Tecnologías de Memoria*

José Enrique Alvarez Estrada

Based on *Memory - Memory Technology*† by

Charles Severance

Kevin Dowd

This work is produced by OpenStax-CNX and licensed under the
Creative Commons Attribution License 3.0‡

Prácticamente todas las memorias rápidas actuales están basadas en semiconductores. [1] Vienen en dos variedades: *memoria dinámica de acceso aleatorio* (DRAM) y *memoria estática de acceso aleatorio* (SRAM). El término *aleatorio* significa que puede usted acceder a las localidades de memoria en cualquier orden. Se usa para distinguir el acceso aleatorio de las memorias seriales, en las que debe recorrer paso a paso todas las celdas intermedias hasta llegar a aquella en particular que le interesa. Un ejemplo de un medio de almacenamiento *no* aleatorio es la cinta magnética. Los términos dinámico y estático tienen que ver con la tecnología usada en el diseño de las celdas de memoria. Los dispositivos DRAM se basan en carga eléctrica, pues cada bit se representa mediante la carga almacenada por un diminuto capacitor. Dicha carga se fuga en un corto periodo de tiempo, asì que el sistema debe refrescarla continuamente para evitar la pérdida de datos. También el acto de leer un bit de la DRAM la descarga, y debe refrescarse. Y no es posible leer un bit de memoria DRAM mientras se está refrescando.

La SRAM se basa en compuertas, y cada bit se almacena mediante un arreglo de cuatro a seis transistores conectados. Las memorias SRAM retienen sus datos mientras tengan energía, sin la necesidad de ningún mecanismo de refresco.

La DRAM ofrece la mejor tasa precio/rendimiento, así como la mayor densidad de celdas de memoria por chip. Ello significa un menor costo, menos espacio en las tarjetas, menos gasto energético y menos calor. Por otra parte, algunas aplicaciones, tales como la cache y la memoria de video, requieren velocidades mayores, para las que la SRAM resulta más adecuada. Actualmente, puede usted elegir entre SRAM y DRAM a velocidades por debajo de los 50 nanosegundos (ns). La SRAM tiene tiempos de acceso de alrededor de 7 ns a costa de mayor costo, calor, energía y espacio en la tarjeta.

El rendimiento de la memoria está limitado, además de por la tecnología básica necesaria para almacenar los bits de datos, por consideraciones prácticas tales como el acomodo de los alambres en el circuito integrado y las patillas externas para comunicar la información sobre direcciones y datos entre la memoria y el procesador.

*Version 1.1: Mar 22, 2011 9:46 pm -0500

†http://cnx.org/content/m32716/1.3/

‡http://creativecommons.org/licenses/by/3.0/

[1] todavía se usa la memoria de núcleo magnético en aplicaciones donde la "dureza" de la radiación -resistencia a cambios causada por radiación ionizante- es importante.

1 Tiempos de Acceso

La cantidad de tiempo que toma leer o escribir una posición de memoria se denomina el *tiempo de acceso a memoria*. Una cantidad relacionada es el *tiempo de ciclo de memoria*. Mientras que el tiempo de acceso nos dice cuan rápidamente puede referenciar una posición de memoria, el tiempo de ciclo describe qué tan a menudo puede hacer referencia a ella. Suenan como la misma cosa, pero no lo son. Por ejemplo, si usted pide datos de un chip de DRAM con un tiempo de acceso de 50 ns, puede necesitar 100 ns antes de que pueda solicitar datos del mismo chip. Ello se debe a que los chips deben recobrarse internamente del acceso previo. Sin embargo, algunas tecnologías han mejorado el rendimiento cuando se está recuperando datos secuencialmente de una DRAM. En tales chips, los datos que inmediatamente después de aquellos accesados previamente, pueden recuperarse tan rápidamente como 10 ns.

Los tiempos de acceso y ciclo de memorias DRAM comerciales son más cortos que hace tan sólo algunos años, lo cuál significa que se pueden construir sistemas de memoria más rápidos. Pero también ha incrementado la velocidad de reloj de la CPU. El mercado de las computadoras caseras es un buen ejemplo. A inicios de la década de 1980, el tiempo de acceso de una DRAM comercial (200 ns) era menor que el ciclo de reloj de la IBM PC XT (4.77 MHz = 210 ns). Ello significa que la DRAM podía conectarse directamente a la CPU, sin preocuparse por rebasar al sistema de memoria. Pero a mitad de la década de 1980 se introdujeron modelos XT y AT más rápidos, con CPUs cuyos relojes superaban a los tiempos de acceso de la memoria comercial disponible. Había memorias más rápidas para quien estuviera dispuesto a pagar por ellas, pero los vendedores apostaron por vender computadoras que agregaban *estados de espera* al ciclo de acceso a la memoria. Los estados de espera son retrasos artificiales que hacen más lentos las referencias, de forma que la memoria parece empatarse con una CPU más rápida -a cambio de una penalización. Sin embargo, esta técnica de agregar estados de espera comenzó a impactar significativamente el rendimiento alrededor de los 25 a 33 MHz. Hoy en día, las velocidades de las CPU están mucho más arriba que las de la DRAM.

La duración de un ciclo de reloj para las computadoras caseras comerciales ha cambiado de los 210 ns de una XT, a alrededor de 3 ns para una Pentium II a 300 MHz. Pero el tiempo de acceso para una DRAM comercial ha decrecido desproporcionadamente menos -de 200 ns a alrededor de 50 ns. El rendimiento del procesador se duplica cada 18 meses, mientras que el rendimiento de la memoria se duplica aproximadamente cada siete años.

La brecha de velocidad entre CPU y memoria es todavía mayor en el caso de las estaciones de trabajo. Algunos modelos tienen periodos de reloj tan cortos como 1.6 ns. ¿Cómo concilian los vendedores esta diferencia de velocidad entre CPU y memoria? La memoria en la supercomputadora Cray-1 empleaba SRAM que era capaz de mantenerse a la par de su ciclo de reloj de 12.5 ns. Usar SRAM en su memoria principal era una de las razones por las que la mayoría de las computadoras Cray requerían refrigeración líquida.

Desafortunadamente, no es práctico para un sistema de precio moderado confiar exclusivamente en la SRAM como almacenamiento. Como tampoco lo es fabricar sistemas económicos con almacenamiento suficiente usando sólo SRAM.

La solución es una jerarquía de memorias, formada por los registros del procesador, de uno a tres niveles de cache SRAM, una memoria principal DRAM y memoria virtual almacenada en medios tales como los discos. En cada punto de esta jerarquía de memoria se emplean trucos para lograr un uso óptimo de la tecnología disponible. En lo que resta de este capítulo examinaremos la jerarquía de memoria y su impacto sobre el rendimiento.

De cierta forma, con los procesadores actuales de alto rendimiento realizando cálculos tan rápidamente, la tarea del programador de alto rendimiento se convierte en administrar cuidadosamente la jerarquía de memoria. En cierto sentido, resulta un ejercicio intelectual útil pesar que los cálculos simples -tales como la suma y la multiplicación- son "infinitamente rápidos", con el objeto de dar al programador una perspectiva correcta acerca del impacto de las operaciones de memoria sobre el rendimiento global del programa.