UT 3. Subsistema de memoria

Tema 3.3 Mejora de prestaciones de la memoria principal

A. Doménech, J. Duato, P. López, V. Lorente,
A. Pérez, S. Petit, J.C. Ruiz, S. Sáez, J. Sahuquillo

Departamento de Informática de Sistemas y Computadores
Universitat Politècnica de València

Índice

Bibliografía

John L. Hennessy and David A. Patterson.
Computer Architecture, Fifth Edition: A Quantitative Approach.
Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 5 edition, 2012.

Índice

Conceptos

La memoria principal atiende las peticiones de la cache y del subsistema de Entrada/Salida.

Objetivo

Desde el punto de vista de la cache el objetivo es reducir la penalización por fallo (PF, *miss penalty*).

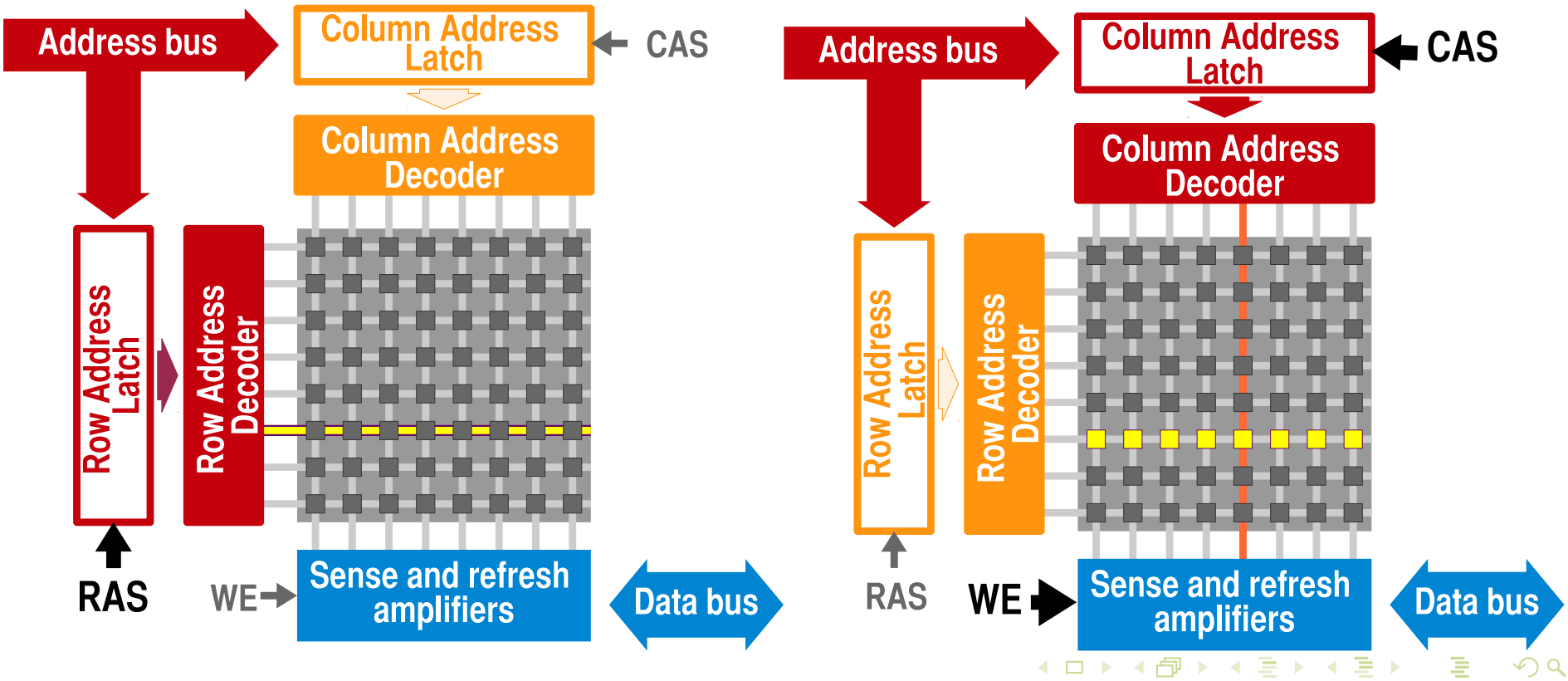
Medida de prestaciones

Si se accede a sólo un dato: $\rightarrow PF =$ Tiempo de acceso a la memoria
Si se accede a múltiples datos (p. ej, un bloque de cache de tamaño B palabras): $\rightarrow PF = L + \frac{1}{B_w} B$

- L, Latencia: Tiempo para satisfacer el primer acceso.
- B_W, Ancho de banda: Número de palabras transferidas por unidad de tiempo.

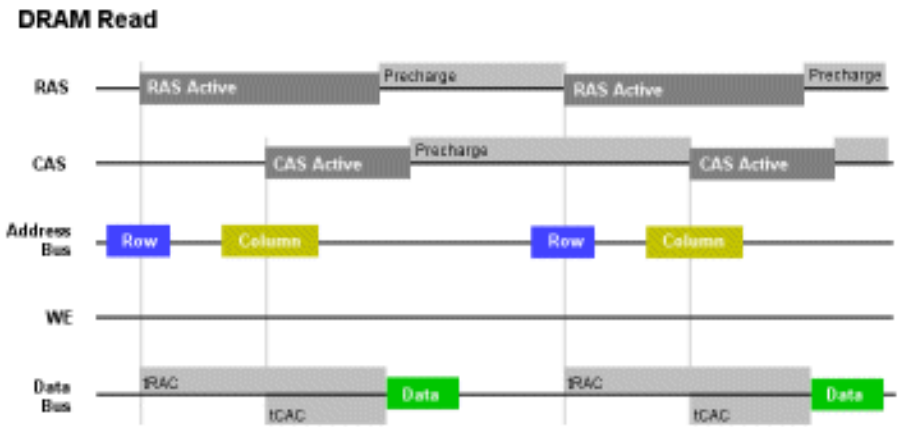
Evolución de la tecnología de DRAM

- DRAM tradicional. Debido a las restricciones en el número de pines, primero se transmite la dirección de fila (validada con la señal RAS), y luego la de columna (validada mediante la señal CAS).



Evolución de la tecnología de DRAM (cont.)

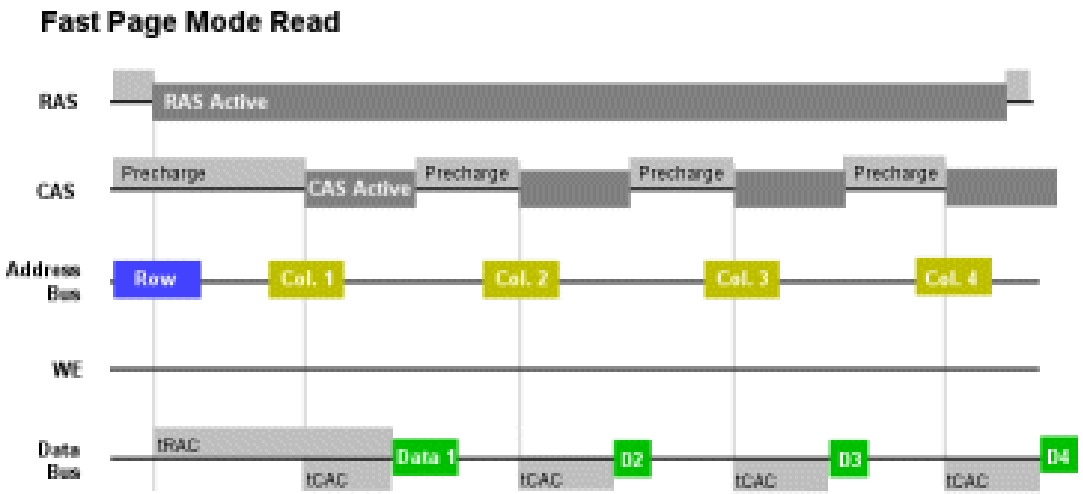
Temporización:



- En cada acceso a la memoria lee o escribe una palabra pero, internamente, se lee y refresca una fila entera.
- Después de acceder a una palabra, el siguiente acceso no puede comenzar mientras no se complete el ciclo de memoria $\rightarrow$ la precarga refresca y cierra una fila.

Evolución de la tecnología de DRAM (cont.)

- *Fast page mode*. Si se añade un buffer de fila, el acceso a otras palabras de la misma fila será más rápido.
- Cuando la fila está disponible, se puede leer (o escribir) una secuencia de varias direcciones de columna.
- Cada acceso sólo requiere una dirección de columna.



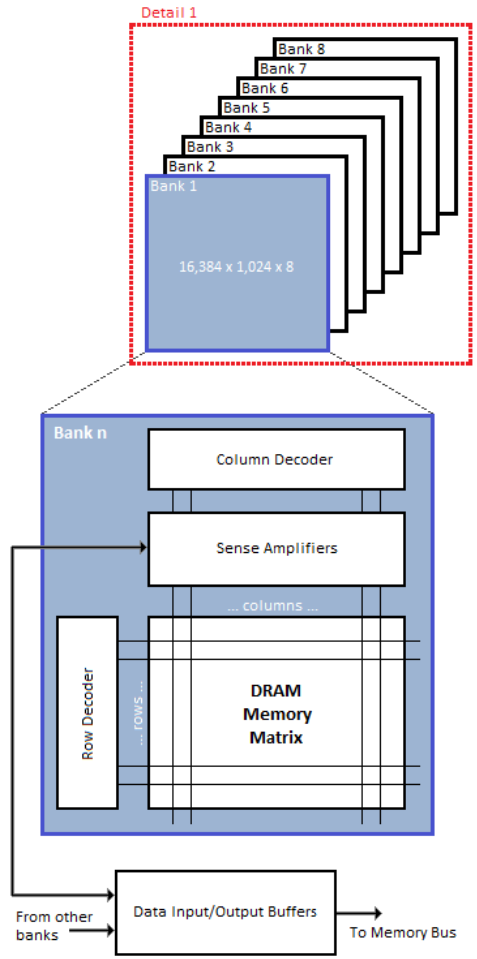
Tecnología DRAM actual

Características de la DRAM síncrona (SDRAM):

- Son síncronas:
  - La señal de reloj se envía a la memoria.
  - La frecuencia de reloj está definida por el controlador de memoria.
  - Los tiempos se miden en ciclos. El número de ciclos de reloj necesarios en cada operación (enviar la dirección, acceder y transferir datos) se leen de una ROM del módulo SDRAM para configurar el controlador de memoria.

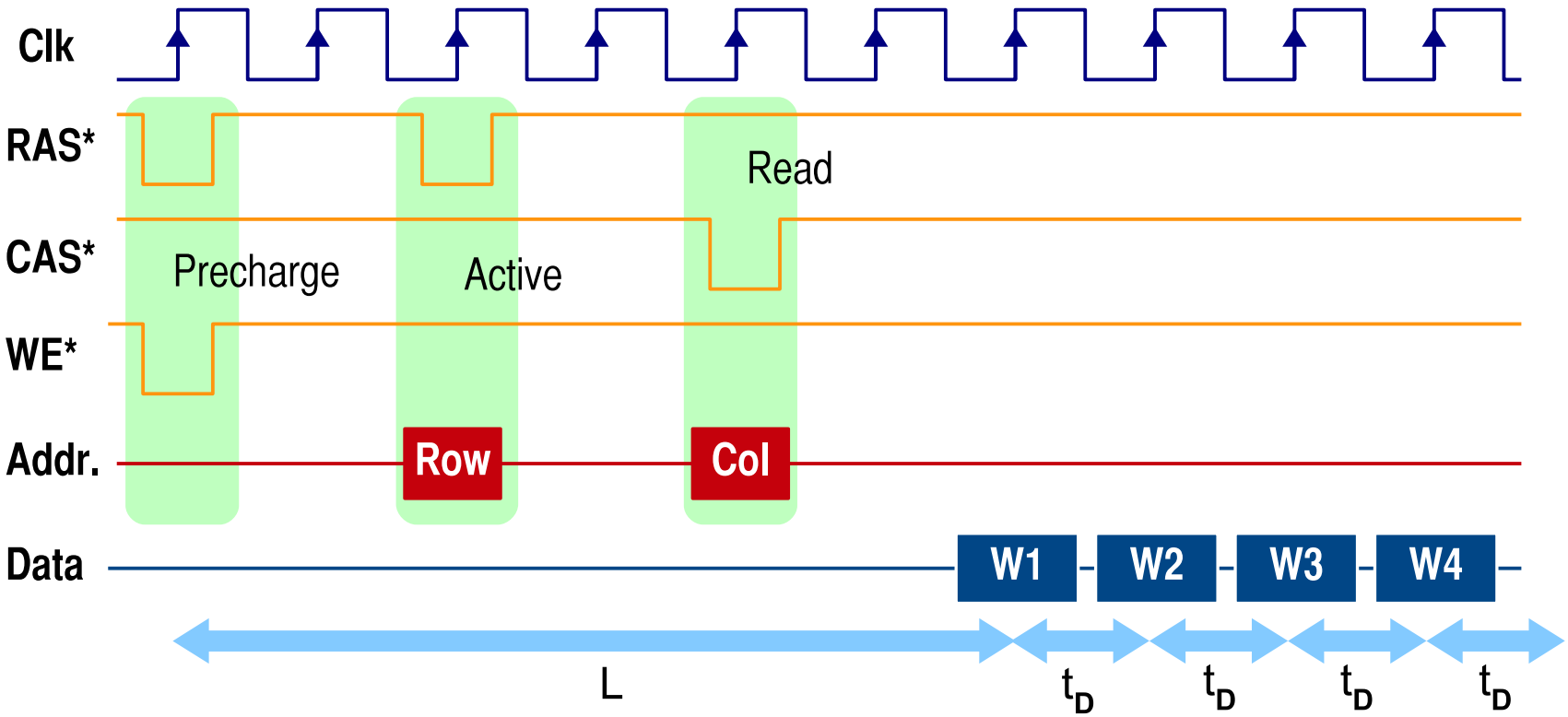
Tecnología DRAM actual (cont.)

- Cada chip SDRAM se organiza internamente como uno o más bancos.
  - Cada banco es una matriz de memoria.
  - Una vez se activa una fila en un banco, se puede leer o escribir cualquier columna.
- Modo ráfaga: La SDRAM usa un contador autoincrementado y un registro de modo para fijar una secuencia de direcciones de columna siguiendo al primer acceso a la fila.



Tecnología DRAM actual (cont).

Cronograma de lectura de una SDRAM



Tecnología DRAM actual (cont).

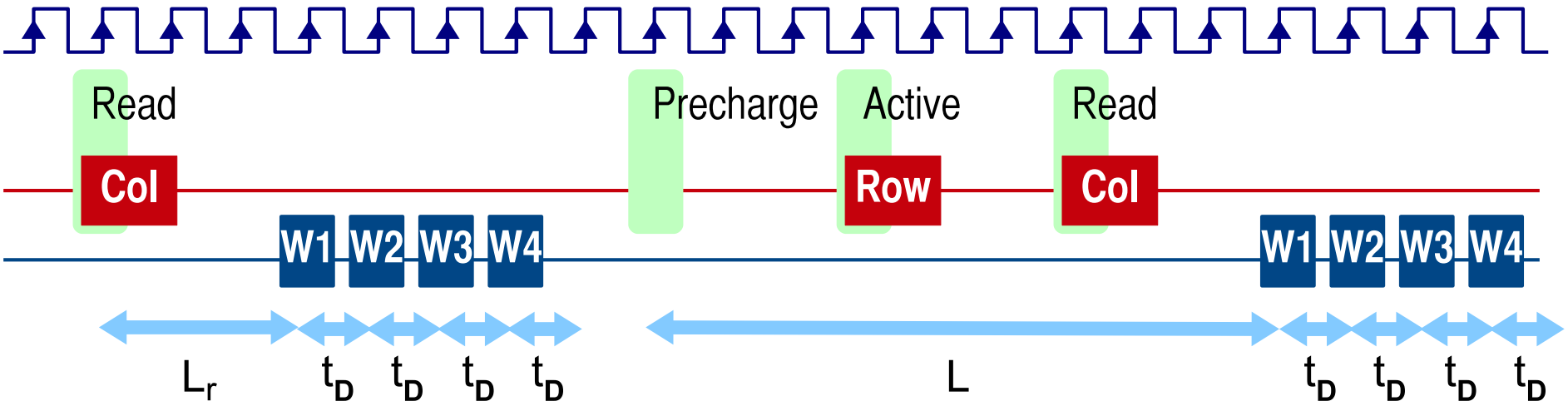
Comentarios:

- Los accesos a bloque de cache o de disco se hacen en modo ráfaga.
- Aumentar la frecuencia de reloj incrementa la velocidad de transferencia pero no reduce el tiempo de acceso L a la primera palabra de la ráfaga.
- El tiempo de acceso L depende de los parámetros de temporización de la memoria y se especifica con un número entero de ciclos de reloj (el menor número de ciclos que permiten completar la operación). Debido al redondeo, una mayor frecuencia de reloj puede alargar el tiempo de acceso.

Tecnología DRAM actual (cont).

- Pueden estar abiertas tantas filas como bancos haya en el chip.
- El tiempo de acceso a bloques consecutivos depende de si la fila está abierta o no. Esto produce variaciones en la penalización por fallo de cache.

La fila está ya abierta: Hay que abrir la fila:



Modelo simple de la memoria

Parámetros genéricos de la memoria:

- L: Latencia o tiempo de acceso (tiempo para leer la primera palabra).
- t_D: Tiempo para transferir una palabra.
- B_w: Ancho de banda del bus, medido en palabras/s.

Consideremos un tamaño de bloque de B palabras de memoria. La penalización por fallo PF es:

$$PF \text{ (en segundos)} = L + t_D \cdot B = L + \frac{1}{B_w} \cdot B$$

Modelo simple de la memoria (cont.)

Expresado en ciclos de reloj:

- f: Frecuencia de reloj del bus.
- L_c: Latencia medida en ciclos a la frecuencia f. $L_c = L \cdot f$
- B_{wc}: Ancho de banda del bus en palabras/ciclo a la frecuencia f. $B_{wc} = B/f$

$$PF \text{ (en ciclos)} = L_c + \frac{1}{B_{wc}} \cdot B$$

$$PF \text{ (en segundos)} = PF \text{ (en ciclos)}/f$$

Modelo simple de la memoria (cont.)

Teniendo en cuenta que la fila abierta puede no ser la deseada:

- Sea *ML (memory locality)* la probabilidad de que un fallo de cache pida un bloque de memoria que pertenece a una fila abierta (activada).
- Si la fila está abierta, el tiempo de acceso se acorta ya que la fila correspondiente está almacenada en un buffer.
- Sea L_r la latencia o tiempo de acceso reducido Sea L_{rc} el valor de L_r medido en ciclos a la frecuencia f.

La penalización promedio por fallo de cache será:

$$PF \text{ (en segundos)} = L \cdot (1 - ML) + L_r \cdot ML + \frac{1}{B_w} \cdot B = PF \text{ (en ciclos)}/f$$

$$PF \text{ (en ciclos)} = L_c \cdot (1 - ML) + L_{rc} \cdot ML + \frac{1}{B_{wc}} \cdot B$$

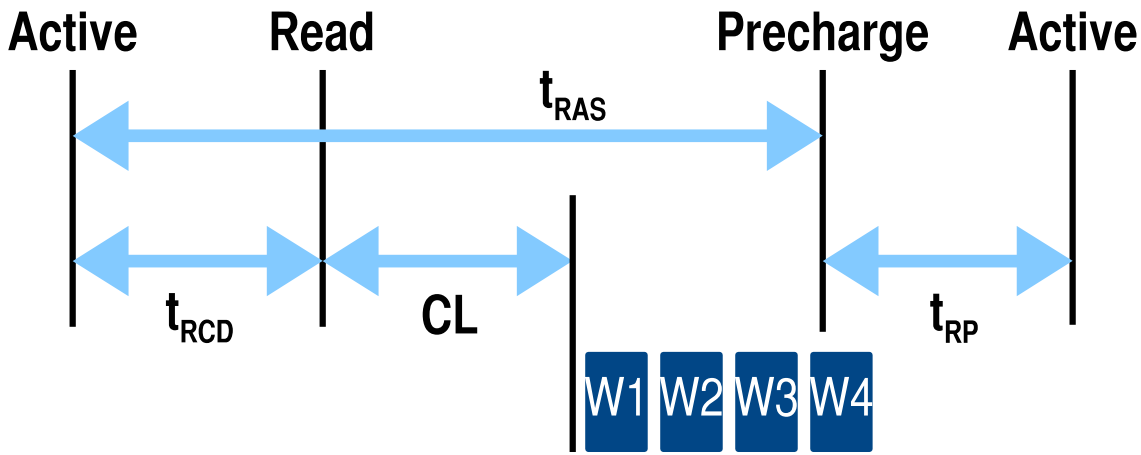
Cálculo de L y L_r

La temporización de las SDRAM se define por la frecuencia de reloj y cuatro parámetros de temporización separados por guiones. Por orden de aparición, son:

- CL: El mínimo tiempo (en ciclos) entre el envío de la dirección de columna a la memoria y el comienzo de la ráfaga.
  Es el tiempo necesario para leer el primer bit de memoria desde una DRAM en una fila ya abierta.
- t_{RCD}: El mínimo número de ciclos entre la apertura de una fila de memoria y el acceso a una columna.
  El tiempo para leer el primer bit de memoria de una DRAM sin una fila activa es $t_{RCD} + CL$.

Cálculo de L y L_r (cont.)

- t_{RP}: El mínimo número de ciclos entre la orden de precarga y la apertura de la fila siguiente.
  El tiempo para leer el primer bit de memoria de una DRAM con la fila equivocada abierta es $t_{RP} + t_{RCD} + CL$.
- t_{RAS}: El mínimo número de ciclos entre la activación de un banco y la orden de precarga. Este es el tiempo necesario para refrescar internamente la fila, y se solapa con t_{RCD}. Este cuarto parámetro no siempre aparece en la especificación de la temporización.

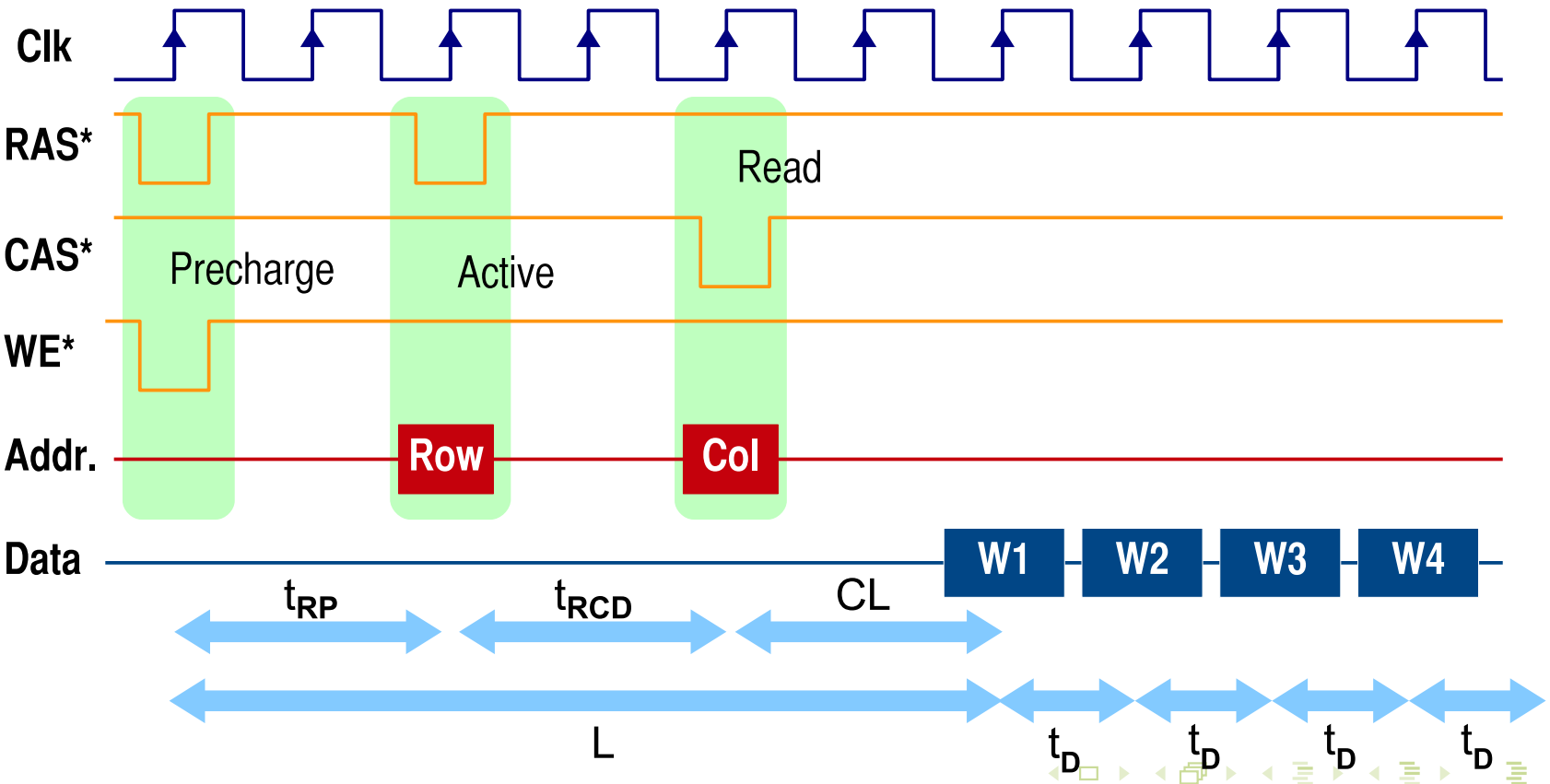


Cálculo de L y L_r (cont.)

Cálculo de L y L_r a partir de los parámetros de temporización:

$$L_c = t_{RP} + t_{RCD} + CL \text{ ciclos}$$

$$L_{rc} = CL \text{ ciclos}$$



Ejemplo

Sea una memoria HyperX KHX1600C9D3/4G, SDRAM DDR3-1600 (reloj de bus a 800 MHz, y transfiere dos datos por ciclo) de 512M x 64-bit (4GB). Puede funcionar con una latencia 9-9-9-27 a 1,65V.

Calcular la *PF* para un tamaño de bloque $B = 8$, suponiendo $ML = 0$.

$$L_c = t_{RP} + t_{RCD} + CL = 9 + 9 + 9 = 27 \text{ ciclos}$$

Al ser DDR (veáse la página 27), el ancho de banda es $B_{wc} = 2$ pal/ciclo. Por tanto, para $ML = 0$:

$$PF \text{ (en ciclos)} = L_c + \frac{1}{B_{wc}} \cdot B = 27 + 8/2 = 31 \text{ ciclos}$$

$$PF \text{ (en segundos)} = PF \text{ (en ciclos)}/f = 31/800 \text{ MHz} = 38{,}75 \text{ ns}$$

Ejemplo (cont.)

A título de comparación, si fuera SDR, ($B_{wc} = 1$), tendríamos que:

$$PF \text{ (en ciclos)} = L_c + \frac{1}{B_{wc}} \cdot B = 27 + 8/1 = 35 \text{ ciclos}$$

$$PF \text{ (en segundos)} = PF \text{ (en ciclos)}/f = 35/800 \text{ MHz} = 43{,}75 \text{ ns}$$

Índice

Técnicas de mejora de prestaciones de la SDRAM

$$PF = L \cdot (1 - ML) + L_r \cdot ML + \frac{1}{B_w} \cdot B$$

Se puede reducir PF reduciendo L y B e incrementando B_w y ML. Sin embargo:

- L supone (con mucho) la mayor contribución a PF.
- L se mantiene aproximadamente constante al incrementar $f \rightarrow$ salvo efectos del redondeo, L_c se incrementa linealmente con f.
- ML depende, sobre todo, de las pautas de acceso a la memoria, pero también del número de bancos.

Técnicas de mejora de prestaciones de la SDRAM (cont.)

Técnicas para mejorar las prestaciones de la memoria principal:

- Incrementar la anchura del bus: $B \downarrow$
- Incrementar B_{wc} transfiriendo datos en ambos flancos (ascendente y descendente) de la señal de reloj, manteniendo la f constante: $B_w \uparrow$
- Incrementar la frecuencia f de reloj manteniendo B_{wc} constante: $B_w \uparrow$
- Aumentar el número de bancos de memoria: $ML \uparrow$
- Implementar varios controladores de memoria. Aunque esto no reduce la penalización por fallo (si las direcciones no están entrelazadas), permite atender concurrentemente varios fallos de cache. También aumenta el número total de bancos de memoria: $ML \uparrow$

Aumento del ancho del bus de memoria

Al ensanchar el bus de memoria, se puede transferir más de una palabra a la vez $\rightarrow$ El número de transferencias se reduce.

Ejemplo: PF con la DDR3-1600 con un bus el doble de ancho

Hacen falta la mitad de las transferencias. Por lo tanto, el tamaño de bloque medido en la nueva palabra de memoria es $B' = \frac{B}{2} = \frac{8}{2} = 4$.

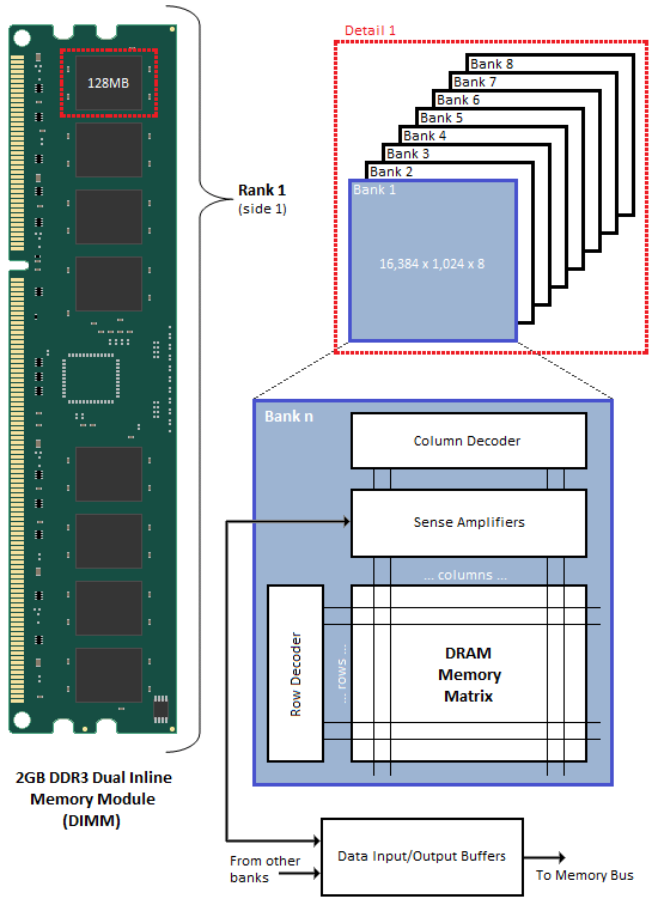
$$PF \text{ (en ciclos)} = L_c + \frac{1}{B_{wc}} \cdot B' = 27 + 4/2 = 29 \text{ ciclos}$$

Limitaciones

Dado que L_c es (con diferencia) la mayor contribución a PF, se ahorra poco ensanchando el bus (la PF con el ancho de bus inicial era de 31 ciclos, página 20).

Aumento del ancho del bus de memoria (cont.)

- Especialmente por razones de fabricación (número de pines), la anchura de los buses de memoria actuales es de 64 bits (8 Bytes).
- Las memorias se organizan en módulos *Dual Inline Memory Module* (DIMMs) con uno o más *ranks*.
- Un rank está compuesto por los chips necesarios para completar 64 bits.

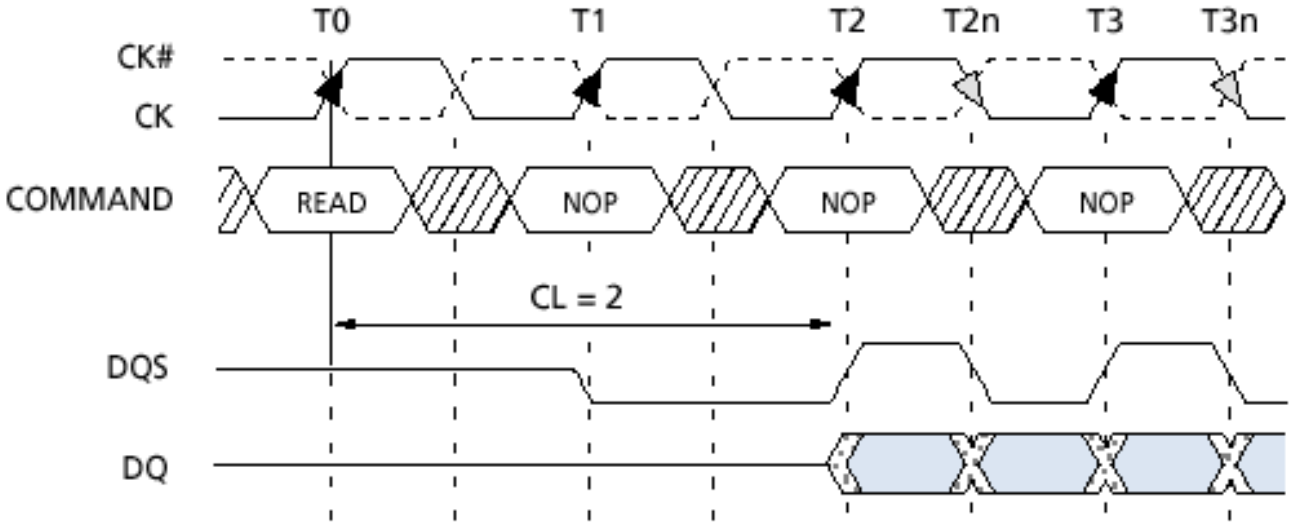


DDR: Double Data Rate

- Idea simple: Transmitir datos en ambos flancos (ascendente y descendente) de la señal de reloj.
- El bus trabaja a la misma velocidad, pero el ancho de banda se duplica.
- La frecuencia máxima de las señales no cambia respecto a la convencional, SDR *(Single Data Rate)*, así que se puede implementar sin tener que mejorar la tecnología.
- Internamente, se duplica el número de columnas accedidas (*2n-prefetch*) y se duplica el número de líneas que conectan los bancos de memoria con el bus de datos $\rightarrow$ la frecuencia interna de la memoria no cambia.

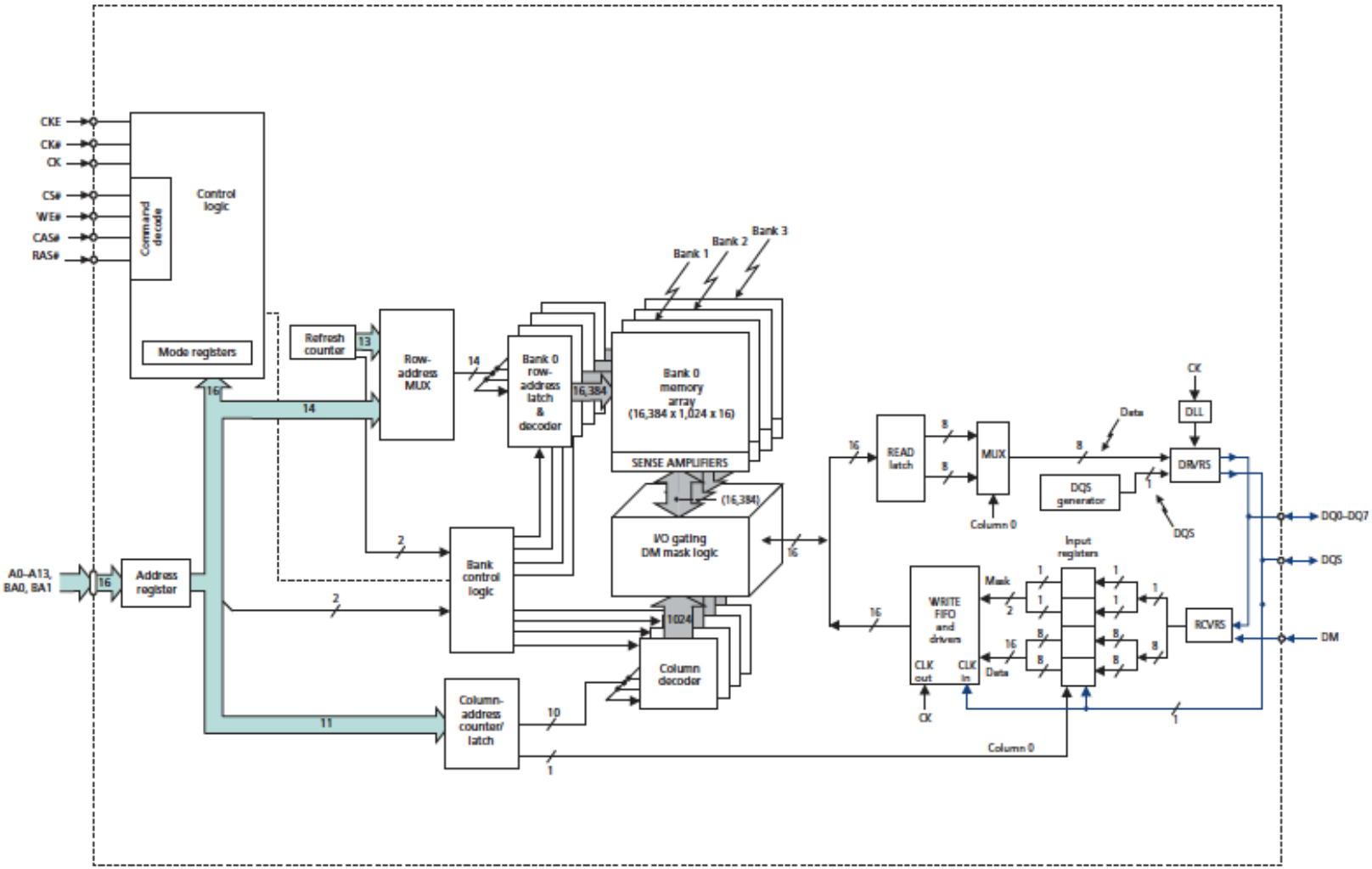
$\rightarrow$ Se duplica B_{wc}.

DDR: Double Data Rate (cont.)



DDR 1Gbit

Figure 4: 128 Meg x 8 Functional Block Diagram



Incremento de la frecuencia de reloj del bus

- La frecuencia de reloj del bus se ha incrementado a lo largo del tiempo y también ha bajado el voltaje para reducir su consumo de potencia.
- Se han desarrollado varias técnicas para mantener la integridad de las señales a frecuencias de reloj más altas (transmisión diferencial, uso de terminadores, cambio de los buses por enlaces punto a punto . . .).
- JEDEC ha estandarizado las frecuencias de reloj y los voltajes. Los valores estándar son:
  - DDR.
    - 2.5V para frecuencia de bus hasta 166 MHz y 2.6V para 200 MHz.
    - Velocidades pico de transferencia de hasta 3200 MB/s
    - Hasta 1 Gb por chip.

Incremento de la frecuencia de reloj del bus (cont.)

- DDR2. Se duplica el número de columnas accedidas respecto a DDR (*4n-prefetch*), modificando también el número de líneas que conectan los bancos de memoria con el bus de datos.
  - Reduce el voltaje a 1.8V
  - Incrementa la frecuencia hasta 533 MHz.
  - Velocidades pico de transferencia de hasta 8533 MB/s.
  - Hasta 4 Gb por chip.
- DDR3. Se duplica el número de columnas accedidas respecto a DDR2 (*8n-prefetch*), modificando también el número de líneas que conectan los bancos de memoria con el bus de datos.
  - Reduce el voltaje a 1.5V
  - Incrementa la frecuencia hasta 1066 MHz.
  - Velocidades pico de transferencia de hasta 17066 MB/s.
  - Hasta 16 Gb por chip.

Incremento de la frecuencia de reloj del bus (cont.)

- DDR4. Mantiene *8n-prefetch*. Agrupa los bancos en grupos, que también pueden direccionarse. Cada grupo puede accederse independientemente de los demás (y del estado de los mismos). Se reemplazan los buses de memoria por canales con enlaces punto a punto.
  - Reduce el voltaje a 1.2V
  - Incrementa la frecuencia hasta 1200 MHz.
  - Velocidades pico de transferencia de hasta 19200 MB/s.
  - Hasta 16 Gb por chip (hasta el momento).

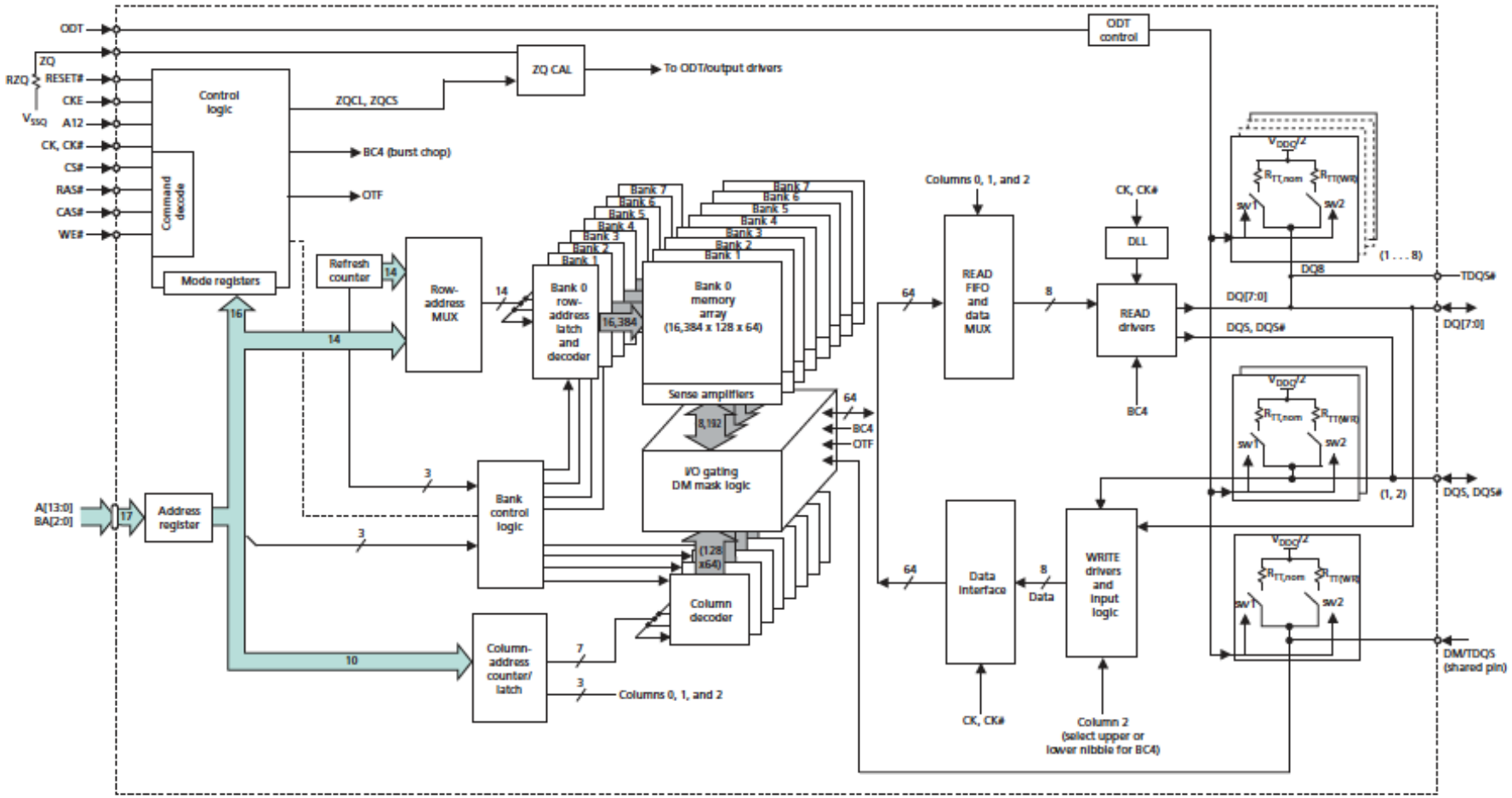
Incremento de la frecuencia de reloj del bus (cont.)

Las notaciones para especificar prestaciones son:

- DDRn-*xxxx*, donde *xxxx* indica la velocidad de transferencia en Mtransferencias/s.
  - Frecuencia de reloj del bus: $f = xxxx/2$ MHz.
  - Ejemplos:
    - DDR-400 funciona a 200MHz y suministra 400 Mtransferencias/s = 3200 MB/s.
    - DDR3-1600 trabaja a 800MHz y suministra 1600 Mtransferencias/s = 12800 MB/s.
- PCn-*yyyy*, donde *yyyy* es el ancho de banda del bus en MB/s.
  - Frecuencia de reloj del bus: $f = yyyy/(8 \times 2)$ MHz.
  - Ejemplos:
    - PC-3200 suministra 3200 MB/s (200 MHz x 8 bytes x 2 (DDR)).
    - PC3-12800 suministra 12800 MB/s (800 MHz x 8 bytes x 2 (DDR)).

DDR3 1Gbit:

Figure 4: 128 Meg x 8 Functional Block Diagram



Incremento de la frecuencia de reloj del bus

Ejemplo: Latencias "más rápidas" no opcionales del estándar JEDEC

Nombre estándar	Bus clock (MHz)	Temporización CL-t_{RCD}-t_{RP} (ciclos)	L_c (ciclos)	L (ns)	PF (ns) ($B = 8$)
DDR-400A	200	2.5-3-3	8.5	42,5	62,5
DDR2-800C	400	4-4-4	12	30	40
DDR2-1066E	533,33	6-6-6	18	33,75	41,25
DDR3-800D	400	5-5-5	15	37,50	47,50
DDR3-1066E	533,33	6-6-6	18	33,75	41,25
DDR3-1600H	800	9-9-9	27	33,75	38,75
DDR3-2133L	1066,67	12-12-12	36	33,75	37,50
DDR4-1600K	800	11-11-11	33	41,25	46,25
DDR4-2133P	1066,67	15-15-15	45	42,19	45,94
DDR4-2400R	1200	16-16-16	48	40	43,33

Incremento de la frecuencia de reloj del bus (cont.)

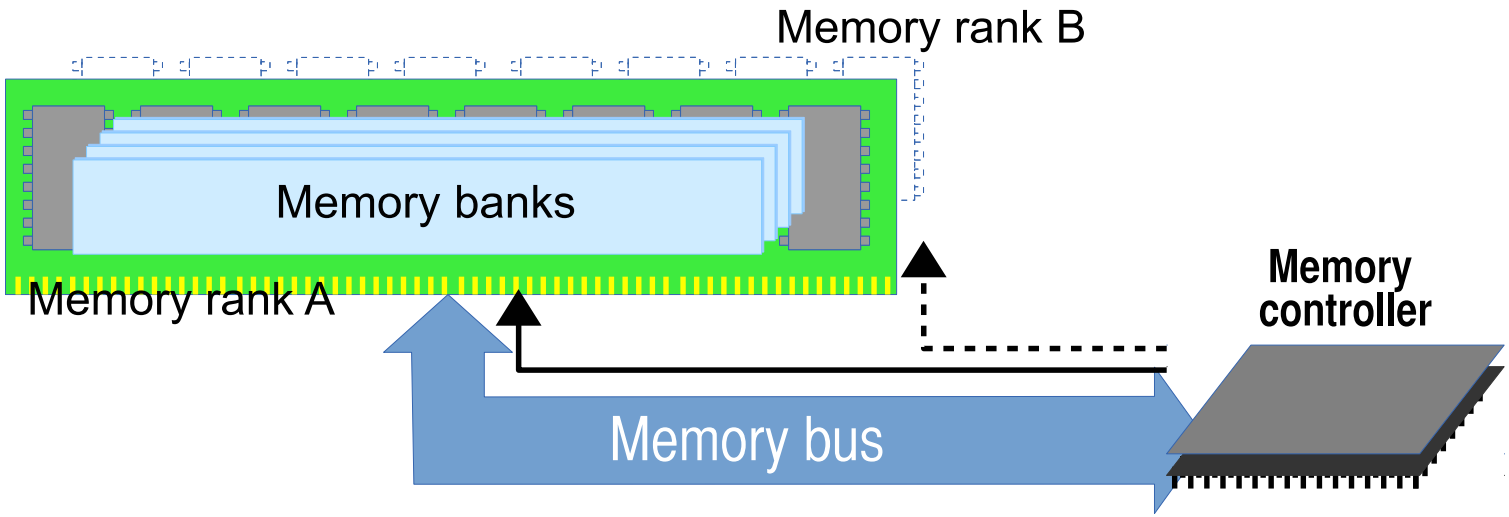
- La latencia (L) se redujo de DDR a DDR2, pero creció ligeramente de DDR2 a DDR3 y de DDR3 a DDR4.
- De hecho, a partir de DDR2, la latencia L y la penalización por fallo (PF) casi siempre aumentan con cada nueva generación (por ejemplo DDR2-800C vs. DDR3-800D; DDR3-2133L vs. DDR4-2400R).
- Pero aumenta la capacidad de las memorias, el soporte para más núcleos y reducen el consumo.

Incremento del número de bancos de memoria

Los chips SDRAM actuales implementan gran número de bancos (típicamente ocho o dieciseis). Las razones son:

- Pueden estar abiertas varias filas (una por banco), incrementando *ML*.
  - Para acceder a una fila abierta distinta, la dirección del banco se suministra junto con la dirección de columna.
  - Acceder a una fila abierta diferente es igual de rápido que acceder a la misma fila.
- Para una capacidad de memoria dada, incrementar el número de bancos reduce el tamaño de banco, reduciendo la latencia.
- Bancos más pequeños implica decodificación más rápida.
- El diseño de un banco se puede replicar, simplificando el diseño del chip de SDRAM.

Incremento del número de bancos de memoria (cont.)



Incremento del número de controladores de memoria

- Los actuales procesadores de altas prestaciones implementan múltiples controladores de memoria.
- Cada controlador de memoria implementa uno o dos canales para acceder a uno o más ranks de DIMMs SDRAM.
- Cada DIMM implementa múltiples bancos internos, como se menciona en la diapositiva anterior.
- El número total de filas que pueden estar abiertas simultáneamente es el número de controladores de memoria por el número de canales por controlador por el número de ranks por canal por el número de bancos por rank.
- Es necesario disponer de un número mayor de filas abiertas para minimizar conflictos cuando múltiples núcleos inician accesos concurrentes a la memoria. De esta manera, muchos accesos a la memoria afectarán a una fila abierta, maximizando *ML*.

Incremento del número de controladores de memoria (cont.)