ARQUITECTURA E INGENIERÍA DE COMPUTADORES

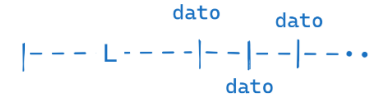
Tema 3.3

Miquel Gómez Corral Ing. Informática ETSINF, UPV

Tema 3.3

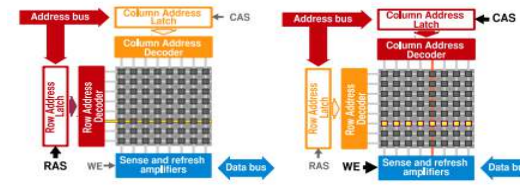
Mejorar prestaciones memoria principal

Cuando la caché accede a memoria principal la PF (*Penalización por fallo*) es $PF = L + \frac{B}{Bw}$. Un tiempo de latencia + el tiempo en enviar los datos:



MATRICES DE CERDAS Y ECUACIÓN PF

La memoria no es un array recto, son un conjunto de matriz cuadradas, por lo que la dirección de memoria es qué matriz, que fila y que columna.



Para acceder se activan ciertas señales: RAS para la fila, se lee la fila entera que contiene el bloque, se almacena en el buffer, y luego se transmite la columna (señal CAS) y luego todo al controlador de memoria…

Tipos de memoria

Síncronas (*el reloj indica cuando se hacen las cosas*), **SDR SDRAM**: una transferencia por ciclo. **DDR SDRAM**: **Dos transferencias por ciclo**. **SDR 8 BYTES POR CICLO, DDR 16 BYTES POR CICLO DE BW**.

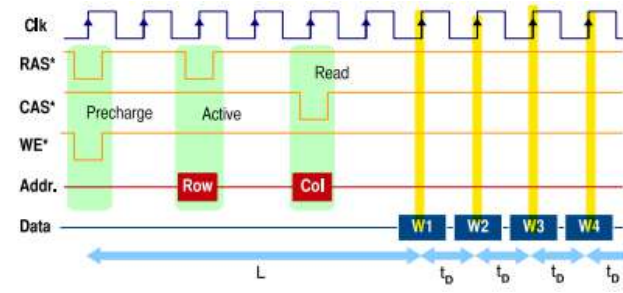
Bw: Para saber cuál es el ancho de banda (el Bw) es: $Transferencia\ por\ bloque = \frac{Tamaño\ bloque}{Ancho\ bus}$

- Para transferir un bloque de 64B por un bus de 64 bits: 4 ciclos si es DDR, 8 si es SDR

Precarga: Si el bloque requerido no está en el **row buffer**, debe cerrarse la fila (lectura destructiva) y abrir la nueva.

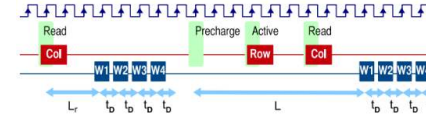
Órdenes del controlador

1. **PRECHARGE**: cerrar fila
2. **ACTIVATE**: abrir fila
3. **READ**: leer columna desde el row buffer
4. **WRITE**: Cuando le llegan los datos al procesador.



Ejemplo: Cronograma de lectura memoria SDR SDRAM. 4 transferencias de 8B en modo ráfaga.

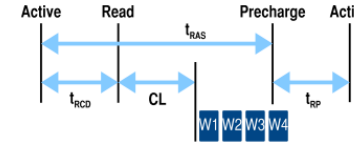
Fila ya abierta: En caso de que **la fila en la que quieres leer** ya esté abierta por transacciones anteriores en vez de cerrarla y abrirla reutilizar lo que ya tienes y SOLO cambias la columna.



Latencia reducida: $Lr = Tlectura\ (columna) = CL$

Latencia larga: $L = Tprecarga + Tabrir\ fila + Tlectura(columna) = TRP + TRD + CL$

Parámetros Temporales



- **CL (CAS Latency)**: Nº de ciclos entre envío de columna y el comienzo de ráfaga.
- **tRCD (RAS to CAS Delay)**: Nº de ciclos entre la apertura de la fila y el acceso a una columna.
- **tRP (RAS Precharge)**: Nº de ciclos entre la precarga y abrir la siguiente fila.
- **tRAS (Active to Precharge Delay)**: Nº mínimo de ciclos entre la activación de una fila y la precarga.

Calculo PF con L y Lr

$$PF = L \cdot (1 - TAbf) + L_r \cdot TAbf + \frac{B}{Bw} ciclos\ de\ bus$$

TAbf. Es la tasa de aciertos en el buffer de fila: $TAbf = \frac{Bloques\ accedidos\ con\ filas\ abierta}{Bloques\ accedidos\ totales}$

Unidades: En **segundos**, en **ciclo del procesador** y en **ciclos de memoria**, so **poner todo igual**.

La Latencia te la dan y la B y la Bw también, pues solo te queda poder mejorar la TAbj:

$$PF = \left(L \cdot (1 - TAbf) + L_r \cdot TAbf + \frac{B}{B_w}\right) \frac{1}{f_{mem}}\ segundos$$

$$PF = \left(L \cdot (1 - TAbf) + L_r \cdot TAbf + \frac{B}{B_w}\right) \frac{f_{cpu}}{f_{mem}}\ ciclos\ de\ CPU$$

Políticas del Controlador

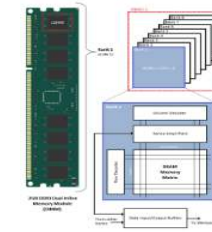
La política de planificación del controlador pot aumentar TAbf: Hacer cada vez que accedes, la fila ya esté abierta.

FCFS: **First come first served**. Utiliza una cola FIFO para su implementación. Hardware simple.

FR-FCFS: **First ready - first come first served**. Prioriza primero **first ready** (*bloque en el row-buffer*) y sobre estas (si las hay) FCFS, si no las hay aplica FCFS sobre las encoladas. *LAS **FILAS QUE YA ESTÉN ABIERTAS Y SOBRE LAS QUE SE QUIERA ACCEDER PASAN ANTES** QUE LAS QUE ACCEDEN A FILAS CERRADAS PARA APROBECHAR, si no pues FIFO.*

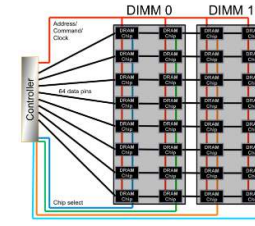
ORGANIZACIÓN DE LA MEMOIRA

Dentro de una memoria ram hay chips, **cada chip** está hecho por bancos **los cuales** tienen filas y columnas. *El dividir cada chip en bancos mejora la TAbf xq pueden tener una fila abierta por cada bank.*



Buffer de fila/rango: Si hay 8 chips de 8 bits cada uno, tengo filas de 64 bits, **entre todos hacen el buffer fila**. Cada 64 bits tienes un rango, si tienes 16 chips x 8 bits tienes 2 rango.

Luego cada chip tiene su buffer local de fila. El rango es justo eso, juntar a la vez una fila de cada chip: DIMMS. *Cada Ram puede tener diferentes rangos. Y tiene **TANTAS FILAS ABIERTAS COMO BANCOS TENGAN LOS CHIPS**.*



Fila: Las celdas (columnas) que se acceden a la vez en un banco de un chip del rango.

Canal: Conjunto de DIMMs conectados a un mismo "bus" del controlador.

NOTACIÓN MÓDULOS DIMM

DDR n xxxx: **n** = **generación**. **xxxx** =**Velocidad** = 2 trans/ciclo · frecuencia -> **Frecuencia** = **xxxx / 2**.

DIVIDIR el numerito ENTRE 2 para sacar la frecuencia.

PC n-yyyy: **Bw**: bus de MBytes/s = 8 Bytes/Trans · 2 transf/ciclo · f·ciclo/seg = 16 · frecuencia memoria.

DIVIDIR el numerito ENTRE 16 para sacar el ancho de banda.

INTERPRETACIÓN DE LA DIRECCIÓN

Canal, Rango, Chips, Banco, Fila, columna y cantidad de elementos en cada coluna.

- **Cantidad de filas abiertas**: $canales \cdot rangos \cdot bancos = 2^{2+2+4} = 256$
- **Nº de Bytes**: $filas \cdot canales \cdot rangos \cdot bancos \cdot columnas \cdot bytes\ en\ bus/columna = 2^{15+2+2+4+9+3}$

fila (15 bits)	canal (2 bits)	rango (2 bit)	banco (4 bits)	columna (9 bits)	bytes en bus (3 bits)