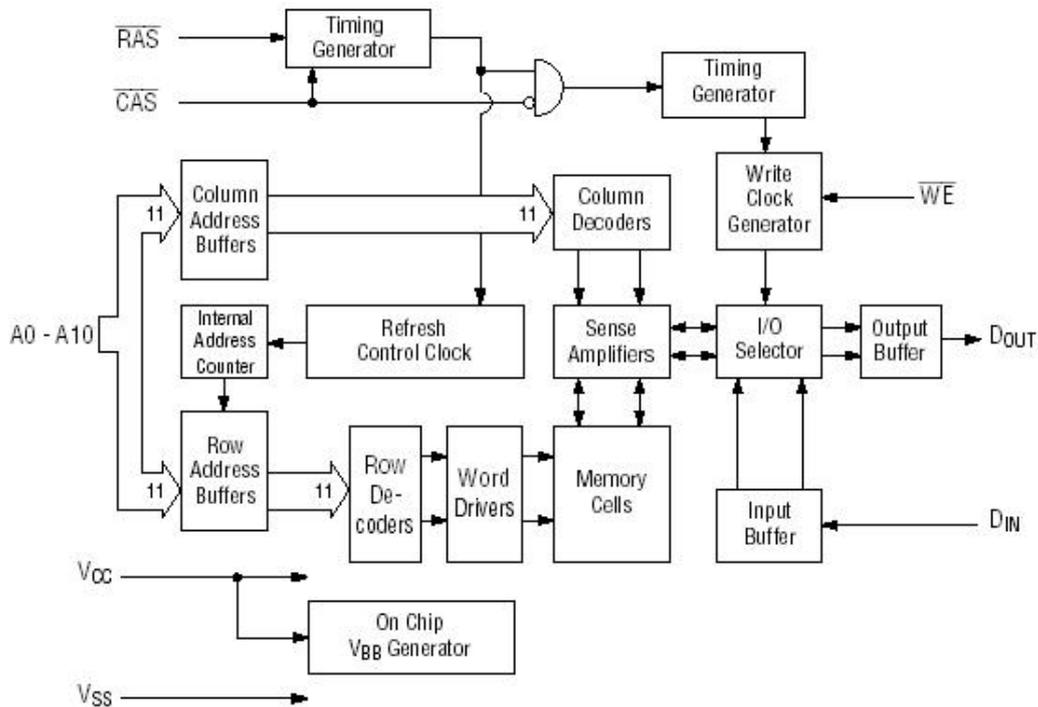
La particular forma de almacenar la información le dan ciertas características propias a estos dispositivos.

Por un lado la lectura es destructiva: al "leer" se quita la carga almacenada lo que genera una corriente por el transistor que es lo que en definitiva se detecta. La forma de mantener la información que contenía la memoria es grabando nuevamente el mismo valor. Por suerte para los diseñadores de sistemas que quieran utilizar estas memorias el propio circuito interno de los chips de este tipo se encarga de este trabajo (cada vez que se lee se hace simultáneamente una operación de escritura).

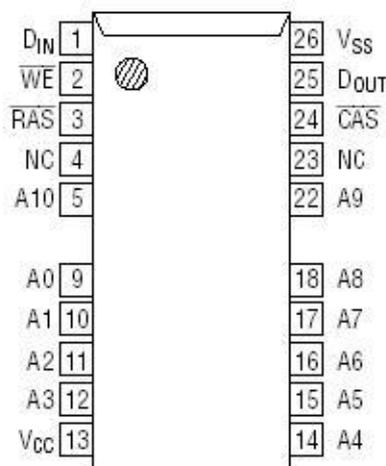
Por otro lado la carga almacenada se va perdiendo con el tiempo. Dado que ningún dispositivo es perfectamente no conductor, siempre existen las denominadas "corrientes de fuga" que hacen que luego de un cierto período de tiempo la carga acumulada desaparezca (y con ella la información almacenada). Para remediar esto las memorias deben ser "refrescadas", es decir se debe periódicamente grabar nuevamente la información para que ésta no se pierda. Dada la naturaleza del procedimiento de lectura de la memoria, esta "regrabación" se logra simplemente leyendo la memoria. Esta lectura periódica debe ser realizada con un período menor al tiempo de pérdida de la carga (tiempo de descarga del "capacitor parásito" que almacena la información).

Teniendo como objetivo primario minimizar el tiempo requerido para refrescar un chip completo de DRAM, y de paso disminuir la cantidad de "patitas" necesarias para el circuito integrado (en el pasado este fue un problema técnico a resolver, dado que los primeros encapsulamientos disponibles tenían un número acotado de pines), estas memorias están organizadas en forma matricial y la dirección se presenta separada en fila (row) y columna (column).

El siguiente es un diagrama de bloques de una DRAM de 4 Mbits (en particular la MSM514100D de OKI Semiconductor):



El circuito integrado tiene el siguiente "pinout" (disposición de las entradas y salidas en las conexiones del encapsulado):



Como vemos tanto en el diagrama de bloques como en el pinout, esta DRAM tiene 11 bits de entrada de dirección. Es claro que con 11 bits de dirección no es posible direccionar directamente los 4 Mbits que dispone la memoria (para ello se requerirían el doble, o sea 22).

El punto es que, como se dijo, estos chips no reciben la información de dirección toda junta sino que reciben primero la información de la fila y luego la información de columna, usando los mismos bits de entrada de dirección en momentos distintos. Para ello poseen los registros internos RAB (Row Address Buffer), también denominado RAR (Row Address Register) y CAB (Column Address Buffer), también denominado CAR (Column Address Register), los cuales son cargados con la dirección presente en los bits A_n en el flanco de las señales RAS (Row Address Strobe) y CAS (Column Address Strobe).

La gran ventaja de este diseño es que manteniendo un valor de columna (CAB) incambiado y cambiando sucesivamente todos los valores de fila (RAB) se consigue refrescar toda la memoria. Esto es porque cada vez que se accede a una fila en realidad se están "leyendo" todos los bits de la fila simultáneamente (luego el circuito selecciona el correcto usando la información de columna). Con esto se reduce drásticamente el tiempo necesario para hacer el "refresh". Para m bits de dirección en vez de hacer 2^m lecturas hacemos $2^{m/2}$. Para el ejemplo donde $2^m = 11$, resulta que en lugar de hacer 2^{22} (4.194.304) hacemos solamente 2^{11} (2.048)

En el caso particular del chip mostrado se dispone de una facilidad adicional de RAS-refresh, en el cual alcanza con activar el RAS para que el circuito interno incremente automáticamente la fila de forma de ir haciendo el refresh.

9.4 Parámetros

En las memorias RAM hay distintos aspectos que caracterizan el comportamiento de la misma, su aplicabilidad a una función específica y la posibilidad de ser insertada en determinado circuito. Los más importantes son:

- capacidad (en bits)
- organización (cuantas palabras de qué tamaño)
- tiempo de acceso en lectura (cuanto tiempo pasa desde que se le presenta una dirección y las señales de control que correspondan hasta que presenta el contenido en la salida)
- tiempo de acceso de escritura (cuanto tiempo pasa desde que se le presenta una dirección, el dato a escribir y las señales de control que correspondan hasta que la memoria adopta el nuevo valor y está en condiciones de iniciar un nuevo ciclo de escritura)
- tipo de tecnología (estática ó dinámica, en definitiva si requieren o no de un circuito de fresco)
- capacidad de retener la información sin estar alimentadas de energía

Los tiempos de acceso en algunos casos son presentados en base a la frecuencia de reloj a la que pueden trabajar. La frecuencia de reloj está muy ligada a la velocidad de transferencia de información que puede tener una memoria sincrónica (aunque, como veremos, no siempre la relación es lineal ya que juegan otros factores arquitectónicos de la solución).

9.5 Variantes Tecnológicas

A continuación veremos algunas variantes tecnológicas que existen de los tipos de memoria básicos vistos.

9.5.1 SRAM

Es la tradicional memoria basada en flip-flops ya mencionada. Tiene menores tiempos de acceso (en general) que la memoria basada en tecnología dinámica. Tiene el inconveniente de ocupar mas espacio (requiere más transistores) y ser más cara.

9.5.2 DRAM

Es la memoria dinámica original. Hoy en día ya no se construyen chips DRAM "clásicos", ya que hay sido superados por distintas mejoras al diseño. Son más económicas que las SRAM, pero más lentas, en particular por el uso de la técnica de direccionamiento matricial en base al CAS y al RAS (columna y fila por separado).

9.5.3 FPM DRAM

Una de las primeras mejoras al diseño DRAM fue el denominado **Fast Page Mode**, donde se optimizaba al acceso, permitiendo múltiples CAS para el mismo RAS. De esta forma se mejoraba sensiblemente el tiempo de acceso. El nombre se debe a que cada fila se puede asociar a una página de un libro: una vez que encontré una "palabra" (un bit) en una "pagina" (fila del chip) las siguientes palabras (bits) las encontraré más rápido si están en la misma página (fila).

9.5.4 EDO DRAM

Las DRAM **Enhanced Data Out** son una mejora sobre las FPM, consistente en acelerar el acceso al próximo bit, iniciando una lectura por adelantado al bit contiguo al accedido en un momento dado. De esta forma se ahorra la espera por la circuitería interna de detección de la carga de almacenamiento.

9.5.5 BEDO DRAM

Las **Burst EDO** son una mejora, relativamente menor, sobre las EDO, para cuando las lecturas son en modo de ráfaga (burst).

9.5.6 SDRAM

Las **Synchronous DRAM** utilizan un reloj para marcar los tiempos de los ciclos de lectura o escritura y mantener en sincronismo la memoria con el resto del sistema (en particular con la CPU). Este sincronismo le permite mejorar los tiempos de acceso respecto a las memorias EDO. En la actualidad todas las memorias DRAM son del tipo sincrónico.

Las SDRAM se sub-clasifican en función de la frecuencia del reloj para la que están diseñadas, para lo que se utiliza una clasificación propuesta por Intel en su especificación del computador tipo PC:

PC-66: reloj de 66 MHz

PC-100: reloj de 100 MHz

PC-133: reloj de 133 MHz

9.5.7 RDRAM

Las **Rambus DRAM** son la propuesta de un fabricante para diseños de DRAM de alta frecuencia (y por lo tanto altas velocidades de transferencia). Los chips de memoria están implementados de forma de utilizar ambos flancos del reloj a los efectos de sincronizar las transferencias. Las frecuencias de reloj disponibles van desde 300 MHz (equivalentes a 600 MHz), 350 MHz (equivalentes a 700 MHz), 400 MHz (equivalentes a 800 MHz), 533 MHz (equivalente a 1066 MHz) hasta 600 MHz (equivalente a 1200 MHz).

La tecnología que utilizan estas memorias fue diseñada por la empresa Rambus que licenció sus patentes a distintos fabricantes, entre ellos Intel. Los problemas legales y los juicios que inició contra distintos fabricantes de memorias (reclamando que la tecnología DDR desarrollada mas tarde por otros "copia" conceptos de su diseño y por tanto infringen

sus patentes) son tan o incluso más conocidos que sus productos.

9.5.8 DDR SDRAM

Estas memorias utilizan ambos flancos del reloj para realizar las operaciones, de allí que reciben el nombre de **Double Data Rate** (Transferencia de Datos Doble). Son un desarrollo estándar realizado por un conjunto de fabricantes para enfrentar el diseño patentado de Rambus.

Por razones de mercadeo los nombres comerciales de este tipo de memorias han sido bastante confusos, en particular porque hacen mención a la "frecuencia equivalente" que tendría una memoria SDRAM "clásica" que solo hiciera una operación por ciclo.

Las versiones disponibles comercialmente son:

DDR-200: reloj de 100 MHz (equivale a 200 MHz, también denominada PC-200)

DDR-266: reloj de 133 MHz (equivale a 266 MHz, también denominada PC-266)

DDR-333: reloj de 167 MHz (equivale a 333 MHz, también denominada PC-333)

DDR-400: reloj de 200 MHz (equivale a 400 MHz, también denominada PC-400)

9.5.9 DDR2 SDRAM

Son la evolución tecnológica de las DDR, con un diseño pensado en aumentar la frecuencia de trabajo. Son técnicamente incompatibles (trabajan a otro voltaje de alimentación y poseen un encapsulado totalmente distinto).

Las variantes disponibles son:

DDR2-400: reloj de 200 MHz (equivale a 400 MHz, también denominada PC2-3200)

DDR2-533: reloj de 266 MHz (equivale a 533 MHz, también denominada PC2-4200)

DDR2-667: reloj de 333 MHz (equivale a 667 MHz, también denominada PC2-5300)

DDR2-800: reloj de 400 MHz (equivale a 800 MHz, también denominada PC2-6400)

DDR2-1066: reloj de 533 MHz (equivale a 1066 MHz, también denominada PC2-5300)

DDR2-1200: reloj de 600 MHz (equivale a 1200 MHz, también denominada PC2-9000)

9.5.10 DDR3 SDRAM

Son el siguiente paso en la evolución tecnológica de las DDR, con mayor la frecuencia de trabajo y menor consumo (basado en un voltaje de trabajo menor). También son técnicamente incompatibles con las DDR2 y las DDR.

Históricamente las variantes disponibles eran:

DDR3-800: reloj de 400 MHz (equivale a 800 MHz, también denominada PC3-6400)

DDR3-1066: reloj de 533 MHz (equivale a 1066 MHz, también denominada PC3-8500)

DDR3-1333: reloj de 667 MHz (equivale a 1333 MHz, también denominada PC3-10600)

DDR3-1600: reloj de 800 MHz (equivale a 1600 MHz, también denominada PC3-12800)

Actualmente se construyen con frecuencias equivalentes de 1866 MHz, 2133 MHz, 2400 MHz, 2666 MHz, 2800 MHz, 2933MHz, 3000 MHz, 3100 MHz y 3200 MHz.

9.5.11 DDR4 SDRAM

Son un paso más en la evolución de las SDRAM, con un voltaje de trabajo menor y frecuencias mayores a las DDR3, que permiten un menor consumo y una mayor performance, llegando a frecuencias equivalentes de 4000 MHz.

9.5.12 VRAM

Para los controladores de video de los computadores hasta hace un tiempo se solía utilizar un tipo especial de memoria (ya sea basada en tecnología SRAM o DRAM), que se denomina Video RAM y cuya característica principal es contar con "doble puerta". Es decir son dispositivos que pueden ser leídos a la misma vez que escritos (en direcciones distintas). Esto es útil para poder mantener el refresco de la información en el monitor a una velocidad constante (leyendo) a la misma vez que el programa actualiza la información a desplegar en la pantalla (escribiendo).

Este tipo de memoria son como los "latch" pero organizados en múltiples palabras. En la actualidad la tendencia es a utilizar memorias GDDR4 / GDDR5 (GDDR = Graphics DDR), basadas en memorias DDR3, con un sistema de acceso simultáneo a dos páginas, con lo que simula ser "dual port" como la VRAM.

9.5.13 NVRAM

La **Non Volatile RAM** es una memoria del tipo RAM que puede conservar su información aún cuando el sistema donde está insertada queda sin alimentación eléctrica. La forma tradicional de implementar este tipo de memorias es con un SRAM de bajo consumo y una batería (ó pila) que le dé alimentación mientras no se dispone de la alimentación eléctrica general del sistema.

9.5.14 CMOS RAM

Estas memorias en realidad son un nombre que popularizó la arquitectura PC para un tipo de NVRAMs construidas con chips de bajo consumo basados en tecnología CMOS (Complementary Metal Oxide Semiconductor). En los PCs esta memoria no volátil se usa para almacenar la configuración básica del computador.

9.6 Módulos de Memoria

Los computadores modernos implementan sus sistemas de memoria principal en base a módulos de memoria. Estos módulos son pequeñas tarjetas de circuito impreso que disponen de los chips de memoria en cantidad y organización adecuada para el tamaño de palabra soportado por la arquitectura del procesador utilizado. Estas tarjetas de circuito se insertan en zócalos (memory sockets) previstos en las placas principales (motherboards) de los computadores a tales efectos.

Algunos ejemplos de estos módulos son:

SIMM:

Estos fueron los primeros módulos en imponerse en la industria del PC. La sigla significa **Single In-line Memory Module**, ya que si bien la placa de circuito tenía contactos en ambas caras estos eran en realidad redundantes. Existieron en dos variantes:

- 30 contactos ó pines (para buses de memoria de 8 bits)



- 72 contactos ó pines (para buses de memoria de 32 bits)



DIMM:

Fueron la evolución de los SIMM asociados con la aparición de las SDRAM y las arquitecturas de memoria de 64 bits (aun para procesadores de 32 bits). La sigla significa **Dual In-line Memory Module**.

Existen variantes con distinta cantidad de contactos.

- 168 contactos ó pines (utilizado con memorias SDRAM)



- 184 contactos ó pines (utilizado con memorias DDR)



- 240 contactos ó pines (utilizado con memorias DDR2)



Los módulos DDR3 también utilizan 240 contactos, aunque como son incompatibles eléctricamente, tiene la "mueca" de identificación (cortes en forma de "media luna" que poseen en ambos extremos y en el medio) en otro lugar para evitar confusiones, mientras que los DDR4 utilizan 288 contactos.

También existen alternativas a cada uno de los tipos anteriores con menor cantidad de contactos (logrando módulos más pequeños físicamente) diseñados para equipos portátiles (Notebooks, Palms, etc.).

A su vez existen distintas denominaciones de acuerdo al tipo de chip SDRAM que utilicen y en particular a la velocidad de transferencia que poseen (expresada en Mbytes/seg.), por ejemplo PC3-12800 (construidas con chips DDR3-1600)

También existen variantes respecto al voltaje de alimentación (5 Volts, 3.3 Volts, 1.5 Volts) y respecto a si son "buffereados" (también se los denomina "registrados") ó no. Los módulos "registrados" disponen de circuitería adicional que amplifica las señales entre el modulo de memoria y el controlador de memoria de la placa principal, dotando de mayor estabilidad y confiabilidad al circuito. Ambas características se reconocen por la posición de las muescas de codificación.

RIMM:

Son los módulos utilizados para memorias con tecnología Rambus. También hay variantes con distinta cantidad de contactos.

- 184 contactos ó pines (utilizado en sistemas de memoria con buses de 16 bits)



- 232 contactos ó pines (utilizado en sistemas de memoria con buses de 32 bits)

