

5.3 ADVANCED DRAM ORGANIZATION

Uno de los cuellos de botella más críticos al usar procesadores de alto rendimiento es la interfaz con la memoria interna principal. Esta interfaz es la vía más importante en todo el sistema informático. El bloque básico de construcción de la memoria principal sigue siendo el chip DRAM, como lo ha sido durante décadas; hasta

Table 5.3 Performance Comparison of Some DRAM Alternatives

	Clock Frequency (MHz)	Transfer Rate (GB/s)	Access Time (ns)	Pin Count
SDRAM	166	1.3	18	168
DDR	200	3.2	12.5	184
RDRAM	600	4.8	12	162

Recientemente, no se han producido cambios significativos en la arquitectura DRAM desde principios de los años setenta. El chip DRAM tradicional está restringido tanto por su arquitectura interna como por su interfaz con el bus de memoria del procesador.

Hemos visto que un ataque al problema de rendimiento de la memoria principal DRAM ha sido insertar uno o más niveles de memoria caché SRAM de alta velocidad entre la memoria principal DRAM y el procesador. Pero la SRAM es mucho más costosa que DRAM, y expandir el tamaño de la caché más allá de cierto punto produce rendimientos decrecientes.

En los últimos años, se han explorado una serie de mejoras en la arquitectura básica de DRAM, y algunas de ellas ya están en el mercado. Los esquemas que actualmente dominan el mercado son SDRAM, DDR-DRAM y RDRAM. La Tabla 5.3 proporciona una comparación de rendimiento. CDRAM también ha recibido considerable atención. Examinamos cada uno de estos enfoques en esta sección.

Synchronous DRAM

Una de las formas más utilizadas de DRAM es la DRAM síncrona (SDRAM) [VOGL94]. A diferencia de la DRAM tradicional, que es asíncrona, la SDRAM intercambia datos con el procesador sincronizado a una señal de reloj externo y se ejecuta a la velocidad máxima del procesador / bus de memoria sin imponer estados de espera.

En una DRAM típica, el procesador presenta direcciones y niveles de control a la memoria, lo que indica que un conjunto de datos en una ubicación particular en la memoria debe leerse o escribirse en la DRAM. Después de un retraso, el tiempo de acceso, la DRAM escribe o lee los datos. Durante el retardo de tiempo de acceso, la DRAM realiza varias funciones internas, como activar la alta capacitancia de las líneas de fila y columna, detectar los datos y direccionar los datos a través de los búferes de salida. El procesador simplemente debe esperar a través de este retraso, reduciendo el rendimiento del sistema.

Con acceso síncrono, la DRAM mueve los datos dentro y fuera bajo el control del reloj del sistema. El procesador u otro maestro emite la instrucción y la información de dirección, que está bloqueada por la DRAM. La DRAM luego responde después de un número determinado de ciclos de reloj. Mientras tanto, el maestro puede realizar otras tareas con seguridad mientras la SDRAM está procesando la solicitud.

La Figura 5.12 muestra la lógica interna de 64-Mb SDRAM de IBM [IBM01], que es típica de la organización SDRAM, y la Tabla 5.4 define las diversas asignaciones de PIN. La SDRAM emplea un modo de ráfaga para eliminar el tiempo de configuración de la dirección y el tiempo de precarga de la línea de fila y columna después del primer acceso. En el modo de ráfaga, una serie de

los bits de datos se pueden desconectar rápidamente después de acceder al primer bit. Este modo es útil cuando todos los bits a acceder están en secuencia y en la misma fila de la matriz que el acceso inicial. Además, la SDRAM tiene una arquitectura interna de múltiples bancos que mejora las oportunidades para el paralelismo en el chip. El registro de modo y la lógica de control asociada es otra característica clave que diferencia las SDRAM de las DRAM convencionales. Proporciona un mecanismo para personalizar la SDRAM para satisfacer las necesidades específicas del sistema. El registro de modo especifica la longitud de ráfaga, que es el número de unidades de datos independientes alimentadas de forma síncrona al bus. El registro también permite al programador ajustar la latencia entre la recepción de una solicitud de lectura y el comienzo de la transferencia de datos.

La SDRAM funciona mejor cuando transfiere grandes bloques de datos de forma seriada, como aplicaciones como procesamiento de texto, hojas de cálculo y multimedia. La Figura 5.13 muestra un ejemplo de operación SDRAM. En este caso, la longitud de ráfaga es 4 y la latencia es 2. El comando de lectura en ráfaga se inicia al tener CS y CAS bajos mientras mantiene RAS y WE en alto en el borde ascendente del reloj. Las entradas de dirección determinan la dirección de la columna de inicio para la ráfaga, y el registro de modo establece el tipo de ráfaga (secuencial o intercalado) y la longitud de la ráfaga (1, 2, 4, 8, página completa). El retraso desde el inicio del comando hasta cuando los datos de la primera celda aparecen en las salidas es igual al valor de la latencia CAS que se establece en el registro de modo.

Ahora hay una versión mejorada de SDRAM, conocida como SDRAM de doble velocidad de datos (DDR-SDRAM) que supera la limitación de una vez por ciclo. DDR-SDRAM puede enviar datos al procesador dos veces por ciclo de reloj.

Rambus DRAM

RDRAM, desarrollado por Rambus [FARM92, CRIS97], ha sido adoptado por Intel para sus procesadores Pentium e Itanium. Se ha convertido en el principal competidor de SDRAM. Los chips RDRAM son paquetes verticales, con todos los pines en un lado. El chip intercambia datos con el procesador a través de 28 cables de no más de 12 centímetros de largo. El bus puede direccionar hasta 320 chips RDRAM y tiene una capacidad de 1.6 GBps. El bus especial RDRAM entrega información de dirección y control usando un protocolo asíncrono orientado a bloques. Después de un tiempo de acceso inicial de 480 ns, esto produce la tasa de datos de 1.6 GBps. Lo que hace posible esta velocidad es el propio bus, que define las impedancias, el reloj y las señales con mucha precisión. En lugar de estar controlado por las señales RAS, CAS, R / W y CE explícitas utilizadas en las DRAM convencionales, una RDRAM recibe una solicitud de memoria sobre el bus de alta velocidad. Esta solicitud contiene la dirección deseada, el tipo de operación y el número de bytes en la operación.

La Figura 5.14 ilustra el diseño de RDRAM. La configuración consta de un controlador y varios módulos RDRAM conectados a través de un bus común. El controlador está en un extremo de la configuración, y el otro extremo del bus es una terminación paralela de las líneas de bus. El autobús incluye 18 líneas de datos (16 datos reales, dos paridad) que ciclan al doble de la velocidad del reloj; es decir, se envía 1 bit en el borde inicial y siguiente de cada señal de reloj. Esto da como resultado una velocidad de señal en cada línea de datos de 800 Mbps. Hay un conjunto separado de 8 líneas (RC) utilizado para las señales de dirección y control. También hay una señal de reloj que comienza en el otro

extremo del controlador que se propaga al extremo del controlador y luego regresa. Un módulo RDRAM envía datos al controlador de forma síncrona al reloj al maestro, y el controlador envía datos a un RDRAM de forma síncrona con la señal del reloj en la dirección opuesta. Las líneas de bus restantes incluyen un voltaje de referencia, tierra y fuente de alimentación.

DDR SDRAM

SDRAM está limitado por el hecho de que solo puede enviar datos al procesador una vez por ciclo de reloj del bus. Una nueva versión de SDRAM, conocida como SDRAM de velocidad de datos doble, puede enviar datos dos veces por ciclo de reloj, una vez en el borde ascendente del pulso de reloj y una vez en el borde descendente.

DDR DRAM fue desarrollado por la Asociación de Tecnología de Estado Sólido de JEDEC, el cuerpo de ingeniería de semiconductores de la Alianza de Industrias Electrónicas.

Numerosas empresas fabrican chips DDR, que son ampliamente utilizados en computadoras de escritorio y servidores.

La Figura 5.15 muestra el tiempo básico para una lectura de DDR. La transferencia de datos está sincronizada con el borde ascendente y descendente del reloj. También se sincroniza con una señal de luz de datos bidireccional (DQS) que es proporcionada por el controlador de memoria durante una lectura y por la DRAM durante una escritura. En implementaciones típicas, DQS se ignora durante la lectura. Una explicación del uso de DQS en escrituras está más allá de nuestro alcance; ver [JACO08] para más detalles.

Ha habido dos generaciones de mejoras en la tecnología DDR. DDR2 aumenta la velocidad de transferencia de datos aumentando la frecuencia operativa del chip RAM y aumentando el buffer de captación previa de 2 bits a 4 bits por chip. El buffer de captación previa es un caché de memoria ubicado en el chip RAM. El búfer permite que el chip RAM tenga a disposición los bits que se colocarán en el bus de datos lo más rápido posible.

DDR3, introducido en 2007, aumenta el tamaño del búfer de captación previa a 8 bits.

Teóricamente, un módulo DDR puede transferir datos a una velocidad de reloj en el rango de 200 a 600 MHz; un módulo DDR2 se transfiere a una velocidad de reloj de 400 a 1066 MHz; y un módulo DDR3 se transfiere a una velocidad de reloj de 800 a 1600 MHz. En la práctica, se logran tasas algo menores.

Cache DRAM

Cache DRAM (CDRAM), desarrollado por Mitsubishi [HIDA90, ZHAN01], integra una pequeña memoria caché SRAM (16 Kb) en un chip DRAM genérico.

La SRAM en el CDRAM se puede usar de dos maneras. En primer lugar, se puede utilizar como un verdadero caché, que consiste en una serie de líneas de 64 bits. El modo de caché de la CDRAM es efectivo para el acceso aleatorio ordinario a la memoria.

La SRAM en el CDRAM también se puede utilizar como un búfer para admitir el acceso en serie de un bloque de datos. Por ejemplo, para actualizar una pantalla de mapa de bits, la CDRAM puede captar previamente los datos de la DRAM en el búfer SRAM. Los accesos posteriores al chip dan como resultado accesos únicamente a la SRAM.

Alumno: Alejandro Adrián Macías Pantoja.