Memoria de un Computador

Nociones Básicas

Juan Esteban Vergara Sánchez

Despartamento de Ingeniería Electrónica y
Telecomunicaciones
Universidad de Antioquia
Medellín
Septiembre de 2020

Índice

1. Introducción

El concepto de memoria en computación data aproximadamente de la década de 1940, cuando las primeras formas de almacenamiento eléctrico y electrónico aparecieron gracias al creciente uso de las llamadas maquinas de Turing. Conceptos que en la actualidad parecen extraños como la memoria de tambor magnético [1] eran las mejores tecnologías disponibles en la época; sin embargo, incluso hoy en día el concepto base es el mismo: usar algún mecanismo para representar unos y ceros, o bits, que a su vez forman bytes (conjuntos de 8 bits) para guardar información y procesarla.

Con el paso de los años y el creciente avance tecnológico que la humanidad experimenta desde hace décadas, se produjo un salto desde las formas tempranas de almacenamiento electrónico hacia las formas de almacenamiento actuales, que es comparable en magnitud al que se dio al pasar de de las pinturas en las paredes de cavernas realizadas por los primeros humanos a los libros impresos que distribuyeron en masa con la invención de la imprenta en los años 1400's [2], pero en una fracción diminuta de tiempo a comparación de esta ultima.

Los equipos de computo se han vuelto una parte absolutamente esencial en estilo de vida moderno, y por tanto, es importante estudiar su funcionamiento básico, tanto dentro de la comunidad ingenieril (para efectos de este curso) como dentro del público en general. Los métodos seguidos para esta investigación consisten en dar respuesta a cuatro preguntas que se corresponden con cada una de las secciones del documento, haciendo uso de la información suministrada por el docente así como también información extraída de fuentes en linea.

2. Definición de Memoria en Computación

En computación, la memoria se define como el espacio donde los datos que van a ser procesados y las instrucciones requeridas para completar esta tarea son almacenados, haciéndola una parte esencial de cualquier sistema informático, siendo tan importante como la CPU misma, tanto en sistemas convencionales como en sistemas embebidos. La memoria guarda esta información en forma de bytes en estructuras llamadas celdas. Cada celda tiene una dirección única, que varia entre 0 y el tamaño de la memoria menos uno [3]. Existen diferentes clases de memoria, cada una con un tipo de celda y una velocidad distintas, pero exploraremos esto con más detalle en la siguiente sección.

La palabra memoria es confundida con frecuencia con la palabra almacenamiento, que si bien está ampliamente relacionada, es distinta. El almacenamiento hace referencia al espacio en el que los datos son guardados a largo plazo; esto es, el disco duro. No obstante, algunas formas de almacenamiento serán exploradas en la sección 2 debido a su amplio uso y a que son el componente base de la llamada memoria virtual.

Existen 2 grupos dentro de los que se puede clasificar la gran mayoría de dispositivos de memoria y almacenamiento:

- **Memoria o Almacenamiento Volátil:** Se caracteriza por su alta velocidad y por el hecho de que ante un corte en su alimentación, todos los datos almacenados se pierden sin posibilidad de ser recuperados.
- **Memoria o Almacenamiento No volátil:** Se usa para almacenar datos a largo plazo debido a que incluso tras cortar de alimentación, los datos persisten; pudiendo llegar a durar décadas inalterados.

3. Tipos de Memoria

Existe una extensa variedad de tipos de memoria, muchos ya obsoletos gracias al paso de los años y a la aparición de nuevas y mejores tecnologías; es por esto que solo se explorarán las clases de memoria que aun se mantienen relevantes y se usan hoy en día, y se excluirán las clases desusadas y poco comunes tales como PCRAM, PCME, HBM, GDDR SDRAM, DDR SDRAM, SDRAM, ReRAM, NVRAM, etc.

- **Memoria Cache:** Un tipo de memoria volátil de muy alta velocidad y costo de producción que se encuentra en pequeñas cantidades en el interior del procesador, y que se usa para almacenar instrucciones o datos de uso frecuente, e instrucciones o datos que están a punto de ser ejecutados o procesados. Hay 3 tipos que se diferencian por su velocidad, siendo L3 la más lenta, L1 la más rápida y L2 un intermedio. También es conocida como SRAM o Static Random Access Memory (Memoria Estática de Acceso Aleatorio). Cada uno de sus celdas está compuesto por un circuito conocido como flip-flop.

- **Memoria RAM** (Random Access Memory - Memoria de Acceso Aleatorio): Es un tipo de memoria volátil más rápida que la memoria flash, pero mas lenta que la memoria cache. Sus datos pueden ser accedidos en cualquier orden, de ahí el nombre de memoria de acceso aleatorio. Trabaja en conjunto con la memoria cache para suministrar datos a la CPU para que puedan ser procesados. También es conocida como DRAM o Dynamic Random Access Memory (Memoria Dinámica de Acceso Aleatorio). Cada uno de sus celdas está compuesto por un transistor y un condensador. Necesita ser refrescada constantemente o de lo contrario sus datos son eliminados debido a la descarga de dichos condensadores.

- **Memoria de Video o VRAM (Video RAM):** Es una variante de la DRAM usada exclusivamente en tarjetas de video, para asistir al procesador grafico o GPU de la misma forma en que la RAM convencional asiste a la CPU.

- **Memoria ROM (Read Only Memory – Memoria de Solo Lectura):** Es un tipo de memoria no volátil que una vez escrita no puede ser modificada, pero si leída. Ejemplos de ella son el chip que almacena la BIOS de cualquier equipo, o los CD's/DVD's de una sola escritura.

- **Almacenamiento Mecánico:** Es un tipo de almacenamiento no volátil que utiliza una serie de discos magnéticos que rotan a altas velocidades para almacenar información.

- **Memoria Flash:** Es un tipo de almacenamiento no volátil que puede ser escrito y leído de manera rápida y simple. Ejemplos de ella son las memorias USB, los discos duros SSD o los discos duros M.2 NVMe. Tambien es conocido como almacenamiento "sólido".

- **Memoria Virtual:** Es un sector del disco duro (mecánico o solido) dedicado a soportar de manera temporal datos en ejecución que se utilizan con baja frecuencia o que ocupan espacio innecesario en algún momento determinado, pero que aún así podrían ser requeridos en algún momento. La memoria virtual, lógicamente, será más rápida en un disco duro SSD (o incluso NVMe M.2) que en uno mecánico.

4. Mecanismos de gestión de Memoria

La memoria en un computador es gestionada por el llamado controlador de memoria, que se encarga de comunicar las instrucciones de la CPU, interviniendo en cada transferencia de información de y hacia la memoria. También, establece la frecuencia con que se realizan las operaciones a través de su reloj, que se encuentra en el orden de los MegaHertz (1MHz = 1000000Hz). Dicho controlador de memoria puede encontrarse en el interior del procesador o en la placa madre, cerca de DIMMs de memoria y la CPU.

Para describir la forma como se maneja la memoria en un computador, veamos un ejemplo:

Supongamos que realizando un proceso dado la CPU requiere un dato que se encuentra en RAM. Para obtenerlo, realiza una petición al controlador de memoria, que a su vez envía dos cosas a la RAM: la dirección en la que dicho dato se encuentra, a través del bus de datos, y una señal de control para indicar que está a la espera de datos. Al recibir la petición la RAM envía cualquier dato que se encuentre en dicha dirección hacia el procesador a través del bus de datos. Este último puede transmitir 64 bits de información de forma simultanea.

Supongamos ahora que después de procesar el dato la CPU necesita devolverlo a la RAM. Para ello, realiza nuevamente una petición al controlador de memoria, quien lo envía a través del bus de datos, junto con la dirección en las que este va a ser escrito (a través del bus de direcciones). Acto seguido envía una señal de control a través del bus de control indicándole a la RAM la tarea que debe realizar.

Cabe destacar que todas las posiciones de memoria que no estén siendo utilizadas en un momento dado, o cuyos datos ya hayan sido utilizados son borradas para ahorrar espacio.

Dado el caso que un no se encuentre cargado en la memoria RAM, pero si esté almacenado en disco, la CPU envía una solicitud a un controlador especial ubicado en la placa madre que se encarga de controlar los dispositivos de almacenamiento para que cargue el dato deseado a la RAM y así poder accederlo. Si en cierto momento el procesador desea guardar información cargada en la RAM en el disco duro, enviará una instrucción al controlador de discos para que escriba una copia exacta de las direcciones de memoria RAM en las que se encuentra dicha información en determinado sector del disco, y luego eliminará la instrucción y la información de la memoria para ahorrar espacio.

Pero, donde entra el cache en todo esto? Gracias a complejos algoritmos, el procesador detecta las instrucciones que se usan con mas frecuencia, y las almacena en cache; así, siempre que se necesita un dato o instrucción, primero se revisa si este se encuentra en cache, y de ser así, una gran cantidad de tiempo es ahorrada. En el ejemplo anterior, antes de realizar cualquiera de las peticiones mencionadas, la CPU revisó en su memoria cache.

La forma como se gestiona la memoria cache es bastante simple: Los datos de uso mas frecuentes se guardan en L1, los que se usan con menos frecuencia o que no pudieron ser almacenados en L1 se guardan en L2, y los que se usan con aún menos frecuencia o no pudieron ser almacenados en L2 se guardan en L3 [4].

5. Sobre la Velocidad de la Memoria

Como se observó en la sección 2, los diferentes tipos de memoria varían de ampliamente en su velocidad; esto se debe a varios factores, como por ejemplo la diferencia en los mecanismos con los que cada tipo de memoria almacena los datos, o algo tan trivial como la distancia física de la memoria a la CPU [5]. Para

visualizarlo mejor, veamos un ejemplo: supongamos que a modo de experimento se envía el mismo dato o instrucción hacia la CPU (3GHz) desde la memoria cache, la memoria RAM, y el disco duro, y se mide el tiempo que este tarda en alcanzarla. Al momento de organizar los resultados del experimento en una tabla, para dar algo de perspectiva, se añadieron 2 columnas: la distancia que viaja la luz en el vació en el tiempo especificado, y, si estiráramos un ciclo del reloj de la CPU desde 1ns a 1s, el tiempo que tomaría realizar dicha acción.

Acción	Tiempo	Distancia	Tiempo Estirado
1 ciclo de reloj (3GHz)	0.3 ns	10.1 cm	1 s
Acceso A Cache L1	0.9 ns	30.4 cm	3 s
Acceso A Cache L2	2.8 ns	85.3 cm	9 s
Acceso A Cache L3	12.9 ns	3.9 m	43 s
Acceso a RAM	70 - 100 ns	21.3 m - 30.4 m	3.5 min - 5.5 min
Disco duro NVMe M.2	7 - 150 µs	2.1 km - 45.7 km	2 h - 2 días
Disco duro mecánico	1 - 10 ms	304 km - 3048 km	11 días - 4 meses

Figura 1: Tabla de resultados [6]

La inmensa diferencia en la velocidad de cada tipo de memoria produce efectos secundarios indeseados, como por ejemplo, que de manera bastante frecuente la CPU se encuentre parada durante una gran cantidad de ciclos de reloj a la espera de las instrucciones que se encuentran en la memoria RAM, o de información almacenada en disco. La pregunta más obvia sería ¿Porque no usar memoria cache para la RAM y el disco? Existen varias razones por las que esto no es posible: primero, sería demasiado caro debido al elevado precio de la memoria cache, además de esto, ocuparía más espacio ya que las celdas de cache son físicamente mas grandes que las de RAM o disco, y, finalmente, seria totalmente impráctico, ya que el cache es un tipo de memoria volátil, por lo que nuestros datos solo durarían hasta que se apague el equipo. Sin embargo, la gran velocidad de la memoria cache, y la forma como es gestionada, puede producir un gran aumento en el rendimiento de una equipo, y reducir la aparición de cuellos de botella [4].

Existe otra medida que se usa para describir la velocidad de la memoria llamada latencia. Hace referencia al tiempo que esta se tarda en leer la información de una dirección de memoria. Esto puede llegar a tardar varios ciclos de reloj (reloj del controlador de memoria), lo que añade varios nanosegundos al tiempo de transferencia y hace que los tiempos ideales teóricos no sean posibles [7].

Referencias

[1] n.a. Timeline of computer history. [Online]. Available: https://www.computerhistory.org/timeline/memory-storage/

[2] ——. ¿quiÉn inventÓ la imprenta? — johann gutenberg. [Online]. Available: https://www.imprentaonline.net/quien-invento-la-imprenta

[3] ——. Computer - memory. [Online]. Available: https://www.tutorialspoint.com/computer$_f$*undamentals*/*computer*$_m$*emory.htm*

[4] ——. Cpu cache explained - what is cache memory? [Online]. Available: https://www.youtube.com/watch?v=yi0FhRqDJfo

[5] L. Stewart. Why is cache memory faster than ram? [Online]. Available: https://www.quora.com/Why-is-cache-memory-faster-than-RAM

[6] n.a. Compute performance – distance of data as a measure of latency. [Online]. Available: https://formulusblack.com/blog/compute-performance-distance-of-data-as-a-measure-of-latency/

[7] A. Salazar, *Taller - Nociones de la memoria del computador*. Universidad de Antioquia.