Capítulo 9

Evolución

MEMORIAS

Sumario

Nuestra cobertura de la evolución de la memoria guarda cierta similitud con el seguimiento que dimos para el microprocesador. Primero haremos una síntesis por décadas desde su nacimiento, y posteriormente, detallaremos por generaciones a partir de la llegada del PC.

SECCIÓN 9.1

Semblanza histórica por décadas

Los 50. El punto de partida de los núcleos de ferrita

◀ 1.1

Para repasar la evolución histórica de los chips de memoria desde su nacimiento, debemos remontarnos hasta los años 50. Entonces aparecían las primeras memorias electrónicas, conformadas por hileras de núcleos o anillos de ferrita con un diámetro de un milímetro cuadrado y ensartados por miles de alambres a modo de una tupida red. Estas memorias ocupaban un enorme tamaño y liberaban gran cantidad de calor. El ENIAC, considerado el primer computador de la historia, se fabricó en 1946 con 4 Kbytes de memoria de este tipo, que ocupaba varios metros cuadrados (para entendernos, cada Kbyte consumía el espacio de un armario ropero).

anillos

1.2 ▶Los 60. La revolución de los circuitos integrados

condensadores

La llegada de los circuitos integrados supuso un gran salto cualitativo en la fabricación de memorias a mediados de los años 60. La celda básica de memoria pasa a ser una minúscula carga eléctrica cuyo signo determinaba el bit de información, almacenado en un pequeño condensador de unos 50 femptofaradios [1] y acompañado de un transistor que hace las labores de conmutador a la hora de conectarlo a la línea de datos.

avances

Esta nueva celda de memoria posibilitó una gran reducción de la tasa de errores, el tiempo de acceso, el espacio físico ocupado, el consumo de potencia, y sobre todo, el coste final del producto. En los diez años que transcurrieron entre 1965 y 1975, fiabilidad y velocidad aumentaron en un factor 10, espacio y consumo se redujeron en un factor 100, y el coste disminuyó en un factor 1000. La tecnología de semiconductores impulsó con mucha más fuerza la fabricación de microprocesadores diez años más tarde, pero fueron precisamente estos años de adelanto de que disfrutó la memoria los que le permitieron disponer de cierta ventaja inicial.

1.3 ▶Los 70. La evidencia de que el procesador progresa más rápido

velocidad

Tras el nacimiento del microprocesador en 1971, la situación cambió muy rápidamente. Los microprocesadores doblan su velocidad cada año y medio según la tendencia vaticinada por Gordon Moore y que ha perdurado hasta nuestros días. La memoria, en cambio, va a necesitar de una década entera para doblar su velocidad, y así mantendrá su peregrinar en décadas sucesivas (ver figura 9.1.a).

pág. 5

tamaño

Mención aparte merece el tamaño. Estamos en la antesala del nacimiento del PC, época en la que la capa software apenas se encontraba desarrollada: Los computadores eran incompatibles entre sí y los programas se desarrollaban a título particular para una arquitectura, cargándose directamente desde cinta magnética o tarjetas perforadas. Los sistemas operativos eran tremendamente primitivos, no existiendo la necesidad de disponer de grandes cantidades de memoria.

1.4 ▶Los 80. Emerge la memoria caché

Como consecuencia de su evolución tan dispar, los microprocesadores dan alcance a la memoria a principios de los años 80, y a partir de ahí van dejándola cada vez más atrás, haciéndose ésta paulatinamente más lenta con respecto al procesador.

Desgraciadamente, una elevada potencia de proceso no sirve de nada si no va acompañada de un sistema de memoria que sea capaz de proporcionar los datos e instrucciones a una velocidad similar. Los ingenieros de computación lo saben, y tratan de reducir la diferencia en velocidad entre ambos sistemas. Pero se topan con dos grandes obstáculos: Uno de índole estructural y otro de índole comercial.

circuito RC vs. transistor

1. **Estructuralmente**, los circuitos de memoria principal se basan en la carga y descarga de condensadores (uno por cada celda de un bit), mientras que los microprocesadores se basan en la conmutación de los transistores. El tiempo de respuesta de un circuito RC (resistencia-condensador) es muy elevado si se compara con el tiempo de propagación de las señales eléctricas por los transistores, y además, el circuito RC resulta mucho más difícil de optimizar. Esto deja muy poco margen de maniobra para futuras mejoras, mientras que en el microprocesador las señales pasan por millones de transistores y siempre existe al menos una forma de mejorar el camino crítico que determina la máxima frecuencia de trabajo. No es de extrañar, pues, que la latencia de los circuitos de memoria haya evolucionado de forma

[1] 50×10^{-15} faradios.

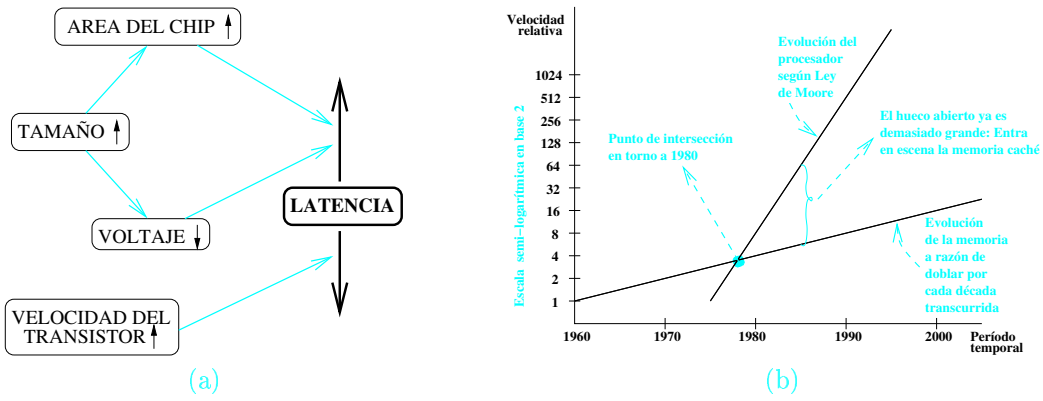


FIGURA 9.1: (a) Influencia de las distintas variables físicas de un circuito de memoria dinámica en su latencia. A lo largo del tiempo, tanto el tamaño en Mbytes como la velocidad de los transistores han ido aumentando, afectando a la latencia de las celdas de memoria de forma contrapuesta. Los efectos se han contrarrestado de forma que la evolución de la latencia de estos circuitos muestra una tendencia casi plana. (b) Comparativa en la evolución de la velocidad de memoria principal respecto al procesador en los últimos 30 años.

tan plana en estos últimos veinte años, mientras que con la frecuencia del microprocesador se hayan hecho verdaderas maravillas en ese mismo período.

2. **Comercialmente**, la capa software demanda memorias cada vez más grandes, y la industria se concentra en la fabricación de chips con mayor capacidad. En la evolución del tamaño, los fabricantes van a tener mejor fortuna, ya que desde este momento y hasta nuestros días, la memoria ha conseguido siempre satisfacer las necesidades del software a un ritmo que sorprendentemente sigue también la Ley de Moore (la capacidad del chip de memoria se dobla cada año y medio). Sin embargo, esta evolución va a afectar negativamente sobre la velocidad de la memoria por dos motivos básicos (ver figura 9.1.b): `tamaño`

   a) En el diseño de circuitos hay una máxima que reza "cuanto más grande, más lento". Para el caso de una malla de circuitos RC, el retraso en su respuesta crece cuadráticamente con el tamaño de la malla de circuitos, con lo cual cada año y medio la memoria se volvería cuatro veces más lenta por la simple evolución de su tamaño.

   b) Conforme el tamaño de los circuitos crece, el voltaje de alimentación ha ido disminuyéndose para mantener el consumo de corriente y la temperatura en niveles moderados. Y desgraciadamente, este menor voltaje también incrementa el tiempo de respuesta del circuito RC.

Si la capacidad de respuesta de un condensador queda muy lejos de lo que se le exige, y además se ve negativamente influenciado por el incremento en el tamaño de la memoria, la única salida al problema de la velocidad consiste en crear una memoria estructuralmente diferente que sea capaz de reducir el tiempo de respuesta a la vez que prometa una evolución más ágil que pueda seguir la pauta marcada por el microprocesador. `alternativa`

La solución no era fácil, pues en los últimos veinte años habían fracasado decenas de nuevas tecnologías, como las memorias de película magnética, de haz electrónico y de haz óptico, frente a las continuas mejoras tecnológicas de la memoria dinámica tradicional. Por entonces ya se sabía que tanto el coste de fabricación de un producto como su tasa de fallos disminuía al aumentar la experiencia en su fabricación, y esto, conocido empíricamente como **curva de aprendizaje**, creaba `curva de aprendizaje`

una doble barrera frente a la introducción de nuevas tecnologías de memoria.

relevo tecnológico

El único ejemplo de relevo tecnológico que había presenciado el mercado era el reemplazo del núcleo de ferrita por la memoria de semiconductores, circunstancia que algunos observadores atribuyeron al hecho de que esta última había avanzado a lo largo de una pronunciada curva de aprendizaje antes de ser utilizada en el contexto de las memorias. Aprendiendo de esta observación, la ciencia apostó por utilizar como recambio los mismos elementos que tan buen resultado habían dado en el microprocesador: Los transistores. Así, los condensadores fueron sustituidos por grupos de puertas lógicas en las que realimentando las salidas de unas a la entrada de otras se conseguía retener la información, siendo necesario el empleo de entre 6 y 8 transistores por cada nueva celda básica de memoria.

memoria caché

El nuevo circuito de memoria se denominó **memoria caché**, entrando en escena a principios de los años 80 como un nuevo integrante de la arquitectura de un PC. Pero esto tampoco representaba una invención, puesto que los registros internos del microprocesador fueron siempre celdas de memoria que se implementaron de esta forma, e igualmente se fabricaron así los registros de almacenamiento de muchas calculadoras en los años 70. Incluso remontándonos más atrás, desde el mismo origen de la memoria de circuitos integrados en los años 60 conocemos de la existencia de chips de memoria de este tipo, entonces referenciados como RAM estática (*Static RAM* - SRAM, que por cierto, estudiaremos en su vertiente actual en el capítulo 11).

pág. 115

avances

La caché supuso simplemente la oportunidad de disfrutar de una memoria unas diez veces más rápida que la memoria dinámica basada en condensadores a costa de pagar más dinero por ella y reducir su capacidad. Esto, que años atrás no compensaba porque el procesador era más lento que la memoria, ahora cobraba todo su sentido, y con ello, el inconveniente del coste fue poco a poco diluyéndose a medida que mejoraba la tecnología y se extendía su uso.

dualidad

Por otro lado, el recorte de la capacidad que trajo consigo la caché no era sostenible con la evolución de la industria del software, conflicto que se resolvió creando una arquitectura del computador en la convivieran ambos tipos de memoria: La memoria de condensadores dotará al sistema de una elevada capacidad de almacenamiento, mientras que la memoria caché actuará de aceleradora ubicándose más cerca del procesador.

1.5 ▶Los 90. La jerarquía de memoria como continuación de tendencia

En los años 90, la memoria caché se convirtió en un elemento indispensable para el rendimiento del computador. Surgió entonces el concepto de jerarquía de memoria, que iría madurando en esta década a medida que la velocidad del microprocesador se fuese desbocando y el software continuara demandando memorias de mayor capacidad. Ante esta situación, la arquitectura se recompone:

tamaño

- La memoria principal responderá a las demandas de tamaño sin problema, pero se estancará en velocidad (la negativa repercusión del tamaño se compensa aquí con las sucesivas mejoras tecnológicas al nivel de integración de circuitos, tal y como indicamos en la figura 9.1.a).

pág. 5

velocidad

- La caché hará justo lo contrario: Seguirá el camino marcado por la velocidad del procesador, pero no el aumento de tamaño de su compañera.

Con este panorama, el sistema pronto se descompensa y la caché se ahoga en su intento por mantener el aluvión de datos que llegan de memoria principal hacia el procesador. La respuesta que ofrece la arquitectura del computador consiste en situar cachés adicionales en el camino hacia memoria principal, sacrificando velocidad paulatinamente al tiempo que se aumenta su tamaño con objeto de equilibrar estas dos variables y eliminar los cuellos de botella. La jerarquía de memoria del computador continúa creciendo en nuestros días, tal y como muestra la tabla 9.1. No

pág. 7

Elemento constitutivo	Nivel de la jerarquía	Tamaño	Latencia
Transistores	Banco de registros	32 a 512 registros	1 ciclo
Puertas lógicas	Caché L1 integrada	8K-64K insts/datos	1 ciclo
Puertas lógicas	Caché L2 integrada	128K-1Mb unificada	2-3 ciclos
Puertas lógicas	Caché L3 interna	1Mb-8Mb unificada	5-7 ciclos
Circuito RC	Memoria principal	16Mb-256Mb	10 ciclos
Sustrato magnetizable	Disco duro	Varios Gigabytes	5-12 milisegs.
Sustrato magnetizable	Cinta	Varios Terabytes	Algunos segs.

TABLA 9.1: La jerarquía de memoria de un PC actual.

niveles

obstante, se aprecia una tendencia a colocar más niveles en dirección al procesador (cachés integradas de primer y segundo nivel), y menos en dirección a la memoria principal, hasta tal punto que en la arquitectura de 2003 la caché externa ha desaparecido de la placa base.

La década actual: Al amparo del interfaz

◀ 1.6

En la actualidad, la memoria principal se sitúa en el nivel central de la jerarquía de memoria, con las diferentes cachés y el banco de registros por delante en dirección al procesador, y el disco duro por detrás como soporte adicional para la memoria virtual. Incluso existe a veces un nivel más exterior donde se ubica el soporte para las copias de seguridad de los datos, rol antiguamente ocupado por la cinta magnética y más recientemente usurpado por la grabadora de CD-ROM.

cambios

De esta manera, la memoria principal ya no dialoga con el procesador, sino con la caché, y además, hace de interfaz hacia el sistema de entrada/salida (discos, vídeo, sonido, ...). La dependencia que el sistema tiene ahora de una memoria caché se refleja tanto en el creciente número de niveles integrados dentro del microprocesador como en su mayor tamaño.

síntesis

En síntesis, podemos decir que la memoria caché ha evolucionado arquitecturalmente, afinando aquel principio físico de los circuitos integrados "más grande, más lento", y beneficiándose además de las mejoras tecnológicas en la distancia de integración de los transistores. En cambio, la memoria principal, basada en el condensador como elemento constituyente, se ha quedado al margen de estas mejoras, confiándose mucho más a los logros conseguidos en su vertiente lógica: En concreto, en sucesivas optimizaciones de su interfaz de diálogo que han sabido explotar la anchura del bus y la línea de caché por un lado, y la localidad de referencia de los programas por el otro. Tras la lectura del capítulo 10 tendremos más clara esta visión.

☛ pág. 13

SECCIÓN 9.2

Evolución por generaciones

referencias

Para describir al nivel del diagrama de bloques la evolución que ha seguido el sistema de memoria de un PC nos apoyaremos en las cuatro últimas generaciones de microprocesadores. Nuestras referencias serán pues un procesador 80486 en la cuarta generación, un Pentium en la quinta, un Pentium II ó III en la sexta, y un Pentium 4 ó K7 en la séptima.

Cuarta generación

◀ 2.1

El 80486 presenta buses de datos y direcciones de 32 bits. La estructura de memoria de un 80486 resulta interesante como precursora de los Pentium, pues muchos de sus rasgos fueron más tarde adaptados a ellos.

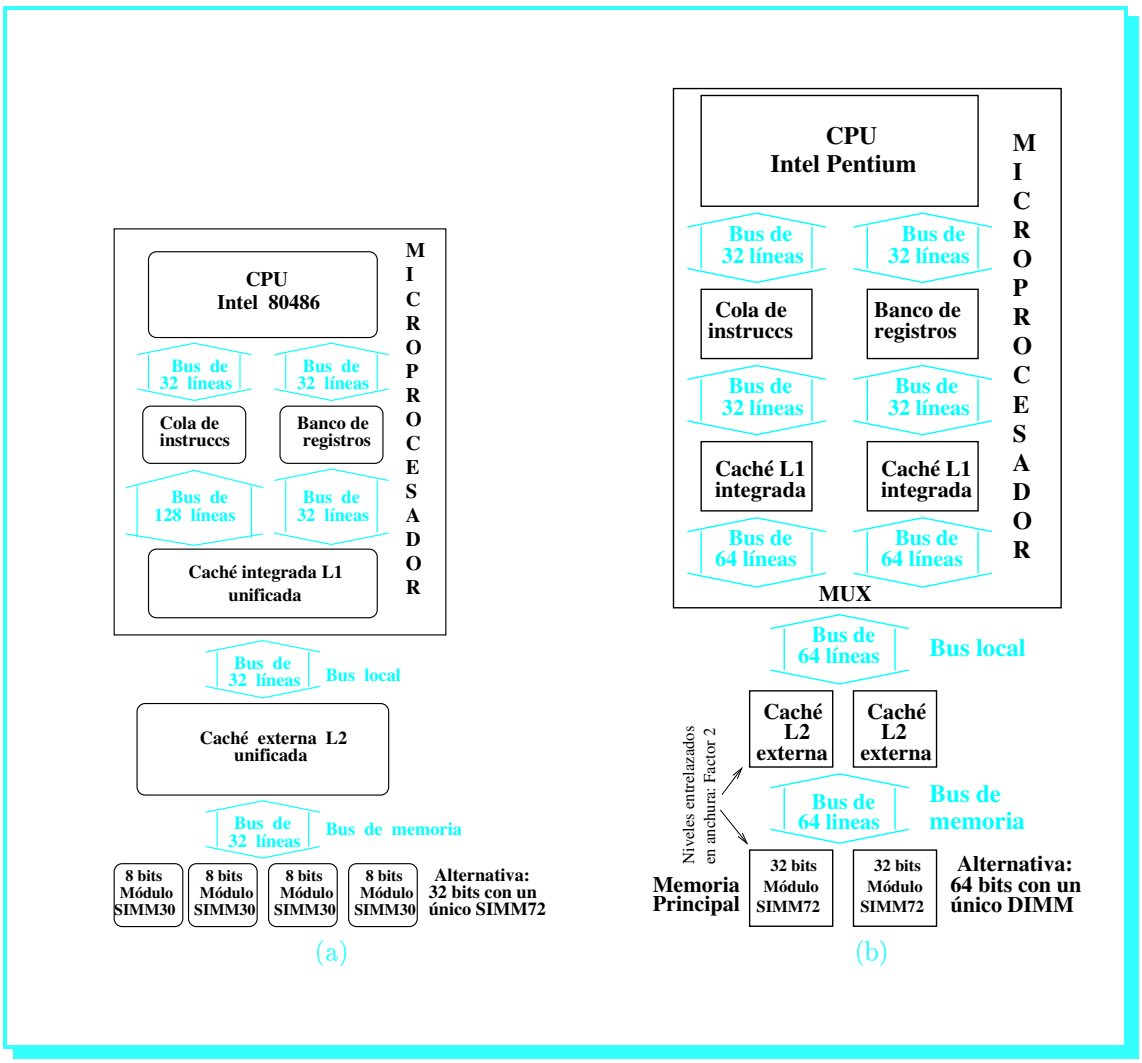


FIGURA 9.2: El sistema de memoria y buses que montaron algunos microprocesadores para PC, desde sus niveles más internos hasta llegar a memoria principal: (a) Intel 80486. (b) Intel Pentium.

pág. 28
pág. 29
SIMM
pág. 30
L1 y L2

Cuando apareció el 80486, la memoria DRAM más avanzada que había en el mercado se presentaba en el formato SIMM de 30 contactos (ver foto 10.3.a y figura 10.4), el cual suministraba únicamente 8 bits de datos. Dado que la placa base tenía una anchura de bus de 32, eran necesarios 4 módulos SIMM entrelazados en anchura para componer un banco de memoria. Posteriormente, apareció la memoria SIMM72 (ver foto 10.5), que posibilitó la implementación de bancos de un solo módulo con anchura igual a la del bus.

Con respecto a la caché, se disponía de una L2 externa implementada en la propia placa base mediante chips de anchura 32 bits, y dentro ya del microprocesador, la caché L1 con un par de particularidades:

1. Era una caché unificada (alojaba tanto datos como instrucciones), en parte debido a que el 80486 carecía de un carácter fuertemente segmentado.
2. La transferencia interna de instrucciones desde aquí a la cola de instrucciones se realizaba por un bus interno de 128 bits que coincidía con la anchura de la línea de caché, permitiendo un fácil llenado de la cola. Por la parte de los datos, la transferencia con el banco de registros y la ALU se realizaba mediante buses internos de 32 bits.

La figura 9.2.a muestra todo el sistema de memoria del microprocesador, tanto en su parte interna como en la organización externa de la placa base.

Quinta generación ◀ 2.2

Con el Pentium llegó el bus de datos de 64 bits, con lo que la memoria principal se montó con bancos de dos módulos SIMM72 de 32 bits. Cuando salió la memoria DRAM en formato DIMM de 168 contactos, muchas placas base incluyeron zócalos DIMM que permitían elegir al usuario entre un banco con los dos módulos SIMM72 entrelazados en anchura o un único módulo DIMM donde el entrelazado en anchura se realizaba internamente al nivel de chip.

DIMM
elección

Las primeras placas base que incluyeron soporte para ambos formatos traían dos bancos:

dos formatos

- El banco 0, que se podía montar con los zócalos SIMM0 y SIMM1 o el zócalo DIMM0.
- El banco 1, utilizado opcionalmente para añadir más memoria llenando los zócalos SIMM2 y SIMM3, o el zócalo DIMM1.

Más adelante, este solape de módulos en el mismo banco desaparecería, pudiendo convivir módulos SIMM y DIMM en una misma placa base.

En lo referente a la caché, la novedad organizacional que introdujo el Pentium fue el entrelazado en anchura de los chips de caché de 32 bits para conseguir la anchura de 64, algo mimético a lo ocurrido con los módulos SIMM72 en la memoria principal, y que también terminaría desapareciendo poco después.

entrelazado

Ya en el interior del microprocesador, nos encontramos con la caché de primer nivel L1 separada para datos e instrucciones, y buses separados para la transferencia de información hacia el interior. Allí nos espera el banco de registros y la cola de prebúsqueda de instrucciones, desde donde la transmisión de datos hacia la CPU también se realiza por vías separadas de 32 bits.

L1 separada

La pregunta que surge en este punto es para qué necesita el Pentium una anchura de memoria de 64 bits si en realidad se trata de un procesador de 32 bits. Responderemos a ella distinguiendo el caso de que se pida una instrucción o un dato a memoria, resolviendo de paso una cuestión ya planteada en el marco de la sexta y séptima generación, donde tanto la anchura de la memoria como la del procesador permanecen en estos mismos valores:

- Si se pide una instrucción, ésta tiene un formato variable en el caso del Pentium (ver figura 4.1). Con instrucciones de 32 bits, la memoria responde proporcionando dos a la vez, una procedente de cada módulo, que viajan juntas por el bus hacia la caché. Recordar que el Pentium es un procesador superescalar de factor dos, por lo que conviene buscar instrucciones de una forma agresiva para evitar que el procesador se quede sin trabajo. Para formatos de instrucción más grandes, el transporte de las instrucciones por el bus será de una por viaje (e incluso puede ser necesario realizar más de un viaje en casos extremos), pero en cualquier caso la información se manipula en múltiplos de 8 bytes, ya que ese es el tamaño del bus por el que viajan, y la línea de caché que los recibe es aún mayor (32 bytes). Desde ahí, la información se distribuye a los dos búfers de prebúsqueda de instrucciones que anteceden a los cauces de ejecución de instrucciones (ver figura 4.2).

  instrucción

  ☛ *Volumen 1*

  ☛ *Volumen 1*

- Si lo que se pide es un dato, cuyo tamaño oscila entre uno y doce bytes, la información se transferirá desde la memoria a la caché de datos también por múltiplos de ocho bytes. Ahora bien, puede que luego la información que se utilice realmente no sea los ocho bytes al completo, sino sólo unos pocos (por ejemplo, cuatro bytes para llenar uno de los registros del banco de 32 bits de propósito general). En ese caso, los demás bytes que han sobrado de la operación de transporte se quedan en caché, en espera de que el procesador los solicite. Esto

  dato

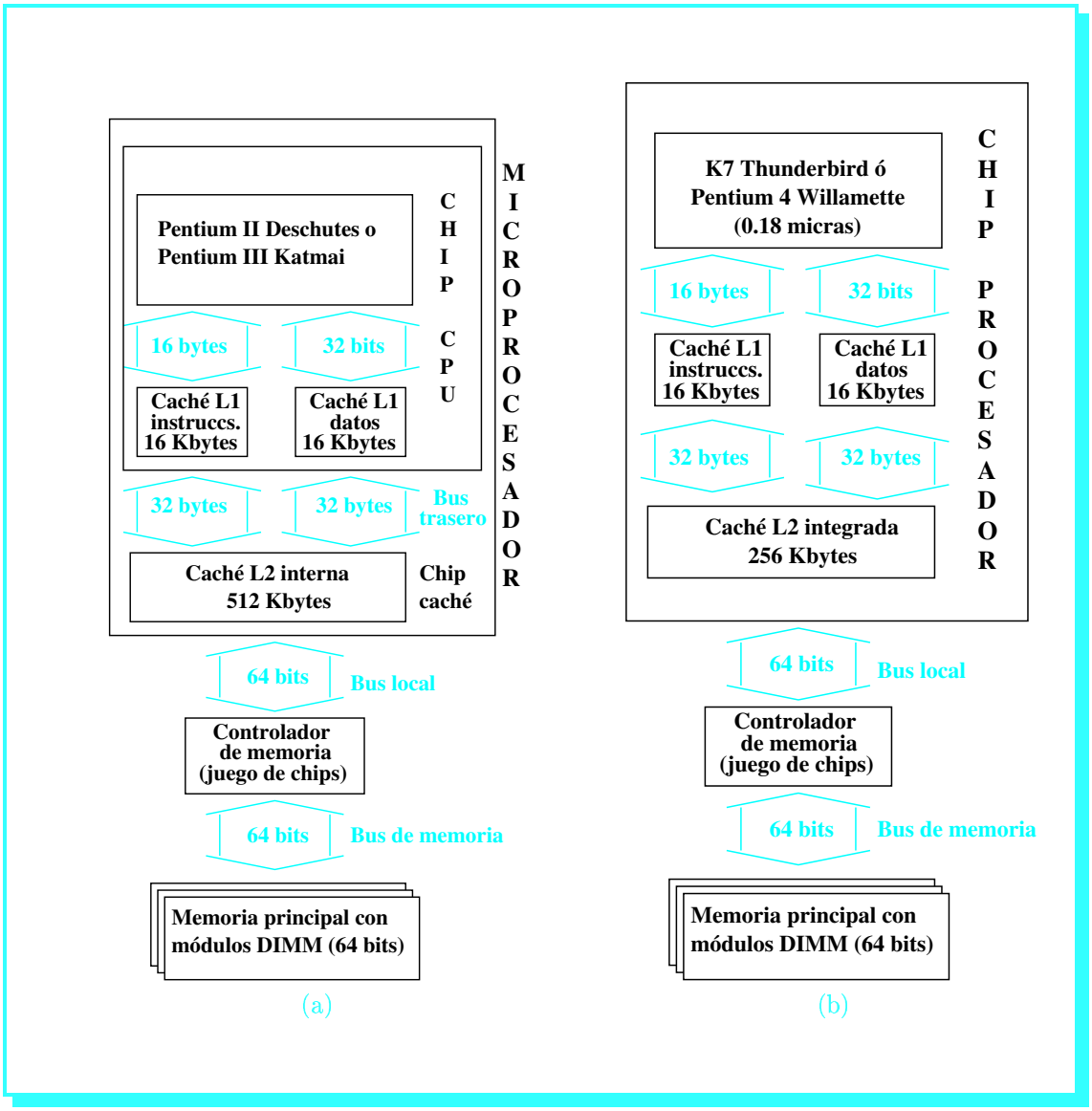


FIGURA 9.3: El sistema de memoria del PC en el marco de la sexta (a) y séptima (b) generación de microprocesadores.

no es baladí, sino algo muy probable dado el principio de localidad espacial que también exhíben las referencias a datos por parte del procesador.

pág. 8
elecciones

El esquema completo de los distintos niveles de memoria y de la dualidad datos/instrucciones puede apreciarse en la figura 9.2.b. Todos los niveles de memoria caché son internos, ya sea al microprocesador o a la placa base, y no van a poder ser elegidos por el usuario. La memoria principal, en cambio, sí nos va a dar la oportunidad de elegir el tipo y tamaño para nuestra configuración. La sección 10.15.2 nos indica las principales directrices para acertar en esta elección.

pág. 98

2.3 Sexta generación

Los módulos SIMM de memoria principal terminaron por extinguirse al final de la quinta generación de microprocesadores. La sexta generación arrancó así con un procesador y unos mó-

dulos de memoria que se encontraban en una anchura común de 64 bits, y el sistema de memoria principal no cambió un ápice durante todo el marco temporal de esa generación.

64 bits

Respecto a la memoria caché, su nivel L2 desapareció de la placa base por la abolición de la caché externa: El procesador de la sexta generación disparó su velocidad abriendo una importante brecha con todo lo que quedaba instalado en la placa base, viéndose multiplicadores incluso por encima de 10 para cubrir el desfase en frecuencia entre uno y otro elemento. Por eso la caché L2 saltó definitivamente a la órbita del procesador, colocándose en un chip aparte pero dentro de la misma placa de circuito impreso que comprendía el microprocesador entendido como producto comercial (ver figura 9.3.a).

caché interna

☛ pág. 10

Séptima generación ◀ 2.4

A nivel organizacional, la memoria principal de la séptima generación no introduce grandes cambios. Se sigue trabajando con módulos de 64 bits de anchura que alimentan a un procesador de su misma anchura, aumentándose el ancho de banda gracias a la frecuencia, donde se introduce un multiplicador de 2x con la llegada de la memoria DDRAM y los módulos de 184 contactos (ver figura 10.7).

DDRAM

☛ pág. 33

Aparece como novedad la memoria RDRAM (ver sección 10.13.6), en la que el bus se estrecha para conseguir un tiempo de ciclo muy bajo y lograr mayores anchos de banda en el contexto de las elevadas frecuencias que ya comienzan a verse en el bus local, el cual deja de ostentar la vitola de cuello de botella del PC.

RDRAM

☛ pág. 73

En relación a la caché, las mejoras en la tecnología de integración, y más concretamente la llegada de las distancias de integración de 0.18 micras en los transistores, precipita la transformación de la L2 interna en integrada, aún a cambio de ciertas concesiones en su tamaño y en el precio final del conjunto.

L2 integrada

La creciente superescalaridad de estos procesadores y el aumento de su frecuencia les hacen depender más de los niveles de caché que del nivel de memoria principal, viéndose ciertas optimizaciones como el doble puerto de acceso en las cachés del K7 o la idea de la caché de traza como forma de optimizar el uso de la caché L1 de instrucciones en el caso del Pentium 4. La figura 9.3.b muestra el sistema general de memoria para un microprocesador de esta generación.

☛ pág. 10

Resumen

La memoria electrónica aparece en los años 50 en forma de anillos de ferrita, experimentando un fuerte auge con la llegada del chip en los 60. A principios de los 70 aparece el microprocesador, cuya velocidad alcanza a la de la memoria a finales de esa década.

nacimiento

En los 80, el auge de la capa software demanda memorias más grandes, que en este período duplican su capacidad cada dieciocho meses, mientras que a la velocidad le lleva la década entera duplicarse. Todo esto motiva la aparición de una jerarquía de memoria con múltiples niveles de caché, seguidos de la memoria principal y el disco.

jerarquía

A comienzos de los 90, el primer nivel de caché se integra en el procesador, y a finales de esa década lo hace el segundo nivel. Durante todo este tiempo, las mejoras se centran en el interfaz entre ésta y memoria principal, así como en explotar la concurrencia entre las vías de comunicación que interconectan todos estos niveles.

interfaz

La tabla 9.2 sintetiza la evolución que hemos narrado a nuestro paso por las últimas cuatro generaciones de microprocesadores.

☛ pág. 12

Generación	Relación anchura memoria principal/procesador y módulos constitutivos de memoria principal			Memoria caché L2		
	Predominante	Alternativa	Tamaño	Tipo	Entrelazado	Tamaño (Kbytes)
4^a	4 x SIMM30	1 x SIMM72	4 Mbytes	No	No	No
5^a	2 x SIMM72	1 x DIMM168	16 Mbytes	Externa	2	256
6^a	1 x DIMM168	2 x SIMM72	64 Mbytes	Interna	1	512
7^a	1 x DIMM184	1 x RIMM184	256 Mbytes	Integrada	1	256

TABLA 9.2: Resumen evolutivo del sistema de memoria de un PC tomando como referencia las cuatro últimas generaciones de microprocesadores.

✍ Cuestionario de evaluación ✍

En las cuestiones que presentan varias respuestas válidas, deberá quedarse con la que considere más exacta y/o completa. Las soluciones a todas las cuestiones se encuentran al final de este volumen.

1 Ordena cronológicamente el nacimiento de los siguientes conceptos

a. La memoria electrónica (ME), la memoria caché (MC), el microprocesador (MI) y el PC.

b. ME, MI, MC y PC.

c. MI, ME, MC y PC.

d. ME, MI, PC y MC.

2 En la década de los 90, la memoria principal ha ganado velocidad gracias sobre todo a las mejoras en

a. La velocidad del transistor.

b. El interfaz de diálogo.

c. La reducción del número de chips que componen un módulo de memoria.

d. Las tres anteriores.

3 La jerarquía de memoria se establece sobre un principio básico. ¿Cuál?

a. Más cara, más rápida.

b. Más rápida, más cerca del procesador.

c. Más grande, más lejos del procesador.

d. Más grande, más lenta.

4 Si la memoria principal de la placa base es la caché del procesador, entonces la caché externa de la placa base es X en el procesador. ¿Quién es X si la analogía se establece en términos de velocidad de acceso a entes de almacenamiento?

a. El directorio caché.

b. Los bancos de registros.

c. La(s) ALU.

d. El conjunto de instrucciones multimedia.

5 ¿Por qué los módulos de memoria SIMM72 (32 bits de datos) pueden montarse aisladamente sobre una placa 80486 mientras que en una placa Pentium deben ir por pares?

a. Porque la memoria no se encuentra entrelazada en el 80486.

b. Porque la anchura del bus de datos externo del 80486 es la mitad que la del Pentium (32 frente a 64).

c. Porque las placas del 80486 funcionaban a la mitad de frecuencia que las del Pentium (33 MHz frente a 66 MHz).

d. Porque la caché L1 del 80486 está unificada, mientras que la del Pentium está separada en datos e instrucciones.

Capítulo 10

Memoria principal

Sumario

La memoria principal es el área de almacenamiento donde se alojan todos los programas que se ejecutan en nuestro PC, tanto las aplicaciones de usuario como los manejadores de dispositivo (drivers) y las estructuras de datos y procesos del sistema operativo.

SECCIÓN 10.1

Etimología

RAM

No son pocos los usuarios que utilizan el término **RAM** para referirse a la memoria principal, quizá por el hecho de que el resto de niveles de la jerarquía de memoria tiene su propio nombre, o porque usan con cierta ligereza un vocablo que desconocen.

aleatoria

volátil

RAM es la abreviatura inglesa de *Random Access Memory* (*Memoria de Acceso Aleatorio*), término que hace referencia a una memoria en la que la palabra a leer/escribir puede seleccionarse libremente indicando su dirección. En la práctica, el término RAM también engloba el carácter volátil de la memoria (pierde su contenido en ausencia de alimentación), frente a la memoria ROM (*Read Only Memory - Memoria de sólo lectura*), cuya información tiene carácter permanente.

todo es RAM

La aleatoriedad en el acceso a memoria es algo presente en casi todas las formas de memoria del PC, a excepción de las memorias magnéticas como la cinta (acceso secuencial) o el disco (acceso por bloques de palabras consecutivas). Su volatilidad, además de a éstas, también excluye a ciertos chips de memoria permanente como la ROM-BIOS en sus variantes de tipo no Flash. Pero exceptuando estos pocos casos, muchos de ellos obsoletos ya, las restantes memorias del PC son todas RAM, desde la memoria principal hasta la caché o el banco de registros.

principal

caché

sinónimos

☛ pág. 16

En la memoria principal del computador, cada celda o bit de información se implementa mediante un minúsculo condensador de unos pocos femptofaradios, siendo el signo positivo o negativo de la carga que almacena lo que determina el valor lógico 0 o 1 de su celda. Esta carga se encuentra en permanente movimiento, tal y como se muestra en la figura 10.1, en contraste con los transistores con que se implementan las puertas lógicas que constituyen la memoria caché, y que exhiben un comportamiento estático. Esta es la justificación de que la memoria principal se conozca también como RAM dinámica (*Dynamic RAM* - DRAM) y la memoria caché como RAM estática (*Static RAM* - SRAM). Sus aspectos diferenciadores se resumen en la tabla 10.1.

☛ pág. 17

comparativa

El coste y espacio físico por celda de las memorias DRAM es bastante inferior al de sus homólogas SRAM, y aunque el tiempo de acceso no sea tan rápido, para cantidades de varios Megabytes es la única alternativa razonable. De hecho, en el mercado actual de la memoria, la DRAM es el producto estrella, siendo sus ventas responsables del 75 % del beneficio obtenido por esta industria. A continuación nos espera un largo camino en su conocimiento.

SECCIÓN 10.2

La operación de refresco

atenuación

Aunque el valor lógico o bit de información de una celda de memoria principal permanezca siempre a 1, su condensador va perdiendo carga internamente a lo largo del tiempo según una

Aspecto	Memoria principal	Memoria caché	Memoria ROM
Naturaleza	RAM dinámica	RAM estática	No volátil
Composición de una celda	Una resistencia y un condensador	De 4 a 6 transistores	Un transistor
Alimentación	Necesaria	Necesaria	Prescindible
Refresco	Necesario	No requiere	No requiere
Direccionamiento	Multiplexación	No mux.	No mux.
Densidad	Elevada	Baja	Muy elevada
Velocidad	Lenta	Rápida	Muy lenta
Potencia disipada	Pequeña	Grande	Muy pequeña

TABLA 10.1: Características de los tipos de memoria aleatoria (RAM) presentes en un computador personal, en los que la memoria principal constituye tan sólo uno de ellos.

curva exponencial negativa. Esto es consecuencia de dos fenómenos físicos:

fuga

❶ El propio condensador tiene una corriente de fuga significativa, lo que conduce a la pérdida de la carga almacenada en unos pocos milisegundos.

compartición

❷ Cuando la celda resulta seleccionada para una operación de lectura, la carga que almacena se comparte con la elevada capacitancia de la línea de datos, que es entre 10 y 20 veces superior, atenuándose por este mismo factor el voltaje que representa la información.

Para evitar la pérdida de la información almacenada en la celda, su carga debe regenerarse periódicamente (ver figura 10.1), y también después de cada acceso a memoria. De esto se encarga el circuito de refresco, que actúa al nivel de fila sobre la malla bidimensional de celdas en que se encuentran organizados los chips, utilizándose un contador interno desde el controlador de memoria para recorrer las filas de manera circular.

pág. 17

regeneración

Ejemplo 10.1: UNA TÍPICA OPERACIÓN DE REFRESCO

Dentro de la séptima generación, lo normal es que el refresco siga una operativa estándar establecida por el JEDEC (Junction Electronic Devices Engineering Council), con un período de refresco de 64 ms. sobre un total de 2K, 4K u 8K filas, lanzándose una operación de refresco desde el controlador a una fila diferente cada 32, 16 u 8 μsg., respectivamente.

En los módulos de memoria para PC, la temporización de refresco (reloj) se emite desde el controlador de memoria, aunque no así el direccionamiento, que se realiza internamente desde los propios chips del módulo en un modo denominado CAS BEFORE RAS en el contexto de las memorias asíncronas (BEDO y sus predecesoras) y AUTO REFRESH en el de las memorias síncronas (SDRAM en adelante). Dentro de éstas últimas, y coincidiendo con la aparición de los modos de bajo consumo en el PC, entra en escena el modo SELF REFRESH, en el que también el reloj se desactiva en la placa base, siendo el módulo autosuficiente para retener los contenidos gracias al empleo de un reloj interno a sus chips. Si nuestros módulos disponen de esta facultad, pueden retener la información incluso con el PC apagado en posición `Stand-By`, lo que a veces es aprovechado en el PC para restablecer de forma casi instantánea el aspecto en que abandonamos su

AUTO REFRESH

SELF REFRESH

Año	Generación	Tipo de módulo	Latencia del módulo	Período de refresco	Unidad de refresco
1996	Quinta	EDO	50 ns.	7 ms.	Filas de 2K
1999	Sexta	SDRAM	10 ns.	64 ms.	Filas de 4K
2002	Séptima	DDRAM	7 ns.	64 ms.	Filas de 8K

TABLA 10.2: Períodos de refresco típicos para memoria EDO, SDRAM y DDRAM de Micron. La latencia es tiempo de respuesta en EDO y tiempo de ciclo en SDRAM/DDRAM.

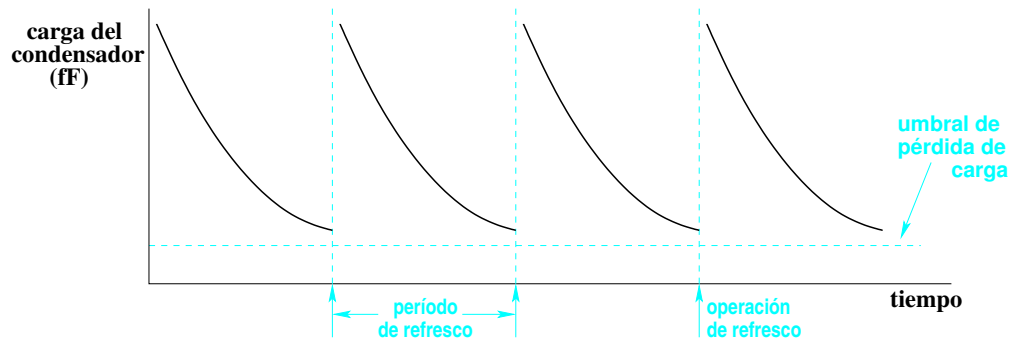


FIGURA 10.1: Evolución de la carga del condensador presente en una celda lógica de un bit en una RAM dinámica. El circuito de refresco de la memoria recarga el condensador con una frecuencia tal que impida a su carga traspasar el umbral de pérdida del valor lógico que almacena (en este caso un '1'). El funcionamiento es completamente análogo para cargas negativas (valor lógico '0'), presentando en este caso la gráfica una simetría de espejo en abcisas con respecto a ésta.

última sesión en uso. En el contexto de los computadores portátiles, estos modos de autorefresco son empleados de forma extensiva para reducir el consumo de energía.

`tiempo de refresco`

El tiempo de refresco del chip de memoria, durante el cual no puede ser utilizado para operaciones de lectura/escritura, ha ido en aumento como consecuencia del creciente número de celdas por chip (ver tabla 10.2), aunque la velocidad de la memoria actúa como atenuante respecto al estorbo que supone la operación de refresco en su operativa. En general, el tiempo en que las celdas son acaparadas para ser refrescadas es inferior al 5 % del total en que quedan disponibles para un acceso, por lo que la incidencia del refresco en el rendimiento del chip no es relevante.

SECCIÓN 10.3

El controlador de memoria principal

`ubicación`

Así como el procesador se divide en unidad de control y unidad de proceso, la memoria requiere un **controlador** que gobierne las operaciones a realizar sobre sus áreas de datos. Este controlador se sitúa en el puente norte del juego de chips (ver figura 10.14), y se encarga de dialogar con los módulos de memoria a través del bus de memoria, que aglutina líneas de direccionamiento, datos y control que discurren por la placa base.

☛ pág. 44

`bus`

`Registered`

En la gama alta de las memorias fabricadas para PC, sobre todo servidores, podemos encontrar módulos de memoria denominados **Registered**, que ya implementan el controlador de memoria en un chip propio anexo a los chips de datos en sí, en una variante que nos recuerda a lo que le

pág. 39 sucede al chip SPD de autoconfiguración (ver sección 10.9).

Los módulos de memoria Registered resultan muy poco frecuentes en las arquitecturas domésticas, y desde aquí no los recomendamos. Primero, porque son incompatibles con los normales, con lo que constituyen una buena ocasión para cerrarse unas cuantas puertas en la interoperabilidad futura de nuestro sistema, y segundo, porque se paga por ellos un sobreprecio sin ganar contraprestación a cambio, ya que el controlador de memoria ya está incluido de serie en el juego de chips de nuestra placa base. Su única ventaja reside en reducir los diálogos entre el puente norte del juego de chips y los zócalos de memoria. Puesto que los módulos Registered apuntan a lo más alto del escalafón en las memorias para PC, la gran mayoría de ellos utiliza para los chips de datos la variante ECC (ver sección 10.5.2.3).

pág. 25

SECCIÓN 10.4

Parámetros de funcionalidad y rendimiento

tamaño

El parámetro que describe la funcionalidad de la memoria es su **tamaño** en Megabytes, magnitud que representa 1024 x 1024 bytes. Una evolución del tamaño y los valores actuales más recomendables pueden consultarse en la sección 10.15.2.3.

pág. 99

Respecto al rendimiento, los parámetros que mejor lo cuantifican son fundamentalmente dos:

latencia

❶ La **latencia**, o tiempo de acceso a memoria, que caracteriza el retardo asociado a la consulta de sus contenidos, y se halla condicionada por la tecnología utilizada en la fabricación de la memoria (ver tabla 9.1). Normalmente, se mide por el **tiempo de respuesta** del chip, el que transcurre desde que éste recibe la dirección de una celda hasta que devuelve el dato que se encuentra en ella.

pág. 7

tiempo de ciclo

En diseños síncronos y segmentados, como la SDRAM y DDRAM, resulta más realista medir la **latencia** utilizando el **tiempo de ciclo** de la memoria, que es el que transcurre entre la aceptación de dos direcciones consecutivas. Visto desde un prisma más global, el tiempo de ciclo marca el ritmo de salida de datos del cauce segmentado de la memoria en consonancia con el bus al que ésta se encuentra conectada.

ancho de banda del bus

❷ El **ancho de banda**, que ya involucra a la velocidad de transporte de datos por el bus de memoria a través de su anchura y frecuencia de trabajo. El ancho de banda de un bus es el producto de estos dos factores, expresándose así su magnitud en $MHz \times bytes = Mbytes/segundo$ [1].

ancho de banda de la memoria

tiempo de servicio

El ancho de banda del bus interviene por partida doble en el ancho de banda efectivo de la memoria ya que determina el número de ciclos de reloj necesarios tanto para enviar la dirección a la memoria como para recibir sus datos. Estos dos sumandos, junto con el correspondiente a la latencia de la memoria engrosan el tiempo de servicio de la memoria, que contabiliza todo el tiempo que realmente transcurre entre que el dispositivo cursa una petición a memoria y le llega el dato solicitado. Dividiendo el volumen de datos servidos por este tiempo, tendremos el ancho de banda efectivo de la memoria en número de palabras transferidas por ciclo de reloj.

[1] Esta fórmula no es matemáticamente exacta, al asumir que 1 Mbyte es un millón de bytes. El hercio utiliza escala decimal y factores 10^3 en sus múltiplos de Kilo- y Mega-, pero el byte emplea escala binaria y factores de 2^{10}, por lo que $1Mbyte = 1024 \times 1024$ bytes $= 1{,}048{,}576$ bytes. El presente capítulo mezcla un puñado de magnitudes binarias provenientes de la capacidad de almacenamiento con otras tantas decimales ligadas sobre todo a la frecuencia, y en este caldo de cultivo, aplicar factores de corrección de manera sistemática supone traspasar la frontera del amor al odio al cálculo matemático. Lamentamos decepcionar aquí a los más puristas, pero en esta ocasión ha podido más el pragmatismo ingenieril del autor que su rigor científico. La anécdota del capítulo 3 profundiza un poco más en la pérdida de exactitud que esta práctica conlleva.

Ejemplo 10.2: Cálculo del ancho de banda efectivo de memoria principal

Calculemos el ancho de banda efectivo en una operación típica de llenado de una línea de caché desde memoria principal. Supongamos que la caché tiene líneas de 8 bytes de datos, y se comunica con memoria principal a través de un bus de memoria de 8 bits y 400 MHz; supongamos también que la anchura de memoria principal es también de 8 bits, mientras que su tiempo de respuesta es de 10 ns.
La penalización por fallo de caché en este caso sería:

- Un ciclo de 400 MHz (esto es, 2.5 ns.) para el envío a la memoria de la dirección (bus de direcciones) y el comando de lectura (bus de control).
- 80 ns. a razón de 10 ns. para leer cada uno de los 8 bytes de memoria.
- 8 ciclos más de 400 MHz (esto es, 8 x 2.5 ns. = 20 ns.) para el envío de los 8 bytes a la caché (bus de datos).

Y el cómputo de los diferentes parámetros quedaría como sigue:

- Tiempo de ciclo de la memoria: 2.5 ns.
- Tiempo de respuesta para la línea de caché: 80 + 20 = 100 ns.
- Tiempo de servicio para la línea de caché: 2.5 + 80 + 20 = 102.5 ns.
- Ancho de banda efectivo: 8 bytes / 102.5 ns. = 74.43 Mbytes/sg.

Tradicionalmente, la latencia ha sido un parámetro más ligado a la caché por afectar directamente a su tiempo de acceso en caso de fallo. En cambio, el ancho de banda se relaciona más con el sistema de entrada/salida, donde los dispositivos transfieren la información por sectores. Pero entre ellos cabe destacar a la propia memoria caché, en la que los diseñadores pueden aprovechar un elevado ancho de banda para aumentar el tamaño de la línea de caché.

`subsistemas asociados`

Analogía 10.1: La cinta de equipajes y los parámetros de la memoria

Para asimilar mejor los conceptos de latencia y ancho de banda podemos fijarnos en la forma en que trabajan las cintas de recogida de equipajes de los aeropuertos. La cinta es el bus de memoria, y la velocidad a que ésta se desliza es su frecuencia. El tiempo que transcurre entre que llegamos a la cinta y aparece la primera maleta es la latencia; a partir de ahí, las maletas se suceden a un ritmo constante, que se corresponde con el tiempo de ciclo, y que los operarios sincronizan con la velocidad de la cinta de la misma forma que se encuentran ligados el tiempo de ciclo y la frecuencia del bus. Nuestro tiempo total de espera es el tiempo de respuesta, y sería el que transcurre entre que llegamos a la cinta y nos hacemos con el equipaje.

Si suponemos que nuestro equipaje sale el primero y se compone de cuatro maletas, el equipaje sería la línea de caché, y cada maleta, una palabra de memoria. El tiempo de respuesta para la línea de caché depende primero de la latencia de la memoria (operarios), después de su tiempo de ciclo (ritmo de colocación de maletas), y finalmente, de la frecuencia del bus (velocidad de la cinta que trae los datos hacia nosotros). El tiempo de servicio es el lapso que transcurre entre que salimos del avión sin equipaje y salimos del aeropuerto cargados de maletas. Aunque didáctica, esta visión del sistema carece de realismo por suponer que estamos solos en el aeropuerto. Todos sabemos que nuestro retraso depende en gran medida del grado de congestión que sufra la terminal de pasajeros. Cuantificar este efecto es algo tremendamente complejo: Lo mismo que le ocurre a la memoria de nuestro PC.

En la cinta de equipaje del aeropuerto	En memoria principal
Tardanza de los operarios	Latencia de la memoria
Ritmo de salida de equipajes por la cinta	Tiempo de ciclo de la memoria
Velocidad de la cinta	Frecuencia del bus de memoria
Nuestro equipaje	La línea de caché
Cada maleta	Una palabra de memoria
Tiempo total en retirar el equipaje	Tiempo de respuesta línea de caché
Lapso entre salir del avión y del aeropuerto	Tiempo de servicio
Congestión de la terminal de pasajeros	Compartición de la memoria y el bus

complicaciones: La arquitectura de un PC actual complica notablemente el modelo de memoria expuesto:

contención
- Externamente, hay que cuantificar la contención producida por el uso compartido de los recursos de memoria desde los diferentes dispositivos interesados en utilizarla de forma simultánea, hecho que se agrava con la moda multimedia. Ahora no es sólo el procesador quien utiliza la memoria, sino también la tarjeta gráfica AGP con su mapa de texturas y el subsistema de entrada/salida con sus múltiples formas de coprocesamiento adicional.

paralelismo
- Internamente, se trata de aprovechar este caudal de peticiones para simultanear el servicio de las mismas empleando alguna de las formas de paralelismo que ya vimos para el procesador, como la segmentación.

simplificaciones Los ocho canales de DMA (acceso directo a memoria) con que cuenta un PC alivian un poco la espera al nivel externo, y las sucesivas mejoras en el interfaz de memoria han aprovechado cada vez más la concurrencia al nivel interno; sendos efectos que cuesta sobremanera contabilizar en un modelo sencillo del rendimiento de la memoria como el que aquí proponemos. Por ello, el núcleo de nuestra vara de medir la memoria quedará conformado por la latencia y el ancho de banda, enriqueciendo nuestro análisis con comentarios relativos a la contención y el paralelismo siempre que alguno de ellos contribuya de forma relevante.

SECCIÓN 10.5

Fiabilidad

errores: En un chip de memoria se pueden producir fallos de dos tipos:

permanentes
- **Permanentes**. Motivados por alguna anomalía física como un exceso térmico, un golpe o una impureza en su fabricación. Atañen a la memoria como a cualquier otro chip.

- **Transitorios**. Provocados puntualmente por interferencias externas. Son más característicos de la memoria y por ello los trataremos aquí de forma específica. `transitorios`

La mayor parte de los errores transitorios en una memoria DRAM actual están provocados por los rayos cósmicos, emisores de radiaciones de elevada energía cuya carga puede modificar la del condensador de una DRAM lo suficiente como para confundir su valor lógico. `rayos`

Ejemplo 10.3: La tasa de errores por rayos cósmicos en memoria DRAM

En el laboratorio IBM Watson Research Center se midió para una muestra de chips DRAM de marcas clónicas una tasa de error de 5950 fallos cada mil millones de horas de funcionamiento al nivel del mar. La misma prueba dentro de una cueva subterránea ubicada bajo 15 metros de roca no produjo ningún error.
Esto quiere decir que en un sistema actual dotado de 32 chips de memoria DRAM se produce un error transitorio cada seis meses debido a la incidencia de rayos cósmicos.

Otros fenómenos que pueden provocar errores transitorios, aunque ya a menor escala son: Inestabilidades o ruido en la línea eléctrica, electricidad estática, configuración incorrecta (ya sea en el interfaz empleado o en la velocidad), interferencias de radiofrecuencia y defectos de temporización. Los dos primeros pueden provocar también errores permanentes. `otras causas`

Además, cada región tendrá su tasa global concreta: Por ejemplo, en zonas de clima estable como Málaga donde las tormentas son infrecuentes, la tasa de errores quedaría por debajo de uno por año, eso sí, con permiso de la paupérrima calidad del suministro eléctrico que padecemos. `subjetividad`

En última instancia, todo depende de para qué utilicemos el PC. En el mío se escribe este libro, y para no echarle la culpa a la memoria sobre las numerosas erratas que tiene, prefiero pagar el 10 % de sobreprecio que supone suscribir una buena cobertura como ECC. A los precios de 2003, supone pagar 6 € por cada 256 Mbytes. `coste`

Quizá después de leer el riesgo 10.1 usted también decida suscribir algún seguro, así que le vamos a mostrar las cuatro coberturas más comunes. De menor a mayor fiabilidad y sobrecoste, las hemos denominado paridad aparente, paridad real, ECC sobre paridad y ECC real. Presentaremos primero los esquemas reales para posteriormente derivar de éstos sus sucedáneos. ☛ *pág. 22* `coberturas`

Paridad

◀ 5.1

Para aumentar la fiabilidad del sistema de memoria, muchas implementaciones comerciales añaden bits de información redundante que permiten detectar los errores producidos durante el proceso de lectura o escritura de la memoria. Estos bits se denominan **bits de paridad**. `bits de paridad`

El bit de paridad acompaña a cada byte de datos, completando la palabra con un cero o un uno, de manera que la información tenga siempre un número par de unos (en el caso de paridad par). De igual forma, si elegimos la convención opuesta de paridad impar, todas las palabras resultarán en un número impar de unos ayudadas por el contenido del bit de paridad. `par` `impar`

Cuando se encuentra una palabra que no cumple con la convención utilizada, se detecta un error y se genera una interrupción no enmascarable (NMI) que llama a una rutina alojada en la BIOS para mostrar en pantalla un mensaje. En el caso del fabricante AMI, este mensaje es “ON BOARD PARITY ERROR ADDR (HEX) = XXXXX”, y tras él se insta al usuario a que desactive la `detección`

interrupción NMI, reinicie el sistema o prosiga con la ejecución. Recomendamos la última opción, pero sólo para salvar nuestras tareas antes de reiniciar, y siempre en un área de memoria nueva (preferiblemente un diskette), nunca sobreescribiendo los ficheros existentes. De lo contrario, incurriremos en el riesgo 10.1.

☛ pág. 22

Riesgo 10.1: Inconsistencias tras un error de paridad

Si el sistema ha detectado un error de paridad y tras informarnos le ordenamos proseguir la ejecución como si nada hubiera sucedido, la mayoría de las veces todo va a ir bien. Pero conviene saber que esto es una temeridad si uno se encuentra realizando tareas donde la exactitud de los datos es un aspecto crítico, e incluso que existe una ínfima probabilidad de que el error se propague sobre otras áreas de datos del dispositivo que sufrió el error. Muchos dispositivos utilizan la memoria principal para guardar información importante para el funcionamiento de sus *drivers*, y otros guardan en ella metadatos, esto es, descriptores que registran cómo están organizados sus datos (las FAT de disco son un buen ejemplo). De ocurrir el error en estas zonas, el dispositivo puede quedar en un estado inconsistente. Por eso, si salvamos la información pendiente sobre un diskette, obtendremos una copia sin haber comprometido la integridad del disco duro.

carencias

La paridad sólo detecta un número impar de errores, pues un número par de errores camufla el dato como correcto (aunque en la práctica esta posibilidad apenas supone el 2 % de los errores).

distinción

Si se opta por adquirir una memoria con paridad, ésta vendrá equipada con chips adicionales que romperán el número potencia de dos que siempre respetan los chips de datos. Por lo tanto, si el número de chips de datos de que dispone el módulo no es potencia de dos, entonces se trata de un módulo con paridad.

coste

pág. 100 ☛

La paridad incrementa el número de celdas teórico en un 12.5 % (un bit por cada ocho de datos), aunque los datos reales del mercado arrojan un sobreprecio del 9 % - ver tabla 10.23.

compatibilidad

La paridad también influye en la compatibilidad e interoperabilidad con otras partes del sistema: Mientras que la memoria con paridad puede montarse sobre cualquier sistema al margen de que su controlador gestione la paridad o no, la memoria sin paridad no funciona sobre controladores con paridad. Si éste es su caso, deberá desactivar la comprobación de paridad en el controlador de memoria, algo que normalmente puede realizarse desde el firmware de la BIOS (consultar la opción Memory Parity/ECC Check del menú BIOS Features Setup - sección 24.3.3 - y su complementaria para las líneas del bus de memoria DRAM Data Integrity Mode en el menú Chipset Features Setup - ver sección 24.3.4).

Volumen 4 ☛

Volumen 4 ☛

5.1.1 Utilización

4ª generación

A finales de la cuarta generación, las configuraciones predeterminadas para PC no incluían paridad, aunque era posible solicitar un sistema con paridad pagando un sobreprecio.

5ª generación

La quinta generación supuso un duro golpe para los sistemas con paridad: El primer juego de chips para Pentium, el 430FX, carecía de paridad; de todos sus hermanos mayores, sólo la incluyó el 430HX, el menos popular de todos ellos. Se produjo entonces una masiva migración de los fabricantes hacia la memoria sin paridad, buscando precios más competitivos para el entor-

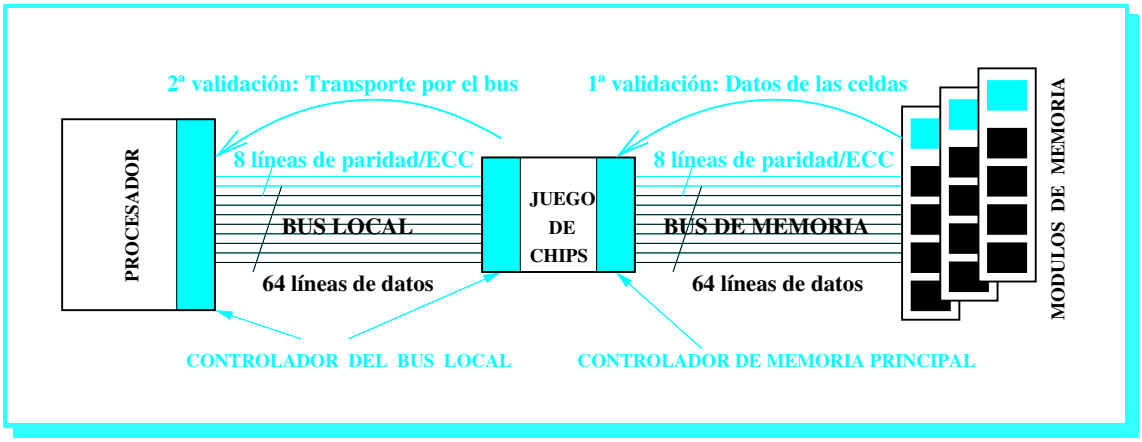


FIGURA 10.2: La verificación de paridad desde los distintos controladores del sistema: En el juego de chips se validan los datos de los chips y su transporte por el bus de memoria, y en el procesador se valida el tráfico por el bus local.

no doméstico, y dejando a las arquitecturas más fiables para los servidores y los equipos de la administración y las grandes corporaciones.

6ª generación

Afortunadamente, la sexta generación cambió ese sentimiento del mercado. Ocurre a veces que la toma de decisiones con mucha antelación sirve para enderezar rumbos equivocados a posteriori: El Pentium Pro comenzó a diseñarse a principios de los 90 cuando la paridad aún gozaba de aceptación, dotándose al procesador con un bus local que en su parte de datos contaba junto a las 64 líneas `D[63:0]` (Data) con 8 líneas adicionales `DEP[7:0]` (Data ECC Protection) que transportaban los bits de paridad hacia el procesador, donde se podía comprobar la integridad en el envío de los datos por el bus. Algunos fabricantes de memoria no quisieron desaprovechar este baluarte, y retomaron los esquemas de paridad. Posteriormente, casi todos los juegos de chips de esta generación también incluyeron el soporte para paridad y ECC.

en bus local DEP

En el lado del procesador hubo poco apoyo, ya que los sucesores `Pentium II` y `III` tuvieron esta opción desactivada por defecto, a pesar de que con ella la paridad puede aprovecharse para validar también las transferencias de datos por los buses. La figura 10.2 refleja esta situación, donde el tramo que comprende el bus de memoria realiza la comprobación en el controlador de memoria validando la integridad de los datos en los chips de los módulos y su transporte, y el siguiente tramo hacia el procesador se valida desde el controlador del bus local.

tramos

5.1.2 Paridad aparente

7ª generación

Con la llegada de la séptima generación (Pentium 4), Intel abolió este esquema, y las líneas `DEP` fueron suprimidas del patillaje. Creemos que una de las claves para justificar esta decisión reside en la fuerte incompatibilidad que acarrea: Piénsese que un procesador con líneas `DEP` obliga a la memoria a suministrar sus valores, y éstos proceden de unos chips de paridad que no siempre están presentes. Para esquivar esta incompatibilidad, los controladores de memoria de aquella época pusieron en práctica una paridad artificial, consistente en generar por su cuenta el bit de paridad que el procesador esperaba recibir. Como consecuencia, se perdía tiempo en generar y posteriormente comprobar la validez de un test que era absolutamente estéril.

artificial

distinción

Esta actitud de los fabricantes se justifica económicamente porque las celdas de paridad en la memoria son mucho más caras que las funciones de generación de estos bits en su controlador. Para distinguir si su memoria es de paridad aparente, debe localizar en sus módulos un pequeño chip que no se parece a los de memoria, y que suele disponer en su lomo del logotipo GSM, el principal fabricante de este generador de paridad artificial.

Esquema de verificación	Módulo SIMM30		Módulo SIMM72		Módulo DIMM168/184	
	Anchura	Incremento	Anchura	Incremento	Anchura	Incremento
Sin paridad	8 bits	0 %	32 bits	0 %	64 bits	0 %
Con paridad	9 bits	12.5 %	36 bits	12.5 %	72 bits	12.5 %
Con ECC	11 bits	37.5 %	39 bits	21.8 %	72 bits	12.5 %

TABLA 10.3: Anchura en bits de los módulos de memoria en función del esquema utilizado para la verificación de sus datos. El incremento del coste es teórico, no valor de mercado.

5.2 ▶ECC

código Hamming

ECC *(Error Correction Code)* es una extensión del mecanismo de paridad cimentado sobre la base matemática del código Hamming. Todas las palabras de la memoria deben pertenecer a este código, compuesto por un campo de datos D y otro de información redundante R obtenido a partir de aquéllos por simples transformaciones lógicas.

corrección

El código quedará definido para que en caso de producirse un error en el `i`-ésimo bit de datos, el campo R alcance precisamente el valor `i`, con lo que cambiando el valor de ese bit habremos corregido el error sin que el usuario lo haya advertido.

5.2.1 Utilización

SEC-DED

El código ECC más utilizado en las memorias se denomina **SEC-DED** *(Single Error Correction - Double Error Detection)* porque permite corregir un error y detectar dos en la misma palabra. En este último caso, se informa al usuario de la eventualidad como ya vimos en paridad.

coste

El error doble podría corregirse ampliando el código con más bits de redundancia, pero su probabilidad de ocurrencia es de tan sólo el 2 % del total de errores de la memoria (el 98 % restante corresponde a errores simples). Puesto que corregir dos errores cuesta más que en el caso simple, la relación beneficio/coste se sitúa claramente en nuestra contra.

uso

La memoria ECC se utiliza sobre todo en el segmento de los servidores y estaciones de trabajo donde la fiabilidad es un aspecto primordial.

distinción

Para saber si su PC dispone de ECC, al nivel del controlador de memoria deberá fijarse en las especificaciones del puente norte del juego de chips, y al nivel de módulo, en su anchura de datos (ver tabla 10.3). Por nuestra propia experiencia, diremos que casi todos los PC domésticos suelen tener controladores ECC, mientras que para los módulos de memoria ocurre justo al contrario: Lo normal es que no lo lleven. Y no nos sorprende que sea así, pues el aumento de complejidad en el chip controlador es irrisorio, mientras que para el módulo sí es apreciable (del 21.8 % en memorias de 32 bits y del 12.5 % en memorias de 64 bits según datos adjuntos en la tabla 10.3).

5.2.2 ECC sobre módulos con paridad

sucedáneo

Al igual que la paridad, ECC también dispone de un sucedáneo introduciendo cierto artificio desde el controlador de memoria. Son sistemas que ofrecen la verificación ECC en el controlador de memoria, pero cuyos módulos de memoria y líneas de bus sólo disponen como bits de redundancia de los bits adicionales definidos en los esquemas de paridad.

la clave

Este esquema híbrido nos sirve para recalcar que el tipo de verificación de errores que se utilice queda determinado por el controlador de memoria mucho antes que por sus módulos. Si el

controlador tiene implementado un amplio repertorio de esquemas de testeo, puede autoconfigurarse en función del módulo de memoria que detecte al otro lado (o esperar su programación desde la BIOS), y emplear las celdas de información redundante para seguir el esquema seleccionado. Para eso es él quien decide tanto el formato de las palabras de memoria cuando las escribe como las comprobaciones a la hora de leerlas.

Lo único que realmente exigirá el controlador al módulo es el espacio de almacenamiento necesario para la información redundante. Pero los esquemas de paridad y los de ECC convergen en este sentido, puesto que el primero crece de forma lineal, mientras que el segundo lo hace de forma logarítmica. Esta tendencia puede apreciarse en la tabla 10.3, donde para ocho bits de datos hay tres de ECC por uno de paridad, mientras que para 64 bits ambos están empatados en ocho bits. El empate propicia que a partir de los módulos DIMM se puedan comprar módulos con paridad para montarlos sobre placas base con ECC y aprovechar ECC lícitamente. Sobre los SIMM, la estrategia que siguieron los controladores ECC para habilitar esto mismo fue realizar la verificación a nivel de banco, esto es, conjuntamente para cada dos SIMM sobre una anchura total de $32 + 32 = 64$ bits y contando con $4 + 4 = 8$ bits adicionales.

convergencia

☛ pág. 24

5.2.3 Registered ECC

Con esta denominación comercial se designan aquellos módulos de memoria que conjugan el esquema ECC con la variante Registered que implementa el controlador de memoria internamente en sus módulos (ver sección 10.3). Con los circuitos de detección y corrección de errores ECC incorporados al propio módulo, los datos salen ya validados hacia el bus de memoria. El controlador de este bus, presente al otro lado en el puente norte del juego de chips, puede a su vez incluir soporte adicional para ECC, pero en ese caso su chequeo únicamente validará el transporte por el bus de memoria.

☛ pág. 17

Los módulos de memoria *Registered ECC* están orientados a los sistemas más comprometidos con la fiabilidad de la información que manejan, y suponen el peldaño de gama más alta en este sentido, no siendo aconsejable su uso entre los usuario de PC medio por su improbable interoperabilidad con placas base más convencionales.

uso

conflictos

SECCIÓN 10.6

Conexión a la placa base

La **forma** en que los chips de memoria se ensamblan a la placa base del PC e interactúan con él es otro de los criterios que podemos utilizar para su clasificación, siendo determinante para mejorar algunos de sus aspectos más importantes en el futuro:

importancia

1. Tamaño. La posible ampliación del número de Mbytes disponibles.
2. Velocidad. La sustitución por otros chips más rápidos.
3. Montaje. La complejidad en la liberación de los chips viejos e inserción de los nuevos.

La interacción física entre la memoria y la placa base ha pasado desde sus inicios por tres **variantes** que describiremos siguiendo un orden cronológico.

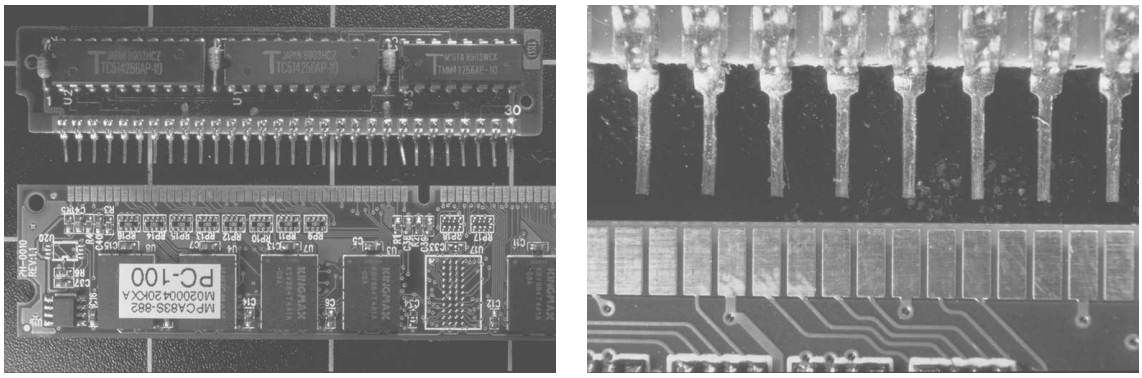
variantes

Púas: SIPP

◀ 6.1

Empaquetado de pines alineados (del inglés, *Single In-line Pin Package*). La memoria se presenta en una placa de circuito impreso o PCB (del inglés, *Printed Circuit Board*), que alberga

descripción



(a) (b)

FOTO 10.1: (a) Aspecto de un módulo de memoria en formato SIPP con sus características patillas alargadas (arriba), en contraste con un módulo de memoria con contactos (abajo). (b) Detalle del patillaje.

un número variable de chips dependiendo del tamaño. Emplea unas púas alargadas para su conexión a un único zócalo longitudinal residente en placa base. La foto 10.1 muestra su aspecto comparado con los contactos de los módulos SIMM/DIMM actuales.

☛ pág. 26

desventajas

Esta variante es la más barata en su fabricación, pero ocupa un mayor espacio en la placa base, se asegura sobre el zócalo de manera poco firme, dificulta la ventilación de los chips de memoria y ofrece un menor grado de resistencia a la corrosión. Por todo ello, resulta poco menos que un hallazgo encontrar memoria SIPP en los computadores actuales. El último vestigio que conocemos de ella lo localizamos en la memoria de 8 bits y 30 contactos para procesador 80386 y 80486, siendo posible convertirlo a módulo SIMM30 soldando los pines a sus respectivos contactos, o más elegantemente, adquiriendo el conversor a zócalo SIMM30 que comercializaron algunas firmas.

6.2 ▶Patillas: DIP

descripción

Siglas también procedentes del inglés, *Dual In-line Package*, o *empaquetado de doble fila*. Se trata de un chip de memoria rectangular con dos filas de patillas dispuestas en paralelo sobre sus dos aristas más largas, y aunque es el formato utilizado por los PC más antiguos (8088), su

vigencia

vigencia llega más cercana a nuestros días que el SIPP:

- En memoria principal, se estuvo utilizando hasta los tiempos del procesador 80386, cuyo controlador de memoria admitía un rango de hasta 1 Megabyte, disponiendo la placa base de 36 huecos (cuatro filas de nueve), de tal forma que cada fila proporcionaba al sistema 256 kilobytes de datos y 256 kilobits de paridad.

aplicación

- En memoria de vídeo, se utilizó hasta la sexta generación con EDO-RAM e incluso SDRAM, gracias a su compacta presentación para cantidades próximas al Mbyte y a la posibilidad de habilitar zócalos individuales para los chips de ampliación posterior. En la séptima generación, la anchura de estos chips crece hasta los 32 bits y se hace necesario migrar al encapsulado QFP primero y al BGA después, donde el patillaje se encuentra ya soldado a la placa de circuito impreso de la tarjeta gráfica. La foto 10.2 muestra las diferencias de estas variantes con el formato SIMM/DIMM, que pasamos a ver seguidamente.

pág. 27 ☛

6.3 ▶Contactos: SIMM/DIMM/RIMM

descripción

Módulos de memoria individual/dual/Rambus alineada (del inglés, *Single/Dual/Rambus*

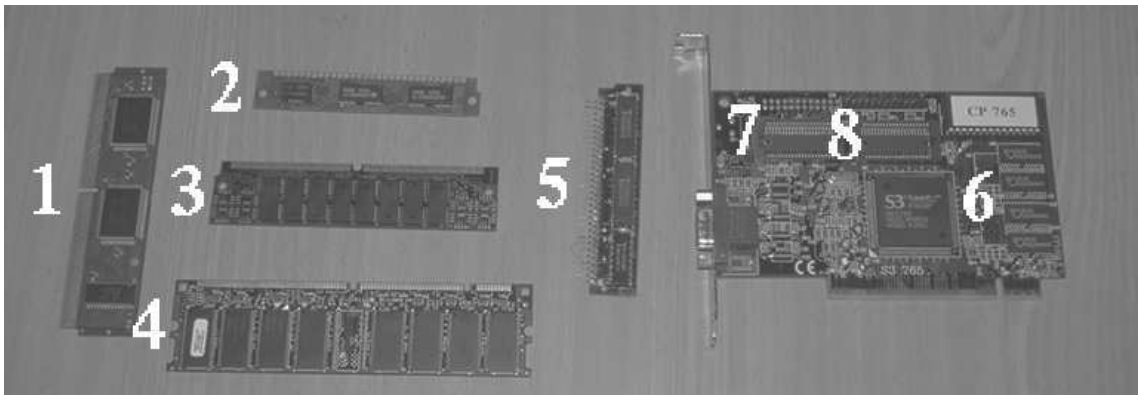


FOTO 10.2: La conexión de la memoria clasificada según su funcionalidad. A la izquierda, (1) Memoria caché según conexión SIMM. Está dotada de tres chips para realizar una ampliación opcional sobre un zócalo COAST dispuesto para tal efecto en la placa base. Esta solución se dió en los sistemas bajo microprocesador 80286, 80386 y 80486, en los que la caché era considerada como un artículo de lujo. (2, 3, 4 y 5) Distintos módulos de memoria principal, donde (2) es un módulo SIMM de 30 contactos tomado de un equipo bajo procesador 80486, (3) es un módulo SIMM de 72 contactos dotado de 10 chips y más común en sistemas bajo microprocesador Pentium, (4) es un módulo DIMM de 168 contactos y 8 chips, más frecuente en sistemas bajo microprocesador Pentium II en adelante, y (5) es un módulo de memoria SIPP de tres chips tomado de un sistema de los años 80. (6, 7 y 8) Memoria de vídeo, conectada por (6) DIP a la tarjeta de vídeo, y (7 y 8) zócalos DIP para su ampliación.

Inline Memory Module). Se trata de láminas de circuito impreso que contienen chips de memoria soldados en número variable, ya sea por una sola cara o por las dos, y que se insertan en unos zócalos dispuestos para tal efecto en la placa base.

`dimensiones`

Los contactos por los que la memoria se comunica con la placa base se disponen sobre una sola arista de la lámina, pero en ambas caras. En los SIMM, los dos contactos simétricos establecidos uno a cada lado de la arista tienen idéntica funcionalidad, mientras que en los módulos DIMM y RIMM existe un aislante intermedio que permite diferenciar eléctricamente uno y otro. Es por esto que el módulo DIMM puede disponer de 168 contactos, más del doble de los 72 que alberga el SIMM más grande (ver sección 10.7), a pesar de que su longitud es tan sólo un 30 % mayor (unos 13.5 cm. en los DIMM frente a 10.5 cm. en los SIMM).

`aplicación`

Desde principios de los años 90, todos los PC utilizan la conexión por contactos para implementar la memoria principal, incluyendo los Mac de Apple: Con SIMM durante la cuarta y quinta generación (ver foto 10.3.a), con DIMM a lo largo de la sexta y séptima generación, y conviviendo ya con los RIMM en ésta última (ver foto 10.3.b). Por todo ello, a partir de ahora dejaremos a un lado los formatos existentes bajo las variantes de púas (SIPP) y patillas (DIP), centrándonos exclusivamente en las conexiones por contactos.

☛ pág. 28

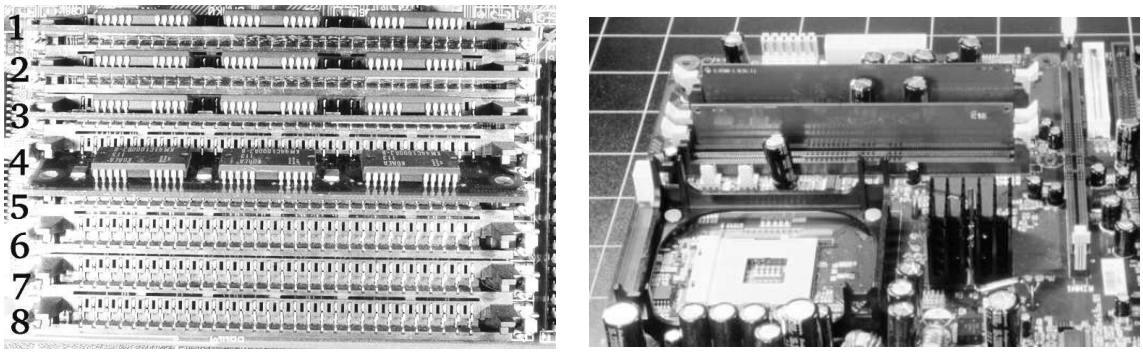
SECCIÓN 10.7

Formato

`anchura`

Cualquier tipo de memoria, ya sea SIPP, DIP, DIMM o RIMM se fabrica en distintos formatos que determinan la anchura de la palabra de memoria (número de bits de datos almacenados en cada una de sus direcciones). La importancia de esta anchura radica en el hecho de que la placa base normalmente sólo acepta una anchura concreta, condicionando las posibilidades de expansión y reutilización de la memoria principal en un futuro. La foto 10.4 muestra algunos

☛ pág. 28



(a) (b)

FOTO 10.3: (a) Detalle del sistema de memoria principal de un equipo basado en el procesador 486 de Cyrix, cuya placa base dispone de dos bancos de memoria principal, cada uno compuesto por cuatro módulos SIMM de 30 contactos. El segundo banco se encuentra vacío (zócalos 5 al 8) y el primero lleno (zócalos 1 al 4), donde hemos desenganchado el cuarto módulo de las presillas laterales con objeto de ladearlo un poco para poder apreciar su composición (tres chips DIP de 16 patillas). (b) Un sistema de memoria genuino de séptima generación con cuatro zócalos RIMM de 184 contactos para Pentium 4.

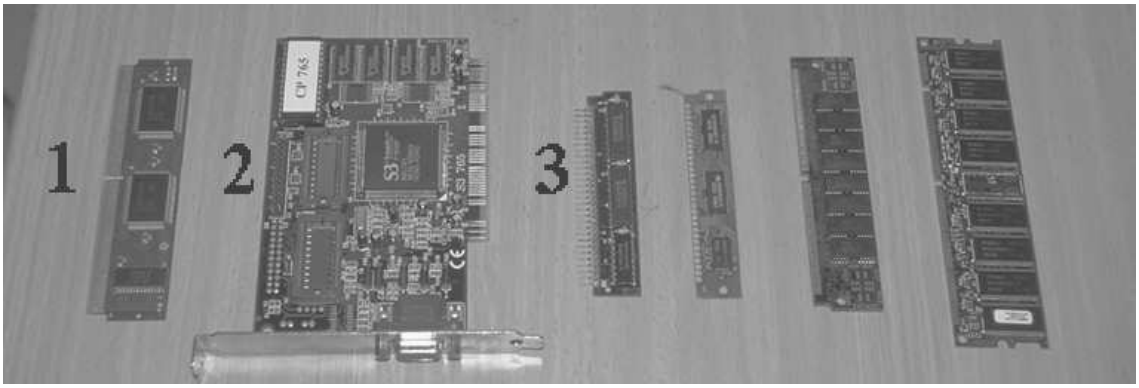


FOTO 10.4: La memoria del sistema clasificada según su formato. (1) Memoria caché. (2) Memoria de vídeo. (3) Memoria principal. Los tres últimos módulos de memoria son el formato SIMM de 30 contactos, el SIMM de 72 contactos y DIMM de 168 contactos.

formatos de anchura, longitud y aplicación muy diversa.

Otras variables que condicionan el formato de un módulo de memoria son el voltaje, el interfaz utilizado, o la presencia de búferes a la salida de las celdas. La figura 10.3 recopila la influencia de todas ellas para situarnos de manera general.

pág. 29

7.1 ▶SIMM de 30 contactos

contactos

En este formato, los contactos se disponen por una sola cara de la placa de circuito impreso del módulo, donde ocho de ellos son de datos (mas uno opcional en caso de implementar la paridad).

chips

Se construyen de dos formas diferentes: Los SIMM más antiguos, dotados de ocho o nueve chips de un bit de anchura dependiendo de la paridad (ver figura 10.4), y los más recientes, construidos a base de dos chips de cuatro bits de anchura para los datos y un chip de un bit opcional de paridad. Los SIMM de tres chips como el de la foto 10.3.a son más fiables, consumen menos y tienen un coste inferior, por lo que resultan más convenientes.

pág. 29

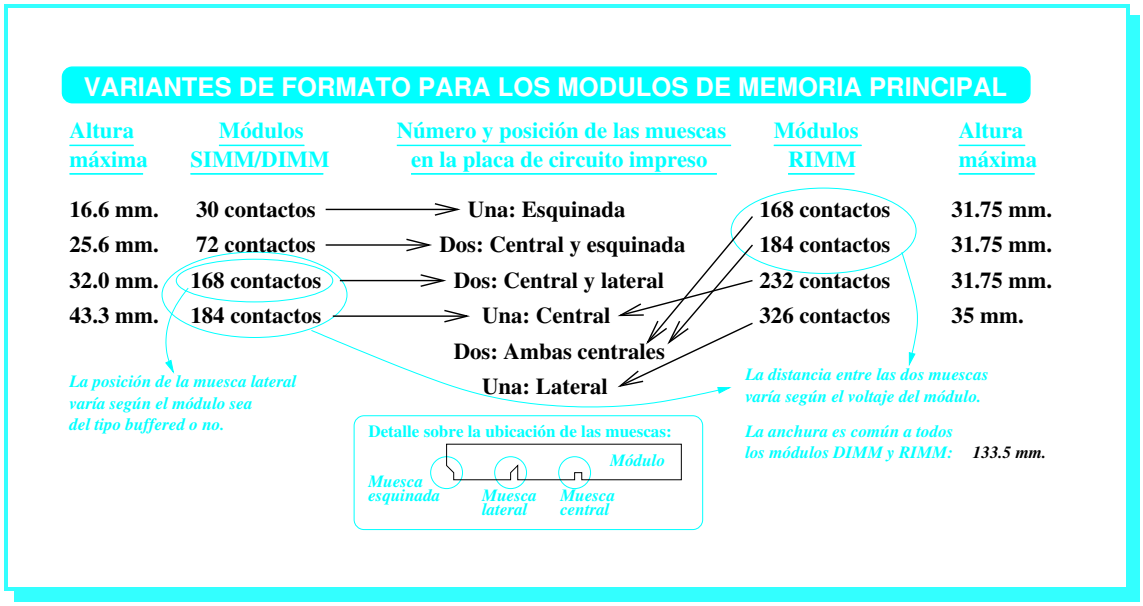


FIGURA 10.3: Variantes de formato para los módulos de memoria principal. Aunque la placa de circuito impreso es muy similar en todos los casos (rectangular muy alargada), existen pequeños detalles que permiten diferenciarlos según el interfaz, voltaje o la presencia de búfers a la salida.

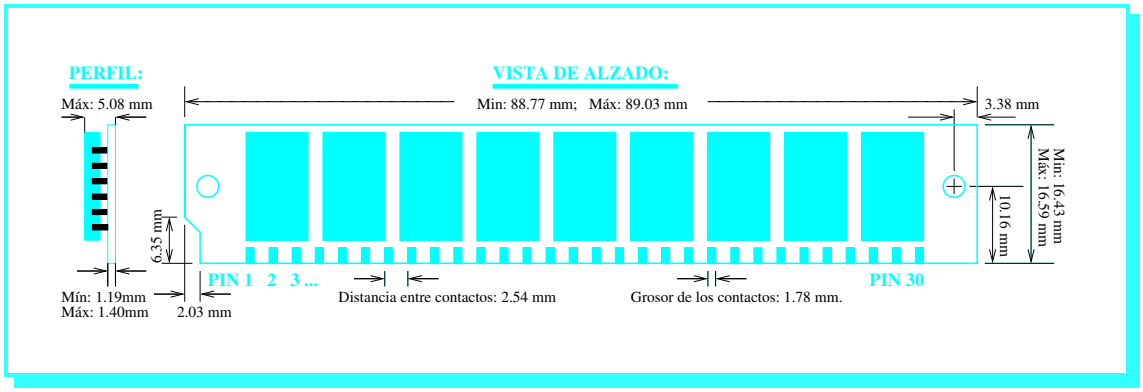


FIGURA 10.4: El formato SIMM de 30 contactos para un módulo de ocho chips de datos mas uno de paridad. Otras variantes disponen de dos o tres chips.

Los SIMM de 30 contactos dispusieron en sus versiones comerciales de un tamaño de 256 Kbytes, 1 Mbyte y 4 Mbytes. Fueron el esquema de memoria principal que siguieron mayoritariamente los procesadores con bus de datos de 8 y 16 bits, y se utilizaron incluso en sistemas con bus de datos de 32 bits, como el 80386 y el 80486.

tamaño

La foto 10.3 ilustra un ejemplo de uso en un equipo basado en el procesador 486 de Cyrix, uno de los últimos en utilizarlo. Con la llegada de la quinta generación de procesadores, el formato SIMM de 30 contactos quedó abolido y su uso quedó relegado a la ampliación de la memoria de algunas tarjetas de sonido.

☛ pág. 28

ejemplos de uso

Apuntaremos para concluir que en los computadores de IBM, como el IBM PS/2 o el XT-286, la función de cada contacto era diferente que en el modelo SIMM estándar, lo que hacía incompatibles sus chips de memoria con el resto de computadores.

incompatibles

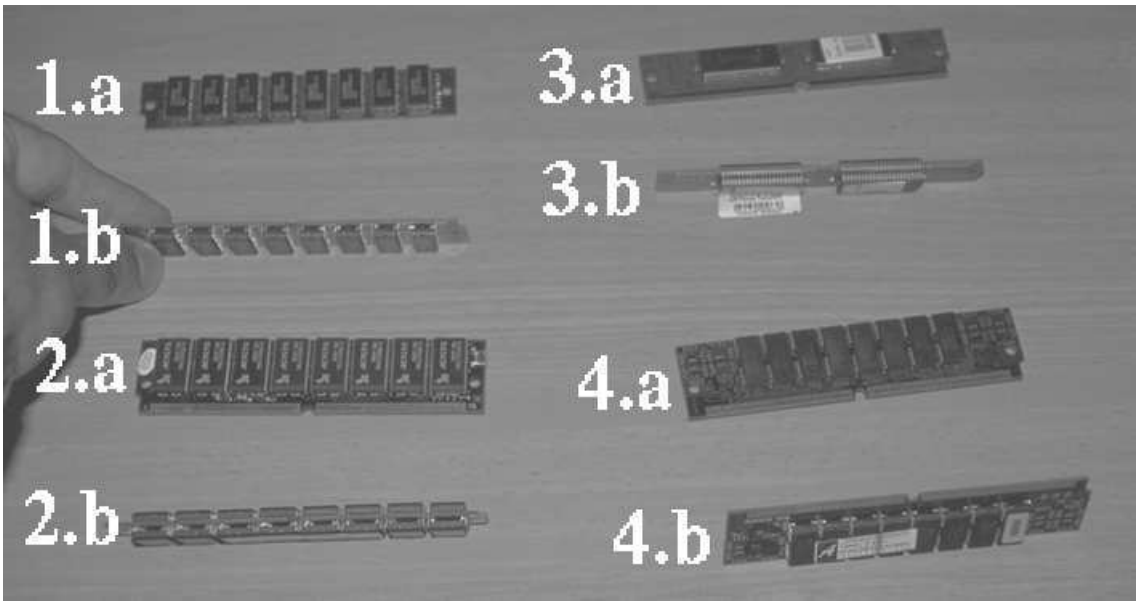


FOTO 10.5: Distintas formas de fabricación para módulos SIMM de 72 contactos. (a) Vistas superiores y (b) perfiles de (1) Un módulo de 8 chips de 1 Mbyte por una sola cara = 8 Mbytes. (2) Un módulo de 8 chips de 1 Mbyte de datos por una cara y cuatro chips de paridad por la otra cara. Capacidad total: 8 Mbytes. (3) Un módulo de 4 chips de 4 Mbytes de datos, dos por cada cara, en disposición transversal. Capacidad total: 16 Mbytes. (4) Un módulo de 12 chips por cada cara, 8 para datos (512 Kbytes) y cuatro para paridad. Capacidad total: 8 Mbytes. De la enorme versatilidad expuesta podemos concluir que no existe ningún tipo de relación entre el número y la disposición de los chips de un módulo y su capacidad total.

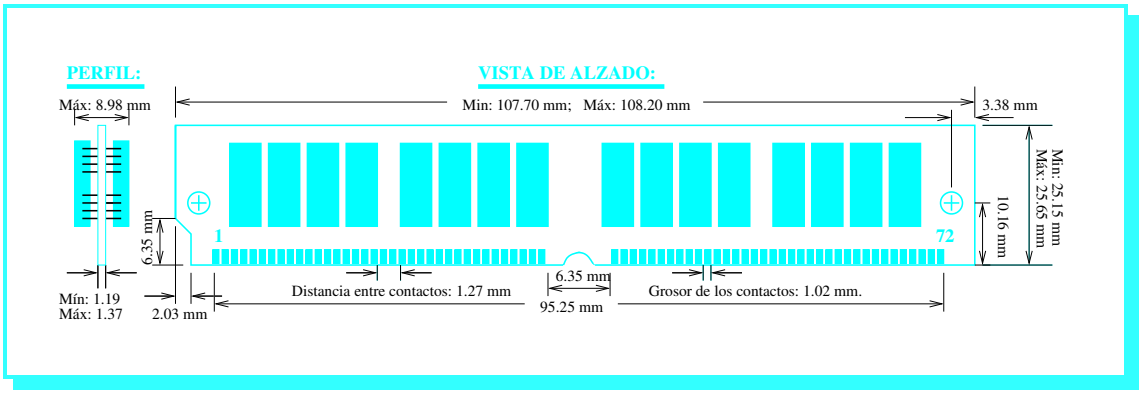


FIGURA 10.5: El formato SIMM de 72 contactos para un módulo de 32 chips (16 por cada cara) sin paridad. Cada chip contribuye con un único bit para formar la palabra de memoria del módulo. La simetría se rompe en el entrante que se habilita en la esquina inferior izquierda.

7.2 ▶SIMM de 72 contactos

descripción

Presentan un aspecto más alargado y una anchura de 32 bits para datos y 4 bits opcionales de paridad, uno por cada byte de datos. Los 72 contactos se disponen en una sola arista de la placa de circuito impreso, pero en sus dos caras, siendo funcionalmente equivalentes dos a dos (esto es, las del anverso se encuentran replicadas en el reverso).

En la foto 10.5 mostramos diferentes variantes comerciales de fabricación de módulos SIMM72, y en la figura 10.5, sus dimensiones más relevantes.

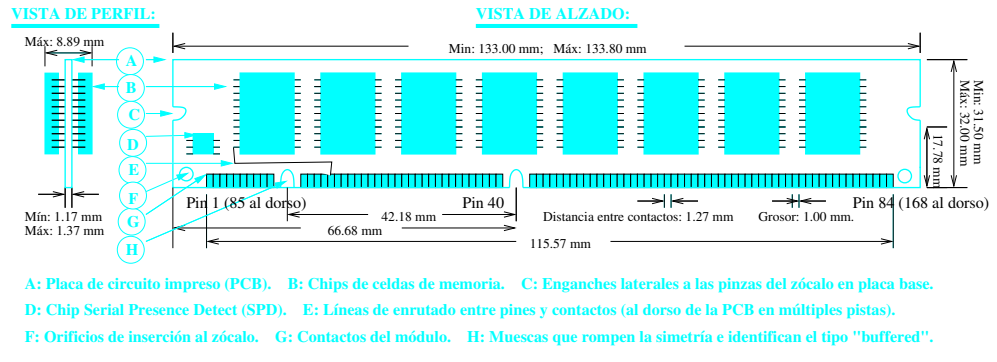


FIGURA 10.6: El formato DIMM de 168 contactos para un módulo de 16 chips, ocho por cada una de sus caras. Los elementos adicionales que se enumeran alfabéticamente en la parte izquierda están también presentes en todos los formatos que veremos a partir de éste, no habiéndolos descrito en más ocasiones para evitar redundancias. A partir de este formato, los cuatro vértices del módulo serán de 90 grados, utilizándose las muescas colocadas en las aristas de los contactos para romper la simetría y diferenciar las múltiples variantes existentes.

Al contrario que en los SIMM30, los contactos de los SIMM72 tienen asignada una única función que ha sido respetada por todas las marcas que conocemos hasta la fecha, evitándose los problemas de compatibilidad entre fabricantes.

compatibilidad

La tendencia en SIMM72 desde sus inicios siguió siendo reducir el número de chips y aumentar su anchura, ya que a las ventajas ya esgrimidas en SIMM30 debemos añadir que se facilita la sincronización interna del módulo.

chips

El formato SIMM de 72 contactos comenzó a utilizarse dentro de la memoria principal a partir del procesador 80486 de Intel, y su uso se propagó hasta las placas del procesador Pentium, e incluso las primeras para Pentium II.

uso

DIMM de 168 contactos

◀ 7.3

Entran en escena con la llegada de los procesadores Pentium: Al doblarse la anchura del bus de datos de 32 a 64 bits, la memoria toma idéntico camino para llenar por completo el ancho del bus, y suministra a la placa base un total de 168 contactos, 84 en cada cara, todos ellos con diferente funcionalidad (ver figura 10.6).

diferente funcionalidad

La memoria DIMM puede verse así como un par de SIMM de 72 contactos implementados conjuntamente, evitándose la sincronización por pares de palabras de datos que requieren los módulos SIMM para llenar los 64 bits del bus de datos del procesador. Por ello, los módulos DIMM suelen ser más rápidos que los SIMM, aunque internamente algunas versiones tengan igual latencia y fabricación que sus homólogos.

velocidad

La forma en la que los módulos SIMM y DIMM se conectan a la placa base es también diferente, pues la distinción eléctrica que DIMM establece entre los contactos situados en cada cara le obliga a utilizar un tipo de zócalo en el que las conexiones a las líneas de la placa base se dispongan a ambos lados, insertándose el módulo en la parte central. Los detalles sobre el montaje de ambos módulos pueden consultarse en la sección 22.6.

conexión al zócalo

☛ Volumen 3

La elección de implementar un banco de memoria mediante dos módulos SIMM o un sólo módulo DIMM dependerá de los zócalos de memoria que traiga nuestra placa base y del tipo

bancos

Formato	1M	2M	4M	8M	16M	32M	64M	128M	256M	512M	1G
SIMM30	✓		✓		✓						
SIMM72	✓	✓	✓	✓	✓	✓	✓	✓			
DIMM168				✓	✓	✓	✓	✓	✓	✓	✓
DIMM184								✓	✓	✓	✓
RIMM184							✓	✓	✓	✓	
RIMM232								✓	✓		

TABLA 10.4: Capacidades de los módulos de memoria comerciales según su formato. El estatus corresponde a finales de 2002 atendiendo al catálogo de Kingston Technology para las tres últimas columnas en las que se esperan aumentos futuros. Por debajo de 1 Mbyte, sólo se comercializó un módulo SIMM30 de 256 Kbytes.

de memoria que estemos interesados en adquirir. Por ejemplo, la memoria SDRAM (ver sección 10.13.4) presenta una operativa de funcionamiento donde es indispensable el formato DIMM, justo lo contrario que le ocurre a la memoria FPM (ver sección 10.13.1), que resulta demasiado antigua para poderla encontrar en DIMM.

☛ pág. 58
☛ pág. 55

chips

Al ser más recientes, lo normal es encontrar los módulos DIMM de 168 contactos implementados con un número reducido de chips que oscila entre dos y ocho.

7.4 ▶DIMM de 184 contactos

El formato DIMM de 184 contactos es el que da cobertura a la memoria DDR-SDRAM, esto es, la SDRAM que proporciona dos salidas de datos por cada ciclo de reloj (ver sección 10.13.5). Los chips del módulo se dividen de esta manera en dos zonas de acceso excluyente, una para los datos que son accedidos en el flanco de subida de la señal de reloj, y otra para los que son accedidos en el de bajada.

pág. 68 ☛

Con esto, las novedades en la asignación de funciones a contactos son las siguientes:

dirs. y datos

❶ La parte de direcciones, datos y paridad conserva el mismo número de contactos que en el formato DIMM168, aunque no se respeta su ubicación, lo que de entrada rompe cualquier opción de compatibilidad entre ambos.

control

❷ La parte de control se amplía con 8 nuevas líneas de enmascaramiento para los 8 bytes de datos accedidos en el flanco de bajada.

reloj

❸ Las tres señales de reloj que gobiernan la salida de datos se duplican para albergar otras tres que proporcionan la misma señal invertida, con objeto de que el módulo pueda construirse con chips DRAM idénticos que respondan funcionalmente en el mismo flanco, independientemente de la submitad de ciclo en la que les toque actuar.

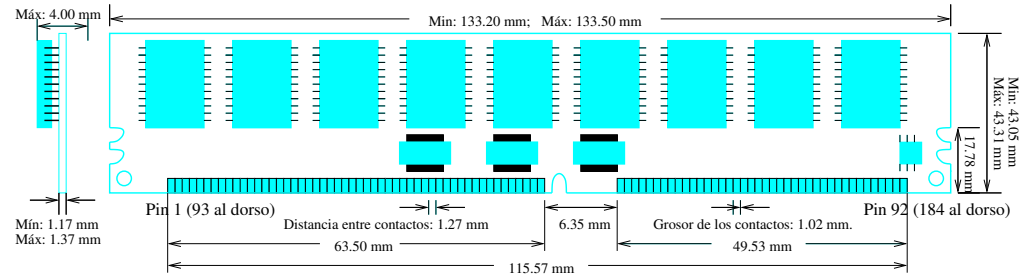
corriente

❹ El suministro de corriente y tierra también ocupa más líneas, con objeto de dispersar las pérdidas I^2R y mitigar el calentamiento a frecuencias más elevadas.

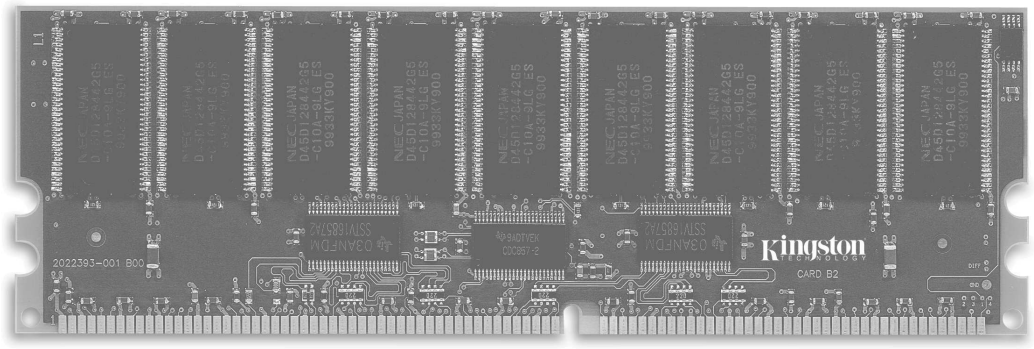
Los 16 contactos adicionales se sitúan en la parte central de la arista del módulo, ocho en cada cara, y dentro de éstas, cuatro por cada lado, prolongando la hilera de 168 anterior desde fuera hacia dentro (ver figura 10.7). Para ello se aprovecha el espacio ya existente, ya que las dimensiones de la placa de circuito impreso son las mismas que en DIMM168. No obstante, cambia el mecanismo para la sujeción a la placa base, al contarse ahora con dos muescas en cada una de las aristas laterales, desplazadas más hacia la parte inferior del zócalo.

pág. 33 ☛

incompatible con DIMM168



(a)



(b)

FIGURA 10.7: (a) El formato DIMM de 184 contactos sobre un módulo de ocho chips de datos mas uno de paridad/ECC, todos ellos dispuestos por la misma cara del módulo (rellenar el reverso es algo opcional para el fabricante). Comparado con el DIMM de 168 contactos, aquí desaparece una de las dos muescas de la arista inferior, y sin embargo, en las aristas laterales, ocurre justo lo contrario. (b) Un módulo DDRAM comercial de Kingston Technology.

Otra salvedad que impide pinchar un módulo DDR-DIMM184 sobre un zócalo antiguo SDR-DIMM168 es que se prescinde de una de las dos muescas que había en la arista de los contactos en SDRAM, perdurando sólo la de la parte central. Esta incompatibilidad es lógica, ya que si los controladores de memoria son diferentes y la función de cada contacto también, es preferible evitar equívocos.

RIMM de 168 contactos

◀ 7.5

Los formatos RIMM aparecen con la llegada de la memoria RDRAM de Rambus (ver sección 10.13.6), cuya velocidad descansa sobre una frecuencia elevada de 800 MHz en detrimento de la anchura para el bus de datos, que inicialmente fue de tan sólo 16 bits.

☛ pág. 73

Este recorte en el canal de datos sugiere la posibilidad de aminorar el número de contactos de la memoria, pero en realidad lo que termina produciéndose es el efecto inverso. Veamos por qué:

distribución de contactos

1. Para trabajar a frecuencias muy elevadas, hace falta dotar de mucha estabilidad a la señal eléctrica transmitida, lo que obliga a colocar líneas individuales de conexión a tierra para aislar cada uno de los contactos del módulo.

tierra

2. Para unificar el comportamiento eléctrico de las líneas, es necesario conectar en serie tanto

los módulos como los chips dentro de él, lo que duplica el patillaje. Por ejemplo, el bus de datos entra por 16 patillas y sale por otras 16 en cualquiera de estos elementos, camino del siguiente eslabón de la cadena que conforman todos ellos (ver figuras 10.32 y 10.33).

pág. 77
pág. 79

Así, el número de contactos dedicados a cada una de las tareas del módulo de memoria (direcciones, datos, control, alimentación) se redistribuye en RIMM respecto a SIMM y DIMM, tal y como hemos reflejado en la tabla 10.5. Se dispara el número de contactos dedicados al direccionamiento y, sobre todo, las conexiones a tierra, que representan casi la mitad del total, alternándose (una sí y otra no) con las señales activas.

pág. 37

conexiones a tierra

dimensiones

Respecto a las dimensiones de este zócalo RIMM, es idéntico en aspecto al DIMM: Negro, alargado de unos 13.5 cm (5.25 pulgadas exactamente), y con 168 contactos dispuestos en sus dos laterales. Ahora bien, la distancia entre contactos se acorta desde 1.27 hasta 1 mm (desde 1 hasta 0.8 mm en el ancho metálico del contacto, y desde 0.27 hasta 0.2 mm en su separación), quedando la parte central del zócalo despejada en previsión de ser poblada en formatos posteriores.

incompatible con DIMM168

Si la funcionalidad es radicalmente diferente en DIMM y RIMM y su compatibilidad imposible, cabe preguntarse por qué se respeta el formato en la placa de circuito impreso (PCB). Recordemos que Rambus sólo diseña y especifica sus memorias, que luego fabrican Kingston, Samsung y Micron entre otros. Cuando un diseñador quiere abrirse paso entre tecnología consolidada (SDRAM) con un producto nuevo (RDRAM), lo mejor para persuadir a los fabricantes de cara a que implementen sus productos es decirles que casi toda la infraestructura de que ya disponen para la vieja tecnología puede ser reutilizada para desarrollar la nueva. De todas formas, la posibilidad de pinchar un módulo SDRAM en otro RDRAM o viceversa se desvanece porque las muescas centrales y laterales definidas en la PCB presentan diferente ubicación en cada caso, e incluso para diferentes voltajes dentro de una misma familia, según detallamos en la sección 10.8.

amortización

pág. 36

7.6 RIMM de 184 contactos

El paso de la SDRAM a DDRAM había provocado la migración del formato DIMM168 al DIMM184 en el año 2000, y estando este último ya consolidado un año más tarde, el RIMM no podía permanecer anclado en los ya obsoletos 168 pines de sus inicios, así que en 2001 Rambus decidió adaptar su formato RIMM al vigente en DDRAM. El nuevo formato coincide en todos los aspectos con el DIMM184 ya visto (ver figura 10.8), con la única salvedad de habilitar una segunda muesca en la arista de los contactos para evitar confusiones, dada la manifiesta incompatibilidad entre ambos. Los dieciséis nuevos contactos no se encuentran conectados al patillaje de los chips de memoria en ninguno de los módulos RIMM que hemos cotejado, y la especificación RDRAM 1.1 de Rambus en 2001 tampoco les asignaba función alguna, por lo que suponemos que la firma la utilizó de nuevo como arma para reaprovechar las mismas placas de circuito impreso ya vigentes para DDRAM.

pág. 35

incompatible con DIMM184

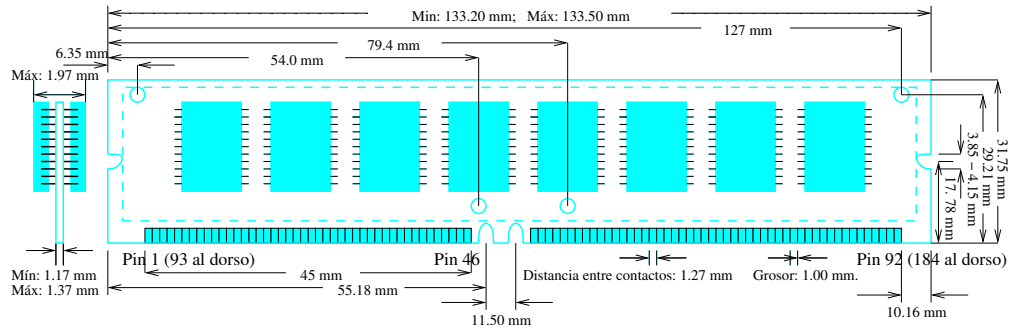
compatible con RIMM168

Todo esto arroja una conclusión meridiana: Todo módulo RIMM de 184 contactos que consiga encajarse sobre un zócalo RIMM de 168 contactos **funcionará perfectamente**. En cambio, si el zócalo en placa base es de 184 contactos y la PCB del módulo es de 168, aún pudiendo funcionar, existen los problemas de fiabilidad derivados de mantener contactos al aire.

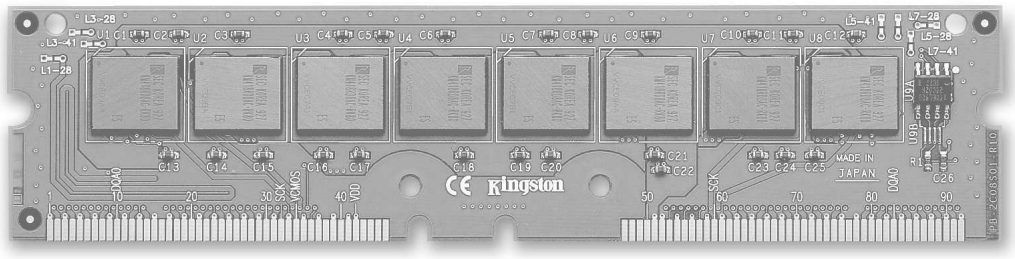
7.7 RIMM de 232 contactos

El formato RIMM de 232 contactos aparece para auspiciar la ampliación del bus de datos de la memoria RDRAM desde los 16 hasta los 32 bits. Las dimensiones del zócalo son coincidentes en todo momento con el formato RIMM184 anterior, pero ahora se termina de rellenar de contactos la parte central de la arista que linda al zócalo de la placa base, tal y como muestra la figura 10.9. Además, se prescinde de la muesca o entrante central, permaneciendo únicamente el que se encontraba más escorado.

pág. 35



(a)



(b)

FIGURA 10.8: El formato RIMM de 184 contactos. En la arista inferior, el módulo dispone de una doble muesca en su parte central que lo diferencia frente a los DIMM. Sus dos laterales presentan una muesca cada uno, ubicada también en su parte central. La línea discontinua del módulo delimita la presencia del disipador de calor que mitiga la temperatura de sus 16 chips (8 por cada cara). (b) Un módulo de memoria comercial de Kingston Technology.

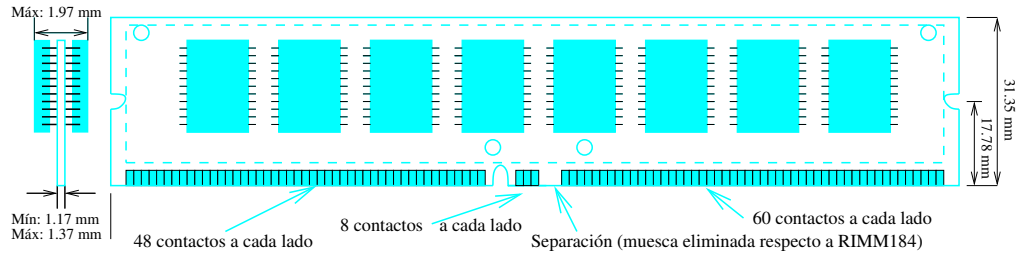


FIGURA 10.9: El formato RIMM de 232 contactos para un módulo de 16 chips, ocho por cada cara. Anchura, altura y muescas laterales son heredadas del formato RIMM184, prescindiéndose únicamente de una de las dos muescas situadas sobre la arista inferior y poblándose más de contactos ésta.

Las memorias de 232 contactos no tienen posibilidad alguna de ser compatibles con las de 184 o 168 contactos, ya que extraen de los chips de memoria el doble de información, y además

incompatible con RIMM184

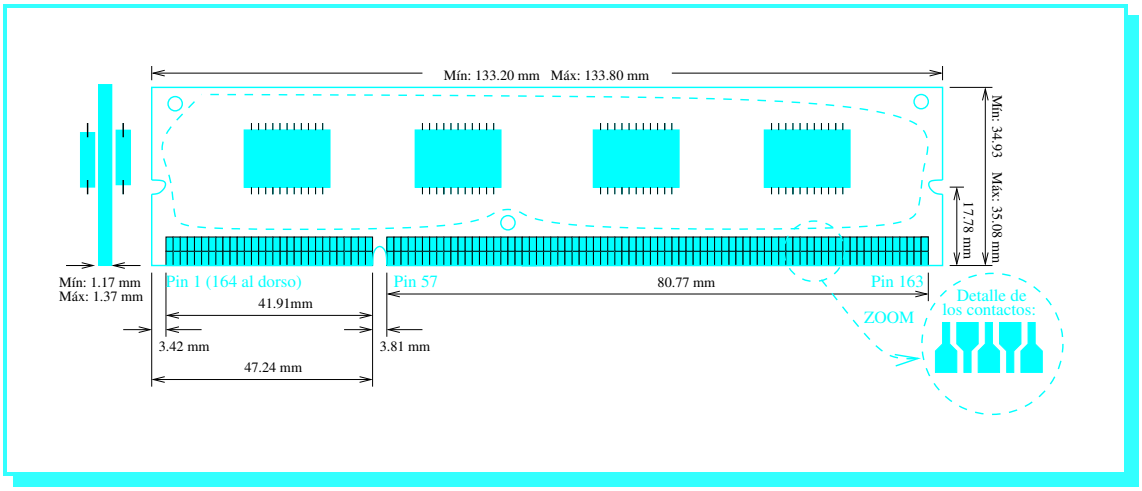


FIGURA 10.10: El formato RIMM de 326 contactos para un módulo de ocho chips de datos, para una palabra de 64 bits, alcanzando por fin la anchura de los DIMM. El crecimiento del número de contactos aboga por aumentar la dimensión horizontal del módulo, pero lo que en realidad crece es la arista vertical. Esto provoca un hacinamiento de contactos que se resuelve encajándolos de forma similar a como procedieron las tarjetas AGP para no aumentar sus dimensiones respecto a las PCI (ver zoom).

redistribuyen la función de los contactos ya existentes. Sin embargo, para la fijación y liberación del módulo al zócalo se sigue preservando una única muesca central en el lateral del módulo, y un mecanismo de pinzas en el zócalo, produciendo la falsa impresión de que todo es intercambiable.

7.8 ▶RIMM de 326 contactos

Este formato aparece para la memoria RDRAM de 64 bits de anchura. Las dimensiones de la placa de circuito impreso son también coincidentes con los módulos RIMM anteriores, pero tenemos dos rasgos distintivos:

`muesca del zócalo`

1. Dispone de una sola muesca o discontinuidad en el zócalo, pero ésta queda desplazada más hacia el lateral del zócalo (en concreto, entre los contactos 56 y 57 del total de 163 que contiene cada cara), con objeto de evitar confusiones con los formatos RIMM anteriores.

`espacio entre contactos`

2. Comprime el espacio entre los contactos, para dejar paso a casi 100 nuevos contactos en el mismo espacio físico que el RIMM de 232 contactos, donde las partidas que más crecen son las dedicadas a datos y conexiones a tierra. Para lograr esta mayor densidad de contactos, se adopta una estrategia de compactación similar a la que utilizaron las tarjetas AGP frente a sus homólogas PCI: Alternar contactos a dos niveles, en forma de botella en el nivel inferior, y de botella invertida en el nivel superior, que se van encajando unos con otros, y que hacen contacto en el zócalo de la placa base a dos alturas diferentes.

La figura 10.10 ilustra esta doble diferencia física, que a simple vista puede pasar inadvertida respecto a los formatos anteriores.

SECCIÓN 10.8

Voltaje

`dual`

Los módulos de memoria principal funcionan desde comienzos de los años 90 bajo un doble

		Número de contactos utilizados para:								
Formato	Año	Bus de dirs.	Bus de datos	Bits de paridad	Control	SPD	Reloj	Aliment.	Tierra	NC
SIMM30	1990	11	8	1	4	0	0	2	2	2
SIMM72	1993	14	32	4	9	5	0	3	3	2
DIMM168	1996	14	64	8	20	3	5	20	18	16
DIMM184	2000	14	64	8	24	3	7	27	23	14
RIMM168	1999	16	32	4	14	3	3	16	74	6
RIMM184	2001	16	32	4	14	3	3	16	74	22
RIMM232	2002	24	48	6	11	3	15	17	108	0
RIMM326	2002	8	96	12	5	3	20	26	156	0

TABLA 10.5: Desglose de los contactos dedicados a cada una de las funciones de la memoria SIMM, DIMM y RIMM. En este último caso, existen el doble de líneas de las esperadas en las partidas de direcciones y datos debido a que la conexión en serie de los zócalos obliga a utilizar una patilla de entrada y otra de salida para cada una de estas funciones. Las iniciales SPD se corresponden con Serial Presence Detect, los bits destinados al reconocimiento automático de la configuración del módulo. Este protocolo lleva su propia señal de reloj, que hemos contabilizado en la partida dedicada al reloj. Finalmente, las iniciales NC de la última columna corresponden a las líneas No Conectadas (o reservadas para el fabricante), que suelen utilizarse para ampliar la funcionalidad en versiones futuras.

voltaje de alimentación:

externo

1. Un voltaje externo, suministrado por las patillas etiquetadas como V_{cmos} o $V_{I/O}$ y utilizado para los circuitos externos, como los búferes para la retención de datos a la salida o el chip SPD donde se guardan sus parámetros de configuración más relevantes. Este ha evolucionado siguiendo los niveles de voltaje CMOS del procesador coetáneo con el módulo de memoria (5 v. hasta la cuarta generación, 3.3 v. en la quinta, 2.5 v. en la sexta y 1.8 v. en la séptima).

interno

2. Un voltaje interno, V_{cc} o V_{dd}, para los chips de celdas en sí, que ha evolucionado más lentamente, moviéndose entre 3.3 voltios en 1993 y 2.5 en 2003. La memoria puede permitirse esta pasividad comparada con el procesador, al no concentrar tantos transistores ni alcanzar frecuencias tan elevadas.

Los módulos implementados con chips por debajo de los 10 Mbytes funcionan en su mayoría con valores de 5 voltios externamente y 3.3 voltios en su nivel de celda interno.

FPM y EDO: 5-3.3 v.

Por encima de 16 Mbytes, suele utilizarse 3.3 voltios tanto externa como internamente, y más arriba, ha ido bajando el voltaje externo y manteniéndose el interno en 3.3 voltios hasta los últimos diseños de SDRAM, incluso de 1 Gbyte en 2003.

SDRAM: 3.3 v.

A partir de ahí no es el tamaño quien mueve el voltaje, sino la velocidad: Las memorias DDRAM de 133x2 MHz y RDRAM de 400x2 MHz utilizan ya voltaje interno de 2.5 voltios, manteniéndose éste para las variantes más rápidas de 200x2 MHz y 533x2 MHz, respectivamente.

DDRAM y RDRAM: 2.5 voltios

No obstante, ambas especificaciones contemplan la posibilidad de los 1.8 voltios (ver figura 10.11), voltaje que ya utilizara internamente Samsung (2.5 voltios externamente) para implementar un prototipo experimental en 1996. En ámbitos más comerciales, sólo hemos visto los 1.8 voltios internamente en la especificación de Rambus para un módulo RDRAM de 256 Mbytes sobre formato de 326 contactos y 64 bits (aún por implementar a fecha 2003) y en un chip de memoria DDRAM de 300x2 MHz fabricado por Infineon (la división de microelectrónica de Philips) para

☛ pág. 39

futuro: 1.8 v.

Año	Tamaño de la memoria (Mbytes)	Voltaje de la memoria Externo	Interno	Distancia de integración del procesador (micras y año)
1980-1989	<1	5 v.	5 v.	0.5 en 1994
1990-1995	2-8	5 v.	3.3 v.	0.35 en 1996
1996-2000	16-64	3.3-2.5 v.	3.3 v.	0.25 en 1998
2000-2005	128-512	2.5-1.8 v.	2.5 v.	0.18 en 2000
2005-2010	$\geq$ 1024	1.5 v.	1.8-1.5 v.	0.10 en 2005

TABLA 10.6: La reducción del voltaje en los chips de memoria principal con el paso del tiempo y el aumento de su capacidad. A la derecha, valores de la distancia de integración en las puertas de los transistores del microprocesador, con los que guarda una similitud fortuita. Dado que el ritmo evolutivo de éstos ha sido tres veces superior, la tendencia se rompe en fechas próximas según nuestras estimaciones de la última fila.

el exigente mercado de las tarjetas gráficas en Abril de 2001. Por ello creemos que los 1.8 voltios internos en memoria principal para PC se consolidarán ya en el próximo lustro 2005-2010.

Riesgo 10.2: CONFUSIONES EN EL VOLTAJE DE LA MEMORIA PRINCIPAL

Según acabamos de ver, los valores de voltaje están bien definidos para cada tipo de memoria, y al disponer cada una de ellas de un zócalo distinto, se minimiza la posibilidad de introducir un módulo de memoria en un zócalo que lo alimente a un voltaje equivocado. El caso más proclive al error queda ya algo lejano en el tiempo. Corresponde a la terna FPM-EDO-BEDO, que compartió el zócalo SIMM72 y disponía de módulos a 5 y 3.3 voltios. Aunque la diferencia parece lo suficientemente ostensible como para provocar fatales consecuencias, algunos fabricantes de solera como Micron advierten en su documentación técnica que sus chips EDO/BEDO de 3.3 voltios toleran los 5 voltios de la antigua FPM sin inmutarse. Sin embargo, también advierten que los chips de 2.5 voltios son bastante más delicados y no aguantan el valor de 3.3 voltios.

Así pues, conviene extremar precauciones cuando nos topemos con placas base con voltaje dual 2.5/3.3 voltios en las que los zócalos con voltaje de 2.5 puedan recibir 3.3 voltios por haber módulos de 3.3 ya conectados a la placa base (ver riesgo 10.3).

☛ pág. 96

evolución

La evolución del voltaje de la memoria principal puede memorizarse si nos ayudamos del capricho del azar, pues sus valores han coincidido con los de la distancia de integración de los transistores del procesador, esto es, progresando de forma discreta a razón del 70 % del valor anterior. Si recordamos la secuencia del transistor en micras, era 0.5-0.35-0.25-0.18-0.13. En voltios, para la memoria tenemos 5-3.5-2.5-1.8, en una clara coincidencia a la que no vemos silogismo científico. La tabla 10.6 refleja este paralelismo, incluyendo estimaciones del autor que romperán esta tendencia en años venideros.

papel

riesgos

El papel del voltaje en la memoria es más importante en su relación con la placa base, puesto que módulos de 3.3 voltios no funcionarán si se conectan a zócalos alimentados a 2.5 voltios, y módulos de 2.5 voltios pueden estropearse si se conectan a zócalos de 3.3 voltios. Esta última posibilidad es la más frecuente, ya que la placa base siempre es la última en evolucionar, en respuesta a reducciones de voltaje por parte del procesador o la memoria.

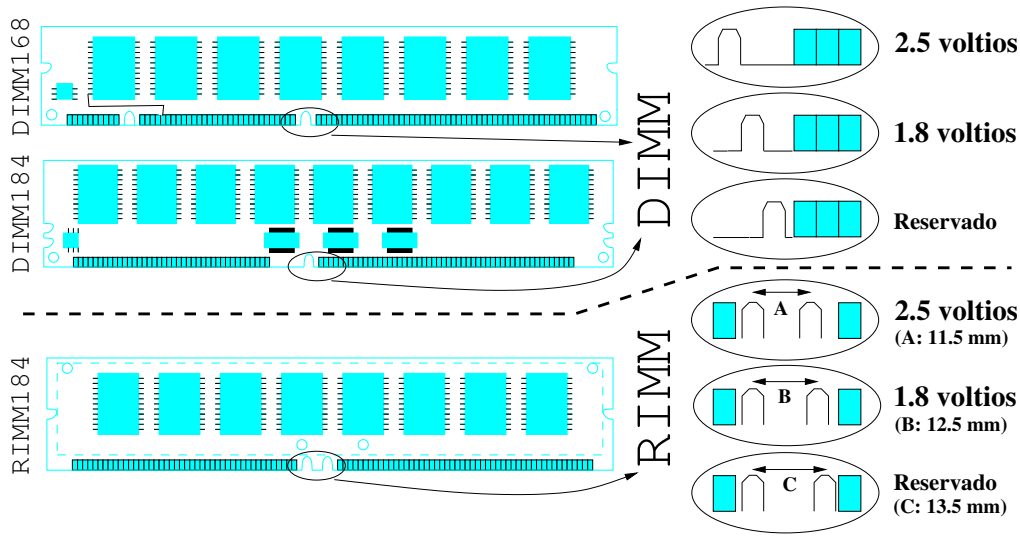


FIGURA 10.11: Cómo distinguir físicamente el voltaje de un módulo de memoria.

Antiguamente, este voltaje explícito correspondía al voltaje externo, obteniéndose el interno a través de una regulación interna al módulo. La expansión del número de contactos en los módulos produjo una sofisticación en el suministro de corriente, permitiendo desglosar la alimentación no sólo por chips individuales, sino incluso por segmentos de éstos (ver tabla 10.5). Como consecuencia, a partir del formato DIMM predominan las líneas V_{dd} sobre las V_{cmos}, y serán las primeras las que caractericen el voltaje del módulo.

distinción

pág. 37

La forma en que los módulos señalizan su voltaje interno es común a los formatos DIMM y RIMM, quedando reflejada tanto en placa base como en la placa de circuito impreso del módulo para evitar confusiones letales:

1. Los zócalos de la placa base suelen tener grabada en su parte central una inscripción en forma de relieve que indica su voltaje. La foto 22.9 muestra este detalle para un zócalo de memoria DIMM.

zócalos

Volumen 3

2. Los módulos presentan en la arista donde se sitúan los contactos una o dos muescas (ver figura 10.3), y será siempre la ubicación de la muesca central la que indique el voltaje del módulo, según detallamos en la figura 10.11.

módulos

pág. 29

SECCIÓN 10.9

Autoconfiguración

Los módulos de memoria SIMM30 debían ser configurados manualmente. Los SIMM72 dedicaban cinco pines a codificar sus dos parámetros principales: Tamaño (1, 2, 4, 8, 16 ó 32 Mbytes) y velocidad (100, 80, 70, 60 ó 50 ns.) en cualquiera de las 30 combinaciones posibles. Conectando cada pin a tierra en el módulo a través de una resistencia se consigue el nivel de voltaje para el cero lógico, y dejando el pin sin conectar, el nivel de voltaje para el uno lógico. Estas señales son luego decodificadas por la lógica de interfaz de memoria de la placa base para ajustar las señales

de 5 pines

de control y direccionamiento automáticamente.

Analogía 10.2: El parámetro ASA de fotografía como forma de autoconfiguración

La forma en que trabajan los pines de autoconfiguración presentes en un módulo de memoria es muy similar a la manera que tienen las cámaras fotográficas de reconocer automáticamente la sensibilidad de la película.
Nada mas cargar la película en su habitáculo interno de la cámara, unos contactos eléctricos pueden leer este dato codificado de la manera que hemos visto en la envolvente cilíndrica del rollo fotográfico. Para ello, la industria del sector desarrolló un código estándar denominado DX que es escrupulosamente seguido por todos los fabricantes.

estándar

En el caso de la memoria para PC, el organismo encargado de formular el estándar es el JEDEC (Junction Electronic Devices Engineering Council), aunque siempre hay fabricantes que se alejan premeditamente de los estándares para forzar la fidelización de los usuarios a su marca en la compra de componentes a posteriori. Hewlett-Packard, Compaq e IBM son tres claros exponentes.

pág. 58
pág. 73

SPD

La llegada de los interfaces síncronos como la SDRAM (ver sección 10.13.4) y la RDRAM (ver sección 10.13.6) multiplicó las posibilidades de los parámetros de configuración e introdujo otros nuevos, lo que provocó el almacenamiento de todos estos valores en una pequeña memoria CMOS denominada SPD (Serial Presence Detect). Esta memoria se utiliza desde la placa base para extraer las características más representativas del módulo, almacenarlas en la BIOS, configurar el sistema automáticamente, e informar al sistema operativo y al usuario de sus prestaciones.

interfaz

El acceso al chip SPD tiene lugar mediante un interfaz serie, esto es, se dedica tan sólo una línea a la extracción de datos. Otra línea se utiliza para el reloj, otra para la alimentación y tres más para el direccionamiento (que determinan el bit a leer/escribir dentro del byte ya seleccionado en una dirección predeterminada).

tamaño

En los chips SPD comerciales incluidos en los módulos de memoria para PC, el tamaño que más se repite es el de 256 bytes, de los que la primera mitad es de sólo lectura y queda establecida por el fabricante en el momento de la manufacturación del módulo, y la segunda mitad es modificable por el usuario y alberga aquellos parámetros configurables por él mismo, normalmente alterables mediante opciones del menú CHIPSET FEATURES SETUP de la BIOS (ver sección 24.3.4).

Volumen 4

SECCIÓN 10.10

Descomposición

Muchas veces la clave para disponer de un sistema de memoria eficiente no está en gastar una cantidad ingente de dinero, sino en saber escoger la placa base y el controlador de memoria adecuados, ya que serán ellos quienes hagan trabajar a nuestros módulos de memoria a pleno rendimiento.

El tener o no un buen gestor de la memoria en la placa base depende de muchas variables:

ancho del bus

1. El bus de memoria debe tener una anchura mínima igual a la de los datos con los que la memoria responde. Esto es algo que hasta ahora se ha cumplido sin problemas, dado que la

anchura del procesador ha ido siempre por delante de la de la memoria.

❷ La placa base debe funcionar a una frecuencia lo más cercana posible a la máxima que toleran los módulos de memoria (ver figura 10.40). Si la placa base es más lenta estamos pagando por unos chips de memoria cuyas prestaciones no aprovechamos, y si es más rápida, el controlador de memoria se verá infrautilizado.

frecuencia de la placa base

☛ pág. 102

Para saber si la memoria que tiene nuestro equipo hace funcionar al sistema a pleno rendimiento y viceversa, muchas placas base son capaces de efectuar comprobaciones internas sobre la máxima velocidad tolerada por ésta, e incluso de introducir estados de espera según el caso. La BIOS de las placas modernas suele disponer de un menú de configuración de memoria donde se nos informa del protocolo de comunicación con la memoria en ciclos de la placa base. La mejor situación se daría cuando leyésemos el mensaje "Zero Wait State Memory" (Memoria de cero estados de espera).

estados de espera

❸ Al margen de las variables más ligadas al hardware, también disponemos de numerosas mejoras al nivel organizativo que irán orientadas a aprovechar el ancho de banda del bus del sistema y la localidad espacial que tengan las referencias a memoria con objeto de mejorar el porcentaje de uso del bus y el tiempo de ciclo de la memoria. Este objetivo es compartido por la organización de la memoria caché en líneas, las cuales, a su vez, condicionan el diseño de la memoria principal y la anchura del bus que une a ambas.

mejoras organizativas

La principal ventaja de actuar al nivel organizativo es que puede influirse decisivamente en el rendimiento de la memoria con un coste próximo a cero. Por tanto, no es de extrañar que hayan proliferado multitud de esquemas organizativos, incluso dentro de una misma arquitectura. La figura 10.12 resume las distintas divisiones organizativas en que se descompone un sistema de memoria principal tipo PC, así como su jerarquía. Describiremos los niveles superiores primero, ya que conforme avancemos hacia abajo en la jerarquía encontraremos mayor complejidad para su entendimiento. La figura 10.13 pone de manifiesto cómo se relacionan las tres dimensiones lógicas en que puede descomponerse el mapa de memoria con los componentes físicos que las implementan.

cómo procederemos

☛ pág. 42

☛ pág. 43

El sistema se compone de bancos

◀ 10.1

En una primera visión del sistema, podemos considerar la memoria del PC descompuesta en **bancos** según una división horizontal y disjunta, esto es, una petición a memoria está en uno y sólo uno de los bancos. Estos bancos pueden tener diferente longitud, pero tienen todos la misma anchura, que suele coincidir con la del bus de datos de la placa con objeto de maximizar el rendimiento de las transferencias con el procesador.

bancos

La utilidad de una descomposición en bancos es múltiple, a saber:

❶ Mientras un banco está funcionando, los demás se encuentran desactivados. Esto permite a los bancos compartir cierta circuitería sin incurrir en conflicto alguno, como los registros de entrada y/o salida de datos, o incluso a un nivel más interno, las líneas de lectura de datos de las celdas.

reutilización

❷ En lugar de un único bloque de memoria, ahora disponemos de un mosaico de bloques de diferente tamaño con los que componer un mapa de memoria que admite un sinfín de combinaciones y formatos comerciales.

flexibilidad

❸ Permite la configuración de un mapa de memoria heterogéneo con direcciones de memoria de diferentes características (por ejemplo, unas más veloces que otras). Bajo este supuesto, el controlador de memoria debe ser capaz de sincronizar la placa base y los módulos en cada caso que se le pueda plantear, conjugando a su vez un voltaje común que no sea nocivo.

heterogeneidad

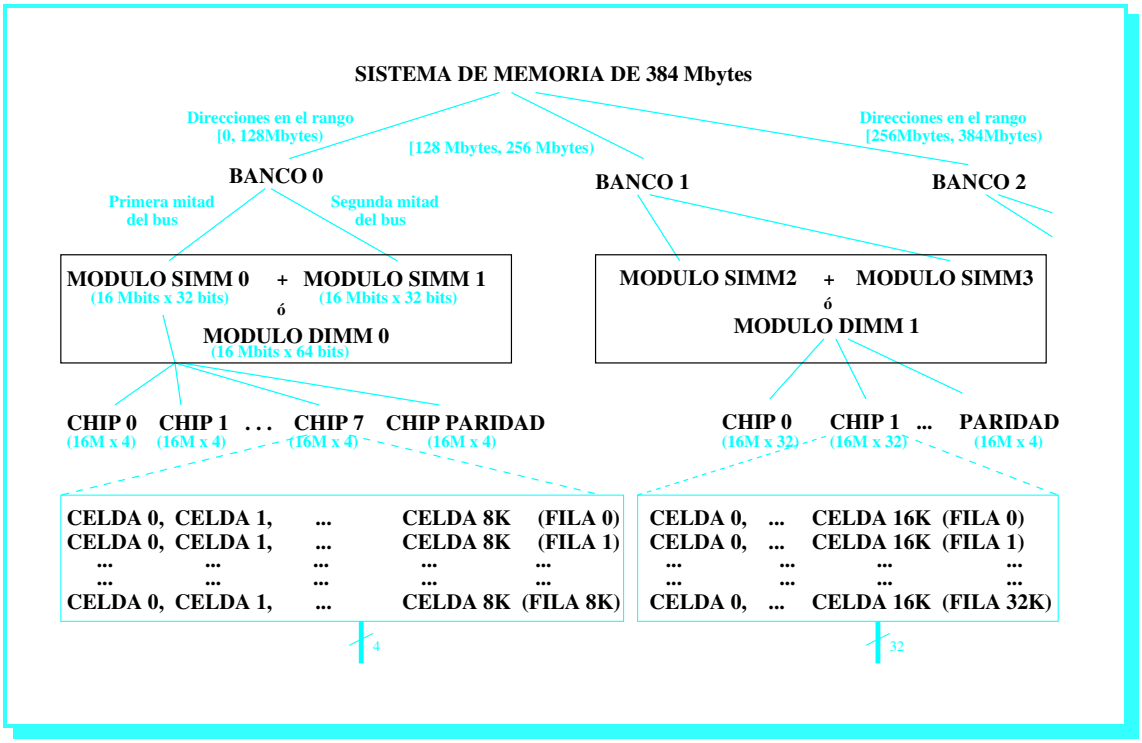


FIGURA 10.12: Las divisiones organizativas en que se descompone el mapa de memoria de un PC actual. El espacio de direcciones se divide en bancos, éstos dividen su anchura en módulos, y éstos a su vez se componen de pastillas o chips con idénticas características. Cada chip se organiza bidimensionalmente con objeto de minimizar el número de líneas de direccionamiento y aprovechar la secuencialidad de referencia del procesador.

tendencia

Dentro de la arquitectura del PC se tiende a aumentar el número de bancos de las placas base para abrir el abanico de posibilidades de configuración tanto cuantitativa como cualitativamente. Así, con la memoria SIMM de 30 contactos, lo normal era tener un único banco. En la memoria SIMM de 72 contactos, el caso más frecuente era dos bancos (cada uno con dos módulos). Y en la memoria DIMM, donde cada banco se implementa con un solo módulo, lo normal es encontrar tres e incluso cuatro bancos independientes.

10.2 ▶ Los bancos se componen de módulos

SIMM/DIMM

Cada banco del sistema de memoria suele descomponerse a su vez en partes iguales o **módulos**, que en el caso de los PC son los populares SIMM o DIMM. El número de módulos que integran un banco vendrá dado por el cociente entre la anchura del banco y la del módulo, dado que todos ellos deben tener idéntica longitud y anchura.

La descomposición de un banco en módulos es el resultado de una carrera perdida por parte de la memoria: La de tratar de seguir la estela del ancho del bus de datos del procesador. La tabla 10.7 resume la evolución de estas dos variables sobre los procesadores de Intel para PC en sus siete generaciones.

pág. 43 ☚

evolución

anchura del bus

1 bit

8 bits

Los primeros PC necesitaban tantos módulos como bits tenía su bus de datos, ya que cada módulo de memoria era en realidad una pastilla DIP de anchura igual a 1 bit. Posteriormente, con la llegada al mercado de los módulos SIMM de 30 contactos, el primer formato que aglutinaba múltiples chips en un solo módulo, la memoria llegó a una anchura de 8 bits, consiguiendo alcanzar al procesador 8088, y necesitando luego de dos módulos para dar servicio a los procesadores de 16

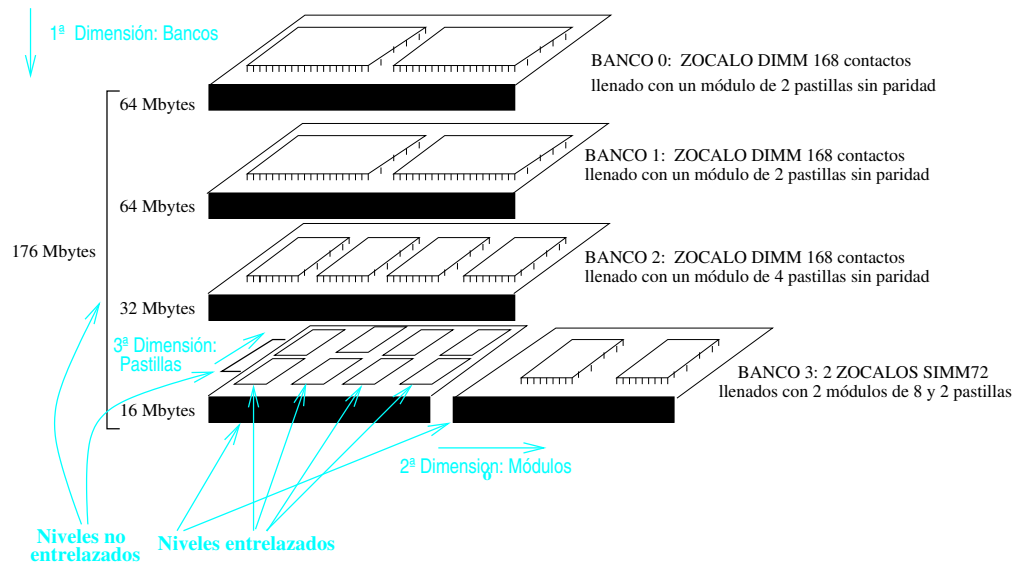


FIGURA 10.13: Vista desde un plano inferior de los zócalos de memoria de una placa base para PC. En ella podemos apreciar las tres dimensiones en que se organizan lógicamente las celdas de memoria: Bancos, módulos y chips.

Generación	Procesador de Intel	Anchura del bus de datos del procesador	Número y tipo de módulos comerciales Recomendable	Alternativo
Primera	8088	8 bits	1 de SIMM30	8 de DIP16
Primera	8086	16 bits	2 de SIMM30	16 de DIP16
Segunda	80286	16 bits	2 de SIMM30	16 de DIP16
Tercera	80386	32 bits	4 de SIMM30	4 de SIPP30
Cuarta	80486	32 bits	1 de SIMM72	4 de SIMM30
Quinta	Pentium	64 bits	1 de DIMM168	2 de SIMM72
Sexta	Pentium II y III	64 bits	1 de DIMM168	No tiene
Séptima	Pentium 4	64 bits	1 de DIMM184	2 de RIMM184

TABLA 10.7: Número de módulos de memoria necesarios en los distintos microprocesadores de Intel para PC dependiendo de la longitud del bus de datos y de la anchura del módulo de memoria comercial utilizado. Por orden de menos a más contemporáneo, un chip DIP16 (esto es, de 16 patillas) tiene una anchura de 1 bit, un SIPP o SIMM de 30 contactos tiene una anchura de 8 bits, un SIMM de 72 contactos proporciona 32 bits, y un DIMM de 168 ó 184 contactos, 64 bits.

bits (8086 y 80286). A raiz del nacimiento del bus de datos de 32 bits (procesadores 80386 y 80486), aparecen los módulos SIMM de 72 contactos, que consiguen de nuevo empatar con el procesador. Finalmente, la llegada del Pentium y el bus de datos de 64 bits impone de nuevo una división en dos módulos SIMM que desaparece con la llegada de los módulos DIMM de 168 contactos y 64 bits de datos.

16 bits
32 bits
64 bits

La obligada sincronización que requieren los módulos de memoria de un mismo banco para llenar la anchura total del bus condiciona a su vez la configuración. Por ejemplo, una arquitectura Pentium provista con zócalos SIMM72 debe venir siempre provista de un número par de zócalos, pues éstos deben llenarse de dos en dos para completar cada banco de memoria. Si dejásemos

condicionantes

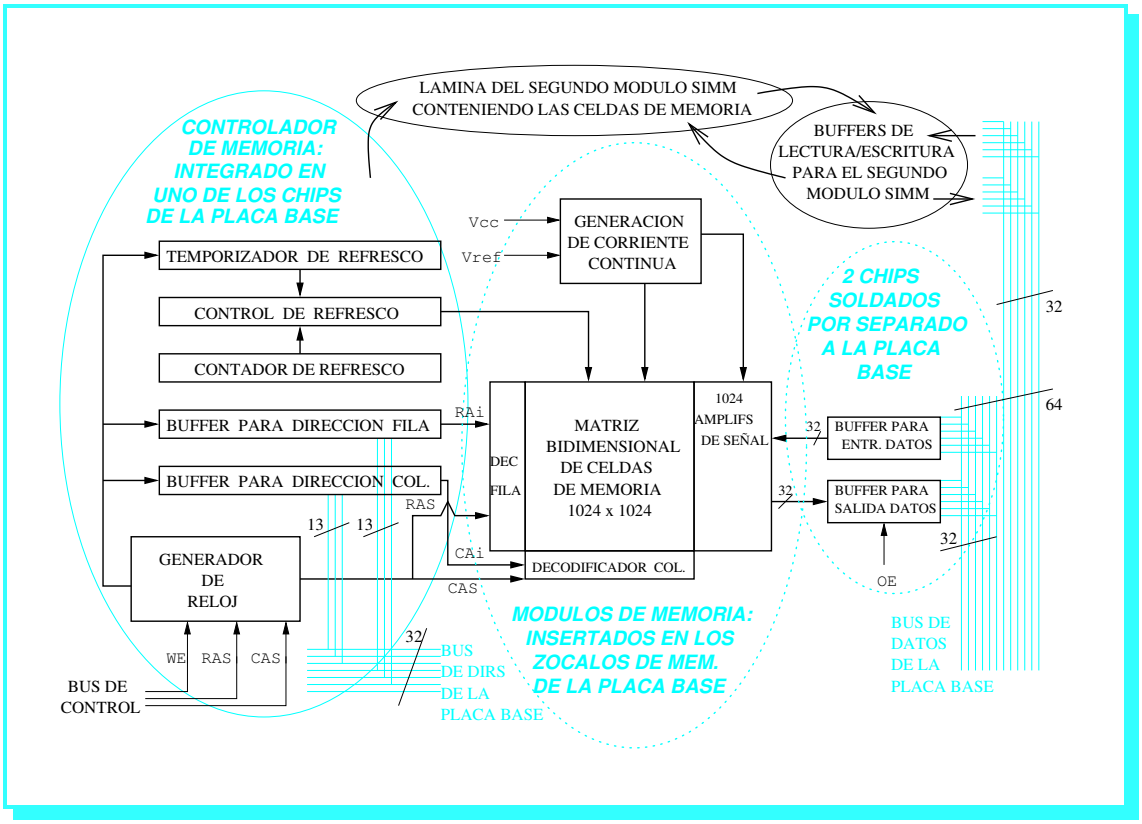


FIGURA 10.14: Diagrama de bloques de un sistema de memoria con su división en módulos. Puede apreciarse cómo la matriz de celdas es el centro neurálgico del mismo sobre el que circundan el resto de elementos. Para situarnos mejor y poder establecer una correspondencia con la visión externa que tenemos del hardware relativo a la memoria, hemos señalizado lo que forma parte del controlador de memoria, así como otros circuitos y registros proporcionados por la propia placa base.

algún módulo SIMM conectado sin su correspondiente pareja, habría accesos de datos que se quedarían en los 32 bits y no podrían llenar los 64 bits del bus que espera la caché para completar el tamaño de su línea de 8 bytes. Además, es deseable que los módulos tengan idéntica velocidad, ya que si uno es más rápido que otro, tendrá que esperar siempre a su pareja, con lo que en la práctica funcionará tan lento como ella.

La figura 10.14 detalla los componentes de un módulo y sitúa a éste en el contexto general de un sistema de memoria.

10.3 ▶Los módulos se componen de chips

pág. 25

Antiguamente, la memoria se adquiría por pastillas o chips cuyas patillas se pinchaban a la placa base de forma independiente (ver conexión DIP en la sección 10.6). Ahora, la unidad de memoria en lo que respecta a su compra y conexión a la placa base es el módulo, que se compone de un número determinado de chips.

apariencia

Los chips pueden apreciarse externamente con sólo echar un vistazo al módulo comercial. Internamente, los chips de un mismo módulo son todos iguales en sus parámetros de más bajo

diferencia

nivel (velocidad, voltaje, temperatura, ...). La única diferencia potencial estriba en que algunos chips pueden alojar bits de datos y otros bits de paridad, lo que puede dar lugar a una anchura

rol del chip

diferente en cada caso (hay un bit de paridad por cada ocho de datos).

Características del chip			Direccionamiento	
Tamaño (Mbytes)	Número de palabras	Ancho de palabra direccionable	Selección de fila	Selección de columna
2	2^{20}	16 bits	Pines 0 al 9	Pines 0 al 9
2	2^{21}	8 bits	Pines 0 al 10	Pines 0 al 9
2	2^{22}	4 bits	Pines 0 al 10	Pines 0 al 10
2	2^{22}	4 bits	Pines 0 al 11	Pines 0 al 9
8	2^{22}	16 bits	Pines 0 al 11	Pines 0 al 9
8	2^{23}	8 bits	Pines 0 al 12	Pines 0 al 9
8	2^{23}	8 bits	Pines 0 al 11	Pines 0 al 10
8	2^{24}	4 bits	Pines 0 al 12	Pines 0 al 10
8	2^{24}	4 bits	Pines 0 al 11	Pines 0 al 11

TABLA 10.8: Características de los chips de memoria DRAM con los que se montaron los módulos comerciales de Micron Technology bajo formato SIMM.

El chip es ahora la unidad de integración en silicio con la que cada fabricante montará sus módulos de memoria. En función de sus dimensiones será necesario conjuntar un número determinado de chips por medio de lógica adicional. La división en chips es por tanto una opción reservada al fabricante, que éste suele aprovechar para tener un mayor grado de flexibilidad en la confección de módulos comerciales. La tabla 10.8 muestra los diversos tamaños que utiliza la empresa Micron para integrar sus chips de memoria DRAM en formato SIMM.

pág. 45

anchura

La anchura del chip determina, por un lado, la unidad de palabra seleccionable al nivel de pastilla y, por el otro, el número de chips necesarios para fabricar el módulo (exceptuando a los módulos RIMM, en todos los SIMM y DIMM su anchura se obtiene como resultado de multiplicar la anchura del chip por el número de chips, ya que actúan todos simultáneamente).

evolución

Históricamente, la anchura del chip ha venido en aumento para dar servicio al mayor ancho de módulos y buses. Así, de la anchura de un bit se pasó rápidamente a la anchura de cuatro bits, denominándose este chip Nibble Mode DRAM (un nibble son cuatro bits). A partir de ahí, la anchura pasó a formar parte explícita del tamaño del chip. Por ejemplo, el primer chip de la tabla 10.8 no se referencia como una pastilla de 2 Megabytes, sino como de 1 Megabit x 16.

La longitud de cada chip (número de palabras) también muestra una clara tendencia al aumento, empujada por la necesidad de conseguir módulos de mayor capacidad al tiempo que se reduce el número de chips por módulo. Así, en los módulos DIMM y RIMM de 2003 con capacidad total de 512 Mbytes solemos encontrar chips de tamaño 64 o 128 Mbytes.

Los chips se componen de celdas

◀ 10.4

tipos de celdas

La **celda** es el elemento que almacena un bit de información, es decir, el átomo de la memoria. En el caso de la memoria dinámica, el cero o uno lógico del bit se registra con una carga positiva o negativa en un minúsculo condensador de 25-30 fF, cuya capacidad puede incrementarse cambiando el dieléctrico (material que separa las placas del condensador) de óxido o nitrido por óxido de tántalo (Ta_2O_5). En función del fabricante también puede cambiar el tipo de celda utilizado para la integración en silicio (llana, surco o loma son las tres denominaciones según su relieve).

amplificación

La diferencia de potencial que resulta en el circuito de memoria como consecuencia de la lectura de esta carga es tan pequeña que obliga a colocar circuitos de amplificación. Esto aumenta la latencia de la memoria, pero es el precio a pagar por utilizar unos condensadores tan minúsculos que permitan integrar millones de celdas en un mismo chip. Esta elección supone en cierta manera priorizar el tamaño frente a la velocidad del chip, algo que forma parte de la filosofía de diseño de las memorias dinámicas y que las distingue de otros diseños como las cachés.

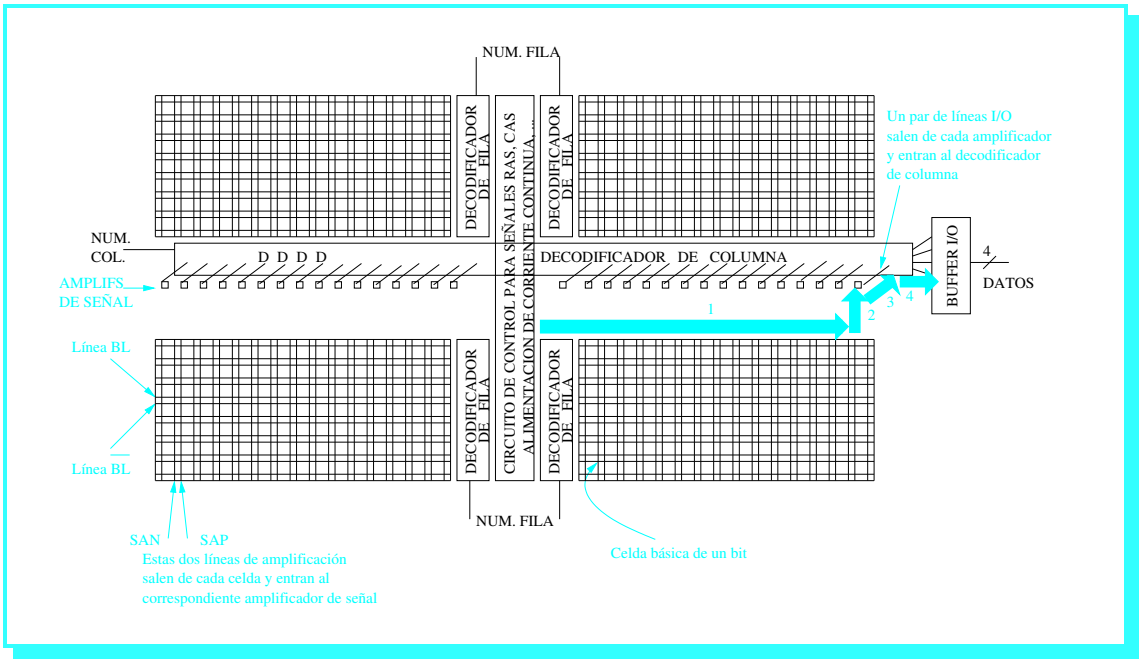


FIGURA 10.15: Arquitectura básica de un chip de memoria, con una malla bidimensional de celdas divididas verticalmente por los circuitos de control y horizontalmente por los amplificadores de señal. Esta disposición o *layout* consigue minimizar la longitud del cableado desde los elementos activos a cada una de las celdas. Respecto a su latencia, cuatro son los tiempos que en ella influyen (decodificación, activación, amplificación y salida), habiéndose marcado con flechas gruesas.

arquitectura

La **arquitectura** de un chip de memoria dinámica con su malla bidimensional de celdas se muestra en la figura 10.15. Las líneas verticales se denominan `I/O` (Input/Output - Entrada/Salida), y sirven para la lectura/escritura de datos de/en las celdas seleccionadas por medio de los decodificadores de fila y columna. Las líneas horizontales se denominan `BL` (*Bit Lines*) y se encargan de introducir una pequeña diferencia de potencial que se amplificará positiva o negativamente en cada celda en función del signo de la carga de su condensador asociado.

☛ pág. 46

líneas

compartir líneas

La diferencia de potencial que aparece en las líneas `BL` es siempre positiva al margen del signo de la carga del condensador. Esto permite a todas las celdas de la misma fila compartir el valor de tensión independientemente del dato que contengan. Por otra parte, dado que sólo se selecciona una fila en cada ocasión, se pueden compartir los circuitos de amplificación de señal para todas las filas, y como en cada columna de la malla hay a lo sumo una celda activa, se puede también compartir la pareja de líneas verticales I/O para todas las celdas de una misma columna.

ubicación de elementos

La lógica de control encargada del protocolo de activación de datos se ubica en el centro del chip partiendo por la mitad la malla de celdas de arriba a abajo para minimizar el cableado. Una estrategia similar se suele adoptar para ubicar los amplificadores horizontalmente, compartiendo las filas de la matriz de celdas para retener temporalmente los datos de la fila solicitada en espera de recibir la columna para sacar los datos al exterior.

número de amplific.

El número de amplificadores de señal se corresponde con la longitud de la fila de celdas, compuesta por una serie de columnas de N celdas, y serán sólo los N amplificadores de estas celdas los que se activen en cada chip para cada acceso, una vez proporcionada la coordenada de columna.

número de chips

Una pregunta que surge al observar el diseño de los chips de memoria es por qué no se implementa éste para abarcar el módulo completo. Si bien minimizar el número de chips favorece la sincronización interna del módulo (y por lo tanto, su velocidad intrínseca), supone también

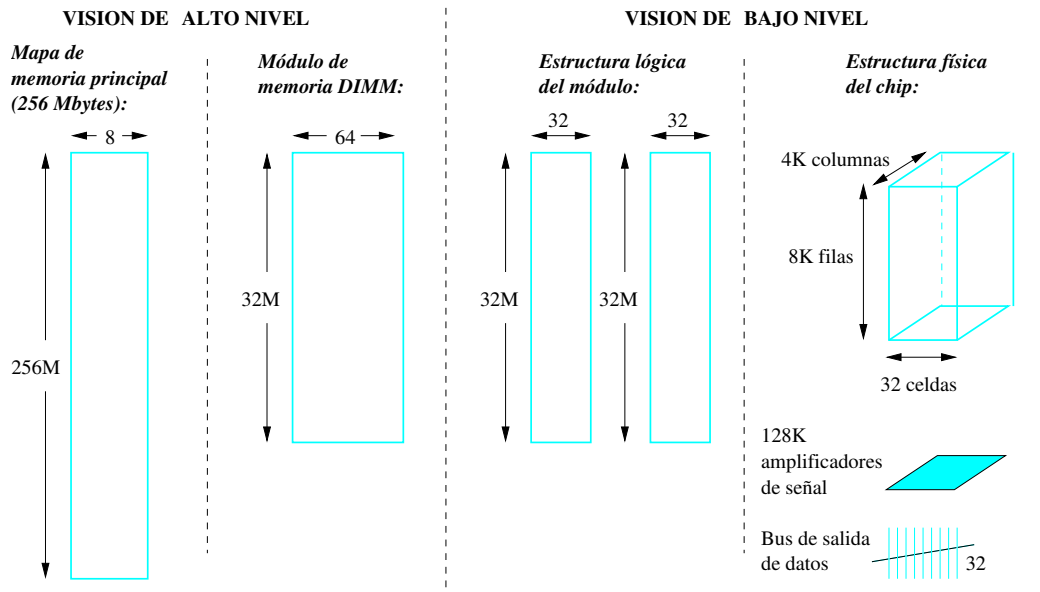


FIGURA 10.16: Descomposición lógica y física de la memoria del PC, desde la concepción del mapa de memoria a que estamos habituados (visión en Mbytes), a la estructura interna a nivel de celdas. Los estados intermedios atañen a los módulos, chips y amplificadores de señal. Nótese que el byte no es utilizado como anchura en ninguno de estos entes, por lo que no resulta adecuado tenerlo como referencia.

aumentar el tamaño de cada chip, creciendo a su vez tanto el número de líneas I/O como el de amplificadores de señal; lo primero aumenta el ruido de las señales y perjudica la fiabilidad del chip, mientras que lo segundo supone un aumento de la potencia consumida y la temperatura.

El ejemplo 10.4 ilustra la descomposición de dos módulos de memoria en chips, celdas y amplificadores, que completamos con la figura 10.16, donde se esquematiza el último caso.

Ejemplo 10.4: DESCOMPOSICIÓN DE MÓDULOS DE MEMORIA SIMM Y DIMM EN CHIPS Y CELDAS

Un módulo SIMM de 16 Mbytes fabricado con 8 chips dispone de:

- 4 Mpalabras de 32 bits en el módulo, organizadas lógicamente en una matriz de 2Kfilas x 2Kcolumnas de 32 bits.
- 4 Mpalabras de 4 bits en cada chip, organizadas físicamente en una matriz de 2Kfilas x 2Kcolumnas de 4 celdas.
- 8K (esto es, 8192) amplificadores de señal en cada chip para retener las 2Kcolumnas de 4 celdas una vez seleccionada la fila, y de los que sólo 4 de ellos sacarán datos al exterior una vez seleccionada la columna.

Un módulo DIMM de 256 Mbytes fabricado con sólo 2 chips contiene:

- 32 Mpalabras de 64 bits en el módulo, organizadas lógicamente en una matriz de 8Kfilas x 4Kcolumnas de 64 celdas.
- 32 Mpalabras de 32 bits en cada chip, organizadas físicamente en una matriz de 8Kfilas de 4Kcolumnas de 32 celdas.
- 128K (esto es, 131.072) amplificadores de señal en cada chip para retener las 4Kcolumnas de 32 celdas, de los que sólo actuarán 32 en cada acceso.

SECCIÓN 10.11

Entrelazado

de factor k

Un **entrelazado de factor** `k` descompone la memoria en `k` bloques de anchura y/o longitud inferior, con objeto de simultanear el acceso a estos `k` bloques a partir de una dirección común.

El parámetro `k` se determina en función del coste hardware del controlador, si bien también revierte positivamente en el rendimiento esperado, obteniéndose una ganancia máxima de factor `k`, siempre dependiente del número de accesos que puedan aprovecharse dentro del sistema.

Analogía 10.3: El entrelazado y la superescalaridad

Si bien la segmentación del procesador puede ser trasladada a la memoria de manera inmediata, la superescalaridad tiene su reencarnación aquí en forma de entrelazado.
Tanto la superescalaridad del procesador como el entrelazado de la memoria son estrategias orientadas a trabajar simultáneamente sobre múltiples peticiones, de cálculo sobre las ALU en el caso del procesador, y de acceso sobre las celdas en el de la memoria.
Así como en el procesador el factor de superescalaridad se encuentra emparentado con el diseño de las capas superiores del hardware, principalmente el conjunto de instrucciones, el factor de entrelazado se haya íntimamente ligado al diseño de la capas superiores de la jerarquía de memoria, en este caso, la memoria caché y su tamaño de línea.
Ahora bien, el entrelazado cuenta con dos argumentos para asumir factores más elevados:

1. No supone un coste tan elevado, ya que no se replica circuitería, sino que se organiza de forma más eficiente la ya existente. Aquí se parece más a la segmentación, ya que se incrementa la complejidad en la unidad de control, no en la de datos.
2. El parámetro al que se encuentra ligado su aprovechamiento admite valores más grandes: Resulta difícil encontrar un código con un centenar de instrucciones independientes que justifique un factor de superescalaridad, pero no una caché con líneas de 128 o 256 palabras de memoria principal.

En la práctica, el secreto de un buen entrelazado está, al igual que en la superescalaridad, en conjugar rendimiento y coste de forma que las capas hardware y software del equipo estén lo más compenetradas posible.

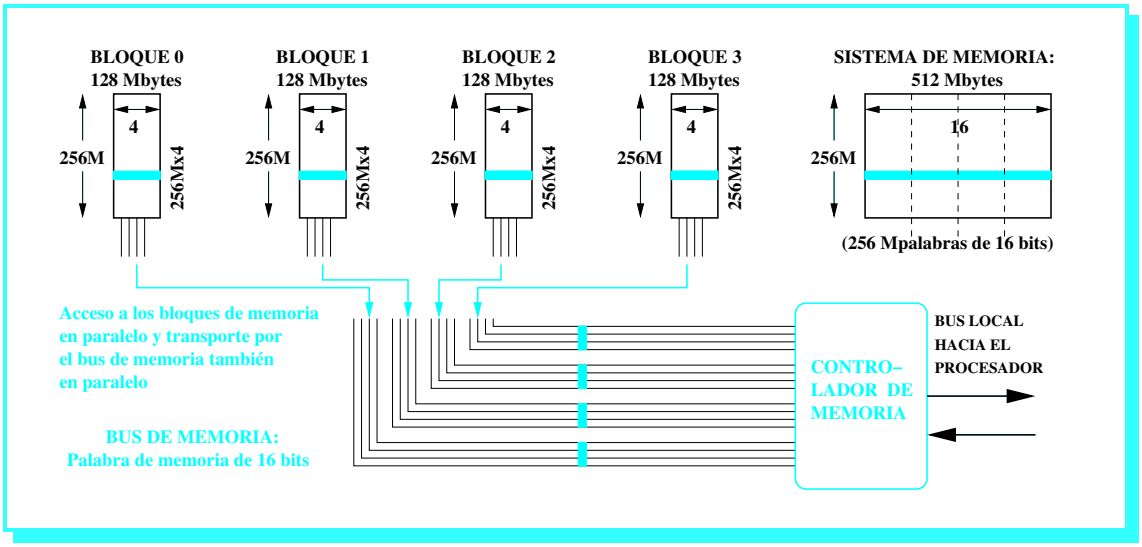


FIGURA 10.17: Entrelazado en anchura de factor $k = 4$ para los bancos de memoria principal.

En la literatura técnica, las unidades en las que el entrelazado descompone la memoria suelen denominarse módulos o bancos, pero nosotros las referenciaremos aquí como bloques de forma genérica para evitar confusiones con nuestros módulos y bancos de memoria del PC, que no necesariamente han de estar entrelazados.

`bloques`

Dimensión

◀ 11.1

La visión lógica de la memoria nos ofrece una sucesión (longitud) de palabras de una anchura determinada, conformando un espacio bidimensional. La descomposición en bloques que lleva a cabo el entrelazado puede actuar sobre cualquiera de estas dos dimensiones.

11.1.1 Anchura

Cuando la palabra de los bloques entrelazados tiene una anchura inferior a la que viaja por el bus de memoria, tenemos un **entrelazado en anchura**. Todos los bloques sincronizan su contribución con una parte de la palabra de memoria, volcando porciones disjuntas al bus, donde viajan en paralelo de forma conjunta (ver figura 10.17).

- Al nivel de banco no es posible aplicar este entrelazado, , ya que por definición la anchura del banco es siempre coincidente con la del bus de memoria.

  `imposible`

- Al nivel de módulo, puede aplicarse en buses de memoria de 32 bits montados con módulos SIMM de 30 contactos hasta la cuarta generación (factor $k = 4$), o más recientemente, en quinta generación con buses de 64 bits montados con una pareja de módulos SIMM de 72 contactos ($k = 2$) para extraer la mitad del bus de cada módulo accediendo a ambos de forma simultánea para conseguir una ganancia de dos. Más recientemente, también se recurre a esta ganancia de dos en las arquitecturas bajo Pentium 4 (bus local de 3.2 Gbytes/sg.) con memoria RDRAM de 16 bits (1.6 Gbytes/sg.), tal y como ilustramos en la figura 6.5.

  `eventual`

  ☛ *Volumen 1*

- Al nivel de los chips que constituyen un módulo, este entrelazado siempre tiene su oportunidad, ya que la anchura de los chips es inferior a la del módulo. En los módulos SIMM y

  `permanente`

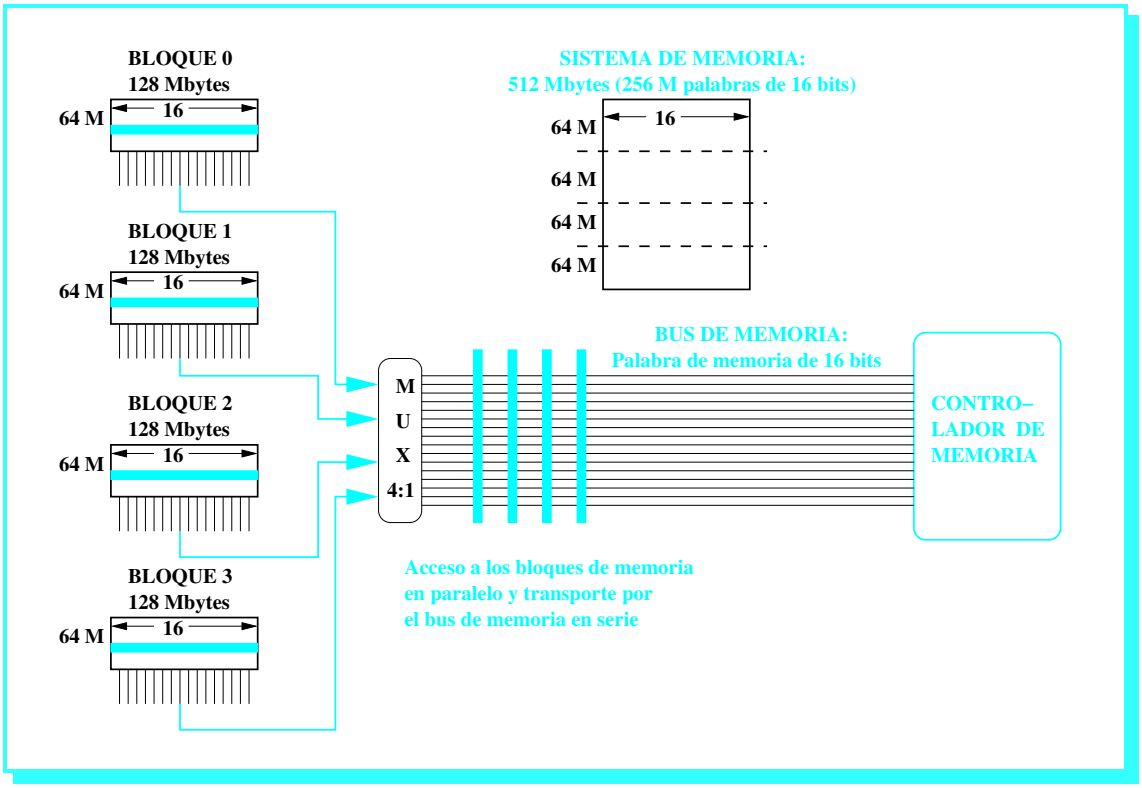


FIGURA 10.18: Entrelazado en longitud de factor $k = 4$ para los bancos de memoria principal.

DIMM, en los que todos los chips contribuyen al unísono para completar la palabra de memoria, se utiliza con factor 2, 4, 8, 16 y hasta 32, desconociendo el controlador de memoria de la placa base de cuántos chips proviene realmente la palabra recibida. En cambio, en los módulos RIMM se sacrifica este entrelazado, pues el controlador dialoga en cada acceso con un único chip del que obtiene todos los datos.

aplicación futura

En la sexta y séptima generación, el entrelazado en anchura se reserva exclusivamente al nivel de chips, ya que los módulos de memoria tienen la misma anchura que el bus de memoria. Más adelante, cuando el bus de memoria se atreva a alcanzar los 128 bits, veremos de nuevo renacer el entrelazado en anchura al nivel de módulo para desdoblar su palabra entretanto no se fabriquen módulos de 128 bits.

11.1.2 Longitud

Cuando el bus de memoria funciona a una velocidad superior a la de la memoria, podemos entrelazar al nivel de palabras completas para simultanear el acceso a todas ellas y secuencializar luego el transporte por el bus. El factor de entrelazado `k` determinará en este caso el número de viajes a realizar por el bus.

Lo hemos denominado **entrelazado en longitud** porque sus bloques se reparten el espacio de direcciones unidimensional (ver figura 10.18). En función de cómo realicemos este reparto, podemos distinguir dos esquemas básicos:

DOS ESQUEMAS:

- de orden superior

- **Entrelazado de orden superior**. Divide la memoria total en `k` bloques de igual tamaño cuyas direcciones contienen todas datos consecutivos de memoria. Esto tiene la ventaja de

simplificar ciertas tareas, como la expansión de la memoria del sistema o el diagnóstico y reparación de rangos de memoria defectuosos.

El problema de esta estrategia es su aplicación, ya que sólo tiene utilidad práctica en entornos multiprocesador, donde cada procesador ejecuta un programa que solicita direcciones de un bloque distinto y puede solapar sus peticiones con los demás. En multiprogramación no es aplicable porque aunque existan múltiples pogramas abiertos, sólo uno de ellos está activo en cada momento, y el flujo de peticiones a memoria proviene únicamente de él.

utilidad

- **Entrelazado de orden inferior**. Asigna de forma circular direcciones consecutivas a bloques de memoria consecutivos, esto es, reparte las palabras de memoria como los naipes de una baraja entre `k` jugadores.

  - de orden inferior

  Este entrelazado permite explotar la localidad de referencia a memoria en sistemas monoprocesador y multiprogramación, optimizando el flujo de datos hacia memoria caché, ya que ésta se estructura en líneas de `N` palabras consecutivas que constituyen su unidad de transferencia con memoria principal. Haciendo coincidir los parámetros `k` y `N`, la utilidad de este entrelazado en la arquitectura PC es extraordinaria.

  utilidad

  - Al nivel de banco, para que el entrelazado en longitud sea posible debe cumplir las condiciones de sus bloques, esto es, todos deben tener el mismo tamaño y sincronizar su velocidad. Llegamos así a un punto en el que debemos sacrificar tanto la flexibilidad en la confección del mapa de memoria como su eventual heterogeneidad, dos de las ventajas que justificaban la organización de la memoria en bancos.

    sacrificio

    Este sacrificio sólo puede justificarlo un gran aumento del rendimiento, pero la misma mejora que estamos tratando de conseguir aquí fue ya aprovechada en los años 80 por la memoria FPM RAM (ver sección 10.13.1) con la estructuración interna de los chips de memoria en filas y columnas: Los datos de una misma fila, que para el procesador son palabras consecutivas en longitud, salen con una celeridad muy superior que se corresponde con el tiempo de ciclo de la memoria.

    ☛ pág. 55

    ya aprovechado

    En consecuencia, no debe extrañarnos que el entrelazado en longitud al nivel de banco haya sido escasamente utilizado (sólo un juego de chips para placa base a lo largo de toda la quinta, sexta y séptima generación lo incluye dentro de su controlador de memoria, el 450NX, y está orientado al segmento de servidores).

    uso escaso

  - Al nivel de módulo, el entrelazado en longitud podría ser útil para multiplicar por `k` la velocidad de salida de datos sin necesidad de mejorar la latencia de sus chips constituyentes, conformando `k` bloques que respondieran en ciclos consecutivos del bus una vez transcurrido el período de latencia inicial común. Sin embargo, a las frecuencias a que se opera hoy en día, una sincronización tan fina resulta mucho más sencilla de establecer al nivel interno de los chips de memoria. El módulo confirma así su rol comercial, delegando las decisiones arquitecturales sobre sus chips.

    delegación

  - A nivel de chip, esta multiplicación de la velocidad en la salida de datos es utilizada por los chips DDRAM y RDRAM con $k = 2$, con el único matiz de que los dos bloques no actúan de forma simultánea, sino desfasados un semiciclo de reloj.

    en la salida

    Dentro del chip, aún se utiliza otro entrelazado en longitud para repartir las filas de las matrices de celdas entre `k` bloques, favoreciendo la implementación de los diseños segmentados como la SDRAM, DDRAM y RDRAM. Esto es consecuencia de que los amplificadores de señal de las filas de un chip actúan durante toda la operativa de columna, que suele dilatarse a lo largo de varias etapas de segmentación, apareciendo dependencias estructurales que se mitigan gracias al entrelazado habilitando varios bloques de amplificadores disjuntos donde podamos simultanear el acceso a las filas demandadas por las operaciones de memoria presentes en el cauce segmentado. Así, en SDRAM y DDRAM, donde la segmentación tiene tres etapas, este entrelazado tiene factor dos o cuatro, y en RDRAM, donde las etapas son siete, alcanza ya el factor 16.

    en las filas

Nivel de entrelazado		Anchura	Longitud (orden inferior)
Bancos		No es posible	No compensa: Ganancia cubierta por la ráfaga de salida
Módulos (según su zócalo de conexión a placa base)	SIMM30	$k = 4$ sobre buses de 32 bits	No se utiliza: Resulta más adecuado delegar su implementación al nivel de chip
	SIMM72	$k = 2$ sobre buses de 64 bits	
	DIMM168	No se utiliza	
	DIMM184	No se utiliza	
	RIMM184	$k = 2$ bajo Pentium 4	
	RIMM232	No se utiliza	
Chips (en la selección de fila)	Asíncronos	No es posible	No se utiliza
	SDRAM	No es posible	$k = 2, 4$
	DDRAM	No es posible	$k = 4$
	RDRAM	No es posible	$k = 16$
Chips (en la salida de datos)	Asíncronos	$k = 2, 4, 8, 16, 32$	No se utiliza
	SDRAM	$k = 2, 4, 8, 16$	No se utiliza
	DDRAM	$k = 2, 4, 8$	$k = 2$
	RDRAM	No se utiliza	$k = 2$

TABLA 10.9: Los esquemas de entrelazado (factor k) utilizados en memoria principal del PC.

prebúsqueda

Otro escenario en el que el entrelazado en longitud de orden inferior podría ser útil es aquel en el que la caché esté implementada con alguna estrategia de prebúsqueda en memoria principal que necesite de la transferencia simultánea de un grupo de líneas consecutivas desde memoria principal. Pero este escenario no es muy aconsejable en la arquitectura PC debido a que la prebúsqueda incide en una mayor congestión del bus local.

Ejemplo 10.5: MEJORAS EN EL ANCHO DE BANDA GRACIAS AL ENTRELAZADO

Tomemos como partida el sistema de memoria descrito en el ejemplo 10.2, donde calculábamos un ancho de banda de 74.43 Mbytes/segundo.

- Con un entrelazado de la memoria en longitud de factor 8, la salida de los datos desde memoria se reduce desde 80 ns hasta sólo 10 ns (los ocho bloques entrelazados responden al unísono en el tiempo de un solo acceso). La llegada de la dirección y el transporte de los datos, en cambio, quedan inalterados. La suma total queda: $2{,}5ns. + 10ns. + 20ns = 32{,}5ns.$ para 8 bytes de datos, lo que arroja un ancho de banda de 234.74 Mbytes/sg.
- Para que el entrelazado se aplicase en anchura con ese mismo factor 8 deberíamos considerar una palabra de memoria de 8 bytes para tener 8 bloques de 1 byte de anchura, y un bus de memoria de 64 líneas para que esta palabra viajara completa por el bus. En ese caso, la salida de los datos se reduce de igual forma que antes hasta los 10 ns., pero además el transporte requiere un solo viaje de 2.5 ns. En total, tenemos: $2{,}5ns. + 10ns. + 2{,}5ns = 15ns.$ para 8 bytes de datos, lo que arroja un ancho de banda de 508.60 Mbytes/sg, casi siete veces superior al de partida.

➤ La vertiente optimista de este escenario está en suponer que los ocho datos de la serie van a ser realmente solicitados a memoria; la pesimista, el no haber considerado la posibilidad de solapar el envío de un dato por el bus con el acceso al siguiente. Esto último ahorraría buena parte del tiempo adicional empleado por el entrelazado en longitud, dejándolo cercano al rendimiento dado por el entrelazado en anchura.

En definitiva, las variantes de diseño que nos ofrece el entrelazado son múltiples, y su aplicación es posible dentro de los tres niveles organizativos de la arquitectura PC: Bancos, módulos y chips. La tabla 10.9 sintetiza todas las posibilidades que hemos comentado. `conclusión`

SECCIÓN 10.12

Concurrencia

Hasta ahora hemos supuesto la existencia de un único sistema de memoria. Sin embargo, todos manejamos ya multitud de procesos de manera simultánea en nuestro PC, y algunos, incluso se atreven con equipos multiprocesador. Los más profanos podrían pensar que el sistema de memoria único que estamos describiendo responde a una arquitectura de memoria primitiva y obsoleta, y no es así. Para matizar todo esto vamos a diferenciar los casos en los que la arquitectura dispone de una única memoria principal de aquellos en los que es realmente un conjunto de sistemas que pueden dar servicio a múltiples peticiones de memoria simultáneas. `complejidad`

Entre las arquitecturas que sustentan un **sistema de memoria único**, existen tres posibilidades, que de menor a mayor complejidad son las siguientes: UNICO:

1. **Monoproceso.** Este es el PC de los años 80, funcionando bajo sistema operativo MS-DOS. `- obsoleto`
2. **Multiproceso.** Este es el PC actual, funcionando bajo Windows o Linux, donde podemos tener muchos procesos trabajando a la vez, y éstos a su vez referenciar a espacios de direcciones disjuntos y lejanos entre sí. `- actual`

   Aunque la capa software produce la ilusión de una ejecución simultánea, en realidad sólo hay un proceso que toma la CPU y usa el bus de acceso a memoria en cada momento, con lo que no es posible servir peticiones de memoria simultáneamente a varios procesos. El sistema operativo conmuta de un proceso a otro, concediendo a cada uno el pleno disfrute del PC durante turnos rotatorios de unos pocos milisegundos. Esta franja de tiempo es efímera para nosotros, provocándonos la sensación de una ejecución concurrente, pero al PC se le hace muy larga, sucediéndose en ella miles de peticiones a memoria. Por ello, desde el punto de vista del hardware, la conmutación de un proceso a otro es un evento improbable, y la ejecución multiproceso puede aproximarse a una monoproceso. `engañoso`
3. **Multizócalo.** Este el PC de la gama servidora, dotado de una placa base multizócalo que alberga varios procesadores . Su arquitectura cuenta con un único bus y un sistema de memoria común, con lo cual sigue existiendo un solo controlador de memoria que sirve las peticiones de todos los procesadores. `- servidor`

   La principal diferencia aquí es que las peticiones sí intercalan en el tiempo direcciones de espacios de memoria disjuntos, aquellos donde residen los procesos en ejecución, uno por cada procesador activo. Por ello, aún siendo un sistema de memoria único, tienen en él cabida esquemas como el entrelazado de orden superior. `factible`

Como computadores que auspician un **sistema de memoria múltiple**, sólo tenemos a las arquitecturas **multiprocesador**, donde la memoria se descompone en múltiples espacios de direcciones, cada uno de ellos con su propio controlador de memoria y sus buses separados para datos MÚLTIPLE:

y direcciones. Por supuesto, cada uno de estos sistemas puede a su vez descomponerse en bancos, módulos y chips como ya hemos visto para el caso de un sistema único de memoria.

coste alto

Este esquema permite dar servicio a múltiples peticiones de memoria aleatorias de manera simultánea, pero el coste de su implementación es muy elevado, lo que lo hace prohibitivo para las arquitecturas domésticas. Suele ser frecuente encontrarlo en los supercomputadores de memoria compartida, donde además se integra con algún sistema de interconexión de bus compartido para poder acceder desde un procesador al sistema de memoria de sus homólogos.

utopía

En sistemas monoprocesador, contar con múltiples sistemas de memoria podría ser útil para simultanear accesos a memoria. Por ejemplo, con cuatro sistemas de memoria, uno de ellos se podría ocupar de llenar una línea de caché de datos, otro podría estar haciendo lo propio con una línea de caché de instrucciones, un tercero podría estar dando servicio al disco duro, y el último podría encargarse de las transferencias hacia la memoria de vídeo. Aunque sencilla sobre el papel, esta posibilidad se torna irrealizable a la hora de sacarle provecho desde la capa software.

SECCIÓN 10.13

Arquitectura e interfaz

Abandonamos la vertiente organizativa de la memoria para adentrarnos en la estructura interna de sus chips y el interfaz que sus módulos establecen con el controlador de memoria principal.

condensador

Los continuos avances en el campo de los microprocesadores demandan tecnologías para la memoria principal que estén a la altura de su velocidad. Aunque la memoria principal se fabrica con tecnología CMOS y anchura de puerta en los transistores muy similares a las del procesador, no debemos olvidar que su *alma mater* no es un transistor, sino un condensador, y esto le ha privado de buena parte de la aceleración que ha arropado al procesador. Esto explica que hoy en día un procesador de 4 GHz se monte sobre una placa base de 533 MHz, operando ocho veces más rápido que memoria principal.

pág. 5

100 veces más lenta

Pero si extrapolamos la gráfica evolutiva del procesador y la memoria a nuestros días (ver figura 9.1), desde principios de los años 80 en que ambas líneas se cruzan, el procesador del 2003 sería 2^{13} veces más rápido que el de entonces, mientras que la memoria sería 2^2 veces más veloz. Comparativamente, el procesador sería unas 2.000 veces más rápido que su memoria principal, o sea, que a un procesador de 4 GHz le correspondería una memoria de 2 MHz, esto es, 100 veces más lenta que la que refleja la realidad del PC a fecha 2003.

pieza clave: el interfaz

Algo no cuadra, y es que las piezas sólo encajarán cuando veamos de qué manera tan decisiva ha ayudado el interfaz. En definitiva, los niveles internos de los chips que conforman un módulo de memoria no han sufrido cambios que justifiquen gran velocidad, pero las formas de diálogo con el procesador y la caché nos han brindado portentosas mejoras, a las que vamos a dedicar buena parte del presente capítulo.

FPM
EDO
BEDO
SDRAM
DDRAM
RDRAM

El punto de partida de las optimizaciones que atañen al interfaz lo constituyó la memoria FPM (Fast Page Mode) a mediados de la década de los 80, momento a partir del cual se encadenaron sucesivas mejoras con las memorias EDO (Extended Data Output) y BEDO (Burst EDO). Con posterioridad, estas memorias se mostraron incapaces de aguantar los buses por encima de 66 MHz, extinguiéndose finalmente con la llegada del bus a 100 MHz. Aparecieron entonces las memorias síncronas (SDRAM), que con un tiempo de ciclo ya por debajo de 10 ns consiguieron entenderse con esa frecuencia de bus y los peldaños superiores de 133 y 166 MHz. Con la DDRAM se consiguió duplicar la frecuencia máxima, alcanzándose los 166x2 MHz. El peldaño superior, que ya se equipara en ancho de banda con el bus de memoria de 533 MHz y 64 bits lo constituye la RDRAM, cuyo último diseño de 2003 presenta 533x2 MHz y 32 bits de anchura.

Mejora en el interfaz	Año	Tamaño	Tiempo de respuesta 1ª palabra	Tiempo de respuesta Línea caché	Tiempo de acceso a fila	Tiempo de acceso a columna	Tiempo de ciclo
FPM	1980	64 Kbit	250 ns	-	180-150 ns	75 ns	150 ns
FPM	1983	256 Kbit	220 ns	-	150-120 ns	50 ns	100 ns
FPM	1986	1 Mbit	190 ns	-	120-100 ns	25 ns	50 ns
FPM	1989	4 Mbit	165 ns	-	100-80 ns	20 ns	40 ns
FPM	1992	16 Mbit	120 ns	-	80-60 ns	15 ns	30 ns
EDO	1995	64 Mbit	90 ns	165 ns	70-50 ns	10 ns	25 ns
SDRAM	1998	256 Mbit	40 ns	70 ns	20 ns	20 ns	10 ns
DDRAM	2001	1 Gbit	40 ns	55 ns	20 ns	20 ns	5 ns
DDRAM	2004	4 Gbit	20 ns	27.5 ns	10 ns	10 ns	2.5 ns

TABLA 10.10: Evolución del tiempo de respuesta y el tiempo de ciclo de la memoria en las tres últimas d'ecadas. El tiempo de respuesta indica la latencia en memorias no estructuradas bidimensionalmente (esto es, variantes no FPM), y a partir de la memoria EDO, sería el tiempo de acceso a fila mas el tiempo de acceso a columna para la primera palabra de la línea de caché, y de tan sólo el tiempo de ciclo para las tres siguientes. Completando la salida de estas cuatro palabras, obtenemos el tiempo de respuesta para la línea de caché.

Ahora bien, al igual que ocurre con los microprocesadores, los usuarios que mueven ficha primero suelen pagar los platos rotos de una bisoña implementación. Deberíamos aprender de ejemplos como la SDRAM, caótica en su comportamiento hasta que no pasaron un par de años desde su nacimiento, para evitar caer en la tentación de comprar una RDRAM de último grito a las primeras de cambio. Hasta llegar ahí, nos espera una larga cadena de eslabones intermedios donde resulta más seguro y económico aferrarse.

precaución

Fast Page Mode RAM (FPM RAM)

◀ 13.1

Nos encontramos frente a la más vetusta y menos sofisticada de las memorias RAM para PC. Para entender su funcionamiento debemos remontarnos al origen de los PC.

origen

Recordemos que el crecimiento en el tamaño de la memoria también sigue la Ley de Moore (duplicar cada 18 meses), y ya desde edades tan tempranas como los años 80, esto provocó un gran incremento del patillaje de los chips de memoria, encareciendo su coste de manera relevante.

patillaje

La primera medida que se aplicó para mitigarlo fue una estructuración de las celdas del chip en una malla bidimensional, lográndose así una reducción del número de patillas de dirección a la mitad. El impacto de esta medida es incluso mayor desde el punto de vista del interfaz con memoria principal, ya que ahora una dirección debe descomponerse en una primera mitad o fila (cuyo suministro se señaliza desde una línea RAS - *Row Access Strobe*), y una segunda mitad o columna que se multiplexa por las mismas patillas que la anterior y se controla desde una línea CAS (*Column Access Strobe*).

bidimensional

RAS

CAS

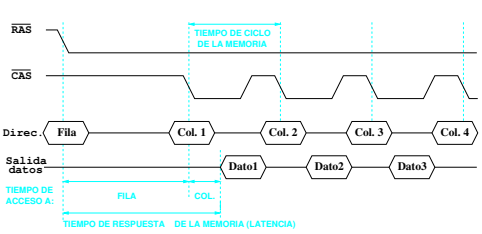
El proceso se completa internamente con la inserción de un circuito que captura el dato procedente de la fila en espera de recibir el acceso a columna, el cual selecciona ya la celda o grupo de celdas del chip que formarán junto a la salida de otros chips la palabra de memoria del módulo.

captura

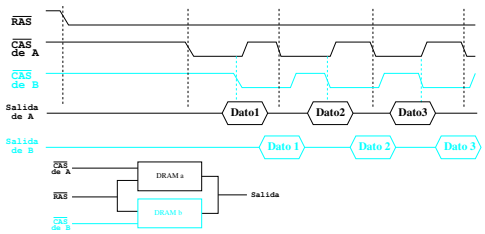
La selección de la columna es bastante más rápida que la de la fila, por lo que cuando varios accesos consecutivos están localizados en la misma fila de celdas, la memoria responde en un tiempo bastante inferior, ya que el dato de la fila se encuentra en un sencillo circuito externo al que sólo es necesario suministrar la columna para extraer cada dato adicional. Esta diferencia llevó a mejorar enormemente el tiempo de ciclo de la memoria frente a su tiempo de respuesta.

simplicidad

Este tipo de memorias se denominaron FPM DRAM (*Fast Page Mode Dynamic RAM* - memo-



(a) FPM convencional.



(b) FPM entrelazada.

FIGURA 10.19: Interfaz de diálogo para una memoria FPM convencional (a) y entrelazada en dos módulos (b). En ésta última se comparte RAS y se desdobla CAS con un desfase de tres ciclos.

localidad

ria dinámica de modo de página rápido), y su peculiar operativa de funcionamiento es muy apreciada por la capa software del equipo: Las mismas propiedades de localidad espacial y temporal que exhíben los algoritmos respecto a la secuencia de peticiones a memoria (y que fundamentan la jerarquía de memoria del computador) pueden ser aprovechadas aquí para reutilizar los datos de una misma fila en sucesivos accesos, y que sea el tiempo de ciclo el que determine la latencia efectiva de la memoria.

predominio

Desde entonces, el tiempo de ciclo cobra mayor relevancia en el rendimiento que el tiempo de respuesta, aunque los fabricantes de memorias no adoptarían éste último como parámetro insignia hasta la llegada de la SDRAM. Este cambio justifica la discontinuidad que aparece a mediados de la década de los 90 en las especificaciones comerciales de la memoria (ver sección 10.14 en general y la tabla 10.18 en particular), pasando súbitamente de estar situadas en torno a los 40-50ns para las memorias BEDO, a quedar entre los 7 y 12 ns. para las SDRAM, cuando sus prestaciones reales están bastante más próximas entre sí. La tabla 10.10 resume la evolución que han experimentado estos dos parámetros en los últimos veinte años.

pág. 89
pág. 89
pág. 55

página=fila

La *página* a que hacen referencia las siglas FPM es en realidad lo que aquí hemos denominado (y denominaremos) **fila**. Preferimos huir del vocablo página para no provocar confusión con la memoria virtual, donde este término se utiliza para referenciar a la unidad de transferencia de datos entre memoria principal y disco.

tamaño fila

Los chips de memoria principal utilizan una fila de celdas cuyo tamaño es igual a la raiz cuadrada del espacio de almacenamiento del chip. Si la raiz cuadrada no es potencia de dos, la fila será la dimensión mayor. Por ejemplo, en un chip de 1 Mbyte (8 Mbits) tendremos una matriz de celdas de 4 Kfilas por 2 Kcolumnas.

latencias

La FPM RAM se comercializó con **latencias** de 70 ns y 60 ns para el acceso a una fila, aunque inicialmente su uso estuvo vetado a procesadores por encima de los 100 MHz o buses con frecuencias de 33 MHz. Esto se debía a que las líneas de datos tenían que reiniciarse *antes* de que la memoria comenzara a volcar los datos de salida de la nueva petición. Puesto que esta inicialización se llevaba a cabo cuando llegaba el flanco de subida de la señal CAS (ver figura 10.19.a) y consumía entre 10 y 15 ns, la próxima transición de bajada de la señal CAS debía retrasarse ese

5-5-5-5

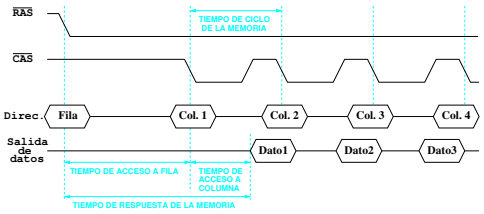
mismo tiempo, conduciendo a una espera de 5 ciclos entre cada dos accesos a la misma columna.

entrelazado

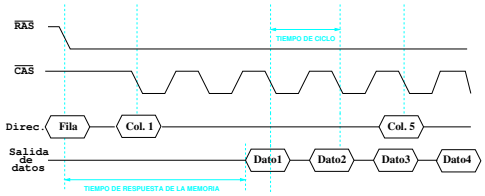
Aprovechando la migración al bus de 64 bits, este problema se solventó entrelazando en anchura los módulos por pares, con objeto de que un módulo respondiera mientras el otro se inicializaba y viceversa. Esto redujo el retardo anterior a tan sólo 3 ciclos, permitiendo su uso en buses con frecuencias de 66 MHz. La figura 10.19.b muestra esta operativa de funcionamiento. La señal RAS se comparte por el par de módulos SIMM72, mientras que la señal CAS de un módulo se desfasa tres ciclos con respecto a su pareja, produciendo una secuencia de latencias de 5-3-3-3-3-3-3-3

5-3-3-3

ciclos de la placa base para una salida de ocho datos consecutivos.



(a) EDO DRAM.



(b) BEDO DRAM.

FIGURA 10.20: Comparativa de temporización para el interfaz de diálogo que definen las memorias EDO (a) y BEDO (b), con la reducción de los tiempos de ciclo y de acceso a columna en ésta última.

La secuencia de temporización anterior determina el rendimiento de estos módulos de memoria, ya que conviene tener presente que un acceso a memoria dinámica siempre se produce porque antes no se ha encontrado el dato correspondiente en la caché, y desde aquí no se pedirá a memoria únicamente ese dato, sino toda la línea de caché de la que éste forma parte. Como el módulo SIMM72 tiene una anchura de 32 bits de datos (4 bytes) y prácticamente la totalidad de las cachés de estos equipos presentan un tamaño de línea de 32 bytes, concluimos que la situación más realista es siempre una petición de *ocho datos consecutivos*, que denominaremos **ráfaga** (del inglés, burst), y ésta va a ser nuestra referencia a partir de ahora para comparar el rendimiento que ofrecen el resto de memorias que aún nos quedan por analizar.

caché

ráfaga

Nótese que en esta secuencia de ocho datos consecutivos hay dos optimizaciones conjuntas:

OPTIMIZ:

1. A nivel de banco, el entrelazado en anchura de los módulos SIMM, que acelera la salida de dos datos consecutivos desde los módulos.

entrelazado

2. A nivel de módulo, la organización bidimensional en filas y columnas de sus chips constitutivos, que para los tres datos que siguen al primero de la ráfaga en cada módulo, simplifican el interfaz de diálogo suministrando únicamente la columna.

matriz

Cuando alcancemos el formato DIMM de 168 contactos y anchura de 64 bits, veremos cambios en estas dos facetas:

CAMBIOS PARA 64 BITS:

1. El entrelazado en anchura ya no tiene sentido al nivel externo del módulo, sino que éste se traspasa al nivel interno de todos los chips que componen una fila de celdas de la matriz bidimensional, siendo necesario el concurso del doble de chips que harían falta sobre el módulo SIMM de 72 contactos.

entrelazado

2. La longitud de la ráfaga de datos quedará reducida de ocho a cuatro, al ser cada dato el doble de grande.

ráfaga

Extended Data Output RAM (EDO DRAM)

◀ 13.2

Para tratar de aminorar el tiempo de ciclo de la memoria, esto es, el que corresponde a la salida de los siete datos que suceden al primero de la secuencia, se ideó un nuevo tipo de memoria que recibió el nombre de **EDO** (*Extended Data Output - salida de datos extendida*). Esta memoria también se denominó memoria de modo hiperpágina.

hiperpágina

La memoria EDO incorpora un **latch** donde los datos de salida permanecen hasta el siguiente

nuevo latch

flanco de bajada de CAS, desacoplando la salida de un dato con la petición del siguiente y evitando la operación de inicialización. Su interfaz de diálogo se muestra en la figura 10.20.a. Esto conduce a un retardo de sólo 2 ciclos para leer cada dato, con lo que se consigue responder a razón de 5-2-2-2 ciclos para la salida de los cuatro datos de 64 bits de la misma fila de un módulo si estamos ante un formato DIMM, y 5-2-2-2-2-2-2-2 si nos encontramos en formato SIMM.

5-2-2-2

El ahorro de un ciclo en el acceso a columna supone en la práctica poder reducir tanto el tiempo de ciclo, t_{col}, como el tiempo de respuesta, $t_{fila} + t_{col}$. Las memorias EDO comerciales se presentaron en dos **latencias** para el tiempo de respuesta: 60ns y 50ns.

velocidades

Dado que el mercado sigue una clara tendencia hacia la operación síncrona entre memoria y procesador, el sincronismo de la memoria EDO (siempre dos ciclos de diferencia entre datos y bajo las mismas transiciones de CLK, RAS y CAS) se convierte en una característica bastante apetecible comparada con sus predecesoras. Esto, unido a la total compatibilidad de la memoria EDO con su precursora la FPM, supuso la rápida aceptación del mercado por esta nueva modalidad como puente hacia la memoria DRAM completamente síncrona, la SDRAM.

sincronismo

13.3 ▶Burst Extended Data Output RAM (BEDO RAM)

La memoria **BEDO** (*Burst EDO - EDO en ráfaga*) trata de aprovechar el hecho de que los datos en una arquitectura de sexta generación se piden por ráfagas de cuatro datos en cada uno de sus dos módulos para mejorar su acceso.

idea

Las diferencias de funcionamiento con respecto a la EDO son básicamente dos:

segmentación

❶ Se sustituye el latch de salida de datos por un registro que permite desacoplar completamente la salida de un dato con la selección de la columna del siguiente. Esto permite segmentar ambas operaciones (*pipelining*), lo que se traduce en una disminución del tiempo de ciclo a partir del segundo dato de la ráfaga, que ahora puede realizarse en un solo ciclo de reloj. La inclusión de este registro introduce un ciclo adicional de CAS para cargar el cauce de segmentación interno, pero no produce retardo adicional en la salida del primer dato: Dado que este ciclo se produce siempre una vez transcurrido el tiempo de acceso a fila, al ser éste mucho mayor que el tiempo de acceso a columna, podemos ocultar éste último anticipando un poco la selección de la columna, tal y como se aprecia en el cronograma de la figura 10.20.b.

pág. 57 ☚

contador de dirección

❷ Se introduce además un contador de dirección que genera internamente las cuatro direcciones de columna de forma automática, siendo necesario presentar externamente tan sólo la primera de ellas. Esto lleva a una simplificación funcional que reduce la secuencia de eventos necesarios para su funcionamiento.

Como resultado de esta doble optimización, la memoria BEDO permite responder a razón de 5-1-1-1 ciclos de reloj para una secuencia de cuatro datos consecutivos de 64 bits (formato DIMM168). La figura 10.20 permite también apreciar la disminución del tiempo de ciclo con respecto a la memoria EDO, a la vez que también se reduce el tiempo de acceso a columna.

5-1-1-1

pág. 57 ☚

13.4 ▶Synchronous Dynamic RAM (SDRAM)

La memoria **SDRAM** o memoria RAM dinámica síncrona trae como novedad el funcionamiento sincronizado con el controlador de memoria inmerso en placa base. El sincronismo presenta múltiples ventajas desde el punto de vista operativo:

sincronismo

VENTAJAS:

temporización

❶ Simplifica la temporización, al no tener que hacerse cargo de retrasos en las señales.

segmentación

❷ Optimiza el rendimiento del cauce segmentado (ver figura 10.22).

pág. 61 ☚

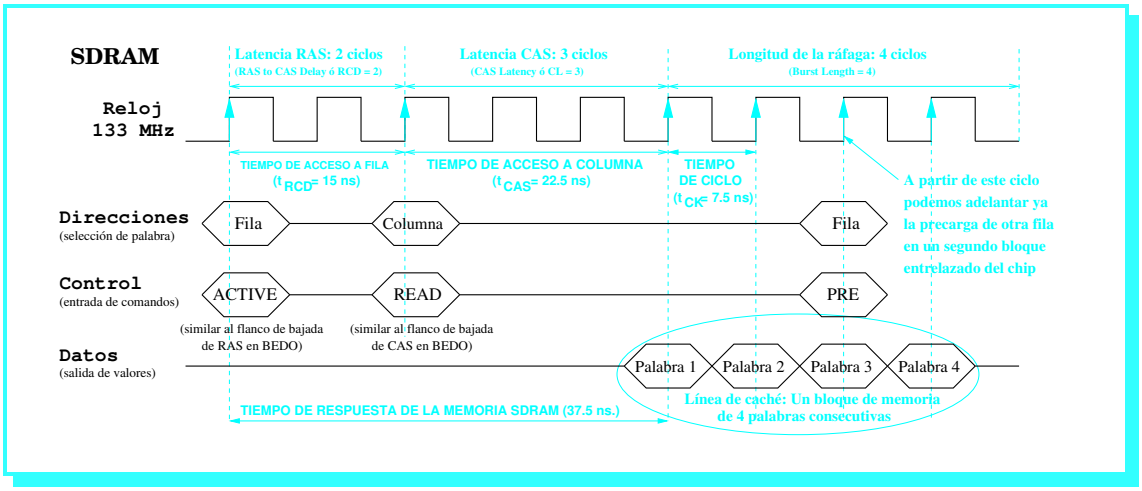


FIGURA 10.21: Interfaz de diálogo con memoria SDRAM, donde puede apreciarse tanto su carácter síncrono como la diferencia entre su tiempo de acceso y su tiempo de ciclo.

❸ Incorpora un doble entrelazado: `entrelazados:`

- En anchura al nivel de módulo, que es lo que en última instancia provoca que sólo podamos encontrar SDRAM en formato DIMM. Esto conduce a un sincronismo total entre el zócalo de memoria y el reloj de la placa base, permitiendo la aceptación de frecuencias más elevadas que la EDO o BEDO (precisamente es la asincronía intermodular inherente a las memorias EDO y BEDO lo que les impide funcionar correctamente por encima de los 66 MHz, aún a pesar de su sincronismo intramodular). `- en anchura`
- En longitud al nivel interno de los chips, con objeto de solapar peticiones de distintas filas en la matriz de celdas. Lo describiremos más adelante (ver sección 10.13.4.5) puesto que antes conviene conocer los detalles del interfaz SDRAM. `- en longitud`

☛ pág. 64

13.4.1 Programación

`programación`

En su fase de inicialización, la placa base programa la SDRAM indicándole los parámetros que establecen la ráfaga a utilizar en todos los diálogos posteriores. El controlador de memoria en placa base (puente norte del juego de chips) toma estos parámetros de la RAM-CMOS, unas veces predefinidos por el fabricante y otros configurables desde la BIOS según indicamos a continuación:

- LONGITUD DE LA RÁFAGA (*Burst Length*). Posibilidades: 1, 2, 4, 8 o una fila completa. Valor predefinido en todos los PC: 4 (número de accesos para llenar una línea de caché). `longitud`
- TIPO DE RÁFAGA (*Burst Type*). Posibilidades: Secuencial (0-1-2-3) o entrelazada (1-0-3-2). Valor predefinido en todos los PC: Secuencial. `tipo`
- MODO DE ESCRITURA (*Write Burst Mode*). Posibilidades: La ráfaga completa o una dirección aislada. Valor predefinido en todos los PC: Ráfaga completa (condicionado también por la manipulación de líneas completas de caché). `modo`
- LATENCIA CAS (*CAS Latency*). Posibilidades: 2 ó 3 ciclos. Puede programarse desde la opción SDRAM CAS LATENCY TIME de la BIOS (ver menú CHIPSET FEATURES SETUP - sección 24.3.4). La sección 10.13.4.7 nos descubre que la segunda opción suele ser siempre más conveniente. `latencia`

☛ Volumen 4
☛ pág. 67

Comando SDRAM	Bus de control				Bus de direcciones		Bus de datos
	CS#	RAS#	CAS#	WE#	BA0-BA1	A0-A12	DQ0-DQN
INHIBIT	1	X	X	X	X	X	X
NO OPERATION	0	1	1	1	X	X	X
ACTIVE	0	0	1	1	Banco	Fila	X
READ	0	1	0	1	Banco	Col.	X
WRITE	0	1	0	0	Banco	Col.	Palabra
BURST TERMINATE	0	1	1	0	X	X	Pal.
PRECHARGE	0	0	1	0	Código		X
AUTO/SELF REFRESH	0	0	0	1	X	X	X

TABLA 10.11: El patillaje de un chip SDRAM agrupado según su relación con el bus de memoria. Se describe además la actividad de cada patilla en función del comando emitido por el controlador SDRAM desde el bus de control. (CS: Chip Select. RAS/CAS: Row/Column Access Strobe. WE: Write Enable. BA: Bank Address. DQ: Data. X: Valor irrelevante. #: Patilla activa a la baja.).

A partir de aquí, todos los datos se extraen bajo la ráfaga establecida, actuando los comandos en el flanco de subida de la señal de reloj de la placa base, tal y como ilustra la figura 10.21.

13.4.2 Tiempos de acceso

operativa

La operativa de funcionamiento más usual en SDRAM comporta el uso del comando `ACTIVE`, seguido de `READ`/ `WRITE` para las ráfagas de lectura/escritura, respectivamente. Junto con `ACTIVE` se proporciona la dirección de fila, y junto con `READ`/`WRITE` la dirección de columna. La operativa básica para esta memoria puede desglosarse en tres partes:

$t_{RCD} = 2$

❶ El tiempo de acceso a fila o latencia `RAS`, cuyo valor más usual es de 2 ciclos de reloj. Se denota por t_{RCD} (RAS to CAS Delay) [2]. Indica el tiempo que tardan en actuar los amplificadores de señal al nivel de la fila de celdas sobre la matriz de cada chip. Es equivalente al tiempo de acceso a fila en EDO y BEDO. De hecho, los chips de SDRAM también disponen de líneas RAS y CAS, aunque éstas van codificadas en el bus de control, habilitándose según el comando que se envíe (ver tabla 10.11).

t_{CAS} =2 ó 3

pág. 61

pág. 66

pág. 67

❷ El tiempo de acceso a columna o latencia `CAS`, valor programable que oscila entre 2 y 3 ciclos. Se señaliza mediante t_{CAS}, dando un tiempo en nanosegundos, o también empleando las siglas CL (CAS Latency), indicando un número de ciclos de reloj. En la sección 10.13.4.3 describiremos la segmentación interna al nivel de columna que disecciona este tiempo, en la figura 10.24 su relación con la frecuencia del bus, y en la sección 10.13.4.7 la forma de elegir este parámetro para lograr una configuración óptima (en general, a mayor frecuencia del bus de memoria, mayor latencia CAS deberá tener su módulo).

ráfaga = 4

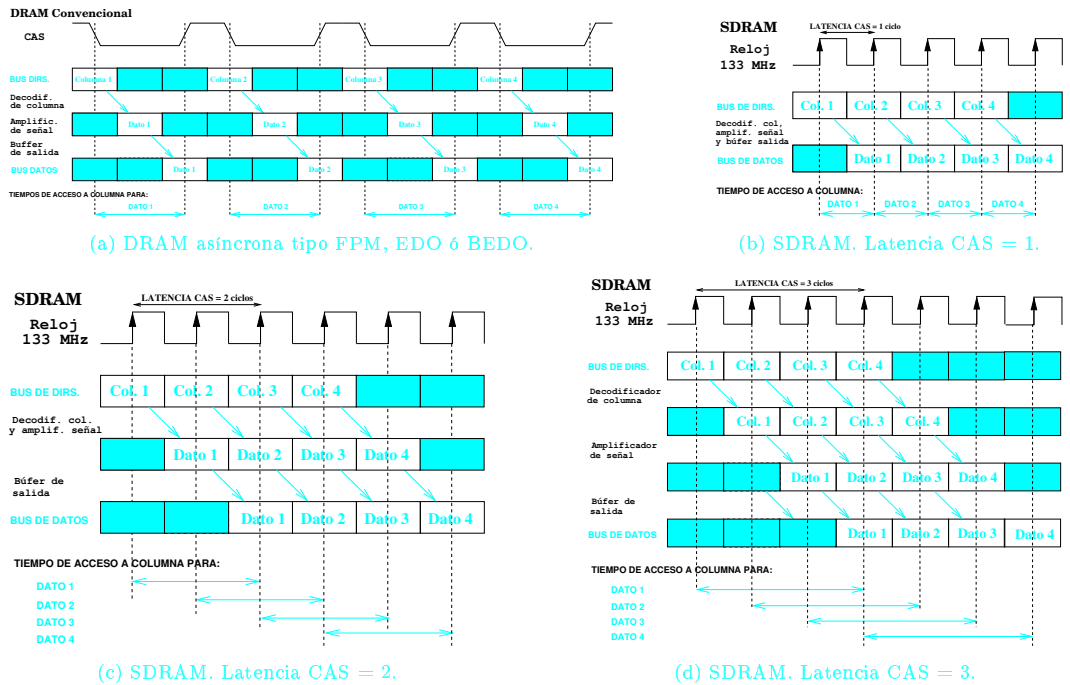
❸ El tiempo de salida para completar la ráfaga, que será siempre de 3 ciclos, ya que se compone de 4 datos y la segmentación posibilita la salida de un dato por ciclo de reloj.

referencias

La suma del tiempo de acceso a fila y columna constituye una vez más el **tiempo de respuesta** para esta memoria, y la inversa de la frecuencia o ritmo de salida de datos, su **tiempo de ciclo** (*Clock Cycle Time*), que se indica mediante t_{CK} [3].

[2] No confundirlo con t_{RAS}, Row Active Time, que indica el tiempo que permanece ocupada una fila durante cada acceso, y que suele estar en torno a los 6 ciclos de reloj.

[3] No confundirlo con Row Cycle Time, t_{RC}, que indica el mínimo lapso de tiempo que debe transcurrir entre dos comandos ACTIVE de diferentes filas del mismo bloque, y que suele ser $t_{RCD} + t_{CAS} + 3 - 1$. No confundirlo tampoco con RAS to RAS Delay, t_{RRD}, que aunque suele ser también de dos ciclos, representa el mínimo lapso de tiempo que debe transcurrir entre dos comandos ACTIVE de bloques distintos.



(a) DRAM asíncrona tipo FPM, EDO ó BEDO.

(b) SDRAM. Latencia CAS = 1.

(c) SDRAM. Latencia CAS = 2.

(d) SDRAM. Latencia CAS = 3.

FIGURA 10.22: El funcionamiento segmentado de la memoria SDRAM. (a) Una memoria DRAM asíncrona, tipo FPM, EDO o BEDO. (b) SDRAM con latencia CAS de 1 ciclo. (c) SDRAM con latencia CAS de 2 ciclos. (d) SDRAM con latencia CAS de 3 ciclos.

Hemos ubicado esta terna de tiempos (t_{RCD}, t_{CAS} y t_{CK}), sobre la figura 10.21, donde se toman latencias RAS y CAS de 2 y 3 ciclos, dando un tiempo de respuesta cinco veces superior al tiempo de ciclo. Si lo que se quiere es una referencia única ligada a la frecuencia, consultar la especificación PC-XXX (ver sección 10.14.2.1).

pág. 59 ☚

pág. 91 ☚

13.4.3 Segmentación

El sincronimo inherente a la SDRAM permite aplicar la idea de la segmentación que ya vimos para el procesador (ver sección 3.3.1) a dos niveles:

DOS NIVELES:

☛ Volumen 1

- En la operativa al nivel de columna, se obtiene una ráfaga de cuatro datos consecutivos en precisos períodos de un ciclo de reloj, con objeto de que la salida se sincronice con el bus de memoria. Si la latencia de CAS fuese siempre de un ciclo, lo tendríamos hecho (sería la idílica situación que se refleja en la figura 10.22.b), pero habitualmente es de 2 ó 3 ciclos, lo que obliga a segmentar en esas mismas etapas para que el ritmo de salida de datos aumente hasta uno por ciclo de reloj. Las etapas de segmentación deberán acogerse a la duración del tiempo de ciclo de la memoria, administrándose según el caso entre las unidades funcionales involucradas según se muestra en la figura 10.23.

- columna

☛ pág. 62

- En la operativa al nivel de fila, conviene observar que el decodificador de fila y la matriz de celdas quedan ociosos una vez ha llegado el comando READ que traslada la actividad al nivel de columna. El bus de direcciones y el bus de control quedan libres también a partir de ese instante, pudiendo ser aprovechados para iniciar un nuevo acceso a memoria mientras se está finalizando el último. Para ello, el controlador de memoria emite un comando

- fila

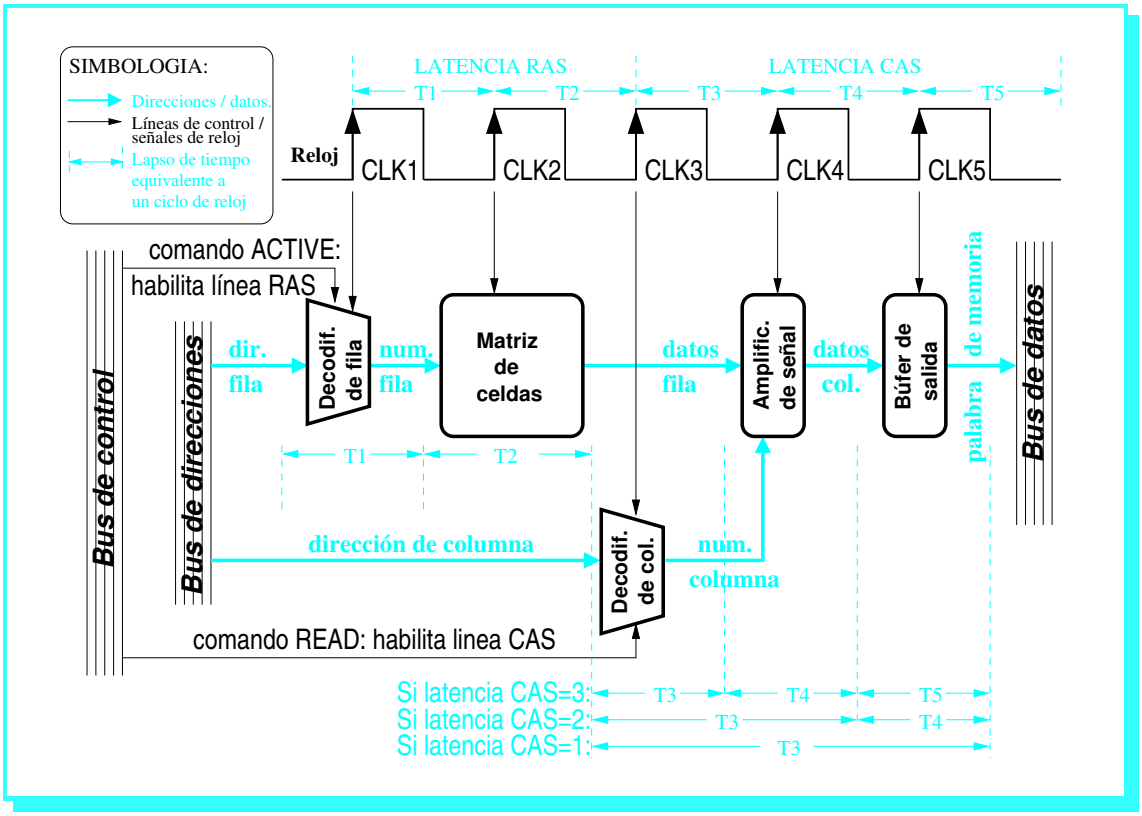


FIGURA 10.23: Segmentación de la SDRAM al nivel de columna para lograr una salida de los cuatro datos que componen la ráfaga en ciclos consecutivos de reloj. Señalamos además todas las unidades funcionales que intervienen en cada ciclo según los valores dados para la latencia de CAS.

PRECHARGE

☛ pág. 59

PRECHARGE, que como máximo puede adelantarse hasta el instante en que sale el tercer dato de la ráfaga por el bus de datos (ver figura 10.21). Esto es así porque el bus de datos no va a quedar libre hasta dos ciclos más tarde, y el mínimo que tardará en necesitarlo este nuevo acceso sería de dos ciclos (uno para la latencia RAS y otro para la latencia CAS en condiciones ideales).

De los nueve ciclos que consume un acceso a la SDRAM en el caso de que la latencia RAS sea dos, la latencia CAS sea tres y la ráfaga tenga una longitud de cuatro, tenemos: Los dos primeros ciclos segmentados inter-acceso para solapar la salida de datos de un acceso con la entrada de la dirección de fila del siguiente, y los siete siguientes segmentados intra-acceso para lograr la salida en ráfaga de cuatro datos dentro de un mismo acceso. Esta última segmentación es más efectiva, no sólo por contar con un mayor número de etapas, sino por tener garantizado el llenado del cauce en todas sus etapas. En cambio, el aprovechamiento de la segmentación al nivel de fila depende de que existan accesos pendientes de ser servidos.

efectividad

análisis t_{RCD}:

La latencia RAS o RCD suele omitirse en las especificaciones técnicas de una memoria SDRAM, lo cual tiene una doble justificación:

t_{RCD} cte.

❶ Suele ser de dos ciclos de forma casi generalizada.

t_{RCD} sin uso

❷ Apenas entra en juego. El creciente tamaño de las memorias y la mayor longitud de fila respecto a la de columna que han utilizado muchos fabricantes aboga por tamaños de fila de hasta 128 Kbits por chip, en los que para una anchura de chip de 4 bits caben hasta 32Kpalabras de memoria. La fila actúa así como una caché más cercana a memoria principal donde puede aprovecharse la localidad de referencia (ver analogía 10.4).

pág. 65 ☛

Tamaño		Acceso secuencial				Acceso aleatorio			
		SIMM		DIMM		SIMM		DIMM	
Area de datos	Unidad de transfer.	FPM 60ns	EDO 60ns	EDO 60ns	SDRAM 10ns	FPM 60ns	EDO 60ns	EDO 60ns	SDRAM 10ns
1 Kbyte	4 bytes	398	403	408	388	298	311	317	292
1 Kbyte	8 bytes	457	459	469	446	454	464	471	447
100 Kbytes	8 bytes	61	61	62	61	61	61	62	61
1 Mbyte	8 bytes	36	38	38	43	36	38	38	43

TABLA 10.12: Comparativa del ancho de banda (en Mbytes/segundo) ofrecido por la memoria principal para sus diferentes tecnologías en función del tamaño de bloque a transferir y la unidad de transporte. La latencia RAS es de 2 ciclos en SDRAM y la latencia CAS es de 3 ciclos.

Dicho de otra forma: Por cómo referencia el software a memoria, lo normal es que el controlador de memoria emplee muchos comandos `READ` por cada `ACTIVE`, ya que la probabilidad de que un nuevo acceso se encuentre en la misma fila que el anterior es muy elevada. Todos estos accesos se inician operando sobre el decodificador de columna, ahorrándose el paso por el decodificador de fila y la matriz de celdas, y recortando dos ciclos al tiempo de respuesta.

`READ/ACTIVE`

Por todo ello, resulta más realista comparar las memorias SDRAM empezando en ese punto, tal y como muestra el firmware de la BIOS a la hora de informarnos de una ráfaga 3-1-1-1 (para CAS = 3) ó 2-1-1-1 (para CAS = 2), y otro tanto ocurre en los catálogos de los fabricantes.

`3-1-1-1`

13.4.4 Rendimiento frente a memorias asíncronas

Contrastando frente a memorias EDO/BEDO, recordaremos que el 5-1-1-1 que éstas indican como mejor salida se toma desde RAS, y que al sumar en SDRAM los dos ciclos de RAS al 3-1-1-1 que parte de CAS, nos queda un marcador igualado de 5-1-1-1 en ambos casos, esto es, 5 ciclos del bus de memoria como tiempo de respuesta y uno solo como tiempo de ciclo. A pesar de lo que pueda parecer, las ventajas de SDRAM no son pocas, ya que:

`frente a EDO`

1. Estamos comparando el mejor supuesto de BEDO (tiempo de ciclo de uno) con el peor de SDRAM (latencia CAS de tres). .
2. Las posibilidades de la SDRAM para robar los dos ciclos de la fila son muy superiores, al contar con las armas de la precarga y la segmentación.
3. La capacidad de la SDRAM para trabajar en buses de memoria de frecuencia mucho más elevada que los que permiten la operativa asíncrona de las FPM, EDO y BEDO las convierten en la única alternativa de futuro.

$t_{CAS} \leq 3$

$t_{RCD} = 0$

`escalable`

Veamos de qué manera se plasman estas mejoras potenciales sobre el rendimiento de aplicaciones reales. Utilizaremos el ancho de banda para comparar las prestaciones de los tipos de memoria principal vistos hasta aquí: FPM, EDO, BEDO y SDRAM.

`ancho banda`

El benchmark ó programa de evaluación utilizado es el `Nbench` 2.1, que opera trasladando bloques de información de un tamaño determinado de unas posiciones de memoria a otras. El tamaño del bloque nos va a permitir de paso analizar el rendimiento de la caché y su influencia en el ancho de banda de la memoria principal.

`benchmark`

El equipo es un Pentium a 166 MHz con placa base P55TVP4 de Asus dotada del chipset 430VX de Intel y que incorpora una caché L2 externa segmentada de 256 Kbytes. Sabemos que es un equipo obsoleto, pero es uno de los pocos que se entiende con los cuatro tipos de memoria que pretendemos comparar.

`equipo PC`

Los resultados obtenidos se resumen en la tabla 10.12:

caché L1

❶ Las dos primeras filas muestran sobre todo el rendimiento de la caché L1 integrada, ya que el bloque de datos a mover cabe íntegramente en los 8 Kbytes que el Pentium incorpora como caché de datos. El ancho de banda resultante de mover los datos de 8 en 8 bytes es considerablemente mayor que de 4 en 4 porque así se aprovecha mejor tanto el ancho del bus del Pentium como el tamaño de su línea de caché. En estos dos casos, el ancho de banda de la SDRAM queda por debajo de la EDO RAM.

caché L2

❷ La tercera fila indica el rendimiento para bloques de 100 Kbytes, que ya se salen de la caché L1 pero quedan dentro de la L2 de placa base, por lo que es el rendimiento de este chip lo que realmente se evalúa. De ahí que los resultados sean independientes del tipo de memoria principal utilizado, e incluso insensibles a la aleatoriedad de referencia. No se refleja en las tablas, pero un aumento de caché L2 hasta los 512 Kbytes tampoco tiene efecto alguno sobre el ancho de banda. Sin embargo, el prescindir de la caché L2 tiene un enorme impacto sobre él, descendiendo hasta los 20 Mbytes/segundo.

memoria principal

❸ La cuarta fila compara el rendimiento de memoria principal. Estos resultados muestran que la memoria EDO tiene el mismo rendimiento en formato SIMM y en formato DIMM, lo cual es evidente a sabiendas de lo que significa el cambio de formato en EDO (doblar el ancho de salida de datos en un módulo DIMM y entrelazar en anchura a nivel de módulo es igual que forzar la inclusión de módulos SIMM por parejas y entrelazar luego al nivel de banco en la placa base). También vemos que SDRAM gana en ancho de banda, consecuencia de haber amortizado ya el coste de la latencia RAS al haberse referenciado a multitud de columnas que comparten una misma fila, pero que su ganancia no es significativa.

En consecuencia, la principal virtud de la SDRAM sigue siendo su escalabilidad a frecuencias elevadas, no compensando su inclusión en buses de frecuencia 66 MHz o inferiores, donde el

margen de seguridad

margen de seguridad que hubo que concederle en sus inicios para paliar su inestabilidad dilapidó las ventajas provenientes de su mejor diseño (entrelazado, segmentación y precarga).

13.4.5 Entrelazado en longitud

Los comandos `ACTIVE` y `PRECHARGE` pueden ser mucho más efectivos si se combinan con un entrelazado en longitud. Veamos en qué puede ayudarnos este ingrediente.

Por un lado, para que la latencia RAS de un comando ACTIVE se amortice desde muchos READ/WRITE, todos ellos deben compartir la misma fila; por otro lado, cuando un comando `PRECHARGE` referencia a una nueva fila, ésta reemplaza a la última que se tenía en los amplificado-

localidad

res de señal, algo nada conveniente, pues por localidad de referencia sabemos que el software va a necesitarla con prontitud.

Para evitar esto, la SDRAM distribuye el total de filas de la matriz de celdas de un chip de memoria en `K` bloques entrelazados en longitud, habilitando `K` filas de amplificadores de señal,

double-sided

uno para cada bloque. Esto provoca a su vez el desdoble de las líneas RAS y CAS por bloques entrelazados, y así, se denomina SDRAM *single-sided* la que cuenta con dos señales RAS y CAS (para estos dos bloques) y SDRAM *double-sided* la que dispone de cuatro.

Ahora, los READ/WRITE que amortizan la latencia RAS son todos los que se encuentran en el subconjunto de `K` filas, y además, la precarga de una fila no fuerza el sacrificio de la anterior,

reemplazo

puesto que puede situarse sobre otro bloque del entrelazado.

Al nivel de chip, la selección del bloque entrelazado sobre el que actúa cada comando se rea-

direccio-namiento

liza desde un par de líneas (`BA0`, `BA1`) de su patillaje, pudiendo considerarse parte del bus de direcciones, ya que actúan conjuntamente con las líneas que multiplexan la selección de fila y co-

lumna, permaneciendo activas en ambos instantes. Como el direccionamiento se gobierna desde el controlador de memoria en placa base, la forma de repartir las filas entre los bloques, y por tanto, el tipo de entrelazado a utilizar en los chips SDRAM no lo deciden ellos, sino que queda en manos del controlador de memoria. Pero como ya adelantamos, el mejor esquema a adoptar en un entorno PC es el entrelazado en longitud de orden inferior con 2 o 4 bloques como máximo (notar que sólo se dispone de dos líneas para direccionar el bloque).

Ya advertimos que la fila de la matriz de celdas se estaba comportando como si fuera una pequeña caché, y con este nuevo ingrediente, estamos en disposición de completar la analogía.

Analogía 10.4: La fila de una SDRAM asume el rol de caché

Una caché trata de acelerar el acceso a memoria delimitando un pequeño subconjunto que concentre la mayoría de peticiones a memoria, para a renglón seguido beneficiarse de una ágil implementación en base a invertir el principio "Más grande, más lento". Aunque nace con una pretensión diferente (ahorrar costes en el patillaje requerido para el direccionamiento del chip), la fila de una matriz de celdas es una delimitación en clara sintonía con esta idea: La llegada del sincronismo y la segmentación interna a la memoria le han puesto en la mano ayudar al rendimiento actuando como un búfer que retiene los accesos más recientes a memoria, precisamente los más referenciados según el principio de localidad que también trata de explotar la caché.

Elemento con que se establece la analogía	Caché en la jerarquía de memoria	Fila en la matriz de celdas DRAM
Unidad física para el bit	Biestable (SRAM)	Amplificador de señal
Unidad de transporte y gestión	Línea de caché de 256 bits	Fila de la matriz de 2048, 4096 ó 8192 bits
Direccionamiento a los datos	Etiqueta en directorio caché	Primera coordenada en la matriz de celdas
Delimitación que captura localidad como bloque	El tamaño de línea	El tamaño de fila
Ente que explota localidad a nivel organizativo	La asociatividad de líneas distribuidas en conjuntos	El entrelazado de filas repartidas en bloques
Rasgos indispensables para optimizar	Sincronización y segmentación	Sincronización y segmentación

La analogía se corrobora a nivel comercial: Los fabricantes de SDRAM emplean sólo dos bloques entrelazados en sus chips fabricados con anterioridad al año 2000, y hasta cuatro bloques en los diseños posteriores, opción que prevalece ya en DDRAM. Justo lo que sucede con las cachés, donde comercialmente se comenzó con un tamaño de conjunto de una o dos líneas, predominando en séptima generación un tamaño de cuatro u ocho. En ninguno de los dos casos se esperan crecimientos mucho mayores en el futuro, en vista del improbable aprovechamiento por parte de una localidad de referencia de unos programas cuya constitución interna apenas ha cambiado en los últimos veinte años.

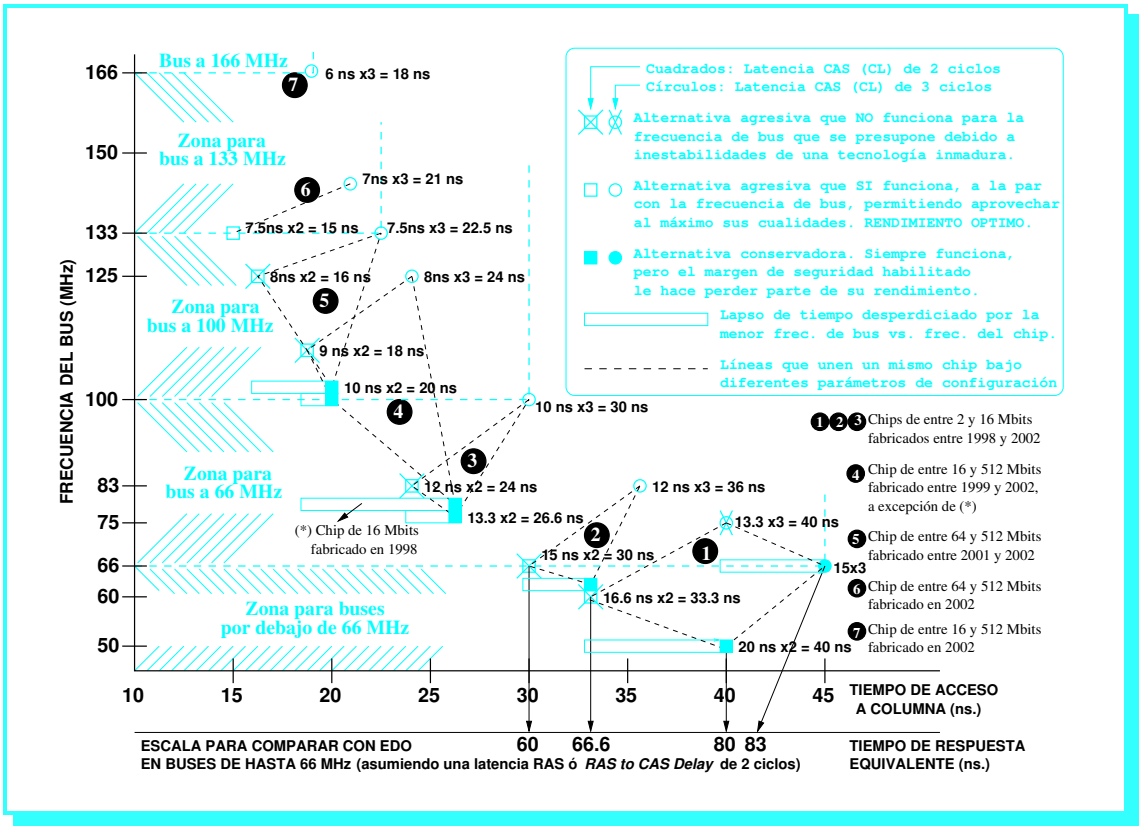


FIGURA 10.24: Los diseños de Micron para SDRAM clasificados según frecuencia de bus (en ordenadas) y tiempo de acceso a columna (en abcisas). Sólo los chips programados con latencia CAS de 3 ciclos soportan la máxima frecuencia, inversa de su tiempo de ciclo, que es su temporización más agresiva. Abajo del todo, comparativa con el tiempo de respuesta de la memoria EDO equivalente para buses de hasta 66 MHz.

13.4.6 Versiones

velocidad

La memoria SDRAM se comercializa con un tiempo de ciclo entre los 15 ns. (1998) y los 6 ns. (2003), siendo ésta última la versión que proporciona pleno rendimiento para un bus de 166 MHz.

variantes

La figura 10.24 ilustra todas estas variantes tomando como referencia los diseños comerciales de Micron en el período 1998-2003. Allí podemos ver cómo las versiones con latencia CAS (CL) de 3 ciclos consiguen el menor tiempo de ciclo y por tanto, serán las que permitirán trabajar al chip a

frecuencia

su máxima frecuencia. Esta frecuencia sería la inversa del tiempo de ciclo en todos los casos si no fuera por el período de inestabilidad que atravesó esta tecnología en sus inicios, obligando a habilitar el margen de seguridad que desperdiciaba parte de su velocidad. El margen de seguridad se fue estrechando conforme aumentaba la frecuencia gracias a la maduración del diseño: En la figura 10.24 vemos que se concede a los chips 1 al 5 para CL=2, pero sólo al chip 1 para CL=3. Se da así la paradoja de que para CL=3 arriesgamos menos en la temporización a pesar de aguantar una frecuencia superior.

margen de seguridad

SDRAM vs. EDO

En la parte inferior del eje de abcisas se incluye una escala comparativa entre el tiempo de acceso a columna de un chip SDRAM y su tiempo de respuesta equivalente para buses de hasta 66 MHz, que es la métrica de rendimiento utilizada por las memorias EDO/BEDO. Ya avisamos de que el tiempo de ciclo en una memoria síncrona resulta entre cuatro y cinco veces inferior al tiempo de respuesta de una memoria EDO equivalente, aunque el rendimiento neto es superior en SDRAM por las ventajas esgrimidas en la sección 10.13.4.4.

pág. 63

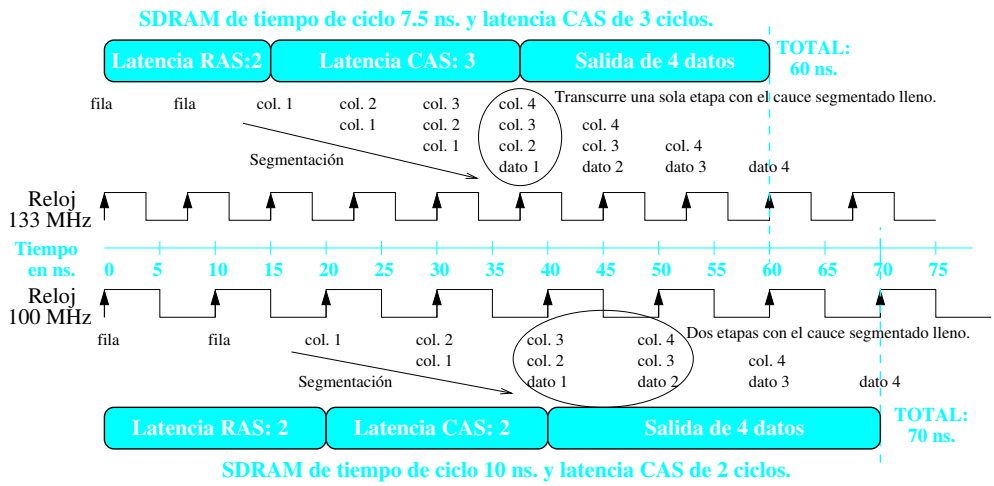


FIGURA 10.25: Comparativa de rendimiento entre una memoria SDRAM de 100 MHz y otra de 133 MHz. La segunda aparenta ser un 33 % más rápida, pero su ganancia neta se queda en sólo la mitad cuando la comparación se efectúa según el tiempo de respuesta de una línea de caché.

13.4.7 Análisis de rendimiento

Otra de las conclusiones que aporta la figura 10.24 es que para aumentar la frecuencia de un chip debemos subir la latencia CAS hasta su máximo de 3 ciclos. En esencia, esto nos descubre que no estamos frente al procesador: Si sobreaceleramos un Pentium 4 de 3 GHz hasta los 4 GHz, el chip es exactamente un 33 % más veloz, pero si hacemos lo mismo con una SDRAM de 100 MHz hasta los 133 MHz, no resultará un chip un 33 % más rápido, ya que ello nos obliga a reestructurar internamente el tiempo, consumiendo más ciclos de un período de reloj más corto. ¿Qué ocurre entonces, ganamos o perdemos? Se intuye que algo ganamos, aunque no tanto como parece.

☛ pág. 66

SDRAM vs CPU

Para dar una respuesta más precisa nos ayudamos de la figura 10.25, donde ilustramos cómo se distribuye el tiempo en sus tres partidas principales (latencia RAS, latencia CAS y salida de ráfaga - ver sección 10.13.4.2) para las dos variantes: SDRAM de 100 MHz con CL=2, y SDRAM de 133 MHz con CL=3. Ambas representan el par de configuraciones válidas del chip comercial número 5 de Micron en la figura 10.24, asumiendo el *RAS-to-CAS-Delay (RCD)* o latencia RAS usual de dos ciclos en SDRAM.

☛ pág. 60

comparativa

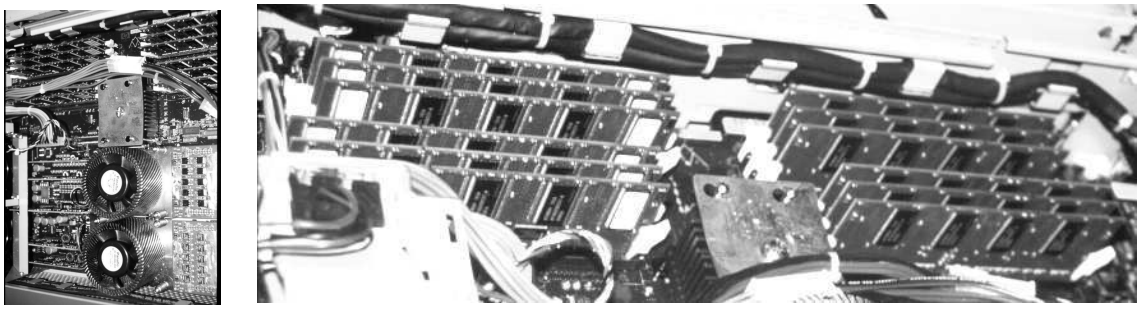
Nuestra vara de medir el rendimiento es en este caso el tiempo que se tarda en servir una línea de caché, la petición de memoria más usual en cualquier PC. La ganancia que refleja la SDRAM de 133 MHz es sólo la mitad del 33 % esperado. Las partidas correspondientes a la latencia RAS y la salida de la ráfaga sí revierten esta mejora, pero la latencia CAS es más lenta aún con el bus más rápido (22.5 ns frente a sólo 20 en 100 MHz), y por eso la ganancia media cae estrepitosamente.

resultado

Llegados a este punto, uno puede acordarse de lo aprendido durante la segmentación del procesador (consultar sección 3.3.1) y recordar aquello de que *una segmentación en N etapas sin dependencias alcanza una aceleración de N*. Extrapolándolo a la segmentación vista para la latencia CAS, un CL de 3 (frente a 2), se traslada a una aceleración de 3 (frente a 2), consumiendo un tiempo inferior. ¿Cómo justificar entonces que salgamos perdiendo? Esgrimimos dos motivos para ello:

segmentación

☛ Volumen 1



(a) (b)

FOTO 10.6: (a) Una plataforma servidor con doble procesador McKinley (el Itanium de 0.13 micras, en la parte inferior) y 1.5 Gbytes de memoria DDRAM (en la parte superior). (b) Detalle del sistema de memoria principal, con 12 zócalos DDRAM de 128 Mbytes.

`llenado`

❶ Se pierde más tiempo en llenar y vaciar el cauce que en explotarlo. De nuevo no estamos frente al procesador, que ejecuta un código de millones de instrucciones. En la memoria las peticiones se realizan de forma atomizada: Por partidas de 4 datos en los que para 3 etapas de segmentación el cauce sólo se llena durante un efímero ciclo de reloj, el único en que funciona a pleno rendimiento.

`balance`

❷ Las etapas de segmentación no están bien aprovechadas. Si para CL=2 el tiempo de acceso a columna es de 20 ns, nada impide que también pueda serlo para CL=3, puesto que el chip es el mismo y sus unidades funcionales tienen en ambos casos el mismo retardo natural. El problema aquí es el pernicioso efecto del reloj digital, cuyo período de 7.5 ns. provoca que para él no existan los 20 ns, sino los 22.5 ns, desperdiciándose más de un 10 % del tiempo en esperar la llegada del siguiente flanco de subida.

`t. servicio`

La caida de la mejora en un 50 % se amortigua un poco si ampliamos nuestra medición al tiempo de servicio de una línea de caché, en el que ya entran en liza tanto el tiempo necesario para enviar la dirección por el bus de memoria como el requerido para recibir la ráfaga de 4 datos. Estas dos operaciones sí revierten en su totalidad la aceleración de la frecuencia del bus hasta los 133 MHz.

`conclusión`

La conclusión a la que llegamos es que un chip SDRAM con CL=2 desaprovecha parte de sus cualidades, porque reprogramándolo con CL=3 podremos apuntar más alto en la frecuencia de bus de memoria, y acabamos de ver que el sacrificio de ese ciclo queda plenamente amortizado.

13.5 ▶Double Data Rate Synchronous Dynamic RAM (DDRAM)

`SDRAM 2x`
pág. 33 ☚

Estamos frente a una SDRAM capaz de responder datos tanto en el flanco de subida como en el flanco de bajada de la señal de reloj, es decir, se trata de una **SDRAM 2x**. La figura 10.7 muestra el aspecto de un módulo de memoria DDRAM, y la foto 10.6 el de un servidor con 1.5 Gbytes de memoria de este tipo.

`origen`

Esta memoria lleva con nosotros desde 1997, momento en que comenzó a ser utilizada en el ámbito de las tarjetas gráficas, campo en el que aún conservaba cierta delantera a mediados de 2002 [4]. Como alternativa para la memoria principal, la DDRAM se topó con el obstáculo del bus local, entonces en 100 MHz, pero tras la llegada del K7 y su bus de 200 MHz (1999), el mercado comenzó a tomar interés en sus propiedades. La propia AMD era parte interesada en su desarrollo, pues el bando enemigo, Intel, apostaba entonces fuerte por la RDRAM en 2000,

[4] 300x2 MHz en un chip experimental de Infineon y 500x2 MHz según pruebas de Samsung.

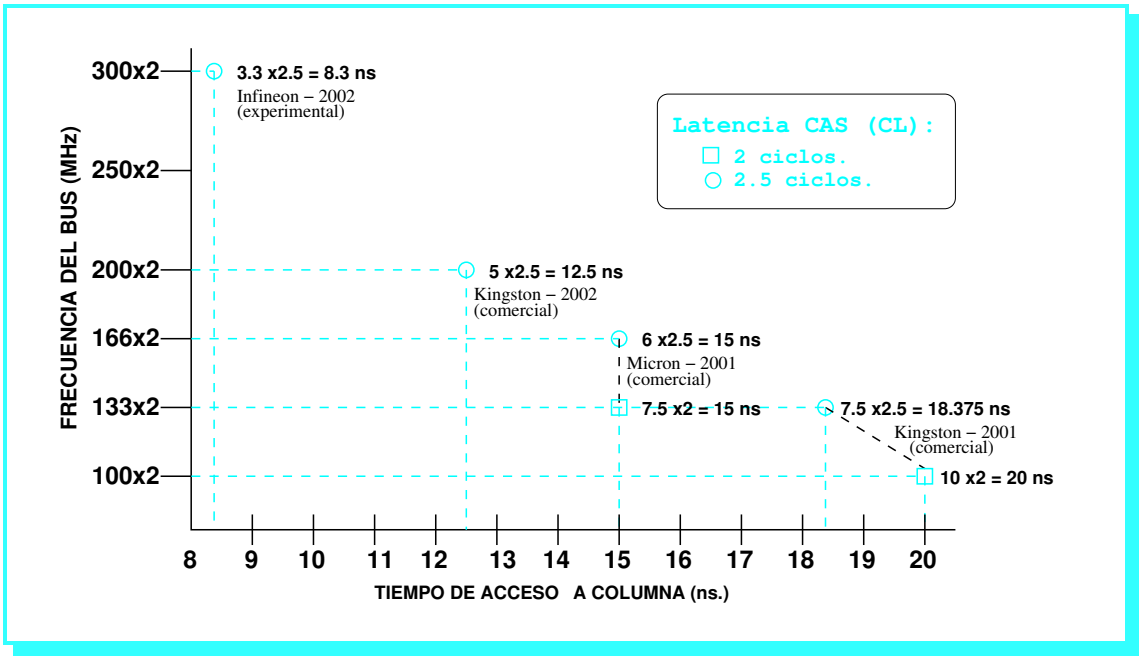


FIGURA 10.26: La DDRAM clasificada según frecuencia de bus (en ordenadas) y tiempo de acceso a columna (en abcisas) para las versiones existentes a finales de 2002. Sólo los chips programados para una latencia CAS de 2.5 ciclos soportan la máxima frecuencia, inversa de su tiempo de ciclo.

la única memoria que pudo montarse sobre los Pentium 4 en su primer semestre de existencia utilizando el juego de chips i850.

La fuerza corporativa de la DDRAM frente a la RDRAM reside en que sus especificaciones no están en manos de un monopolio como Rambus que cobra un canon por su uso, sino que son desarrolladas por el JEDEC (Junction Electronic Devices Engineering Council), el organismo de estandarización de semiconductores que reúne a más de 300 compañías del sector. El mercado del PC, pesetero donde los haya, siempre tuvo claro a quién apadrinar, obligando a Intel a rectificar poco más tarde para dar cobertura a la DDRAM con su juego de chips i845. Este hecho constituyó el espaldarazo definitivo que la DDRAM necesitaba para su posterior desarrollo.

estándar

13.5.1 Rendimiento frente a SDRAM

Decir que una DDRAM es una SDRAM 2x puede llevar al equívoco de creer que es el doble de rápida. Su rendimiento no es tan superlativo, ya que la respuesta en flanco de subida y bajada sólo se aprovecha durante la ráfaga de salida de datos. En las latencias RAS (RCD) y CAS (CL), el ciclo de reloj que se emplea es el mismo que en SDRAM, y aunque CL baja su cota máxima de 3 a 2.5 ciclos, RCD puede subir de 2 a 3 ciclos. En ese caso, para ganar con DDRAM hay que apelar al ciclo de reloj más veloz al que apunta por su mejor diseño y fabricación.

reloj

Para verlo más claro, enfrentamos el tiempo de respuesta de una línea de caché para una SDRAM de 133 MHz (parte superior de la figura 10.25) con el de una DDRAM de 133x2 MHz (parte inferior de la figura 10.27). Ambas llegan a su cita con la salida de datos a los 37.5 ns., y es a partir de ahí cuando la DDRAM cobra ventaja gracias a su desdoble en la salida de datos. La ganancia efectiva de la DDRAM es inferior al 20 % (ahorra 11.25 de los 60 ns. que consume la SDRAM), frente al 100 % de mejora que vislumbraba la frecuencia. Esto se debe a que durante el tiempo de respuesta de la primera palabra de memoria ambas actúan de forma muy similar.

☛ pág. 67
☛ pág. 70

ganancia: 20 %

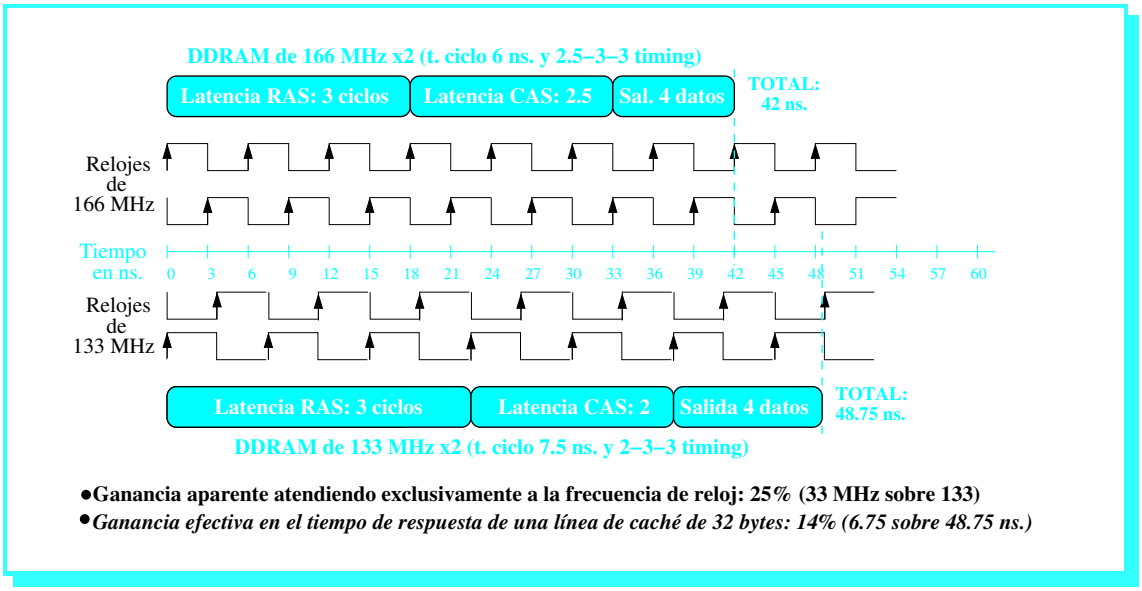


FIGURA 10.27: Comparativa de rendimiento entre una memoria DDRAM de 133x2 MHz y otra de 166x2 MHz. La segunda aparenta ser un 25 % más rápida, pero su ganancia neta se queda en sólo un 14 % cuando la comparación se efectúa según el tiempo de respuesta de una línea de caché.

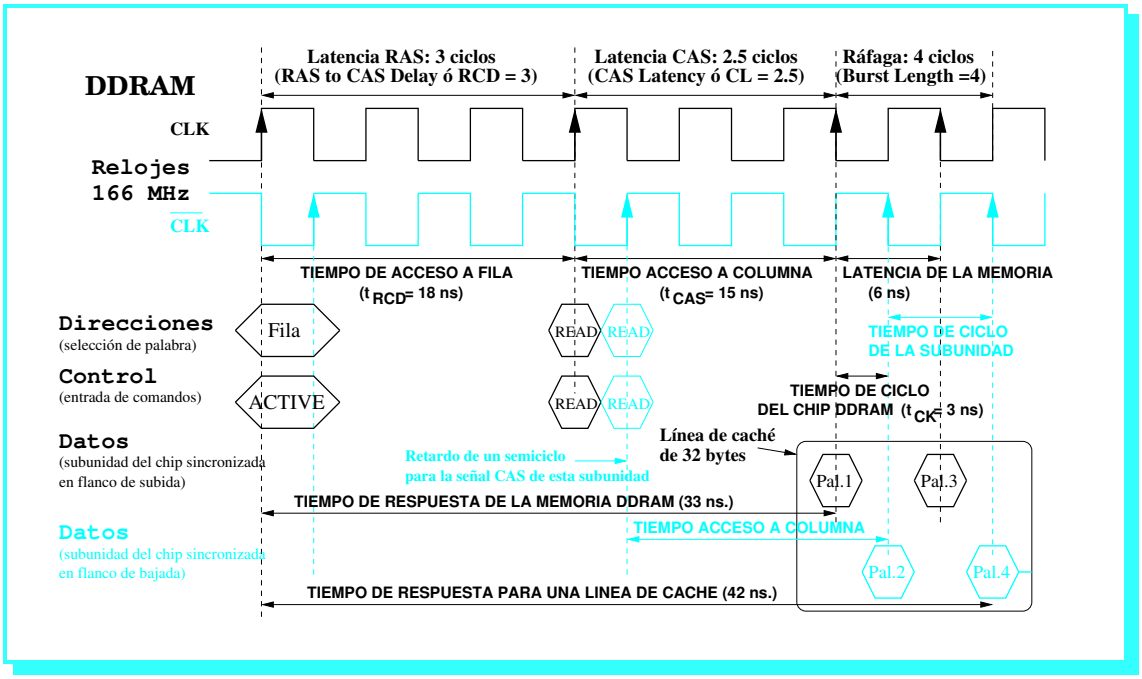


FIGURA 10.28: Interfaz de diálogo con memoria DDRAM, donde podemos apreciar la señal de reloj desdoblada que permite a la salida funcionar bajo semiciclos ó flancos de subida y bajada.

13.5.2 Programación

Los parámetros a configurar en una DDRAM son los mismos que ya conocemos para SDRAM: Latencia CAS, latencia RAS y tiempo de precarga en otro bloque entrelazado del chip. La consolidación de esta terna ha traído como secuela que se abrevie diciendo que una DDRAM es 3-2-2 timing (ver sección 10.14.2.3), donde estos tres números representan los ciclos que definen cada

X-Y-Z timing
☛ pág. 92

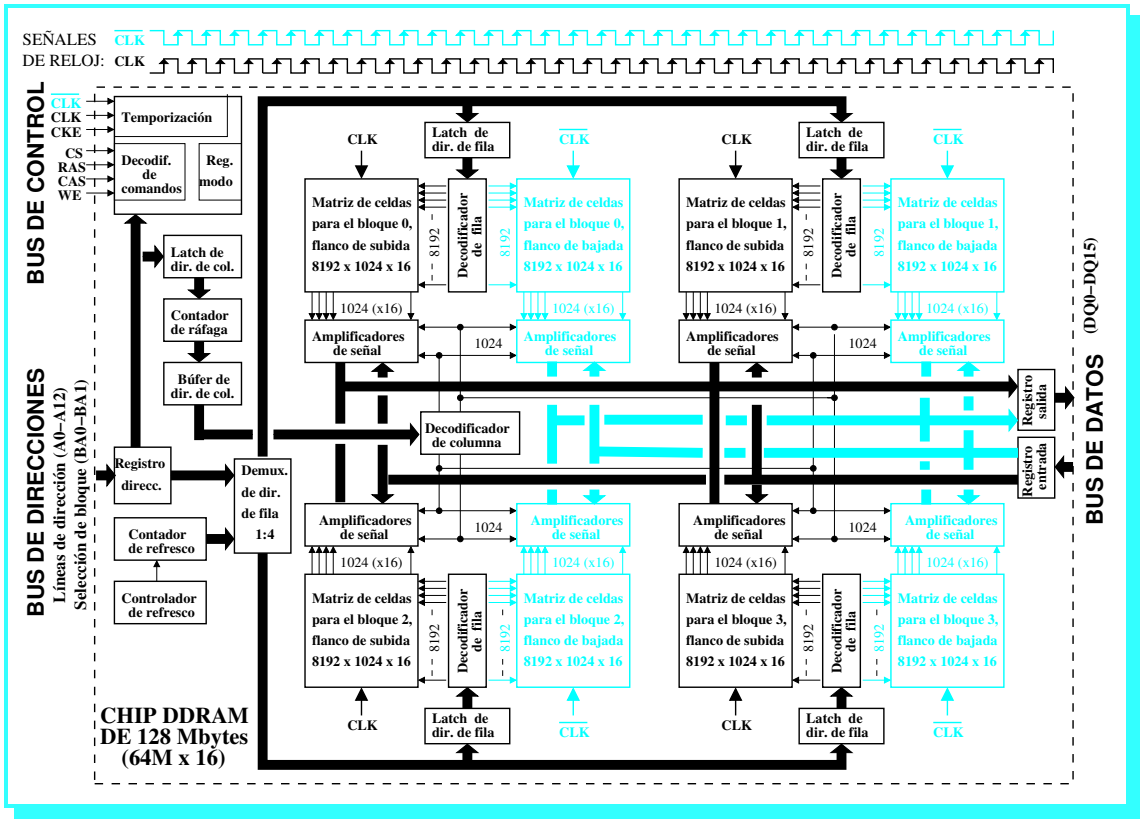


FIGURA 10.29: Diagrama de bloques de un chip DDRAM de 128 Mbytes y anchura 16 bits, dotado de 4 bloques entrelazados en longitud de orden inferior. Dado que el ancho del módulo DDRAM es de 64 bits, son necesarios cuatro chips como éste trabajando conjuntamente para conformar un módulo (de 512 Mbytes). Al prescindir de la parte coloreada nos quedaría el diagrama de bloques de un chip SDRAM de 64 Mbytes, también de 16 bits y 4 bloques entrelazados.

uno de los tres parámetros por ese orden.

De nuevo, todas las unidades funcionales trabajan en el flanco de subida de la señal de reloj CLK, con lo cual, para responder en flanco de subida y bajada se utiliza un segundo reloj, $\overline{CLK}$, cuya señal es inversa del primero. Esto justifica que en la temporización de la DDRAM sea lícito considerar semiciclos de reloj (como en los 2.5 apuntados para CL). Si tomamos un único reloj, pongamos de 166 MHz, y comparamos con SDRAM, la ráfaga de salida de datos sería 3-1-1-1 en SDRAM y 2.5-0.5-0.5-0.5 en DDRAM.

temporización

En la figura 10.28 presentamos el cronograma de funcionamiento para esta memoria, distinguiendo por colores sus dos relojes. Allí observamos que el tiempo de ciclo que marca el ritmo de salida de datos es de 3 ns. Pero si medimos en latencia como hacen los fabricantes, cada subunidad del chip controlada por una señal de reloj (CLK o $\overline{CLK}$) muestra internamente un retardo de 6 ns. Esto explica que en las especificaciones de las casas comerciales y en su etiquetado sea éste último el valor que aparezca precedido del guión que marca el tiempo de ciclo. En definitiva, como parámetro externo del interfaz DDRAM para nosotros, el tiempo de ciclo debe ser 3 ns, lo que también encaja como valor inverso de la frecuencia 166x2 MHz operativa para el chip.

☛ pág. 70

t. ciclo

13.5.3 Arquitectura

La figura 10.29 muestra el diagrama de bloques para un chip DDRAM de 128 Mbytes de ca-

Año del chip	Fabricación en micras	Tamaño en Mbytes	Anchura en bits	Frecuencias, incluyendo ya el factor 2x (timing)
2000/01	0.18	64	4, 8	200MHz (2-2-2), 266MHz (2.5-3-3 y 2-2-2)
2001/02	0.18	64	16, 32	200MHz (2-2-2) hasta 400MHz (2.5-2-2)
2000/02	0.18	128	4, 8	200MHz (2-2-2), 266MHz (2.5-3-3 y 2-2-2)
2002	0.18	128	4, 8	266MHz (2-2-2), 333MHz (2.5-2-2)
2001	0.18	128	16, 32	333MHz (2.5-2-2), 400MHz (2.5-2-2)
2002	0.18	128	16, 32	333MHz (2.5-2-2) hasta 600MHz (2.5-3-3)
2001	0.15	256	4, 8, 16	200MHz (2-2-2) hasta 333MHz (2.5-3-3)
2002	0.15	256	4, 8, 16	200MHz (2-2-2) hasta 333MHz (2.5-2-2)
2002	0.15	512	4, 8, 16	200MHz (2-2-2) hasta 266MHz (2.5-2-2)
2002	0.15	512	4, 8, 16	200MHz (2-2-2) hasta 333MHz (2.5-2-2)

TABLA 10.13: La gama de productos comerciales disponible para los chips DDRAM de Micron.

vs. SDRAM

pacidad. Si nos quedamos sólo con los trazos negros, lo que tenemos es un chip SDRAM de 64 Mbytes, hecho que nos permite sopesar que no son tan grandes las diferencias entre ambas.

relojes

Las dos señales de reloj, CLK y $\overline{CLK}$, actúan sobre matrices de celdas disjuntas cuya respuesta se alterna para producir una salida el doble de rápida que sus latencias individuales. La lógica de direccionamiento para la fila y la columna es común a ambas, sincronizándose de forma conjunta bajo una misma señal de reloj, pero no así las líneas RAS y CAS que entran a cada chip cuando se decodifican los comandos DDRAM.

double-sided

pág. 64

RAS y CAS

En la sección 10.13.4.5 comentamos que en el chip SDRAM existían un par de líneas RAS y CAS diferentes por cada uno de sus dos o cuatro bloques (*single-sided* o *double-sided*, respectivamente). Estas líneas CAS separadas se aprovechan en la arquitectura DDRAM para desfasarlas medio ciclo de reloj cuando se decodifican los comandos DDRAM en aquellas matrices de celdas que conforman la salida para la señal de reloj $\overline{CLK}$. De esta manera, el comportamiento de estas matrices es mimético al de la otra mitad del chip, que responde en CLK, mientras que la lógica de direccionamiento para la fila y la columna es común a ambas mitades, sincronizándose de forma conjunta bajo una misma señal de reloj.

relojes

13.5.4 Versiones

La tabla 10.13 describe la gama de productos disponibles por parte de Micron a finales de 2002. Los saltos se suceden de 66 en 66 MHz debido a que el reloj de la placa base ha venido progresando a pasos de 33 MHz, valor al que debemos añadir el multiplicador 2x característico de la DDRAM. La punta de velocidad para esta tecnología son los 600 MHz, punto en el que recogerá el testigo la nueva memoria DDR-II (ver sección 13.3.3).

pág. 165

efecto CAS

El salto tecnológico hasta las 0.15 micras ha permitido crecer en velocidad y capacidad a la DDRAM, en clara sintonía con lo que le ocurre al procesador. También se aprecia que la latencia CAS se resiste mucho más al descenso que su homóloga RAS, y que se ve más perjudicada por la aceleración que por el incremento de tamaño.

13.5.5 Análisis de rendimiento

Para analizar el rendimiento de la DDRAM en sus diferentes variantes de programación y entrelazado, utilizaremos la especificación *X-Y-Z timing* (ver sección 10.14.2.3). Una comparativa en términos de frecuencia y ancho de banda bajo la especificación PC-XXXX e incluyendo ya a la RDRAM, nos la ofrece conjuntamente la sección 10.14.2.2 y la figura 10.40.

pág. 92

pág. 91

pág. 102

Latencia RAS (RCD)	3	3	3	2	2	2	2
Latencia CAS (CL)	3	3	3	3	2	2	2
Factor de entrelazado	1	2	4	1	1	2	4
Gestión y bases de datos	100	N/D	N/D	100.32	100.64	N/D	N/D
Gráficos y juegos 3D	100	N/D	N/D	100.85	101.61	N/D	N/D
Ancho de banda con CPU	100	104.36	112.50	103.58	116.00	118.96	125.19
Ancho de banda con FPU	100	104.07	120.33	103.90	119.25	124.20	136.56

TABLA 10.14: Rendimiento de la memoria DDRAM 133x2 MHz en sus diversas variantes. El índice empleado es proporcional al rendimiento, habiéndose normalizado a 100 en el caso más desfavorable, que actúa de punto de partida en todas las optimizaciones (N/D = No disponible).

Hemos escogido una serie de equipos con procesador de 800 MHz, 128 Mbytes de memoria DDR PC-2100 de Micron con etiqueta -7 (que denota un tiempo de ciclo de 3.5 ns. para nosotros) y 133x2 MHz de frecuencia, y una serie de cuatro placas base: Una con juegos de chips KX133 de VIA bajo K7 Thunderbird de AMD, y tres con juego de chips de Intel para su Pentium III Coppermine: 440BX, Apollo Pro 133 y 133A. Las pruebas se realizaron para el benchmark Winstone 99 1.2 como representativo de los programas de gestión, el benchmark 3DMark 2000 para el software gráfico de juegos 3D (consume buena parte de su tiempo realizando operaciones de renderizado) y el benchmark SiSoft Sandra 2000 para medir las transferencias de datos con la CPU y la FPU. Los números reflejan la media obtenida sobre la serie de cuatro plataformas, a excepción de los correspondientes al entrelazado en longitud a nivel de chip, que pertenecen únicamente al PC equipado con K7.

equipo PC

benchmark

Los resultados obtenidos se muestran en la tabla 10.14. Allí podemos apreciar cómo la reducción de 3 a 2 ciclos en la latencia RAS apenas produce mejoras, mientras que esa misma reducción en CAS es bastante más efectiva, lo que confirma la gran cantidad de casos en los que la DDRAM es capaz de esquivar la latencia RAS mediante la reutilización de los datos de una misma fila y/o la precarga en un bloque diferente. La mejora respecto a una latencia CAS de 2 ciclos no llega al 10 % salvo que la aplicación haga un uso muy intensivo de la memoria, con lo cual, si el sobreprecio a pagar es superior a ese 10 %, es preferible descartar la mejora.

latencias

Respecto al entrelazado, el incremento del rendimiento es superior en la transición de 2 a 4 bloques que en el paso inicial de 1 a 2. Creemos que en SDRAM este incremento estaría más equilibrado en ambos casos, pero en DDRAM el trasiego de datos es superior y las oportunidades de precarga en otro bloque se convierten en un gran baluarte.

entrelazado

Conclusión: Dado que la fluctuación en los parámetros RCD y CL no es muy grande de cara al rendimiento neto de la memoria, una vez más se recomienda sacrificar éstos en favor de apurar la máxima frecuencia que admita el funcionamiento del chip. Respecto a las posibilidades de entrelazado interno, las mejoras son más significativas, pero prácticamente todos los diseños DDRAM del mercado disponen de un factor 4 de entrelazado fijo, por lo que aquí no existe disyuntiva alguna de programación.

conclusión

Rambus Dynamic RAM (RDRAM) ◀ 13.6

La percepción de interactividad por parte de un usuario multimedia pasa ineludiblemente por disponer de un rápido acceso a memoria, y estas necesidades contrastan sobremanera con el lento peregrinar de la tecnología de memoria.

Los esfuerzos por mejorar la memoria DRAM convencional han quedado en un mero aprovechamiento de la integración en silicio para escalar el reloj de la SDRAM hasta los 200 MHz y desdoblar su señal con multiplicadores de 2x (DDR) y 4x (DDR-II - ver sección 13.3.3).

frecuencia

☛ pág. 165

anchura

La otra solución para incrementar el ancho de banda consiste en aumentar la anchura del bus hasta las 128 líneas. Esta idea, tan sencilla sobre el papel, desborda el patillaje del chip, arruina la sincronización a elevadas frecuencias y complica el enrutado de las líneas por el espacio físico de una placa base actual.

alternativas

Ante semejantes escollos en la vertiente evolutiva más lógica para la memoria, se buscaron alternativas en la concepción de la arquitectura interna de los chips DRAM con objeto de dotarlos de mejores propiedades eléctricas. Las variantes que quedaban eran reducir la frecuencia o acortar la anchura de los datos, y ambas parecían contraproducentes. La primera lo es a todas luces, pero la segunda no: Supone dar un paso atrás para tomar impulso y encadenar varios hacia adelante. Con un bus estrecho, la sincronización es mucho mejor, los problemas de ruido eléctrico y enrutado se minimizan, y la frecuencia tiene licencia para volar hasta valores estratosféricos. Así han ganado batallas buses estrechos como USB y FireWire frente a alternativas anchas como el PCI de 64 bits, especificado mucho tiempo atrás y rechazado por el mercado por caminar por una senda equivocada.

Rambus

Esta misma idea aplicada a la memoria es lo que abandera RDRAM (Rambus Dynamic RAM), un diseño de Rambus, compañía fundada en 1990 cuyo primer golpe de efecto llegó en las navidades del 97, cuando batió el record de ventas con Nintendo 64, la video-consola basada en tecnología RISC. Desde entonces, la compañía ha mantenido el liderato del ancho de banda de la memoria, llevando a cabo numerosas mejoras para minimizar su única debilidad: La elevada latencia que antecede a la transferencia de datos.

13.6.1 El bus de memoria

Las prestaciones de la memoria RDRAM se sustentan sobre dos pilares básicos:

arquitectural
Canal

❶ A nivel arquitectural, un nuevo bus denominado Canal Rambus Directo que apuesta por la transmisión a frecuencias muy elevadas bajo dos premisas fundamentales:

- La separación de los buses de datos, direcciones y control para minimizar los efectos provocados por el ruido eléctrico.
- El estrechamiento del bus de datos, inicialmente hasta sólo 16 bits (aunque luego ampliado), con objeto de favorecer la sincronización entre las líneas.

eléctrico
RSL

❷ A nivel eléctrico, la adopción de una nueva especificación de señales denominada RSL *(Rambus Signaling Level)*, que permite operar a frecuencias de 400 MHz en flanco de subida y bajada, esto es, 400x2 = 800 MHz.

En los módulos SDRAM, cada línea tiene un enrutado diferente y dependiente de dos variables internas: El número de chips presentes en la placa de circuito impreso del módulo, y su espacio de almacenamiento (ver figura 10.30.a). Como consecuencia de ello, el retraso de cada tipo de línea (datos, direcciones y control) se degrada de forma completamente diferente cuando aumenta el número de módulos (ver figura 10.31), lo que complica el margen de temporización del sistema. Como además los módulos DIMM se conectan en paralelo al bus de la placa, cada DIMM presenta o bien una fuerte carga capacitiva o un largo enrutado que le impide un funcionamiento correcto a elevadas frecuencias.

pág. 76
pág. 77

En cambio, en un sistema de memoria RDRAM los módulos se conectan en serie formando una cadena (ver figura 10.30.b), lo que hace que la carga de las patillas esté desacoplada del resto y presente idénticas propiedades eléctricas. Así, conforme vamos añadiendo más memoria, la carga se incrementa por igual con independencia del tipo de línea de que se trate, siendo ésta otra de las claves para la escalabilidad de la memoria RDRAM.

pág. 76

escalabilidad

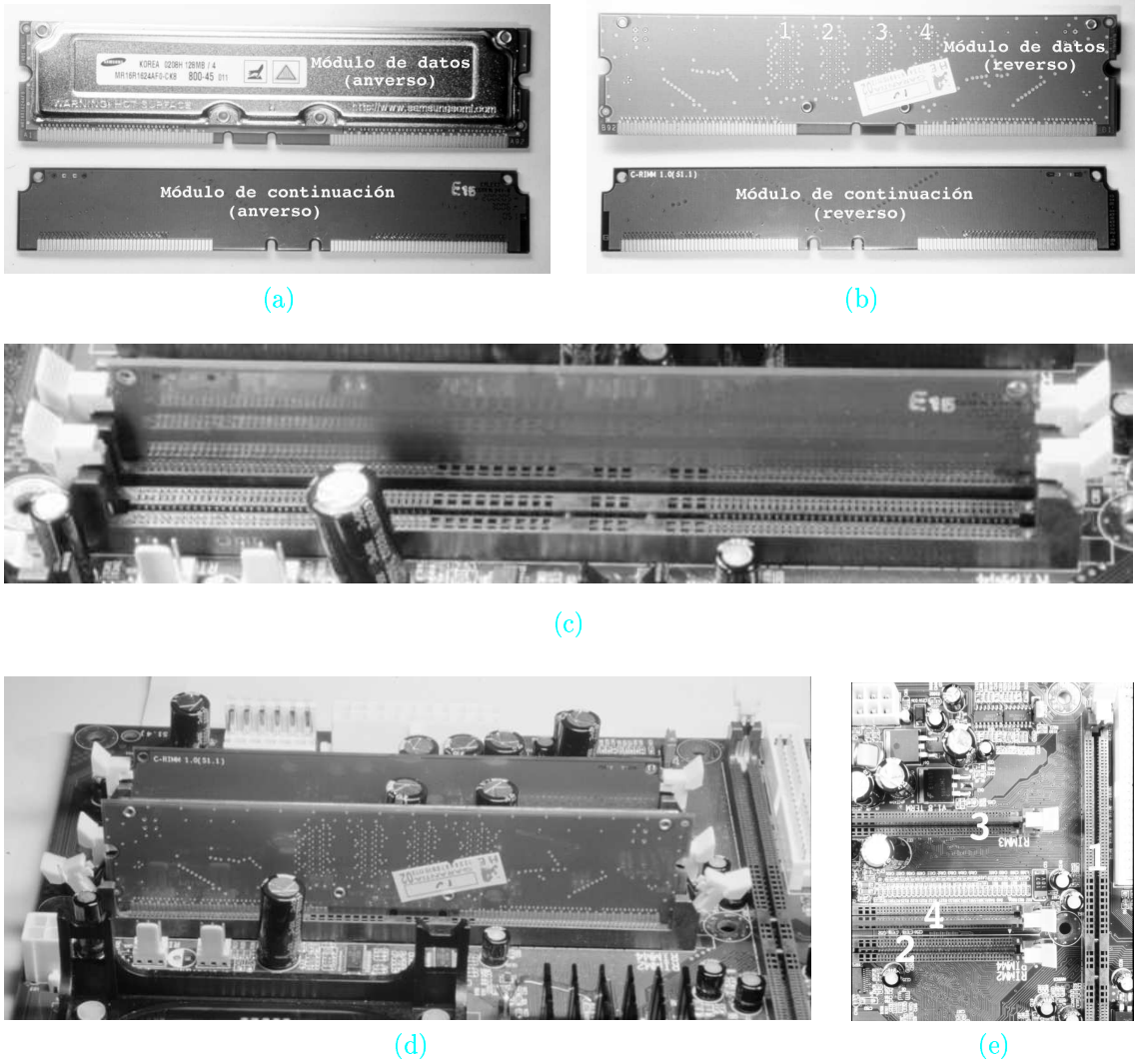


FOTO 10.7: Los elementos de la memoria RDRAM: (a) Envés y (b) revés para los módulos de datos (arriba) y continuidad (abajo) de Samsung, ambos de 184 contactos. La presencia de un delgado disipador de calor nos impide ver los chips en el envés, delatándose en el revés cuatro unidades al apreciarse las microsoldaduras para su patillaje. (c) Dos zócalos RIMM vacíos, uno con el correspondiente módulo de continuación y el otro desprovisto de él. (d) Aspecto de una placa base para un sistema de memoria RDRAM con cuatro zócalos estructurados en dos bancos, el primero libre y y el segundo ocupado. (e) Vista cenital del sistema numerando sus zócalos.

13.6.2 Módulo y zócalo RIMM

El nuevo zócalo de la memoria RDRAM se denomina RIMM (Rambus Inline Memory Module). Tanto las dimensiones del módulo como sus fijaciones al zócalo de placa base son similares a las que ya conocemos desde la llegada de los módulos DIMM (ver sección 10.7.6). Esto permite abaratar los costes de fabricación por la reutilización de las mismas plantas de ensamblaje de componentes.

dimensiones y fijaciones

pág. 34

La foto 10.7 muestra los módulos y zócalos para un típico sistema de memoria RDRAM. A pesar de su aspecto muy similar a DIMM, cualquier aspiración a la interoperabilidad entre ambos resulta inútil. Desde la frecuencia de reloj hasta la funcionalidad del patillaje resultan radicalmen-

interoper.

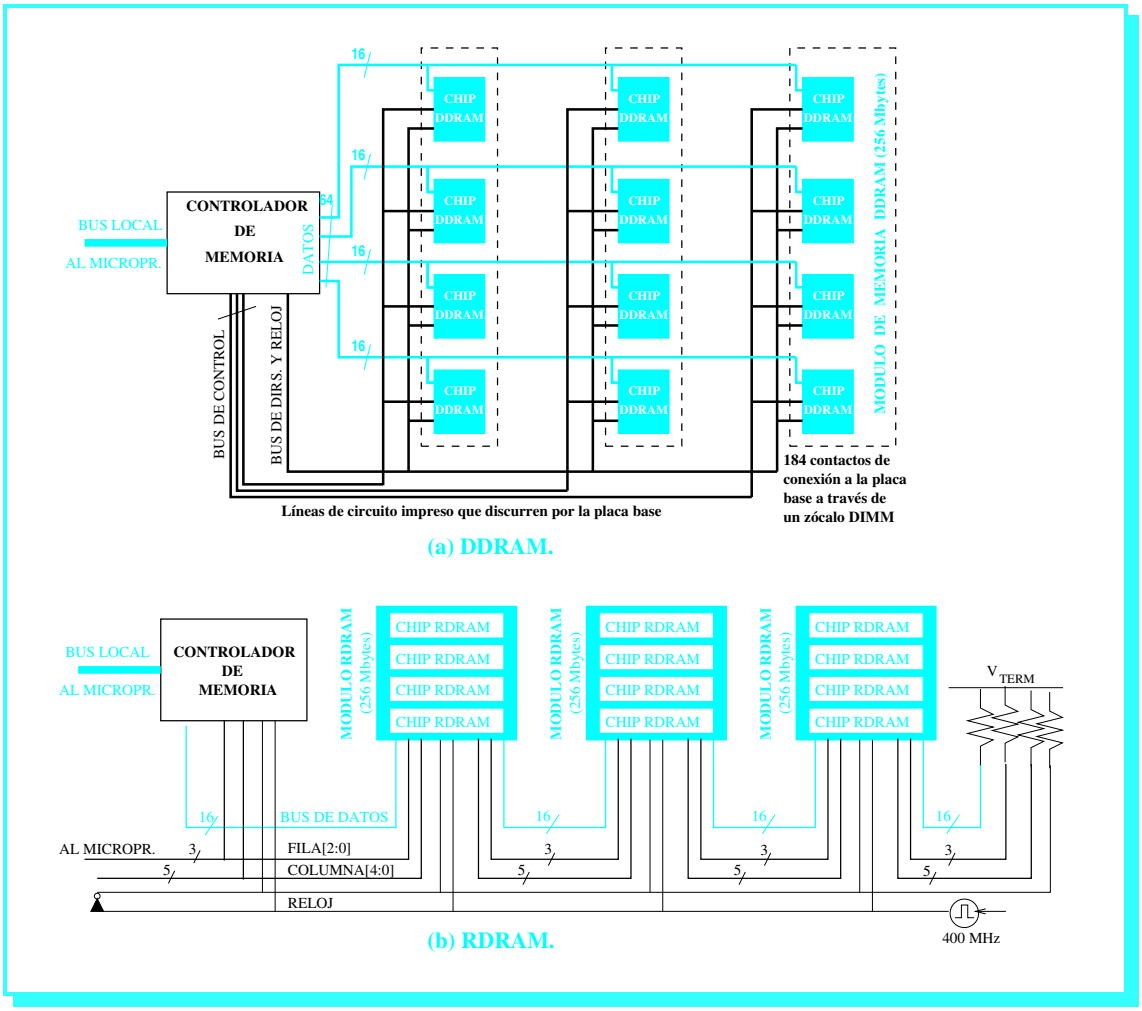


FIGURA 10.30: Comparativa en la conexión y entrelazado de módulos (a) DDRAM en paralelo, y (b) RDRAM en serie para un sistema de memoria principal dotado de 768 Mbytes con 3 módulos de 256 Mbytes y 4 chips cada uno.

te diferentes en RDRAM y DDRAM.

distancia

Su elevada frecuencia de funcionamiento obliga a los zócalos RIMM a situarse a pocos centímetros del procesador, ya que el margen de fluctuación que se permite a estas señales es bajísimo, y el retardo que éstas sufren es directamente proporcional a la distancia recorrida.

pág. 77
continuadores

Las placas base suelen disponer de entre dos y cuatro zócalos RIMM, formando una cadena en serie que se extiende hasta el controlador de memoria, con terminación paralela RSL (Rambus Signaling Level) en el otro extremo según indicamos en la figura 10.32. Si el usuario decide no llenar todos los zócalos RIMM, entonces tendrá que insertar continuadores del canal para mantener la integridad de las comunicaciones.

generador de reloj

El generador de reloj se posiciona en un chip dedicado ubicado justo al final del zócalo RIMM más lejano al controlador (ver figura 10.32).

pág. 77
SPD

Los módulos RIMM también incluyen el típico chip SPD (*Serial Presence Detect*) que implemente la autoconfiguración y proporcione sus parámetros a la BIOS durante el proceso de encendido del sistema. Esto salvaguarda la compatibilidad de todos los módulos RDRAM al margen del fabricante y de la capacidad del chip.

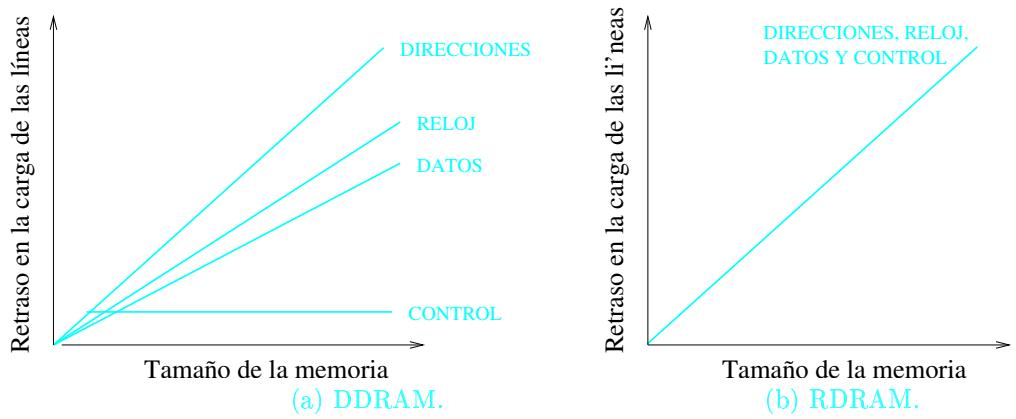


FIGURA 10.31: Evolución que sigue la carga de las diferentes señales de una memoria con respecto a su capacidad en Mbytes. La gráfica de la izquierda representa el comportamiento de la memoria DDRAM, mientras que la de la derecha corresponde a la RDRAM y refleja la unificación de las propiedades eléctricas de todas sus patillas y contactos que le permiten aspirar a frecuencias mucho más elevadas (escalabilidad).

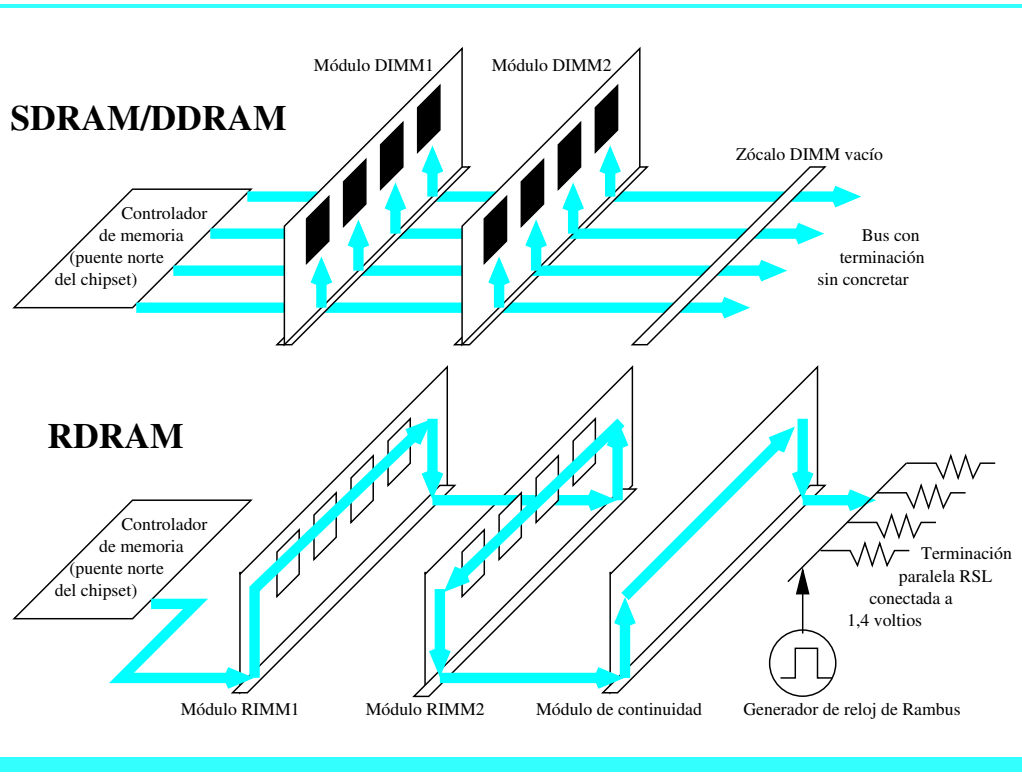


FIGURA 10.32: Relación del bus de memoria con sus módulos en DDRAM frente a RDRAM.

13.6.3 Fabricación y coste

En la memoria SDRAM, el zócalo DIMM se conecta con el controlador de memoria a través de largas trazas donde el fenómeno de la reflexión de señales eléctricas incide notablemente, más

trazas

aún porque las líneas del bus no finalizan limpiamente en el extremo opuesto al controlador (ver figura 10.32). En la memoria RDRAM, en cambio, el montaje sobre un zócalo RIMM se comporta en todo momento como si el módulo se hubiese soldado directamente a la placa base, eliminando cualquier tipo de reflexión y ofreciendo a la vez una baja inductancia que posibilita un funcionamiento mucho más uniforme a elevadas frecuencias.

exigencias

Para hacer eso posible, la manufacturación de un módulo RIMM es tremendamente exigente: Requiere mínimas diferencias en la longitud de la traza de sus líneas, el grosor de éstas, y la distancia de separación entre pistas, lo que indudablemente encarece su coste.

área de integración

pág. 79

El nivel eléctrico y la presencia de un gran bus interno de 144 líneas (128 si no dispone de paridad/ECC - ver figura 10.34) incurren en una complejidad adicional que desemboca en un área de integración en torno al 20 % superior al de un chip DDRAM de similar capacidad.

encapsulado

Vol.5 en Web

Y si el chip sale más caro por materia prima, aún se encarece más si nos adentramos en su proceso de fabricación. Su complejidad inherente obliga a integrar el chip bajo encapsulado micro-BGA (*Ball Grid Array*, con microsoldaduras sobre la base del chip - ver sección 34.3.1) o su equivalente más actual CSP (*Chip Scale Packaging*), ambos bastante más caros que el TSOP (*Thin Small Outline Package*, con patillaje dispuesto sobre las aristas laterales) que puede ser utilizado como alternativa barata en DDRAM.

validación

Además, la inmaculada perfección que se exige al comportamiento eléctrico de la RDRAM hace que pocos chips de una oblea pasen la validación en la horquilla superior del rango de frecuencias, lo que aboca a malvender la mayor parte de ellos.

Ejemplo 10.6: Porcentajes de validación de los chips RDRAM

Los datos de que disponemos corresponden a la fase inicial de fabricación de la RDRAM, cuando se integraba a 0.22 micras bajo tres frecuencias candidatas: 300x2, 350x2 y 400x2 MHz. No obstante, los consideramos extrapolables a la situación actual, en que se integra a 0.15 ó 0.13 micras exigiendo frecuencias de 400x2, 533x2 y 600x2 MHz.
Acomodando nuestra información a la campana de Gauss típica del proceso de validación de chips, el porcentaje de chips que pasaría las pruebas de validación a cada una de las frecuencias de trabajo sería el siguiente:

- A 400x2 MHz: 28 %
- A 350x2 MHz: 40 %
- A 300x2 MHz: 28 %
- Chips defectuosos: 4 % restante.

Sólo para contrastar con SDRAM, allí el porcentaje de chips que superaba las pruebas a la frecuencia de 133 MHz (la máxima para esa misma distancia de integración y marco temporal) era superior al 70 %.

agravantes

A los argumentos anteriores, todavía hay que sumar dos agravantes: Primero, el coste del disipador de calor que debe incluir el módulo de memoria para esparcir con presteza el calor generado en los puntos más tórridos de sus chips constituyentes. Segundo, el coste adicional de la placa base que auspicie los módulos, a la que se exigen idénticas filigranas eléctricas en las líneas y contactos metálicos relacionados con los zócalos RIMM.

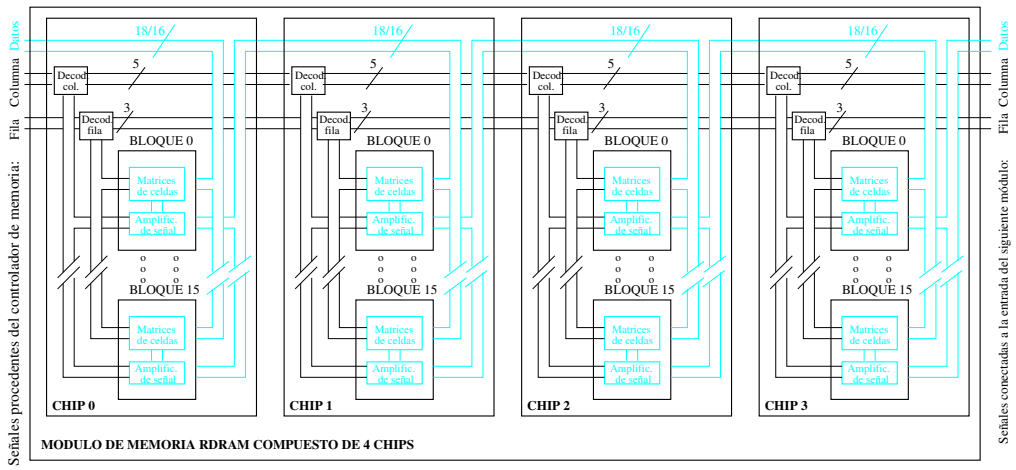


FIGURA 10.33: Esquema de un módulo de memoria RDRAM, con su característico bus estrecho de 16 líneas (18 si la memoria dispone de paridad).

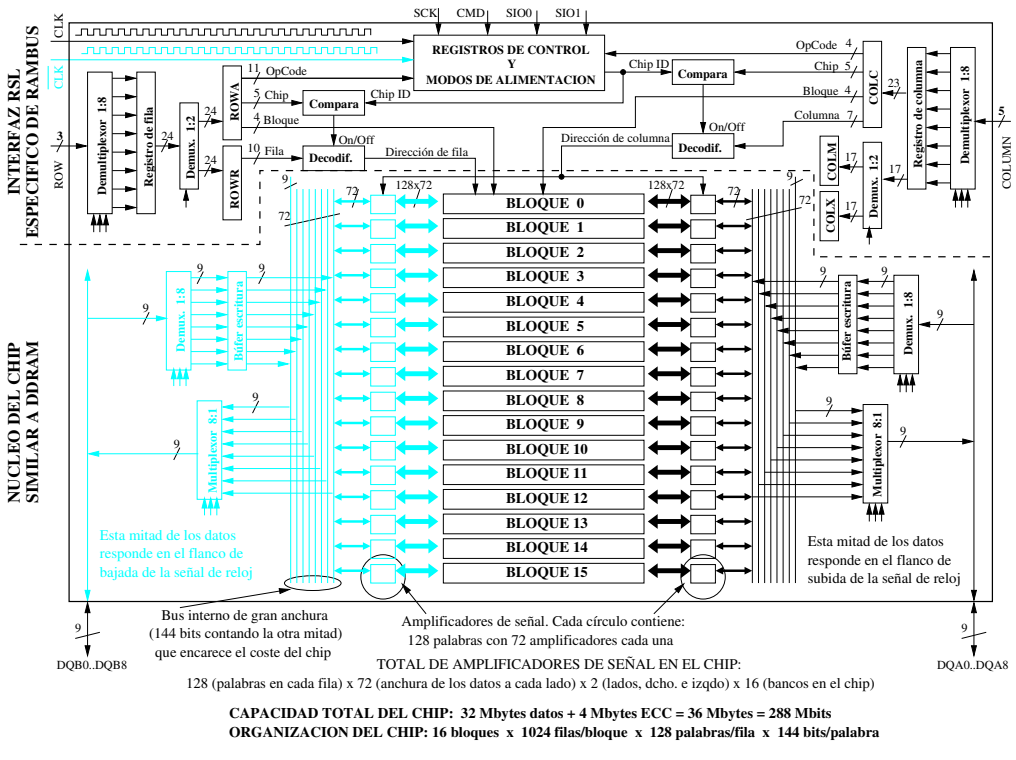


FIGURA 10.34: Diagrama de bloques de un chip RDRAM de 32 Mbytes con paridad/ECC. Puede apreciarse el sucesivo desdoble del interfaz hasta conformar un sinfín de líneas internas.

13.6.4 Arquitectura

El controlador de memoria RDRAM se comunica con sus módulos de memoria (y éstos a su vez con sus chips) por medio de buses muy estrechos, de 16 líneas de datos y 8 líneas de dirección, 3 para la fila y 5 para la columna (ver figura 10.33). Dentro de cada uno de los chips, estos buses se desdoblan en ocho para proporcionar internamente las respectivas vías de comunicación:

bus de datos

1. Un bus de datos interno, de 128 líneas, que se llena a razón de 8 viajes de 16 bits (uno cada 1.25 ns en el caso de la versión base de 400x2 MHz, que completa el llenado en lo que sería un ciclo de 100 MHz).

bus de dirs.

2. Un bus de direcciones, de 24 líneas para la fila y 40 para la columna, que se utilizan parcialmente para emitir los comandos de control y códigos de operación al chip.

dualidad

Esto explica la dualidad de esta arquitectura, cuyos chips se encuentran internamente atestados de líneas de comunicación, pero que atendiendo a su aspecto externo presentan pocas necesidades de patillaje tanto a nivel de chip como de módulo. En la sección 10.7.6 vimos que se utilizan un total de 184 contactos para éste último, pero la tabla 10.5 revela que la mayor parte de ellos se dedican a dos claras prioridades en entornos de elevada frecuencia:

pág. 34
pág. 37

temperatura

- Distribuir la alimentación de forma muy tamizada para evitar corrientes elevadas, y por consiguiente, reducir la potencia disipada para aliviar la temperatura.

estabilidad

- Aislar separadamente cada línea activa utilizando referencias a tierra individuales con objeto de dotar de gran estabilidad a las señales.

selección del chip

Aunque todas estas líneas entran de forma conjunta a cada uno de los chips del módulo RDRAM, un campo del bus de control transporta el identificador del único chip que actúa para resolver el acceso. Sendos comparadores a derecha e izquierda de los registros de control permiten al resto de chips desactivarse y no interceder en el uso del bus con aquél.

número de chips

El número de chips es algo que se deja a elección del fabricante, con una limitación máxima de 32 chips (el campo que identifica el chip en el bus de control dispone de 5 bits). Respecto al mínimo, si un módulo RDRAM suministra 16 bits y puede ser implementado mediante un único chip, podríamos pensar en usar 8 chips para multiplicar la anchura de la memoria hasta 128 bits. Pero así es precisamente como trabajan la SDRAM y la DDRAM, soluciones que se muestran mucho menos escalables que ésta.

direccionamiento: -bloque -fila -columna

Dentro ya del único chip activo, se decodifica el número de bloque proveniente de otro campo del bus de control para enviar la señal de activación al único bloque que interactúa con el bus de datos interno de 144 líneas. Este bloque acepta la coordenada de fila para volcarla íntegramente a los amplificadores de señal, una mitad a cada uno de sus dos laterales si tomamos como referencia la figura 10.34. Sobre estos dos grupos de amplificadores actúa la coordenada de columna para seleccionar la palabra que sale al bus de datos a razón de ocho viajes de 18 bits (16 si no existe paridad/ECC), cuatro bajo CLK por la derecha y otros cuatro bajo $\overline{CLK}$ por la izquierda en pulsos alternos.

pág. 79

llenado por pares en Pentium 4

En determinadas arquitecturas, como por ejemplo el Pentium 4, la placa base obliga a que los zócalos RIMM se llenen por pares de módulos RDRAM. Esta imposición no tiene nada que ver con el funcionamiento interno de la memoria, sino que está relacionada con el hecho de que el ancho de banda en un módulo RDRAM de 16 bits y frecuencia base 400x2 MHz es justo la mitad que en el bus local del Pentium 4. Así, la placa base realiza un entrelazado externo en anchura por parejas de módulos para lograr equilibrar el ancho de banda en ambas partes. La figura 6.5 ilustra la necesidad de este conexionado por pares, que ya no es necesario cuando se utilizan módulos RDRAM de 32 bits.

Volumen 1

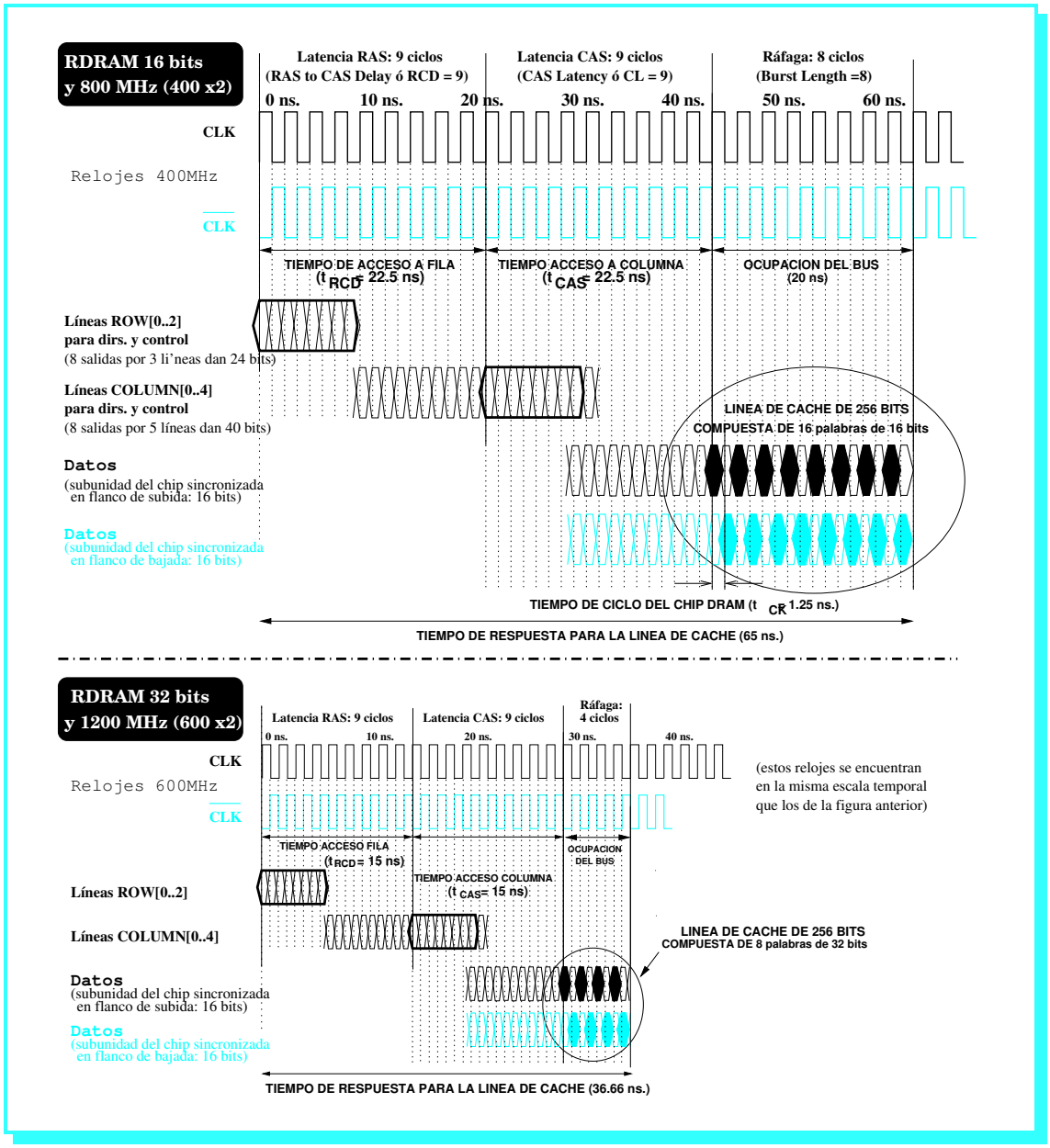


FIGURA 10.35: Interfaz de diálogo con memoria RDRAM, donde podemos apreciar algunos elementos ya familiares de la DDRAM, como la selección de fila y columna y la señal de reloj desdoblada que permite a la salida funcionar bajo semiciclos ó flancos de subida y bajada.

13.6.5 Similitudes con los diseños precedentes

A pesar de su originalidad, la RDRAM presenta algunas similitudes con diseños anteriores:

matriz de celdas

❶ El esquema de direccionamiento dentro de las matrices de celdas es el mismo que se viene repitiendo ya desde la memoria FPM DRAM.

paridad/ECC

❷ La versión de módulos RDRAM con paridad dedica un bit adicional por cada byte de datos. El controlador puede también implementar la corrección de errores ECC de 16 bits sobre una palabra de 128 bits, con lo que ambas operaciones pueden implementarse sin aumentar

el número de chips del módulo. No obstante, suscribir paridad o ECC aquí está menos justificado que en SDRAM/DDRAM, ya que su tasa de errores es inferior a la de aquéllas.

direcciones

❸ Existe un modo de funcionamiento en el que las direcciones de columna pueden ser aleatorias dentro de la fila, en consonancia con lo que ya le ocurría a la SDRAM.

2x

❹ La forma en que se consigue responder en flanco de subida y bajada de la señal de reloj es mimética al caso de la DDRAM, aunque no está de más recordar que en este caso la imitadora es esta última por ser el diseño más joven de todos. Los matices de direccionamiento y salida de datos pueden observarse más lúcidamente en el cronograma para esta memoria, que adjuntamos en la figura 10.35.

pág. 81 ☚

segmentación

❺ Los chips RDRAM también implementan una segmentación interna que solapa estas tareas al igual que en SDRAM/DDRAM, permitiendo una continua ocupación del bus interno durante los 10 ns que dura la salida de datos. El protocolo RDRAM tiene además un control directo sobre todos los recursos de fila y columna concurrentemente con las transferencias de datos, de ahí el calificativo de "directo" con que se apoda el canal. La segmentación interna consta de siete etapas: Transporte de fila, decodificación de fila, transporte de columna, decodificación de columna, acceso a celda, transferencia de datos y escritura en búfer.

❻ Un módulo de memoria RDRAM también puede realizar la precarga y la selección de una nueva fila de celdas concurrentemente con operaciones de columna sobre otra fila ayudándose del entrelazado en longitud al nivel de fila, repartiendo éstas entre los bloques siguiendo un esquema de orden inferior para aprovechar las propiedades de localidad en las referencias a memoria. Ahora bien, en RDRAM el factor de entrelazado es más agresivo (16 frente a 4 en DDRAM y SDRAM), y con él aumentan las oportunidades para solapar estas latencias. Esto es coherente si recordamos que el entrelazado al nivel de las filas de un chip se articulaba para mitigar las dependencias estructurales en los amplificadores de señal del cauce segmentado, y que en RDRAM tenemos siete etapas de segmentación frente a tres en SDRAM, lo que aumenta tanto la posibilidad de conflictos como el grado de concurrencia de los accesos. Por eso, el control interno en RDRAM admite tres modos de precarga y el procesamiento de hasta cuatro peticiones simultáneas que involucren a bloques entrelazados diferentes, con las que se logra superar el 95 % en la utilización de los buses.

entrelazado

☞ **Ejemplo 10.7:** El entrelazado en la memoria RDRAM frente a SDRAM/DDRAM

Consideremos un PC dotado de 768 Mbytes de memoria principal compuesta de 3 módulos de 256 Mbytes y 4 chips cada uno.
En caso de optar por memoria SDRAM/DDRAM, tenemos el esquema de la figura 10.30.a, donde se entrelazan en anchura los cuatro chips del módulo, y en longitud los cuatro bloques del chip que se reparten las filas siguiendo un esquema de orden inferior.
En caso de disponer de memoria RDRAM, tenemos el esquema de la figura 10.30.b, donde se prescinde del entrelazado en anchura al nivel de los chips de cada módulo pero en cambio se disponen 16 bloques entrelazados en longitud y de orden inferior para las filas al nivel interno de cada chip.

❼ Para permitir optimizaciones al nivel interno de cada módulo RDRAM, el número de chips por módulo y el tamaño de la fila en sus matrices bidimensionales de celdas también se

Frecuencia (MHz)	Anchura (bits)	Ancho banda (Mbytes/sg.)	Latencia (ns.)	Núm. de contactos	Fecha de lanzamiento comercial por Kingston
300x2	16	1200	53	168	Abril de 1999
350x2	16	1400	50	168	Abril de 1999
400x2	16	1600	45, 40	168/184	Abril de 1999
533x2	16	2100	35, 32, 30	184	Junio de 2002
600x2	16	2400	N/D	184	Estimado para 2003
667x2	16	2666	N/D	184	Estimado para 2004
400x2	32	3200	45, 40	232	Octubre de 2002
533x2	32	4200	35, 32, 30	232	Octubre de 2002
600x2	32	4800	N/D	232	Estimado para 2003
667x2	32	5333	N/D	232	Estimado para 2004
667x2	64	10666	N/D	326	Mínimo 20005

TABLA 10.15: La gama de módulos RDRAM disponible en 2003. La fecha de lanzamiento corresponde a la disponibilidad de los módulos fabricados por Kingston Technology, una de las firmas pioneras del mercado. El módulo de 64 bits ya dispone de especificación, pero a fecha 2003 aún falta un trecho antes de que se produzca su disponibilidad comercial (N/D = Aún no disponible).

han dejado a elección del fabricante. Pero al igual que en SDRAM/DDRAM parecen existir ciertos valores de consenso, como los que apuntamos en el ejemplo 10.8.

Ejemplo 10.8: DESCOMPOSICIÓN DE LOS MÓDULOS RDRAM DE 128 MBYTES

La configuración que más se repite en los módulos RDRAM de 128 Mbytes es la siguiente:

- 4 chips de 32 Mbytes (256 Mbits) cada uno.
- 16 bloques entrelazados en longitud dentro de cada chip.
- Cada bloque compuesto de 16 Mbits estructurados en matrices de 1024 filas de 128 palabras de 128 celdas cada una.

13.6.6 Versiones

Aunque inicialmente existió una remesa de chips de 300x2 y 350x2 MHz, la primera RDRAM que adquiere cierta presencia en el mercado es la de 400x2 MHz. A partir de ahí, la frecuencia sobrepasa el gigahercio para trabajar a 533x2, 600x2 y 667x2 MHz, según se indica en la tabla 10.15.

`frecuencia`

La anchura de los chips y módulos RDRAM se mantuvo estable durante unos años en los 16 bits, siendo ésta una de sus características distintivas. A mediados de 2002 comenzaron a ver la luz los primeros diseños con anchura de 32 bits, y la especificación de 64 bits se encuentra ya finalizada y no tardará en engendrar modelos comerciales.

`anchura`

ancho de banda

El ancho de banda es la gran baza de la memoria RDRAM. Proporcionando 16 bits cada 1.25 ns en su versión base de 400x2 MHz se consigue un ancho de banda total para la memoria de 1.6 Gbytes/sg. Para la versión de 32 bits y 600x2 MHz contemporánea para 2003, ese ancho de banda se triplica hasta los 4.8 Gbytes/sg.

latencia

La latencia es la espada de Damocles de la RDRAM. Las latencias RAS y CAS internas más comunes son de dos ciclos, algo mejores que las postuladas por la DDRAM, pero la idiosincrasia de ese bus de Rambus que tanto engrandece el ancho de banda va a pasar una elevada factura a la latencia. Las penalizaciones que debemos reseñar son fundamentalmente dos:

1. El estrecho bus de direcciones (sólo actúan 3 líneas `Row` durante la latencia RAS y 5 líneas `Column` durante la latencia CAS) requiere su desdoble sucesivo durante 8 semiciclos para proporcionar toda la información que requiere el interfaz de direccionamiento y control (24 y 40 bits, respectivamente). Esto retrasa 4 ciclos las latencias RAS y CAS a su comienzo.
2. La conexión encadenada de los chips por el bus hace que las señales eléctricas lleguen pronto a los primeros chips y tarde a los últimos. Para lograr una sincronización común que unifique la respuesta de la memoria con independencia del chip con el que dialoga el controlador, éste programa un retardo variable para sus discípulos que involucra de forma conjunta al viaje de ida (dirs.) y de vuelta (datos) con el que se garantiza que los valores procedentes de memoria estarán siempre disponibles al margen de su procedencia. El retardo máximo puede ser de 2, 3, 4 ó 5 ciclos dependiendo de la frecuencia del módulo y la latencia de sus chips constituyentes, aunque tanto para 400x2 como para 533x2 MHz la especificación de Rambus recoge 6 supuestos y 3 de ellos requieren un retardo de 3 ciclos. Al ser este valor la mediana y la moda de la muestra, lo hemos tomado como valor representativo en todos nuestros análisis. Así pues, esta partida retrasa 3 ciclos las latencias RAS y CAS a su finalización.

Contabilizando la suma total, las latencias RAS y CAS ascienden a nueve ciclos cada una: Cuatro por el desdoble, dos por la latencia en sí y tres por la sincronización común. Estos nueve ciclos para cada partida son los que aparecen en nuestro cronograma de la figura 10.35 como valores más característicos para la RDRAM.

pág. 81

Conviene aclarar que aunque estemos haciendo consideraciones acerca del bus, no estamos contabilizando en ningún caso el tiempo de transporte por el mismo, Dicho de otra manera, no hemos evaluado el tiempo de servicio de la RDRAM, sino su tiempo de respuesta, el mismo parámetro de rendimiento que estudiamos para SDRAM y DDRAM, lo que nos otorga licencia para una comparativa lícita.

13.7 Comparativa: DDRAM frente a RDRAM

13.7.1 Analítica

Volumen 1

Ya avisamos cuando comparamos el Pentium 4 frente al K7 (ver sección 6.6) de que el principal problema era el sesgo de la vara de medir, pues una vez quedan al descubierto las virtudes y defectos de cada diseño, resulta fácil pervertir la comparativa para dar el ganador que prefiramos.

vs. benchmark

Este riesgo no es exclusivo de una comparativa analítica, sino que atañe también a los resultados de un *benchmark*. Nuestra predilección por la primera no se sustenta sólo sobre el rigor académico que se nos presupone, sino en otro hecho más relevante si cabe: El *benchmark* puede encubrir multitud de trampas, mientras que la analítica enseña siempre las cartas con las que se juega.

el parámetro

En el caso que ahora nos ocupa, hay que buscar el parámetro de rendimiento más realista para una memoria trabajando dentro de una arquitectura PC. Elegir el tiempo de respuesta para una

Año / lín. caché	Tipo de memoria	Configuración más representativa para ese año: Ancho	Frecuencia	Lat. RAS	Lat. CAS	T (ns.)	Ganador y porcentaje
2000 / 32 bytes	RDRAM	16 bits	400x2 MHz	9 ciclos	9 ciclos	65 ns.	SDRAM (7.69 %)
	SDRAM	64 bits	133 MHz	2 ciclos	3 ciclos	60 ns.	
2003 / 32 bytes	RDRAM	32 bits	533x2 MHz	9 ciclos	9 ciclos	36.66 ns.	RDRAM (12.69 %)
	DDRAM	64 bits	166x2 MHz	2 ciclos	2.5 ciclos	42 ns.	
2003 / 64 bytes	RDRAM	32 bits	533x2 MHz	9 ciclos	9 ciclos	43.33 ns.	RDRAM (19.75 %)
	DDRAM	64 bits	166x2 MHz	2 ciclos	2.5 ciclos	54 ns.	
2003 / 128 bytes	RDRAM	32 bits	533x2 MHz	9 ciclos	9 ciclos	56.66 ns.	RDRAM (27.35 %)
	DDRAM	64 bits	166x2 MHz	2 ciclos	2.5 ciclos	78 ns.	

TABLA 10.16: Las configuraciones utilizadas para nuestra comparativa RDRAM vs. DDRAM. La columna etiquetada con una "T" denota el tiempo de respuesta para una línea de caché del tamaño indicado en la primera columna junto al año.

línea de caché L2 supone caracterizar fielmente la forma de cursar peticiones a memoria principal en un PC, al tiempo que se involucran conjuntamente latencia y ancho de banda.

No obstante, debemos tener presente un punto débil: Al considerar una petición aislada a memoria, no podemos evaluar la eficiencia en el servicio de peticiones concurrentes. En realidad, esto resultaría tan complejo de incorporar en nuestro análisis, que antes de proponer un indigesto modelo que le de cobertura, preferimos limitarnos a hacer las consideraciones de paralelismo más relevantes como colofón final a nuestra comparativa.

paralelismo

A la hora de elegir los modelos comerciales a enfrentar, hemos seleccionado dos configuraciones típicas del año 2000 y otras dos del año 2003, según hemos reflejado en la tabla 10.16. De esta manera lograremos, por un lado, cierta perspectiva evolutiva al tratarse de versiones en idéntico estadio de progreso para cada tipo de memoria, y por el otro, una perspectiva actual al constituir las configuraciones elegidas para 2003 las más representativas del mercado en el momento de cerrar la edición de este libro. Hemos incluido además tres tamaños de línea de caché L2 para este año, 32, 64 y 128 bytes, ya que aunque 32 era un valor casi inamovible hasta la séptima generación, a partir de ahí el K7 alcanza los 64 bytes en algunos modelos y el Pentium 4 los 128 bytes ya desde sus orígenes, con lo que estos valores son más realistas.

dos configs. típicas

tres tamaños de línea

Comenzando por el año 2000, la configuración de RDRAM proporciona un ancho de banda de 1.6 Gbytes/sg, mientras que su contrapartida en SDRAM sólo llega a 1.06 Gbytes/sg. Pero en el tiempo de respuesta de la línea de caché, la RDRAM pierde la batalla (36.66 ns para RDRAM según apuntamos en el cronograma de la figura 10.35, frente a 42 ns para DDRAM - ver figura 10.28). Esto es debido a su mayor latencia y a que la línea de caché no es suficientemente grande como para sacar provecho del ancho de banda.

2000

☛ pág. 81
☛ pág. 70

Tres años más tarde, la RDRAM ha acortado las latencias RAS y CAS (aunque mantienen el montante de 9 ciclos, la frecuencia es un 50 % superior) y casi triplicado el ancho de banda. En contraste, la DDRAM apenas mejora la latencia (se limita a recortar medio ciclo en la parte CAS), y duplica el ancho de banda con la incorporación del multiplicador 2x. Como resultado, las diferencias se acortan en la latencia y se mantienen en el ancho de banda en torno a un 35 % favorables a la RDRAM (4.2 frente a 2.7 Gbytes/sg.), colocando ya por delante a la RDRAM.

2003

Para tamaños de línea de caché superiores a 32 bytes, los escenarios son más favorables a la RDRAM, ya que aquí lo único que cambia es el tamaño del bloque de datos a transferir, y para esta tarea lo que cuenta es el ancho de banda, el gran bastión de la RDRAM. Las distancias van así ensanchándose a su favor, en torno a un 8 % por cada duplicación de la línea de caché. Las formas en que las memorias DDRAM de 64 bits y RDRAM de 32 bits responden a la hora de servir una línea de caché de 128 bytes a un procesador Pentium 4 fueron ya estudiadas a nuestro paso por el

tamaños de línea

estudio del microprocesador Northwood.

A la hora de evaluar la eficiencia en el tratamiento de peticiones simultáneas, la ventaja de la RDRAM se engrandece todavía más:

1. La concurrencia solapa latencias pero aumenta el uso del bus, lo que desplaza el protagonismo de los parámetros de rendimiento desde la latencia hacia el ancho de banda.
2. La arquitectura de la RDRAM cuenta con mejores bazas para explotar el paralelismo:
   - A nivel de segmentación, cuenta con 7 etapas frente a 3 en SDRAM/DDRAM (para el caso de CL=2).
   - A nivel de entrelazado, porque admite un factor 16 frente a sólo 4 en SDRAM/DDRAM, que además se aprovecha más porque recordemos que el entrelazado oculta la latencia RAS, que es de 9 ciclos en RDRAM frente a 2 en SDRAM/DDRAM.

En conclusión, nuestra comparativa señala a la RDRAM como ganadora, aunque aún debemos apuntalar el análisis con algunos matices de relevancia en el ámbito tecnológico y comercial.

13.7.2 Tecnológica

apuesta:

La primera vez que supimos de RDRAM fue en 1999, cuando nos encontrábamos escribiendo la primera edición de Arquitectura del PC. Quedamos fascinados de sus portentosas facultades, y apostamos por ella. Cuatro años más tarde hemos ganado la apuesta: tecnológicamente, claro, porque comercialmente la hemos perdido de calle.

- ganada

Pero vamos por partes. Tecnológicamente, tenemos una memoria mucho más estable que la DDRAM. Sander Sassen, nuestro mismo colaborador que en la edición de 2001 de este libro nos facilitó sus experimentos de sobreaceleración que demostraron que un Pentium III de 500 MHz podía trabajar a 1 GHz manteniendo a raya la variable térmica, ha conseguido hacer funcionar a 600x2 MHz una memoria RDRAM de 400x2 MHz y 2 Gbytes bajo una placa base con chipset i850 de Intel para Pentium 4. La sobreaceleración es del 50 %, muy superior a lo que podíamos esperar tratándose de la memoria, y la gran estabilidad de las señales eléctricas bajo ese régimen no hace sino corroborar nuestra percepción de que RDRAM es el mejor diseño si se pretende apuntar hacia frecuencias elevadas.

Un experimento similar sobre DDRAM nos hace salir bastante escaldados: La memoria DDRAM 133x2 MHz sobre juego de chips i845 de Intel para su mismo Pentium 4 sólo consigue un peldaño de sobreaceleración hasta 166x2 (un 25 %), tan sólo sobre 1 Gbyte de memoria como máximo, y subiendo la latencia CAS desde 2 a 2.5, algo que como ya analizamos en la sección 10.13.5.5 pág. 72 penaliza casi la mitad de esa mejora.

13.7.3 Comercial

- perdida

Comercialmente, RDRAM es una memoria derrotada. Las previsiones del mercado DRAM para 2003 hablan de una cuota de entre el 10-20 % para DDRAM (y subiendo) y de entre el 0.1-0.2 % para RDRAM (y bajando), esto es, se vende una RDRAM por cada cien DDRAM. Parece toda una paradoja, pero no lo es. Desde aquel 1999 hemos presenciado toda una campaña de descrédito para la RDRAM orquestada por intereses mercantilistas:

fabricantes

- Para los fabricantes, el coste por oblea es muy superior integrando RDRAM que DDRAM.

distribuidores

- Para los distribuidores, el volumen de unidades a repartir es muy inferior en RDRAM.

Característica	DDRAM	RDRAM	Mejor
Fiabilidad			
Tasa de errores	Media	Baja	RDRAM
Estabilidad eléctrica	Baja	Alta	RDRAM
Temperatura			
Consumo de potencia	Bajo	Medio	DDRAM
Calentamiento	Bajo	Alto	DDRAM
Arquitectura			
Segmentación	3 etapas	7 etapas	RDRAM
Factor de entrelazado en longitud	4	16	RDRAM
Rendimiento			
Latencia	Media	Alta	DDRAM
Ancho de banda	Medio	Alto	RDRAM
Tiempo de respuesta (línea caché)	Medio	Bajo	RDRAM
Sobreaceleración	Baja	Alta	RDRAM
Adquisición			
Coste	Medio	Alto	DDRAM
Disponibilidad comercial	Alta	Baja	DDRAM

TABLA 10.17: Comparativa DDRAM-RDRAM sobre los aspectos clave de memoria principal.

- Para los mayoristas, el margen de beneficio es más estrecho en RDRAM. `mayoristas`

- Para los usuarios, el producto sale más caro a igual capacidad frente a DDRAM. `usuarios`

En definitiva, que Rambus pretendía ganar mucho dinero con todo esto, y nos parece bien. Pero que lo haga a costa de toda la cadena comercial, incluido el usuario final, ya no nos parece tan buena idea, máxime cuando Rambus ni siquiera fabrica, pues su labor concluye en la especificación del producto. El sistema en bloque se ha confabulado en su contra, porque no tiene que aceptar sangrantes dictaduras cuando existen alternativas democráticas. Toma nota y cabalga con mesura, Microsoft; te enviamos saludos desde el mundo Linux. `la lección`

13.7.4 Conclusión

La tabla 10.17 resume las principales cualidades de cada memoria. La memoria RDRAM de Rambus nos hubiera permitido disfrutar de una memoria principal al más puro estilo procesador: Elevada frecuencia y gran estabilidad, sólo mediatizadas por la temperatura. Pero no va a ser así, porque el futuro comercial está acomodado a sus espaldas. `al estilo procesador`

Una vez más, lo importante para nosotros no es el ganador, sino la didáctica que encierra el haber asistido al desarrollo de dos escuelas antagónicas cuya finalidad común es el rendimiento. Esperamos que ahora se tenga más claro el papel que juegan la latencia y el ancho de banda en el rendimiento de memoria principal. Después de todo, tanto DDRAM como RDRAM van a desembocar en un punto común, ya que cada una está buscando lo que tiene la otra: Los zócalos de RDRAM de 64 bits ya están especificados (ver sección 10.7.8), y la memoria DDRAM de 1 GHz es el objetivo del consorcio DDR-II y DDR-III (ver sección 13.3.3).

☛ pág. 36
☛ pág. 165

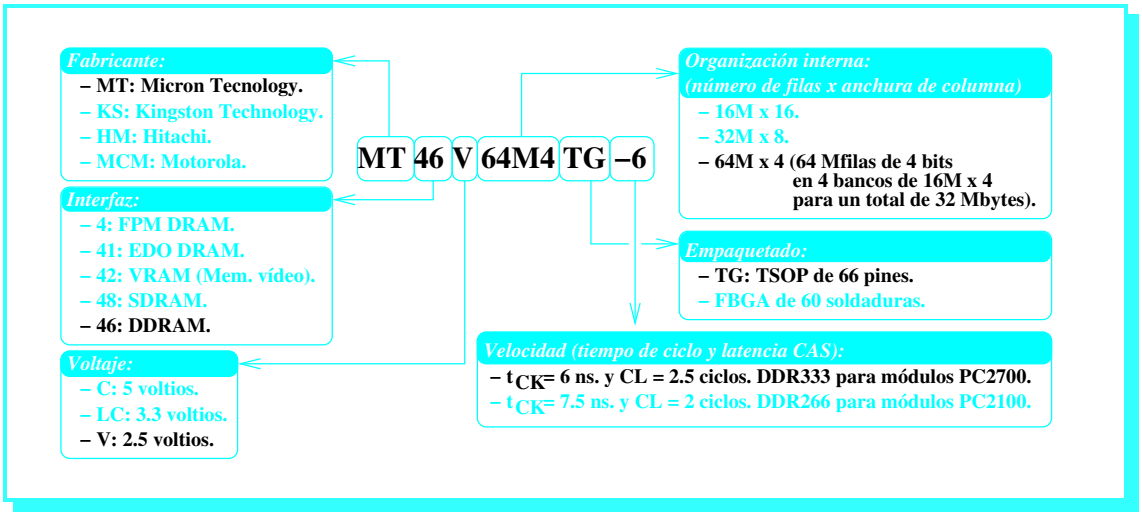


FIGURA 10.36: El etiquetado en el lomo de los chips de memoria según una de las versiones más extendidas, correspondiente al formato empleado por Micron Technology.

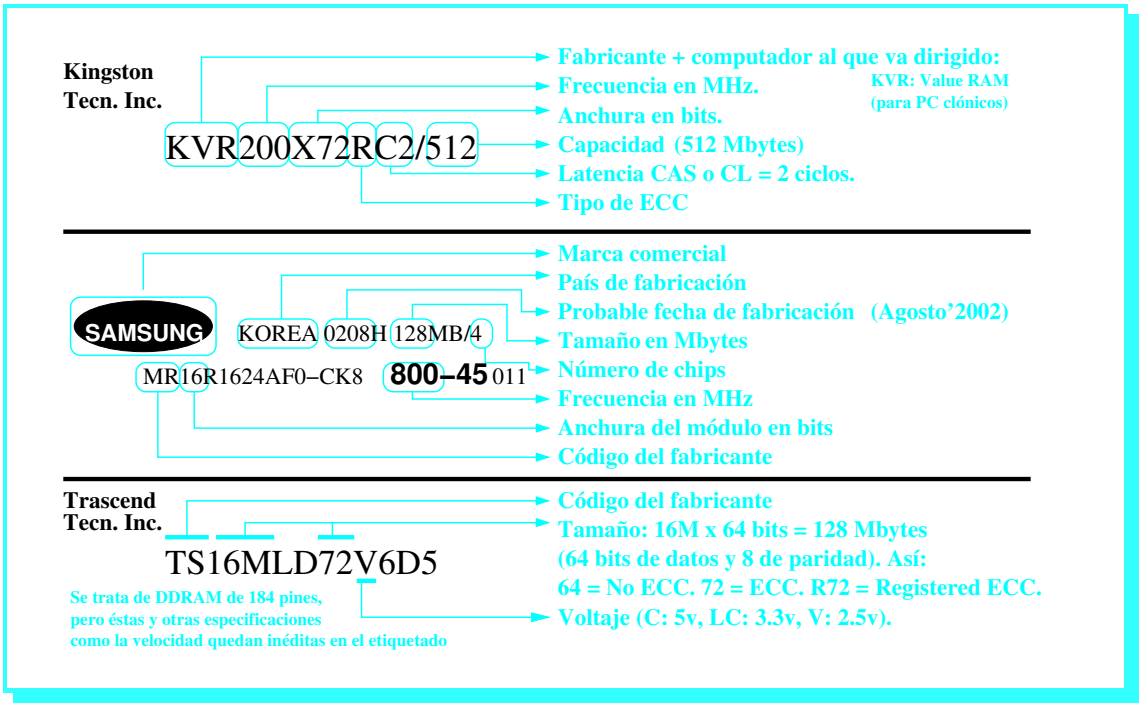


FIGURA 10.37: El etiquetado en los módulos de memoria para tres firmas de distintos segmentos: Kingston, como representante de gama alta, Samsung, como marca de calidad estándar, y Transcend como firma clónica.

Período temporal	Tipo de memoria	Parámetro comercial de referencia	Magnitud implícita	Ejemplos de uso
1990-1996	FPM, EDO, BEDO	Tiempo de respuesta	Decenas de ns.	-7, -6, -5
1997-1999	SDRAM	Tiempo de ciclo	Nanosegundos	-10, -8, -7
1997-1999	SDRAM	Frecuencia	Ninguna	100 MHz, 133 MHz
2000-2001	DDRAM/RDRAM	Frecuencia de bus	MHz	PC-100, PC-133
2002-2003	DDRAM/RDRAM	Ancho de banda	Mbytes/sg.	PC-2700, PC-3200

TABLA 10.18: El referente de rendimiento más utilizado por el mercado para cada tipo de memoria.

SECCIÓN 10.14

Etiquetado y especificaciones

A la hora de referirse a un producto de memoria, el mercado quiere una sola magnitud, y que pueda entenderse sin intrincados tecnicismos. Así se ha tenido al Mbyte como referente del tamaño, aunque el episodio del rendimiento ha sido más tortuoso, con el agravante de que simplificar aquí supone una temeridad, la misma que ya se comete con el procesador y sus MHz.

referente

A nuestro paso por los parámetros de rendimiento de la memoria señalamos la latencia y el ancho de banda como los más determinantes. El mercado comenzó apadrinando únicamente la latencia, y poco a poco se fue dando cuenta de que los tiros van más por el ancho de banda. Lo deseable entonces sería conjugar ambos en un único parámetro, y puestos a ser exigentes, ponderar más el ancho de banda que la latencia. ¿Lo ha logrado el mercado? Sí, aunque después de quince años y un largo culebrón que se resume en la tabla 10.18, y que trataremos de clarificar dado su enorme valor didáctico.

Para los chips

◀ 14.1

Los chips de memoria llevan una inscripción en su lomo que suele hacer las veces de número de serie o *Part Number*. Este **etiquetado** dista mucho de seguir un estándar a pesar de las numerosas recomendaciones hechas públicas por los principales fabricantes del sector, aunque lo habitual es codificar en él atributos como el tipo de memoria, su voltaje, el tamaño y su organización, el empaquetado y la velocidad de los chips. La figura 10.36 muestra el patrón que más se repite con un ejemplo de DDRAM correspondiente al fabricante Micron.

sin estándar
atributos

☛ pág. 88

Hasta la llegada de la memoria RDRAM, el referente para la velocidad de los chips fue siempre la latencia, antecedida por un guión en el número de serie del chip. Este número representaba decenas de nanosegundos de tiempo de respuesta en los diseños asíncronos, y nanosegundos de tiempo de ciclo en los diseños síncronos.

latencia
tiempo de respuesta
tiempo de ciclo

☞ **Ejemplo 10.9:** LA LATENCIA DE LA MEMORIA EN EDO Y SDRAM

En memorias FPM, EDO y BEDO, los chips indican la latencia con una sola cifra a la que hay que agregar un cero para obtener tiempo de respuesta. Los valores empleados por el mercado han sido:

- -7 para FPM (70 ns.).
- -6 y -5 para EDO (70 y 60 ns.).
- -5 y -4 para BEDO (50 y 40 ns.).

En memorias SDRAM y DDRAM, los chips indican una o dos cifras para el tiempo de ciclo, siendo los valores empleados los siguientes:

- -15 (ns. para 66 MHz), -12 (83 MHz), -10 (100 MHz), -8 (125 MHz), -75 (7.5ns. para 133 MHz), -7 (143 MHz) y -6 (166 MHz).
- -75 (7.5 ns para 133x2 MHz), -6 (166x2 MHz), -5 (200x2 MHz) y -33 (3.3 ns. para 300x2 MHz).

Nótese que si en las memorias asíncronas se asume implícito un cero, en las memorias síncronas el punto decimal hay que intuirlo, puesto que no aparece ni en 7.5 ni en 3.3 ns.

primer lío

Es decir, que en una memoria EDO, -7 representa 70 ns. de tiempo de respuesta, mientras que ese mismo -7 en una SDRAM significa 7 ns. de tiempo de ciclo. Para completar el caos, el tiempo de ciclo en las memorias síncronas coetáneas con las asíncronas equivale a la cuarta-quinta parte del tiempo de respuesta (o sea, que una memoria de 100 ns. de tiempo de respuesta se empareja en la operativa SDRAM con un tiempo de ciclo de 20-25 ns. - ver escala en abcisas de la figura 10.24), y ésta última puede además confundirse con una DDRAM de 2.5 ns. cuando aparezca esta memoria a 800 MHz (400x2).

pág. 66

frecuencia

Para distinguir entre tiempo de respuesta y tiempo de ciclo, la primera idea fue aprovechar el sincronismo de la memoria SDRAM para utilizar la frecuencia como escaparate, lo cual suscribía tres ventajas:

1. La frecuencia se extrapolaba directamente del tiempo de ciclo, al ser su magnitud inversa.
2. Se daba al mercado un valor al que estaba sobradamente acostumbrado por ser el parámetro por excelencia del procesador.
3. Se utilizaba un indicador numérico **mejor cuanto más alto**, en consonancia con los índices de rendimiento.

segundo lío

El principal inconveniente de la frecuencia era su ambigüedad, pues podía referirse a los chips de memoria o al bus de memoria, e incluso podía confundirse con la frecuencia de la placa base o del propio procesador. Para apuntillar el caos, los chips SDRAM de 66, 100 y 133 MHz sólo pueden conectarse a un bus de esa frecuencia si se programan para una latencia CAS de 3 ciclos, ya que de reducir ésta a 2 ciclos nos vemos abocados a aumentar el tiempo de ciclo para ensanchar ligeramente el período de reloj, bajando entonces la frecuencia en similar proporción (ver sección 10.13.4.7 y de nuevo la figura 10.24).

pág. 67
pág. 66

Para los módulos

◀ 14.2

En los módulos, el **etiquetado** se simplifica un tanto porque no trascienden tantos aspectos de bajo nivel, aunque podremos descender ese peldaño fijándonos en los chips siempre que el etiquetado del módulo o un eventual disipador de calor no hayan tapado su serigrafía. La figura 10.37 muestra tres ejemplos de **etiquetado** de módulos procedentes de fabricantes bien dispares; todos ellos mantienen información en su página Web acerca de cómo extraer toda esta información a partir del etiquetado.

pág. 88

Web

La frecuencia indicada para el módulo no tiene por qué coincidir con la que calculamos para el chip, ya que eso sólo ocurre en caso de programar el chip para el valor CL (CAS Latency) más conservador (3 ciclos en SDRAM y 2.5 en DDRAM). En realidad, el único **etiquetado** de los tres que no es ambiguo es el de Kingston, porque al indicar tanto frecuencia como latencia CAS para sus chips constituyentes, sabemos que queda margen para programar la latencia CAS subiéndola desde 2 hasta 2.5 ciclos, y así aumentar la frecuencia desde los 200 hasta los 266 MHz para los chips, hecho que se traslada de inmediato a su módulo.

ambigüedad

14.2.1 La denominación PC-XXX

Tras el galimatías anterior, los fabricantes de módulos de memoria tomaron conciencia de dos cosas: Había que unificar el etiquetado al nivel de módulo frente al nivel de chips, y tomar en consideración el ancho de banda frente a la latencia. De esta manera, muchos módulos de memoria comerciales comenzaron a adoptar una especificación propia, pegándole al módulo un adhesivo en el que se rotulaba la cadena de caracteres **PC-XXX**. El valor XXX apuntaba la frecuencia del bus al que debemos conectar nuestro módulo de memoria para conseguir un óptimo aprovechamiento de sus prestaciones.

frecuencia del bus

Los valores CL con que se programaban los chips y el resto de sus cualidades internas pasaban así a un segundo plano: La firma que manufacturaba el módulo compraba una remesa de chips, fijaba su parámetro CL para delimitar su tiempo de ciclo, y luego calculaba la inversa de éste para obtener la frecuencia que trasladaba a la etiqueta PC-XXX. Puesto que el ancho del bus llevaba largo trecho estabilizado en 64 bits (8 bytes), sólo había que multiplicar XXX por 8 para obtener el ancho de banda de la memoria. Por ejemplo, una SDRAM PC-133 tiene un ancho de banda de 1066 Mbytes/sg y obtiene su máximo rendimiento sobre una placa base de 133 MHz.

relación con placa base

En esas estábamos cuando irrumpió la memoria DDRAM, en la que la frecuencia llevaba implícito un multiplicador de dos que confundía este cálculo, y la memoria RDRAM, en la que el ancho del bus ya no era de 64 bits, sino de 16, evolucionando posteriormente hasta los 32 (2002) y los 64 bits (2003). Ante este embrollo, lo mejor era indicar de forma explícita el ancho de banda.

tercer lío

14.2.2 La denominación PC-XXXX

Sin previo aviso, la denominación **PC-XXXX** pasó a tener 4 dígitos, y el número a representar una magnitud bien diferente: Ancho de banda en Mbytes/sg. Esa vara de medir es mucho más justa, ya que incluye a la frecuencia, a sus posibles multiplicadores de reloj y a la anchura del bus de forma conjunta, sirviendo para cualquier tipo de memoria: SDRAM, DDRAM y RDRAM. Aunque en ésta última los fabricantes siguen sosteniendo PC-XXX con el significado de frecuencia, creemos que no tardarán en emplear el ancho de banda, habida cuenta de que en el horizonte de la RDRAM aparecen tres anchuras de bus diferentes.

ancho de banda

Nosotros vamos a unificar criterios utilizando PC-XXXX en todos los casos, facilitándonos así la labor de comparar prestaciones. Si el mercado no nos siguiera y continuaran apareciendo módulos de memoria donde XXXX es inferior a 1500, ya sabe que están incurriendo en la ambigüedad de indicarle sólo la frecuencia, y que tendrá que descubrir la anchura en bytes para multiplicar

advertencias

ambos y llegar así a nuestra normalización. Otra advertencia que queremos cursarle es que del producto MHz × bytes no resulta de forma exacta el Mbyte/sg, sino un valor ligeramente inferior. Este redondeo, del que ya advertimos cuando definimos el ancho de banda (ver sección 10.4), queda justificado por lo mucho que simplifica los cálculos. Por ejemplo, para conocer la frecuencia de la placa base con la que se saca el máximo partido a una memoria PC-XXXX, basta dividir por ocho el número XXXX en todos los casos. Veamos por qué:

redondeo
☛ pág. 18

dividir por ocho

- En DDRAM, cuya anchura es siempre 64 bits, dividiendo XXXX por 8 obtenemos la frecuencia del bus de memoria, que es la misma que la de la placa base a lo largo de toda la séptima generación.
- En RDRAM, cuya anchura es de 16, 32 y 64 bits, dividiendo XXXX por 2, 4 y 8 respectivamente, obtenemos de nuevo a la frecuencia del bus de memoria, que debe ser superior en un factor de 4, 2 y 1 a la de la placa base, por lo que si corregimos este factor llegamos en todos los casos a aplicar una única división por ocho, al igual que en la DDRAM.

relación con placa base
pág. 102 ☚

La figura 10.40 refleja la correspondencia entre las especificaciones PC-XXX y PC-XXXX y la frecuencia en placa base. Llegados a este punto, recomendamos moverse en todo momento tomando el ancho de banda como referencia (o en su defecto, la frecuencia), pero nunca derivar hacia tiempos de ciclo, so pena de incurrir en un nuevo caos: En DDRAM, llegaríamos a un tiempo de ciclo para el módulo que no coincidiría con el de sus chips constituyentes, ya que éstos no son el doble de rápidos que los de SDRAM, sino submitades de éstos organizados en bloques internos que responden en semiciclos de reloj de forma alterna.

cuarto lío

14.2.3 La denominación X-Y-Z timing

Hasta el momento hemos establecido ligaduras entre el etiquetado de la memoria y su rendimiento, pero en SDRAM y DDRAM el rendimiento está condicionado por la selección de unos parámetros que configuran el interfaz. En la literatura más técnica, estos parámetros conforman un trío bajo la denominación "X-Y-Z timing", donde:

1. X representa la latencia CAS ó *CAS Latency* (latencia de columna ó CL).
2. Y representa la latencia desde RAS a CAS ó *RAS to CAS Delay* (latencia de fila ó RCD).
3. Z representa la latencia para la precarga de una fila desde otro bloque entrelazado del chip, es decir, cuantifica el beneficio en la latencia de fila gracias al entrelazado.

quinto lío

Los tres parámetros se dan en ciclos de un reloj cuyo período determina la propia terna, lo cual no deja de ser chocante. También resulta confuso que se indique primero la latencia de columna y después la de fila, cuando la secuencia de funcionamiento en la SDRAM sugiere justo lo contrario.

sexto lío

Para colmo, en no pocas ocasiones hemos visto prescindir de la coletilla *timing*, y al emplearse también guiones para separar los valores de la terna de configuración, uno tiende a confundir una memoria DDRAM 2-2-2 con sus valores para la ráfaga de salida de datos. En este sentido, recordaremos que la especificación de la ráfaga forma un cuarteto en el que los tres últimos valores son siempre iguales, mientras que en este trío no tienen por qué serlo.

SECCIÓN 10.15

Diez consejos para elegir la memoria principal del PC

puntos de atención
pág. 93 ☚

Con objeto de dotar a nuestro PC de la mejor memoria principal, hemos seleccionado diez puntos de atención agrupados dentro de tres vertientes principales (ver figura 10.38):

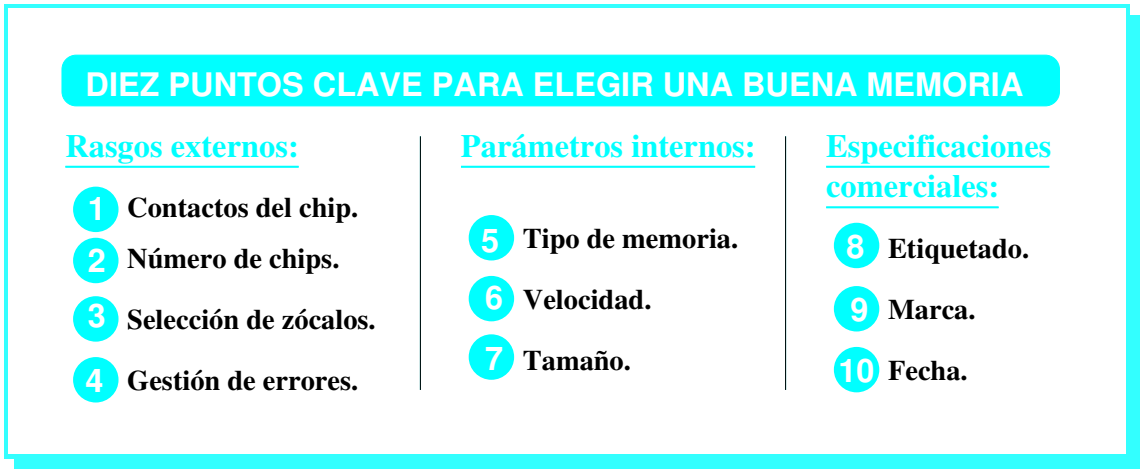


FIGURA 10.38: Nuestra lista de consejos para elegir la memoria pricipal del PC estructurada en sus tres vertientes principales.

- **Rasgos externos.** Agrupan aquellos indicios de calidad que pueden ser reconocidos mediante una sencilla exploración visual. Cuatro cosas tenemos aquí: Los contactos del módulo, sus chips, los zócalos de la placa base y la gestión de errores (paridad y ECC). `externos`
- **Parámetros internos.** Tipo, velocidad y tamaño de la memoria. `internos`
- **Especificaciones comerciales.** El etiquetado y las diferentes denominaciones de los folletos publicitarios, las marcas más recomendables, y la fecha de fabricación. `comerciales`

Nuestra atención recaerá a partir de ahora sobre el módulo DRAM, la unidad comercial de memoria principal disponible en tiendas.

Rasgos externos

◀ 15.1

15.1.1 Los contactos del módulo

El número de contactos del módulo es mayor cuanto más actual es el módulo. 72 contactos entre 1990 y 1996, 168 contactos entre 1997 y 1999, 184 contactos entre 2000 y 2002, y 232 y 326 en 2003 para los nuevos formatos de memoria RDRAM de 32 y 64 bits. `número`

El color y material utilizado para estos contactos también delata la antigüedad del módulo. Antiguamente era frecuente encontrar contactos de latón (plateados), mientras que ahora son casi todos dorados, más recomendables por ser las aleaciones del oro materiales con mejores propiedades para la conducción eléctrica. `color y material`

Además, conviene echar un vistazo al material utilizado en los pines de los zócalos de memoria en placa base, porque si son de diferente material que los del módulo, la tasa de errores aumentará. Esto es así porque el latón se irrita al ponerse en contacto a presión con el oro u otro metal, y el óxido que desprende se adhiere al oro y se endurece, convirtiéndose en una línea de alta impedancia en períodos incluso inferiores a los doce meses. Conjugando esta premisa y la del párrafo anterior tenemos el orden de prelación que refleja la tabla 10.19. `placa base`

☛ pág. 94

15.1.2 Los chips

Estudios de los propios fabricantes de memorias de semiconductores revelan que el número de fallos que se producen en un módulo de memoria principal está más relacionado con el número

Posición en el ránking	Contactos en la placa de circuito impreso del módulo	Contactos en el zócalo zócalo de la placa base
1	Dorados	Dorados
2	Plateados	Plateados
3	Dorados	Plateados
4	Plateados	Dorados

TABLA 10.19: Preferencias sobre los materiales de los contactos en un módulo de memoria y su relación con sus homólogos en placa base.

Velocidad	Anchura de los chips (a mayor anchura, menor número de chips por módulo) 4 bits	8 bits	16 bits	32 bits
100 MHz x 2	128, 256, 512 Mb.	128, 256, 512 Mb.	128, 256, 512 Mb.	No se fabrica
133 MHz x 2	128, 256, 512 Mb.	128, 256, 512 Mb.	128, 256, 512 Mb.	No se fabrica
166 MHz x 2	256 Mb.	128, 256 Mb.	128, 256 Mb.	No se fabrica
183 MHz x 2	No se fabrica	No se fabrica	No se fabrica	64 Mb.
200 MHz x 2	No se fabrica	No se fabrica	No se fabrica	64, 128 Mb.
250 MHz x 2	No se fabrica	No se fabrica	No se fabrica	128 Mb.
300 MHz x 2	No se fabrica	No se fabrica	No se fabrica	128, 256 Mb.
400 MHz x 2	No se fabrica	No se fabrica	No se fabrica	256 Mb.

TABLA 10.20: Correlación entre la velocidad y el número de chips de un módulo de memoria según la gama de productos DDRAM de Micron disponible comercialmente en el segundo trimestre de 2003 (salvo las versiones de 300 y 400 MHz, previstas para fin de año). Cada casillero muestra la gama de tamaños comerciales (en Mbytes) para los módulos de memoria fabricados bajo una velocidad y anchura de chips determinadas. Dado que la anchura del módulo es siempre 64 bits, a mayor anchura en los chips, menor número de éstos por módulo comercial.

FALLOS: de chips empaquetados separadamente que con el número total de bits del módulo. Atribuimos este comportamiento a cuatro razones:

fabricación
1. El proceso de manufacturación es más complejo, con un número creciente de soldaduras, componentes a ser manipulados, y extensión de las líneas eléctricas que los interrelacionan.

sincronismo en DDRAM
2. La sincronización interna entre las celdas de un mismo chip es más exacta que entre las celdas de chips vecinos, aspecto más crítico a regímenes de elevada frecuencia. Con menos chips es más fácil que el módulo en su conjunto cumpla las mismas especificaciones de sus chips constituyentes. Por eso no es de extrañar que los fabricantes escojan menos chips (esto es, chips de mayor anchura) para montar sus módulos de memoria conforme éstos aumentan su velocidad. La tabla 10.20 evidencia esta correlación para el caso de la DDRAM.

retardo en RDRAM
3. En la escuela opuesta, RDRAM, la sincronización anterior no entra en juego, puesto que cada acceso responde desde un único chip. Lo que perjudica aquí es el hecho de que las señales eléctricas atraviesan los chips consecutivamente ya que se conectan en serie al bus, y no en paralelo como la DDRAM. Por lo tanto, el retardo de las señales es proporcional al número de chips, lastrando la frecuencia de trabajo del bus.

tasa de errores
4. La tasa de errores aumenta por dos causas fundamentales: Primero, el enrutado de las líneas es más largo y complejo a mayor número de chips, y por lo tanto, también más propenso a las interferencias electromagnéticas. Segundo, las incidencias provocadas por los rayos cósmicos aumentan para una memoria de mayor densidad, siendo ésta dos veces más sensible al cambio de uno de sus bits cuando tiene cuatro veces más celdas. En el ejemplo 10.10 constatamos que esto perjudica a los módulos con un número elevado de chips.

pág. 95 ☚

Ejemplo 10.10: Influencia del número de chips en la tasa de errores de un módulo de memoria

Supongamos un módulo de 256 Mbytes de memoria principal y tres posibles diseños:

- Módulo A: 16 chips de 16 Mbytes cada uno.
- Módulo B: 4 chips de 64 Mbytes cada uno.
- Módulo C: Un único chip de 256 Mbytes.

Si atribuimos una tasa de errores E al chip de 16 Mbytes, entonces según la relación anterior, el chip de 64 Mbytes tendrá una tasa de errores de 2E, y el chip de 256 Mbytes tendrá una tasa de errores de 4E. La tasa total de fallos en cada caso será la siguiente:

- Tasa de errores de A: 16 chips x (1E errores/chip) = 16E
- Tasa de errores de B: 4 chips x (2E errores/chip) = 8E
- Tasa de errores de C: 1 chip x (4E errores/chip) = 4E

Por lo tanto, el diseño C de un solo chip es dos veces más fiable que B y cuatro veces más que A a igual tamaño en Mbytes.

`coste de fabricación`

Si muchos fabricantes optan por colocar un gran número de chips por módulo es porque les resulta más barato de producir. Recordemos que la fase de verificación se efectúa al nivel de chip, y un solo defecto obliga a sacrificar todas sus celdas. Con pocos chips, cada uno tendrá más celdas, lo que aumentará tanto la probabilidad de que el chip tenga un defecto como la cantidad de memoria a sacrificar por su culpa. De hecho, los fabricantes con solera del mercado ofrecen en torno a un 20 % más baratos los módulos que están muy poblados de chips.

Ejemplo 10.11: Influencia del número de chips en el coste de un módulo de memoria

En Octubre de 2002 consultamos los precios de memoria principal de su fabricante líder, Kingston. Elegimos una tecnología plenamente consolidada como la SDRAM PC-133 de 128 Mbytes para obtener un precio lo más estable posible.
El precio dado por un mayorista de Málaga para un módulo compuesto por 16 chips de 8 Mbytes fue de 18.25 €, mientras que el mismo módulo pero con 8 chips de 16 Mbytes costaba 22.50 €. El sobreprecio a pagar fue del 23 %.

`consejo`

La inmensa mayoría de fabricantes no repercute esta diferencia de precio en su gama de productos, por lo que nuestro consejo aquí es claro: Siempre que le den a elegir a igualdad de precio,

quédese con los módulos de memoria provistos del menor número posible de chips. Probablemente dentro de unos años, este consejo tenga una contraindicación en el aspecto térmico, pero puesto que aquí el procesador va muy por delante, nos enseñará sobrados antídotos al respecto.

15.1.3 Los zócalos

numerosos

Nuestro sistema deberá venir equipado con el mayor número posible de zócalos de memoria, ya que esto aumenta tanto la flexibilidad como la futura expansión de una configuración.

A la hora de pinchar los módulos en sus zócalos de la placa base debemos tener en cuenta las combinaciones permitidas y prohibidas, tal y como detallamos en la sección 22.6.1. Dentro de las que son lícitas, nuestras recomendaciones son las siguientes:

Volumen 3 ☚
sugerencias:

número

1. Utilizar el menor número posible de zócalos, ya que cuantos más queden libres, mayor será la capacidad de expansión futura del PC y la flexibilidad de su configuración. Se trata, por lo tanto, de acaparar la máxima cantidad de memoria con el mínimo número de módulos, eso sí teniendo en cuenta la pérdida de modularidad que esto conlleva, ya que en caso de estropearse un chip, el sacrificio repercutirá sobre su módulo al completo,

   El ahorro de costes que supone, por ejemplo, adquirir un solo módulo de 256 Mbytes en lugar de dos de 128 Mbytes tampoco es significativo, tal y como se desprende de la tabla 10.23. Tan sólo nos hemos encontrado diferencias importantes en tamaños extremos: por ejemplo, en el período inaugural de comercialización de los módulos de 1 Gbyte, su precio era muy superior al de dos módulos equivalentes de 512 Mbytes.

pág. 100 ☚

velocidad

2. En caso de utilizar módulos de diferente velocidad, pinchar la memoria sobre los zócalos de la placa base de forma que los bancos numéricamente más bajos alojen la memoria más rápida. Esto es así porque los sistemas operativos actuales suelen utilizar las posiciones más bajas del espacio de direcciones de memoria para alojar la información y los procesos más críticos para el rendimiento del sistema (vectores de interrupción, controladores de dispositivo, planificador de procesos, ...).

voltaje

3. En caso de utilizar módulos de diferente voltaje, vigilar que la placa base sea capaz de suministrar este voltaje dual.

Riesgo 10.3: Una dualidad de voltaje arriesgada en placa base

Conocemos bastantes casos de placas dotadas de zócalos SIMM de 5 voltios junto a zócalos DIMM de 3.3 voltios que suministran 5 voltios a todo el sistema de memoria en cuanto se llena alguno de los zócalos SIMM. En estas placas base hemos visto morir unos cuantos módulos de memoria DIMM por llegarle 5 voltios a pesar de estar situados sobre zócalos de 3.3 voltios (sólo si son de una marca de solera tolerarán este cambio, tal y como ya indicamos en el riesgo 10.2).

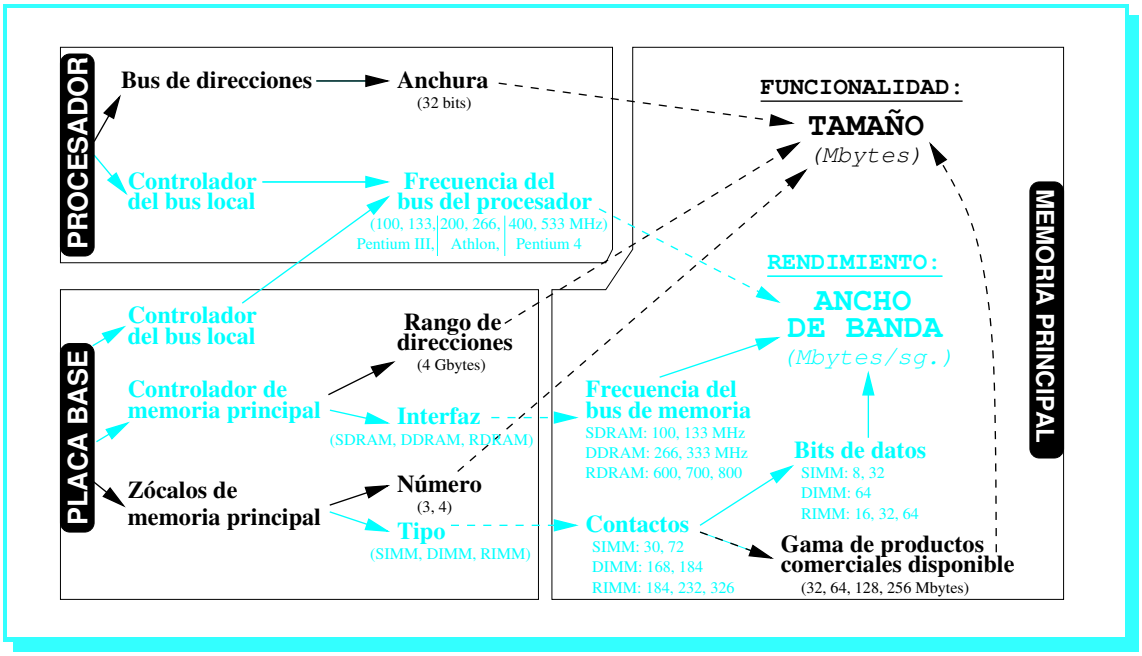


FIGURA 10.39: Influencia del procesador y la placa base en los parámetros internos de memoria principal. Hemos utilizado dos colores para separar aquellos aspectos que influyen mayormente sobre el tamaño y aquellos que repercuten más sobre el ancho de banda. Asímismo, utilizamos trazos discontinuos para marcar las relaciones que condicionan o restringen un aspecto concreto, y trazos continuos para marcar aquellas que lo determinan de forma definitiva.

Zócalo de memoria principal		Interfaces existentes para ese zócalo	
Tipo	Número de contactos	Tipo	Anchura del bus de datos
SIMM	30	FPM	8 bits
SIMM	72	FPM, EDO, BEDO	32 bits
DIMM	168	EDO, BEDO, SDRAM	64 bits
DIMM	184	DDRAM	64 bits
RIMM	168/184	RDRAM	16 bits
RIMM	232	RDRAM	32 bits
RIMM	326	RDRAM	64 bits

TABLA 10.21: Relación entre el zócalo de una memoria y su interfaz y anchura. Puesto que la placa base alberga los zócalos, su elección condiciona sobremanera el tipo de memoria y sus prestaciones.

15.1.4 Detección y corrección de errores

distinción

Si el número de chips del módulo no es potencia de dos, tenemos buenas noticias: La presencia de chips de paridad/ ECC que aumentan la fiabilidad del módulo. Esta cualidad suele encarecer el precio de un módulo de memoria en torno al 10-15 %, mereciendo la pena en función de la seriedad de las tareas a que dediquemos el PC.

interoperab.

Si uno declina esta posibilidad, debe saber que los módulos sin paridad/ ECC imponen ciertas restricciones de interoperabilidad: Pueden no funcionar en las placas base con paridad/ECC que no admitan programar esta cualidad, la cual debe ser desactivada utilizando la opción DRAM DATA INTEGRITY MODE del menú CHIPSET FEATURES SETUP de la BIOS para el controlador en placa base (ver sección 24.3.4), y la opción MEMORY PARITY/ECC CHECK del menú BIOS FEATURES SETUP para las líneas adicionales en el bus de memoria (ver sección 24.3.3).

☛ Volumen 4
☛ Volumen 4

mezcla

Adicionalmente, los módulos sin paridad/ ECC no pueden acompañar en un sistema de memoria a otros con paridad/ECC. En cuanto uno de los módulos lleve esta cualidad, todos sus módulos acompañantes deberán llevarla también (a no ser, claro está, que programemos la desactivación conjunta de todos ellos).

15.2 Parámetros internos

placa base

pág. 97

Entramos a analizar los aspectos internos de la memoria principal. Las restricciones a la hora de elegir e instalar el mapa de memoria de un PC vendrán impuestas por las propiedades de la placa base y del procesador, siendo ella más importante, sobre todo a la hora de lograr un óptimo rendimiento de la memoria. La figura 10.39 sintetiza todas las implicaciones de uno y otro en el tamaño y el ancho de banda de la memoria.

15.2.1 Interfaz y formato

Cada sistema permite la conexión de unos tipos de memoria determinados dependiendo de dos características de su placa base: Los zócalos que utilice y el controlador de memoria de que disponga.

pág. 97

zócalo

- **El tipo de zócalo.** Condicionará la elección de los módulos de memoria principal que se hayan fabricado bajo ese formato. La tabla 10.21 sintetiza los interfaces a los que se encuentra limitado cada tipo de zócalo, lo que a su vez determina la anchura del bus de datos.

controlador

- **El controlador de la placa base.** Determina las opciones de refresco que se pueden aplicar sobre los módulos de memoria. Modalidades de refresco hay muchas, y habrá que asegurarse de que el controlador tiene implementada aquella que necesita el módulo que queremos incorporar al sistema.

Ejemplo 10.12: CONTROLADORES DE MEMORIA INCOMPATIBLES A CAUSA DEL CIRCUITO DE REFRESCO

- En memoria principal de quinta y sexta generación, se producía una incompatibilidad entre la memoria FPM y la EDO por el hecho de que ésta última introdujo novedades en el circuito de refresco, consecuencia de la necesaria reorganización de tareas para su posterior segmentación. Esto provocó que en muchas placas base antiguas no funcionara el tipo de memoria EDO/BEDO.
- En épocas más recientes de séptima generación, en el segmento de los portátiles, el refresco se realiza internamente desde los propios chips de memoria en lugar de ser controlado desde la placa base, lo que también impide su interoperabilidad con la placa base de un PC convencional (al margen de que su formato es de dimensiones más reducidas y no admite tal interconexión).

15.2.2 Velocidad

a la par

De cara a no desaprovechar prestaciones de nuestro equipo, resulta fundamental conjugar correctamente las velocidades de placa base y memoria principal. Si colocamos memoria demasiado lenta, la placa base se verá forzada a introducir estados de espera o *wait states* para lograr la sincronización, si es que finalmente lo consigue. En el extremo opuesto, si la memoria es demasiado rápida, funcionará como una más lenta o incluso puede no funcionar.

controlador

Una vez más, serán el controlador y los zócalos de memoria en placa base nuestros principales puntos de atención aquí. El manual de la placa base suele especificar el rango de velocidad que toleran los circuitos de sincronización y temporización de su controlador de memoria ubicado en el puente norte del juego de chips, e incluso recomendar los valores más aconsejables. En ausencia de esta información, la figura 10.40 nos indica para cada placa base qué tipo de memorias se quedan cortas, largas o a la par con ella, en función de la velocidad de ambas. Sólo resta identificar el tipo de zócalo de que disponemos en placa base (DIMM o RIMM) y el número de contactos para saber si podemos acoplarle la que allí se apunta como más conveniente.

☛ *pág. 102*

zócalo

entrelazado

latencia CAS

Aparte de la frecuencia y el tiempo de ciclo que conforman el ancho de banda, hay que considerar la influencia de otros parámetros: En SDRAM/DDRAM, el entrelazado interno de los chips en un factor 2 ó 4 puede proporcionarnos un rendimiento extra en torno al 10 %, y otro tanto ocurre con la reducción de la latencia CAS desde 3 a 2 ciclos (ver tabla 10.14) si el chip lo admite (esto suele venir especificado en el número de serie del módulo, tal y como se refleja en la figura 10.37). Si no aparece indicado allí, lo más probable es que ostente el valor más elevado, que se corresponde con una latencia CAS de 3 y 2.5 ciclos para SDRAM y DDRAM, respectivamente.

☛ *pág. 73*

☛ *pág. 88*

misma vel.

Y para finalizar, un último consejo: Procure adquirir siempre memoria de la misma velocidad en todos sus bancos, ya que de no ser así, el módulo más rápido dentro de un mismo banco deberá siempre esperar al más lento, con lo cual, aún suponiendo que el controlador de memoria pudiera gestionar este asincronismo (no todos son capaces de hacerlo), la latencia efectiva sería siempre la del módulo más lento, desaprovechando el resto.

15.2.3 Tamaño

La máxima cantidad de memoria principal instalable en un PC viene limitada por la menor de las tres cantidades siguientes:

zócalos

1. El **número de zócalos** que tenga la placa base, multiplicado por el tamaño máximo con que se fabrican los respectivos módulos según se indica en la tabla 10.4. En las placas base de séptima generación de 2001, lo más usual es encontrar 4 zócalos DIMM, mientras que el máximo tamaño que se fabrica para un módulo de memoria en este formato es de 512 Mbytes ó 1 Gbyte según el fabricante, lo que resulta en una cota superior de 2 ó 4 Gbytes, respectivamente.

☛ *pág. 32*

controlador

2. El rango de direcciones que permite manejar el **controlador** de la placa base. Aún siendo imposible generalizar, el valor máximo más común fue de 256 Mbytes en el contexto de la quinta generación, 512 Mbytes para la sexta y 1 Gbyte para la séptima.

procesador

3. La cantidad de memoria direccionable por el **procesador**. Todos los procesadores de quinta, sexta y séptima generación disponen de un bus de direcciones de 32 bits, lo que les permite direccionar hasta $2^{32} = 4$ Gbytes. Normalmente, una parte de ese espacio se corresponde con el direccionamiento de la entrada/salida y el sistema operativo, con lo que el espacio de direcciones de usuario se queda normalmente en la mitad o así, como ocurre en Windows.

valor mínimo

Lo más frecuente es que el valor mínimo que determina el límite de memoria instalable venga dado por el número de zócalos de memoria. Dado que ese límite hardware suele quedar muy por

Aplicación		Sistema operativo		
Demanda	Descripción	Windows'98	Windows'00	Linux
Perfil de usuario: Estudiantes y administración y servicios				
Baja	Procesamiento de textos, correo electrónico, entrada de datos	32-64	64-96	48-80
Media	Comunicaciones por fax, hojas de cáculo y administración de bases de datos (más de dos aplicaciones abiertas a la vez)	64 - 128	64 - 128	48 - 112
Alta	Documentación compleja, contabilidad, aplicaciones gráficas y de presentación de ponencias, conectividad en red	128 - 384	96 - 256	80-240
Perfil de usuario: Domésticos, ejecutivo y analistas				
Baja	Propuestas, informes, hojas de cálculo, aplicaciones gráficas, bases de datos, comunicaciones fax/e-mail, presentaciones	32 - 48	64 - 96	48 - 80
Media	Presentaciones complejas, análisis de ventas y mercados, gestión de proyectos, acceso a Internet	48 - 64	96 - 128	80 - 112
Alta	Aplicaciones estadísticas, grandes bases de datos, análisis técnicos y de investigación, video-conferencias	64 - 128	128 - 512	112 - 512
Perfil de usuario: Ingenieros y diseñadores				
Baja	Diseño de páginas, dibujos con líneas de 2 a 4 colores, manipulación simple de imágenes, gestión de gráficos sencillos	-	96 - 128	80 - 112
Media	Diseño asistido por ordenador (CAD) 2D, presentaciones multimedia, fotoedición simple, desarrollo de páginas Web	-	128 - 512	112 - 512
Alta	Animaciones, edición fotográfica compleja, vídeo en tiempo real, CAD 3D, modelado de sólidos, análisis de elementos finitos	-	256 - 1024	240 - 1024

Cifras para Windows: Estimaciones del fabricante Kingston.

TABLA 10.22: Tamaño ideal en Mbytes de la memoria de un PC del año 2002 en función de la demanda de memoria que muestran las aplicaciones y las necesidades que presentan los sistemas operativos actuales.

Velocidad	128 Mbytes ⟵ Tamaño y fiabilidad ⟶ 256 Mbytes			
	Sin paridad/ECC	Con paridad/ECC	Sin paridad/ECC	Con paridad/ECC
800 MHz	53 €	57 €	112 €	130 €
1066 MHz	105 €	111 €	197 €	208 €
Incremento medio relativo del coste de un módulo:				
❖ Por incorporar paridad:				9 %
❖ Por duplicar tamaño:				103 %
❖ Por acelerar 266 MHz sobre 800 MHz:				83 %
❖ Por cada 1 % de incremento en velocidad:				2.5 %

TABLA 10.23: Influencia de la velocidad, el tamaño y la fiabilidad en el precio de los chips de memoria principal. La tabla indica el coste en euros para cada una de las configuraciones indicadas tomando como base la memoria RDRAM y el fabricante Kingston a fecha Septiembre de 2002 (apenas tres meses después de la aparición de los módulos RDRAM de 1066 MHz).

encima de nuestras necesidades reales, la clave para pagar sólo por la memoria que realmente necesitamos está en la capa software del equipo. Más concretamente, en el sistema operativo que vayamos a utilizar y en nuestro perfil como usuario del equipo. La tabla 10.22 especifica el tamaño ideal de una configuración de memoria para un PC del año 2001 en función de estas dos variables.

☛ pág. 100

En líneas generales, Windows (incluida la version para trabajo en grupo) puede trabajar a partir de 32 Mbytes con la mayoría de aplicaciones, y si éstas son un poco más exigentes con la memoria, ampliaríamos el rango a una horquilla situada entre los 64 y los 128 Mbytes.

En entornos domésticos, Windows 95 funcionaba con un mínimo de 8 Mbytes, pero se necesitaban al menos 16 Mbytes para poder trabajar cómodamente con aplicaciones nativas, experimentándose una mejora significativa en el rendimiento con tamaños de 32 Mbytes y superiores. Para Windows 98 las cantidades anteriores se duplicarían, manteniéndose para el Windows Millenium.

Windows 95

Windows 98
Millenium

En arquitecturas de tipo servidor, Windows NT funcionaba a partir de 16 Mbytes, mostrando una mejora de entre un 30 % y un 40 % para la ampliación a 32 Mbytes, y de hasta el 63 % para 64 Mbytes. La configuración básica se situaría aquí entre los 32 y los 48 Mbytes. Si se trata del Windows NT Server, nuestro punto de partida serían los 64 Mbytes, que se mantendrían para Windows 2000.

Windows NT

Windows NT
Server
Windows 2000

La presencia de determinados periféricos en el hardware de nuestro equipo resulta también bastante reveladora a la hora de determinar las necesidades de memoria de nuestro sistema. Dispositivos como CD-ROM, escáneres y aceleradores gráficos serán un fiel indicador de que nos movemos en un entorno software que hace un uso intensivo de la memoria.

periféricos

Por último, respecto al reparto del tamaño total entre un número mayor o menor de módulos, los datos de la tabla 10.23 muestran que el tamaño incrementa el coste de un módulo de forma lineal, así que pagaremos prácticamente lo mismo por llevarnos dos módulos de 128 Mbytes que uno solo de 256 Mbytes.

número
de módulos

☛ pág. 100

Especificaciones comerciales

◀ 15.3

15.3.1 Etiquetado

A la hora de identificar nuestra mejor opción de compra a partir del etiquetado de la memoria, todo depende de a dónde dirijamos nuestra mirada:

- En los chips encontraremos tiempos de respuesta o de ciclo en los casos más antiguos, o frecuencia en los más recientes (ver tabla 10.18). Para los tiempos, elegiremos el menor valor posible (en la etiqueta impresa en el lomo, es el único valor que viene precedido por un guión). Para la frecuencia, buscaremos un valor cuanto más grande mejor.

☛ pág. 89

- En los módulos, la referencia clave es la etiqueta PC-XXX, que refleja la frecuencia en los casos más antiguos (ver sección 10.14.2.1) y el ancho de banda en los más recientes (ver sección 10.14.2.2), optando siempre por un valor XXX mejor cuanto más grande. La figura 10.40 nos descubre el valor óptimo en función de la velocidad de nuestra placa base.

☛ pág. 91
☛ pág. 91
☛ pág. 102

15.3.2 Marca

En un ránking comercial que establecemos según nuestra propia experiencia, las diferentes marcas quedarían situadas como sigue (ver tabla 10.24):

CALIDAD:

☛ pág. 102

- ❶ Arriba del todo, Kingston como marca de calidad suprema, ya que pocos pueden ofertar garantía de por vida en sus productos (aunque eso sí, con un sobreprecio del 20 % respecto al siguiente escalón).

elevada

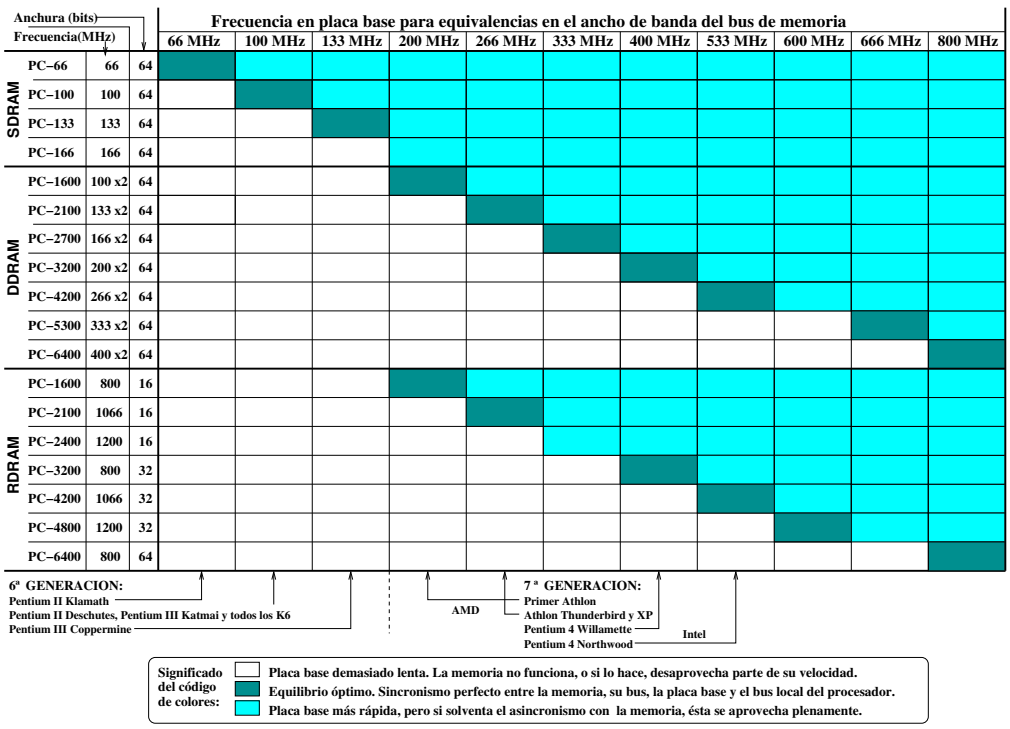


FIGURA 10.40: Información que aporta la etiqueta PC-XXXX de los módulos de memoria: Por un lado, el ancho de banda, la frecuencia de los módulos y su anchura como parámetros más representativos. Por otro otro, la relación con el bus de memoria en placa base y el modelo del procesador como ligaduras más significativas.

Calidad	Fabricante	Iniciales en el lomo de los chips
Elevada	Kingston	KVR, KTC, KFJ, KGW, KTH, KSG, ...
Buena	Micron	MT
Razonable	Fujitsu-Siemens	MB, HYB, F, FJ, FS
	Hitachi	HM
	Mitsubishi	M5M junto a su logo de trébol
	Motorola	MCM junto a su logo de M bajo círculo
	NEC	NEC, PD
	Samsung	KM, MR, SEC
	Texas Instruments	TI, TMS
	Toshiba	TC, TMM
Cuestionable	Firmas clónicas	Etiquetado deficiente

TABLA 10.24: Ránking de calidad en fabricantes de memoria principal sobre las diez firmas más importantes del mercado. La terna de iniciales de Kingston es muy amplia, puesto que las dos últimas letras indican el tipo y marca de computador al que van dirigidos: VR: Value RAM (para PC clónico en general), TC: Para PC y servidores de Compaq, FJ: Para Fujitsu, GW: Para Gateway, TH: Para Hewlett-Packard, SG: Para Silicon Graphics, ...

`buena`

❷ Bajamos un escalón para encontrar a Micron como firma comprometida también con la memoria de calidad, pero algo menos exigente. En Micron es más destacable su amplio abanico de productos.

❸ Descendiendo otro peldaño, situamos a marcas muy conocidas como Fujitsu-Siemens, Hitachi, Mitsubishi, Motorola, NEC Samsung, Texas Instruments y Toshiba, pero que no son especialistas en este sector. `razonable`

❹ El estrato más inferior, ya a cierta distancia, queda conformado por toda la legión de clónicos muy populares por premiar el bajo coste frente a la calidad del producto. `cuestionable`

Por norma general, el etiquetado suele ser más claro y completo cuanto mejor es el fabricante, y ciertamente obtuso en las marcas clónicas. Esto se refleja en la figura 10.37, donde seleccionamos premeditadamente un etiquetado de gama alta (Kingston), otro de gama media (Samsung), y un tercero de gama baja (Transcend).

☛ *pág. 88*

15.3.3 Fecha

No es que la memoria tenga fecha de caducidad, pero conocer que ha sido fabricada recientemente es la mejor manera de asegurar que el chip ha dado pocas vueltas en su larga cadena de distribución y/o no ha sido utilizado por ningún intermediario entretanto. Lo ideal en este sentido es que no sobrepase los dos años de edad. `edad`

Aunque la fecha no siempre suele venir impresa en la serigrafía de los chips, firmas como Fujitsu, Micron, Motorola, NEC y Toshiba sí suelen hacerlo. Por ejemplo, un código 9904 embebido dentro de la larga sucesión de letras y números que conforma un etiquetado indica que la memoria fue fabricada en Abril de 1999. Un ejemplo similar podemos verlo en el etiquetado de Samsung de la figura 10.37, supuestamente perteneciente a un módulo de Agosto de 2002. `código`

☛ *pág. 88*

Resumen

La memoria principal o dinámica (DRAM) es el área donde se alojan los programas mientras se ejecutan en el PC. Sus prestaciones ofrecen una doble vertiente: La funcionalidad, que va ligada al tamaño, y el rendimiento, medido en latencia (retardo en ns. en forma de tiempo de respuesta o tiempo de ciclo) y ancho de banda (transporte de los datos en Mbytes/sg.). `funcionalidad y rendimiento`

El sistema de memoria se estructura en bancos responsables de su tamaño flexible, éstos en módulos de formato y anchura concretos que se enganchan a su zócalo en placa base, y éstos a su vez en chips organizados en matrices bidimensionales de celdas. `estructura`

La memoria principal no ha sufrido grandes cambios al nivel de su celda básica. Sus mejoras, que resumimos en la tabla 10.25, han venido más por la vertiente que han articulado conjuntamente el interfaz de diálogo y la organización interna: `mejoras`

☛ *pág. 104*

❶ Primero, proporcionando una salida de cuatro datos consecutivos o ráfaga para llenar una línea de caché (FPM), optimización que nunca fue consecuencia del entrelazado, sino que guarda relación con la estructura bidimensional de celdas donde la coordenada de columna aprovecha la preselección de la fila. `ráfaga`

❷ Después, segmentando y automatizando este proceso (EDO y BEDO). `segmentación`

❸ Finalmente, desarrollando diseños síncronos (SDRAM y DDRAM) que permitieron adoptar otras estrategias ya usuales en el procesador, como la segmentación, la precarga y el empleo de multiplicadores de reloj. `sincronismo`

Interfaz de memoria	Mejoras aplicadas sobre la arquitectura	Temporización de salida	Rango de frecuencias
FPM	Entrelazado para línea CAS	5-5-5-5 a 5-3-3-3	Hasta 66 MHz
EDO	Latch desacoplando dirección y datos	5-2-2-2	Hasta 66 MHz
BEDO	Latch sustituido por contador interno.	5-1-1-1	Hasta 66 MHz
SDRAM	Segmentación, precarga y entrelazado internos	(2+)3-1-1-1	Entre 66 MHz y 166 MHz
DDRAM	Desdoble sobre la señal de reloj y las matrices de celdas	(2+)2.5-0.5-0.5-0.5	Entre 100x2 MHz y 300x2 MHz
RDRAM	Nueva arquitectura de bus estrecho y gran ancho de banda	(9+)9-0.5-0.5-0.5	Entre 400x2 y 600x2 MHz

TABLA 10.25: Resumen de la temporización en la ráfaga de salida de datos para los distintos interfaces de memoria principal. Entre paréntesis, los ciclos que ahorra el empleo conjunto de precarga y entrelazado. Los efectos benefactores de la segmentación revierten sobre la frecuencia.

Clasificación de los aspectos a considerar			Recomendación
Rasgos externos	1	Contactos del módulo: Cantidad, color y material	Gran número, color dorado y aleaciones del oro
	2	Número de chips del módulo	Reducido, para mejorar sincronización y velocidad
	3	Zócalos: Número y uso	Emplear pocos y no mezclar velocidades ni voltajes
	4	Gestión de errores	Paridad/ECC para mayor interoperabilidad y fiabilidad, pero 10 % más caro.
Parámetros internos	5	Controlador de memoria	Que acepte el interfaz de los módulos y su refresco
	6	Velocidad	Frecuencia a la par con placa base. Latencias RAS y CAS de 2 ciclos, evitar 3.
	7	Tamaño	Fijarse en los zócalos, el controlador de la placa base y el microprocesador
Especificaciones comerciales	8	Etiquetado	PC-XXX para frecuencia y PC-XXXX para ancho de banda. Ambos lo mayor posible.
	9	Marca	Kingston, Micron, marcas globales y clónicas, por ese orden.
	10	Fecha	Lo más reciente posible. Inferior a dos años de antigüedad.

TABLA 10.26: Resumen de las diez recomendaciones para elegir la memoria principal del PC.

A finales de los años 90 surgió una nueva corriente en el diseño de la memoria principal: La RDRAM, que apostaba por la simplicidad de un bus estrecho con objeto de lograr frecuencias muy elevadas. Pero su precio ha supuesto una losa para su despegue, y aunque finalmente ha conseguido cierta cuota de mercado, no ha cumplido con las expectativas depositadas.

RDRAM

Como epílogo, la tabla 10.26 resume las diez recomendaciones más importantes a la hora de seleccionar los módulos de memoria principal de nuestro PC.

La anécdota: ¿Quién se ha llevado mi byte?

A nuestro paso por el estudio del procesador, dilapidamos al que tradicionalmente ha sido su magnitud más representativa en el mercado: El MHz. La historia vuelve a repetirse en el caso de la memoria y su byte, en una lección tremendamente didáctica que hemos querido reservar para el final.

La visión lógica que todo el mundo tiene de su memoria es aquella que se le ofrece con sus, pongamos, 512 Mbytes. Parece sugerírsele así que dispone de 512 Mpalabras de un byte de anchura, y como lo del byte es algo que se da por seguro, se supone implícito y se abrevia diciendo que se tienen *512 Megas*.

Pero seguro que a más de uno no le ha pasado por alto que el byte pinta bastante poco en la arquitectura de memoria principal de un PC. Es más, diríamos que a la hora de entender la composición física de la memoria, el byte estorba más que otra cosa. No tenemos mas que repasar contenidos:

- Los módulos de memoria tienen 32 ó 64 bits de anchura, y esta anchura no hará otra cosa que aumentar en el futuro, separándose cada vez más de los ocho bits representados por el byte.

- Los chips de los módulos tienen una anchura de 2, 4, 8, 16 o 32 bits, siendo ocho el valor menos utilizado de la serie.

- Las matrices de celdas de los chips tienen un número de filas y columnas de varios Kbits, y como la tendencia es a crecer, sus dimensiones estarán también cada vez más lejos de la órbita del ocho. Otro tanto le ocurre a los amplificadores de señal y a los búfers a la salida de las celdas.

¿Dónde está el byte? En ningún sitio, y no aparecerá por mucho que lo busquemos. Como muestra, un botón, o mejor, la camisa entera: Obsérvese la disección realizada en el ejemplo 10.4 para descomponer dos de los módulos de memoria SIMM y DIMM más populares del mercado (el último de ellos ilustrado también en la figura 10.16). Su anchura es de 32 y 64 bits, la de sus chips es de 4 y 32 bits, y la de sus filas y columnas es de 2048, 4096 y 8192 bits. Todos los números son potencia de dos, y el ocho, aún siendolo también, no se utiliza en ningún caso. Otro tanto le ocurre a los módulos RIMM en el caso del ejemplo 10.8, a la triple descomposición de un módulo en sus chips constituyentes que apuntamos en el ejemplo 10.10, y a las dos formas de entrelazado más usuales para módulos DIMM y RIMM que describimos en el ejemplo 10.7.

☛ pág. 47

☛ pág. 47

☛ pág. 83

☛ pág. 95

☛ pág. 82

Mi mejor consejo para no fracasar en el intento de entender la memoria del PC es que vea el byte como un múltiplo del bit, de igual forma que el kilogramo lo es del gramo, que realice la conversión a bits de todos los valores que tengan al byte como referente, y que a partir de ahí se olvide de los bytes y maneje todos los valores en bits. Si necesita múltiplos, acuda al Kbit como kilogramo o al Mbit como tonelada, recordando que estas escalas son binarias y multiplican por 1024 en lugar de por 1000.