



Tema 5: Memorias

Tecnología y Organización de Computadores

Grado en Ingeniería Informática

Grado en Ingeniería de Computadores



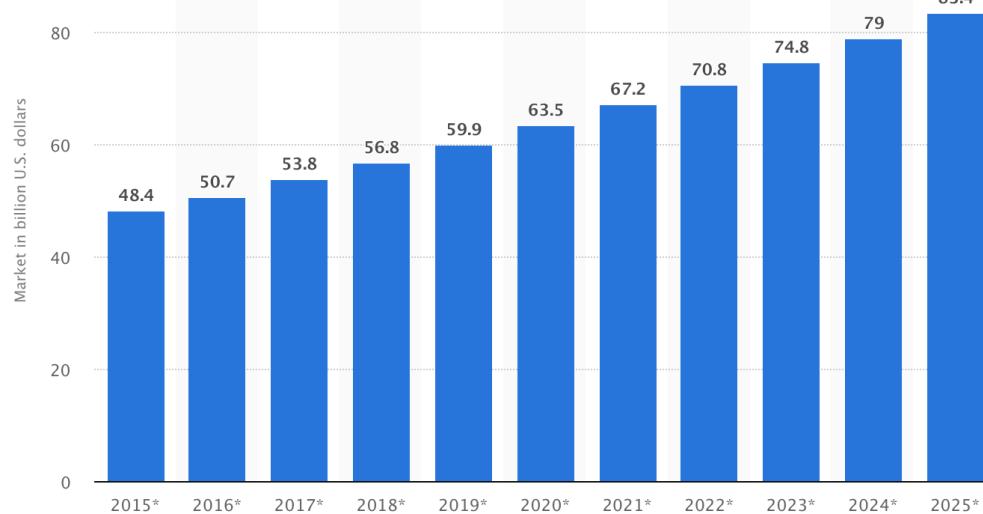
Índice

1. Introducción
2. Clasificación
3. Parámetros
4. Tecnologías de memoria
 1. SRAM
 2. DRAM
 3. SDRAM
5. Memorias BRAM Spartan-3
6. Ejemplo de diseño



¿Dónde hay memorias? Ejemplos

Tamaño del mercado de las memorias DRAM en miles de millones (Fuente: Statista)



SoC X1000:

- 16 KB I/D cache
- 512 KB on-die SRAM

(Fuente: Intel)

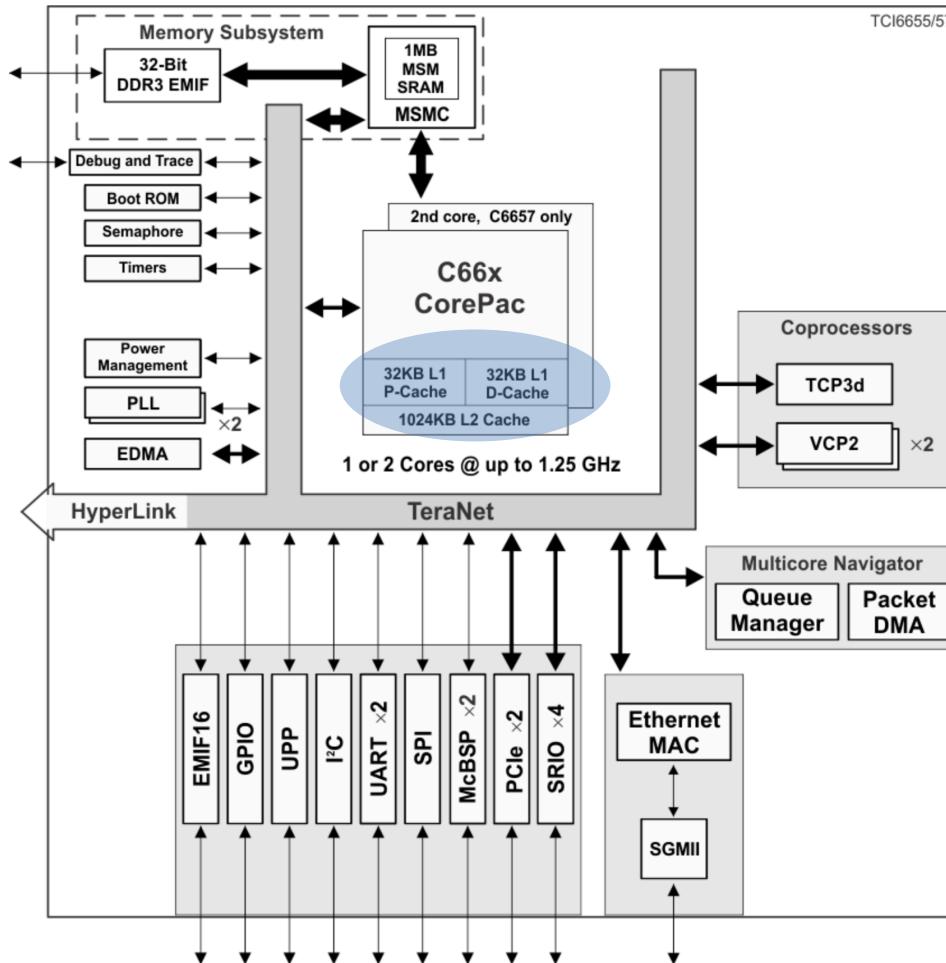


- L2 shared cache: 128–2048 KiB

(Fuente: ARM)

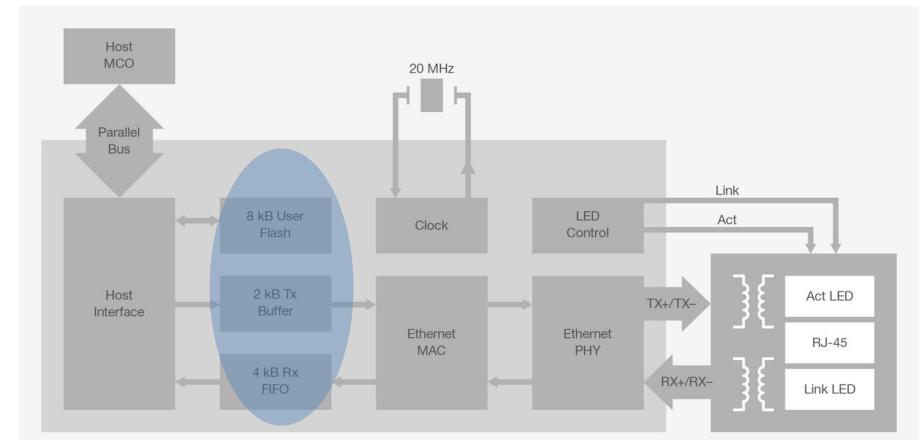


Más ejemplos



DSP C665x (Fuente: Texas Instruments Inc.)

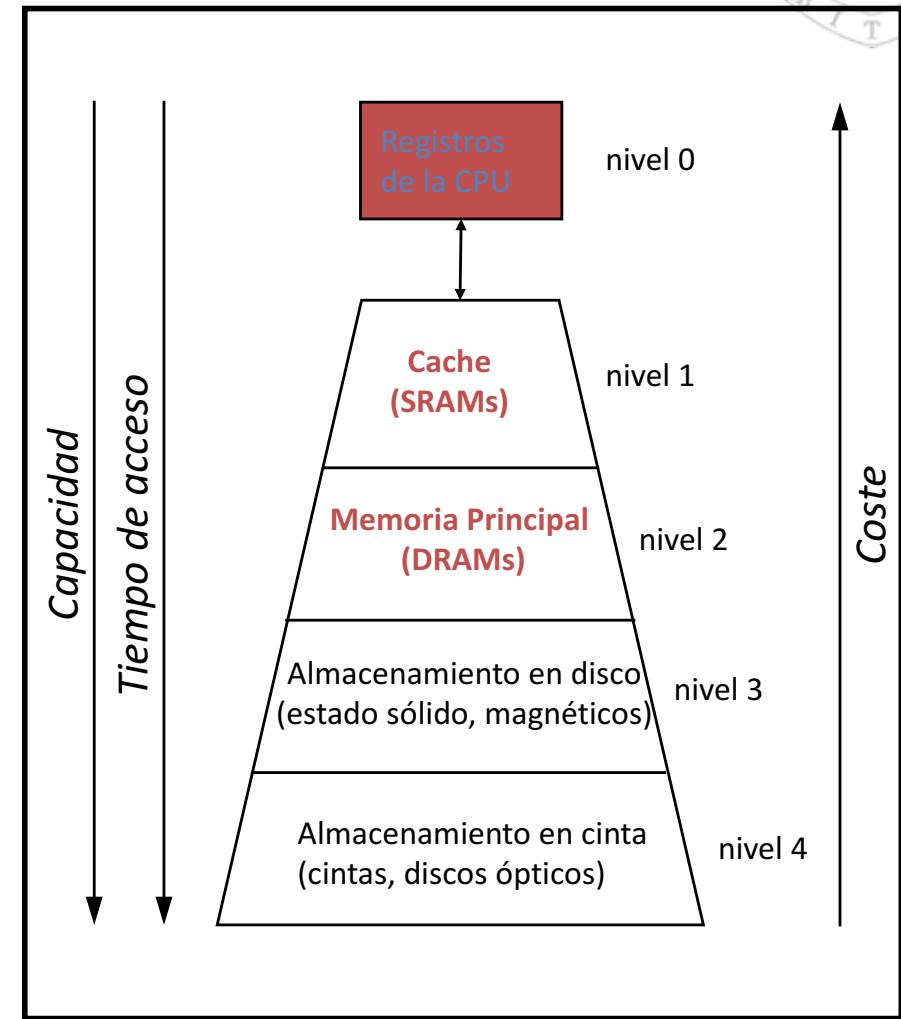
CP220x Ethernet Controller (Fuente: Silicon Labs)





Más ejemplos: el procesador

- Un computador típico está formado por diversos niveles de memoria, organizados de forma jerárquica:
 - Registros de la CPU
 - Memoria Cache
 - Memoria Principal
 - Memoria Secundaria (discos)
 - CDs - DVDs
- El coste de todo el sistema de memoria excede al coste de la CPU
 - Es muy importante optimizar su uso





Método de acceso

- **Acceso directo** (Ej. discos)
 - Los bloques de información se organizan en regiones (pistas)
 - Para leer/escribir un bloque se accede de forma directa a la región específica y dentro de esa región se realiza una búsqueda secuencial del bloque en cuestión (sector)
 - El tiempo de acceso es variable
- **Acceso aleatorio -- random --** (Ej. memoria principal)
 - Cada posición de memoria tiene un único método de acceso cableado físicamente
 - El tiempo de acceso a una posición es independiente de su dirección o de la secuencia de accesos previos
- **Acceso asociativo** (Ej. memoria cache)
 - Es una memoria de acceso aleatorio en la que las palabras no están ordenadas por dirección
 - Cada palabra tiene asociada una marca o *tag* (normalmente almacena su dirección o parte de la misma)
 - Para acceder a una determinada dirección de memoria es necesario comparar la marca a la que se desea acceder con cada una de las marcas de todas las palabras de memoria



Tecnologías de estado sólido

- SRAM: **Static RAM**
- DRAM: **Dynamic RAM**
- SDRAM: **Synchronous DRAM**
- EEPROM: **Electrically Erasable Programmable ROM**
- Nand-flash



Parámetros

- **Tiempo de acceso (Ti)**
 - En memorias de acceso aleatorio
 - Tiempo que transcurre desde que se especifica una dirección de memoria hasta que el dato o bien ha sido almacenado o bien está disponible para su uso.
 - En memorias de acceso secuencial/directo
 - Tiempo que se tarda en situar el mecanismo de lectura/escritura sobre la posición deseada.
- **Tiempo de ciclo de memoria**
 - Aplicable únicamente a memorias de acceso aleatorio
 - Tiempo mínimo que debe dejarse transcurrir entre dos accesos consecutivos.
 - Es algo mayor que el tiempo de acceso (contempla los tiempos de permanencia de las señales en los buses).



Parámetros

- **Ancho de banda o velocidad de transferencia (Bi)**
 - Velocidad a la que se pueden transmitir datos desde/hacia una unidad de memoria
 - En memorias de acceso aleatorio
 - Es igual a la inversa del tiempo de ciclo de memoria
 - En memorias de acceso secuencial/directo
 - Depende del tiempo de acceso medio y de la velocidad de transferencia del dispositivo
- **Tamaño de la memoria (Si)**
 - Nº de bytes que pueden almacenarse en el dispositivo de memoria
- **Coste por byte (Ci)**
 - Coste medio estimado por cada byte de memoria
 - El coste total de un dispositivo de memoria viene dado por $Ci \cdot Si$
- **Unidad de transferencia (Xi)**
 - Unidad de información con las que trabaja un determinado dispositivo de memoria
 - Puede oscilar entre un byte o una palabra, hasta bloques de varios Kbytes.

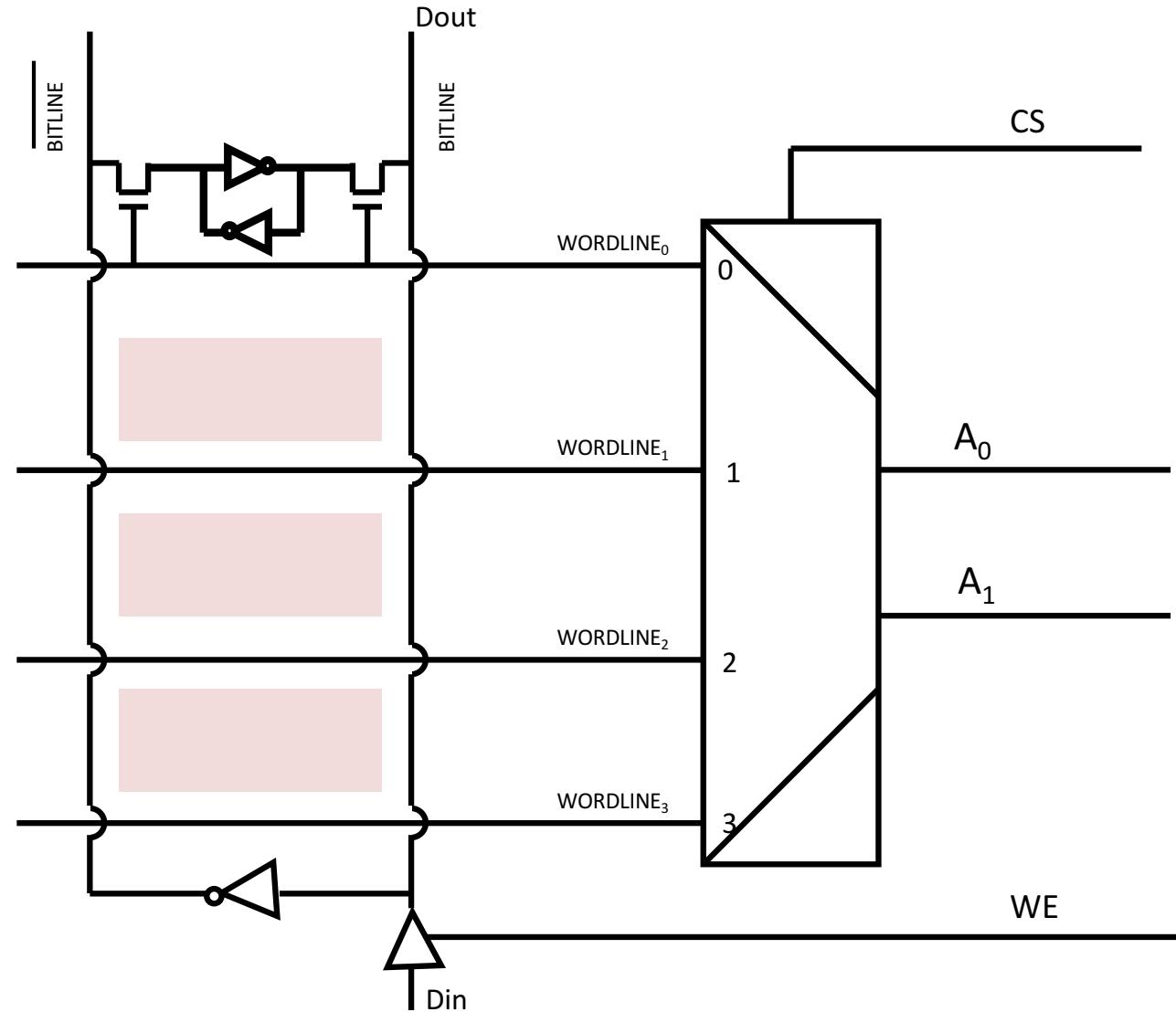


RAM Estática

- **Static Random-Access Memory (SRAM)**
 - Almacena un bit en un biestable
 - El valor permanece mientras la memoria tenga alimentación.
 - La ventaja de las memorias SRAM es su rapidez.
 - El inconveniente es que necesitan 6 transistores
 - No permite alcanzar grandes densidades de integración.

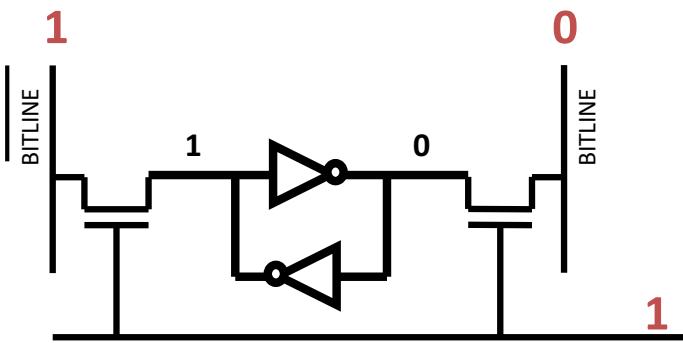
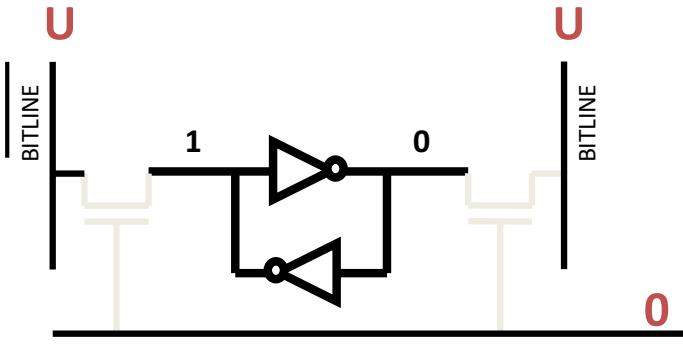


Celda básica





Operación



Cargamos o somos capaces de ver el valor almacenado, si y sólo si, wordline vale '1'

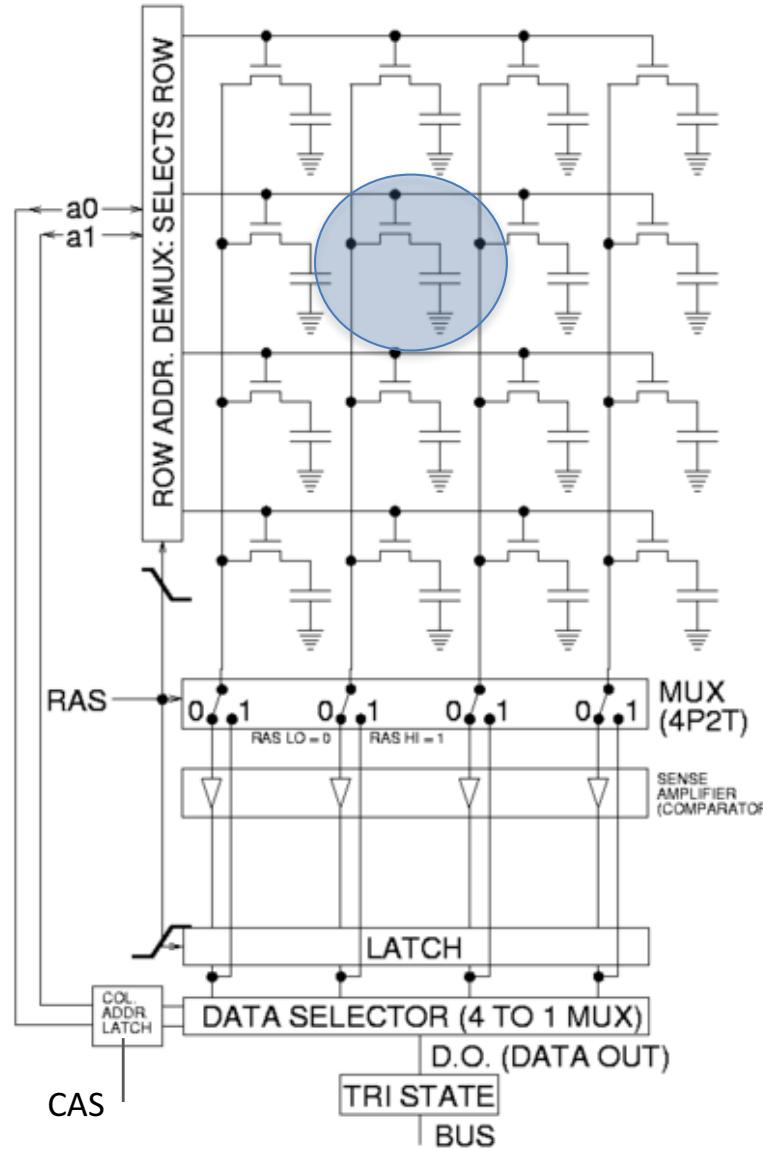


RAM Dinámica

- **Dynamic Random-Access Memory (DRAM)**
 - Almacena un bit en un condensador (cargado – descargado)
 - El condensador pierde voltaje a lo largo del tiempo por lo que tiene que ser *refrescado* periódicamente -> Por eso se dice que es una memoria dinámica
 - La ventaja de las memorias DRAM es su simplicidad, sólo se necesita un condensador y un transistor para almacenar un bit (SRAM necesita mínimo 6 transistores).
 - Esto permite alcanzar grandes densidades de integración (billones de bits en un chip)
 - Es una memoria volátil, pierde rápidamente la información cuando se apaga

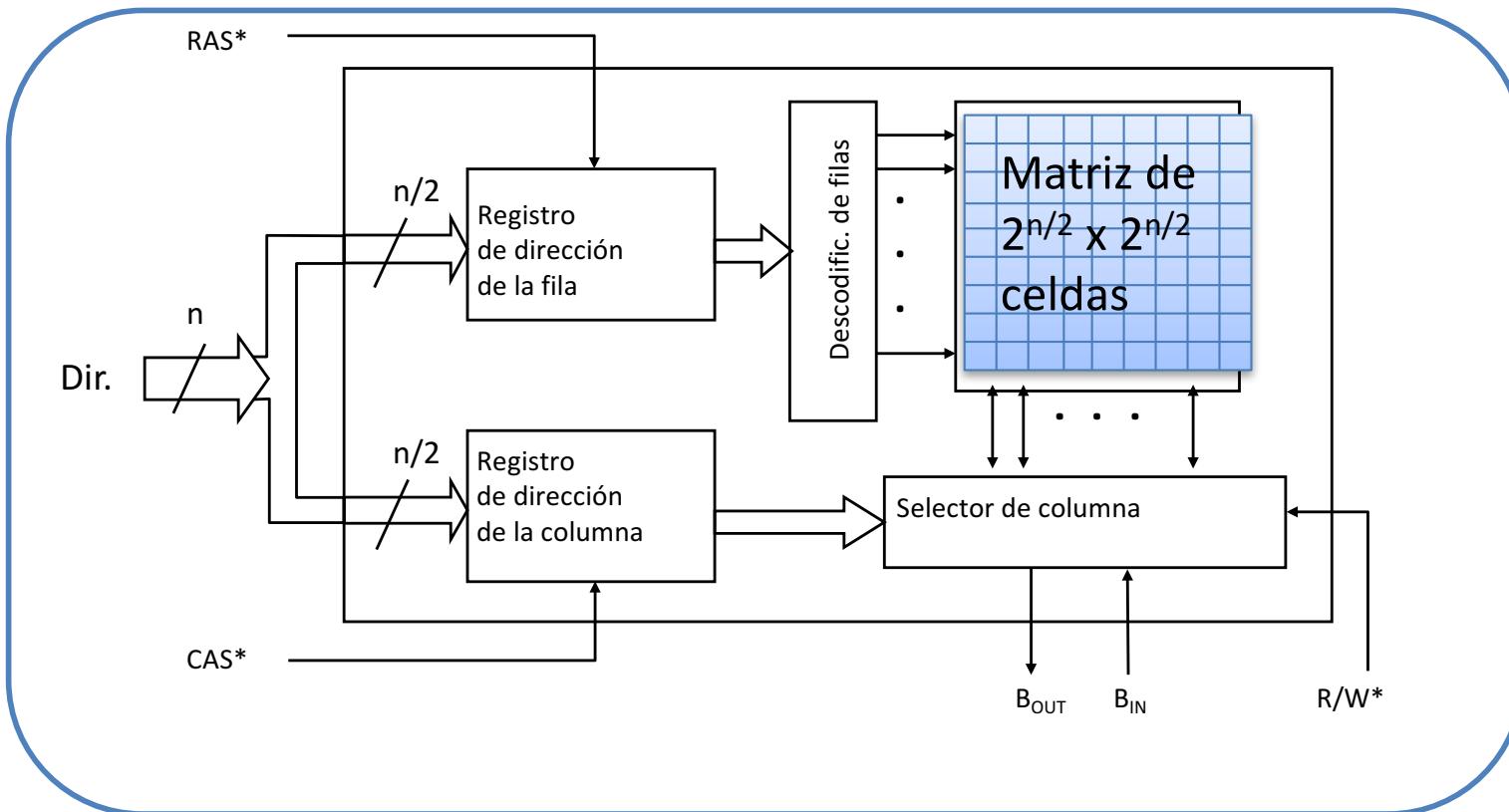


Celda básica





Organización



RAS*: Línea de selección de la fila

CAS*: Línea de selección de la columna



Lectura

1. Los amplificadores están desconectados (sense amplifiers)
2. Las *bit-lines* se pre-cargan a un valor intermedio entre el cero y el uno.
3. La *word-line* de la fila deseada se pone a uno para conectar el condensador a su *bit-line*.
 - El condensador se descarga a diferente velocidad dependiendo de si estaba inicialmente cargado o descargado.
4. Los amplificadores están conectados a la *bit-line* amplificando las pequeñas diferencias de voltaje entre las líneas pares e impares.
 - La diferencia en voltaje entre x y x^* es positiva -> x almacenaba un uno.
5. La salida de los amplificadores se almacenan en *latches*



Lectura

t_{RC} : Tiempo de ciclo de lectura

Tiempo máximo que deben permanecer las direcciones estables a la entrada de la memoria antes de iniciar un nuevo acceso

t_{ASR} y t_{RAH} : Tiempos de estabilización y permanencia de la dirección de fila

Tiempo mínimo que debe permanecer la dirección de fila estable antes y después de la activación de RAS*

t_{ASC} y t_{CAH} : Tiempos de estabilización y permanencia de la dirección de columna

Tiempo mínimo que debe permanecer la dirección de columna estable antes y después de la activación de CAS*

t_{CAS} : Ancho del pulso de la señal CAS*

Tiempo mínimo que debe permanecer CAS* activada

t_{RAS} : Ancho del pulso de la señal RAS*

Tiempo mínimo que debe permanecer RAS* activada

t_{RDC} : Tiempo retardo entre RAS* y CAS*

Intervalo temporal que debe transcurrir entre la activación de RAS* y la activación de CAS*

t_{RAC} : Tiempo de acceso desde la activación RAS*

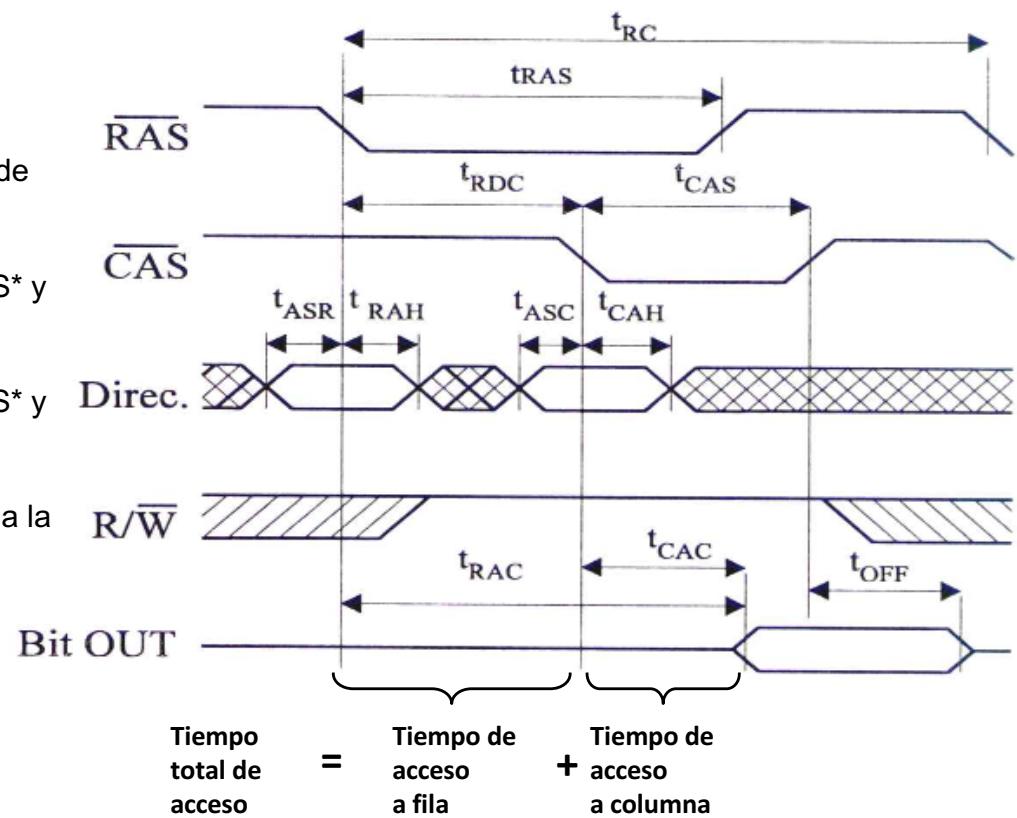
Tiempo máximo que transcurre desde la activación de RAS* y la aparición de datos estables a la salida

t_{CAC} : Tiempo de acceso desde la activación CAS*

Tiempo máximo que transcurre desde la activación de CAS* y la aparición de datos estables a la salida

t_{OFF} : Tiempo de permanencia de los datos

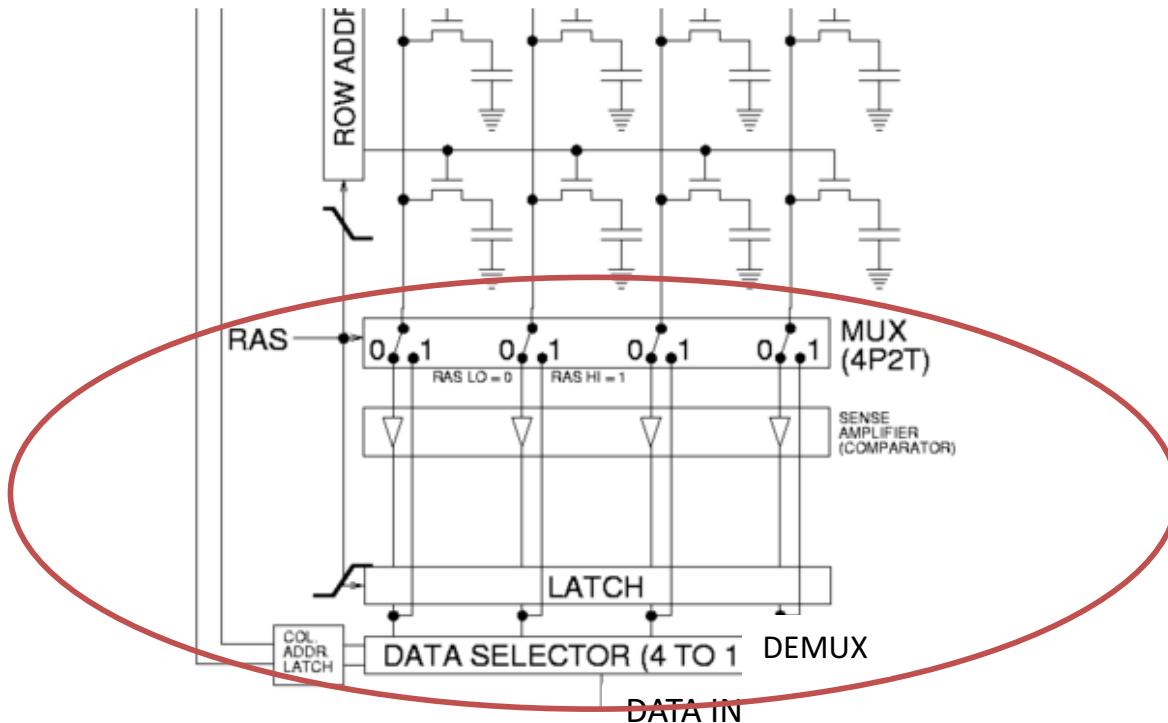
Tiempo mínimo que deben permanecer los datos estables a la salida después de desactivar las líneas RAS* y CAS*





Escritura

- Se fuerza a los amplificadores a un valor cero o uno para cargar o descargar los condensadores de la fila previamente seleccionada





Escritura

t_{RC} : Tiempo de ciclo de escritura

Tiempo máximo que deben permanecer las direcciones estables a la entrada de la memoria antes de iniciar un nuevo acceso

t_{ASR} y t_{RAH} : Tiempos de estabilización y permanencia de la dirección de fila

Tiempo mínimo que debe permanecer la dirección de fila estable antes y después de la activación de RAS*

t_{ASC} y t_{CAH} : Tiempos de estabilización y permanencia de la dirección de columna

Tiempo mínimo que debe permanecer la dirección de columna estable antes y después de la activación de CAS*

t_{CAS} : Ancho del pulso de la señal CAS*

Tiempo mínimo que debe permanecer CAS* activada

t_{RAS} : Ancho del pulso de la señal RAS*

Tiempo mínimo que debe permanecer RAS* activada

t_{RDC} : Tiempo retardo entre RAS* y CAS*

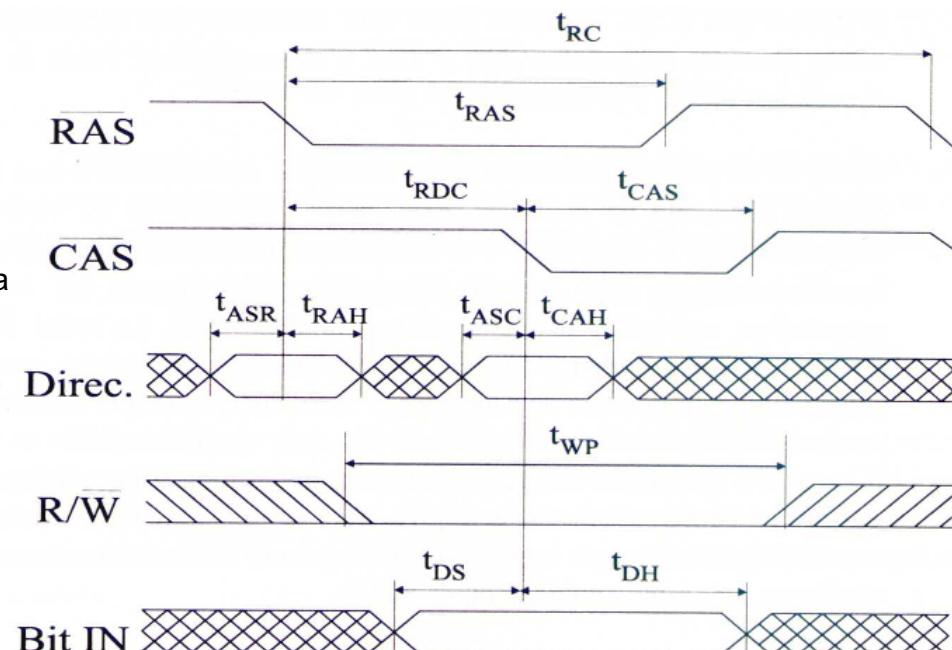
Intervalo temporal que debe transcurrir entre la activación de RAS* y la activación de CAS*

t_{WP} : Tiempo de ancho de la señal de escritura

Tiempo mínimo que debe permanecer la señal de escritura activada ($R/W^*=0$)

t_{DS} y t_{DH} : Tiempos de estabilización y permanencia de los datos

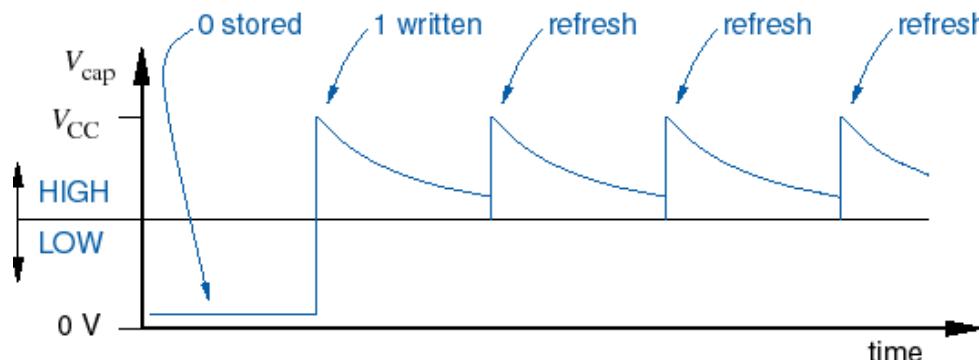
Tiempo mínimo que deben permanecer los datos estables antes y después de la activación de CAS*





Refresco

- Modos de refresco
 - Fila a fila en modo ráfaga, todas las filas en 64 ms
 - De manera escalonada durante 64 ms
 - Los dos modos necesitan un contador que indica cuál es la siguiente fila a refrescar (ese contador está incluido en el chip)





Comparativa SRAM vs DRAM

Memoria	Acceso (B)	Latencia	Coste (\$/MB)	Energía
SRAM embedded	10	~100 ps	1-100	1 nJ
SRAM	100	~ ns	1-10	10 -100 nJ
DRAM	1000	10 -100 ns	0,1	1 – 100 nJ
Disco duro	1000	ms	0,001	100 -1000 mJ



DRAM síncrona

- **Synchronous Dynamic Random Access Memory (SDRAM)**
 - Simplificando, es una DRAM síncrona.
 - Las DRAM clásicas son asíncronas para responder lo más rápido posible.
 - Se suele sincronizar con el bus del sistema
 - La memoria espera el flanco de subida del reloj antes de responder, este reloj se utiliza para hacer funcionar la máquina de estados asociada a esta memoria. Así la memoria puede responder a patrones de acceso complicados.
 - Las memorias SDRAM están internamente divididas en 2 o 4 bancos independientes



Señales de control

- Además del reloj hay 6 señales de control:
 - **CKE** (Clock Enable): con esta señal en baja el circuito funciona como si el reloj estuviera parado. En realidad esta señal se utiliza para enmascarar el reloj que entra en la memoria
 - **CS*** (Chip Select): cuando esta señal está en alta se ignoran el resto de señales (excepto CKE), en caso contrario se pasa a tener en cuenta las siguientes señales descritas.
 - **DQM** (Data Mask): en alta se suprime la comunicación con el exterior; no se escribe en memoria o no se lee si DQM se activa dos ciclos antes del ciclo de lectura
 - Existe una línea DQM por cada 8 bits en cada chip x16 o DIMM.
 - **RAS*** Row Address Strobe
 - **CAS*** Column Address Strobe
 - **WE*** Write enable

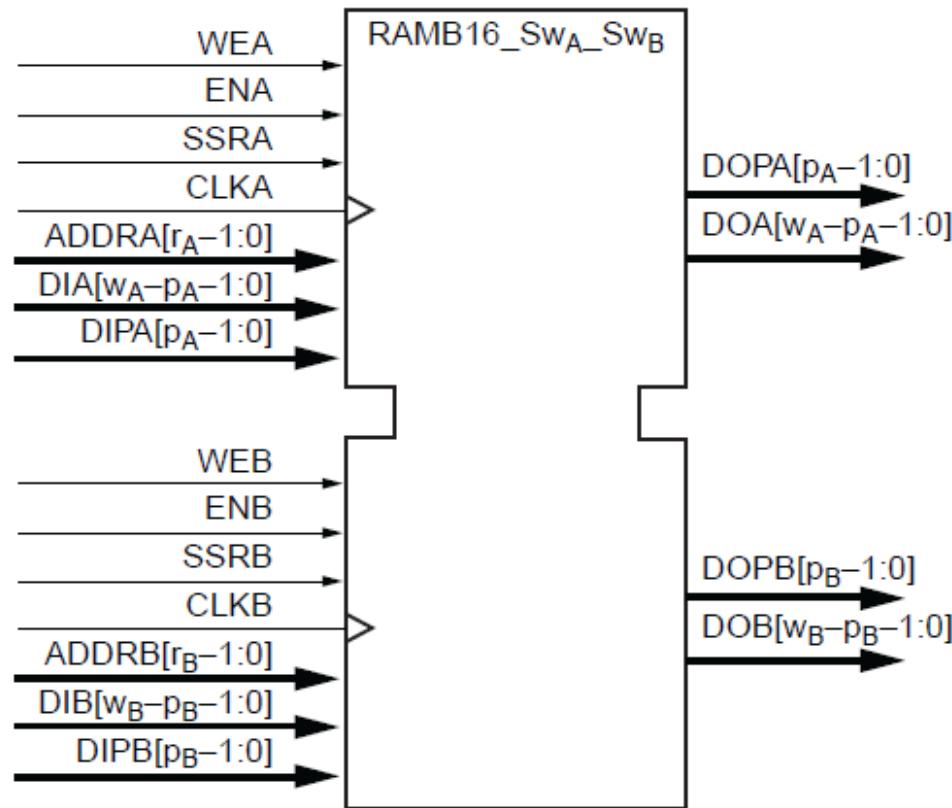


BRAMs

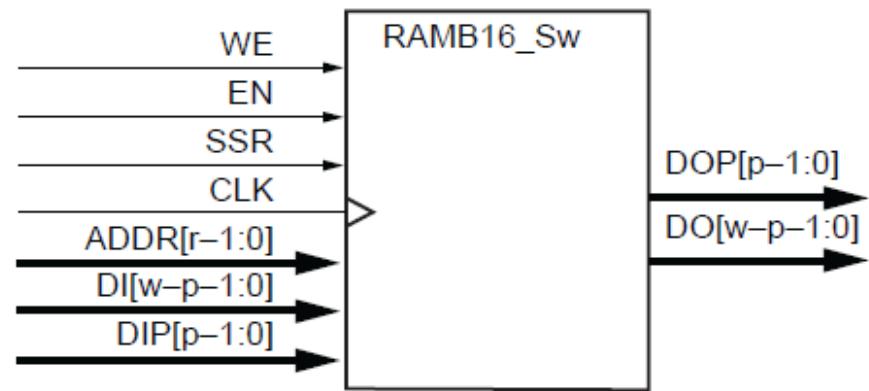
- Muchas familias de FPGAs tienen también internamente bloques de memoria, los cuales tienen unas características particulares:
 - Presentan diferentes organizaciones:
 - Desde 16Kx1 bit hasta 512x32 bits
 - Tipos:
 - Single port
 - Simple dual port
 - True dual port
 - La lectura y la escritura es síncrona
 - Presentan, opcionalmente, bits de control de errores: bits de paridad



Puertos



(a) Dual-Port



(b) Single-Port



Puertos

WE	Permite la escritura
EN	Permite el funcionamiento de la memoria
SSR	Set/reset síncrono
ADDR	Dirección de lectura/escritura
DI	Datos de entrada
DIP	Paridad de los bytes de los datos de entrada
DO	Datos de salida
DOP	Paridad de los bytes de los datos de salida





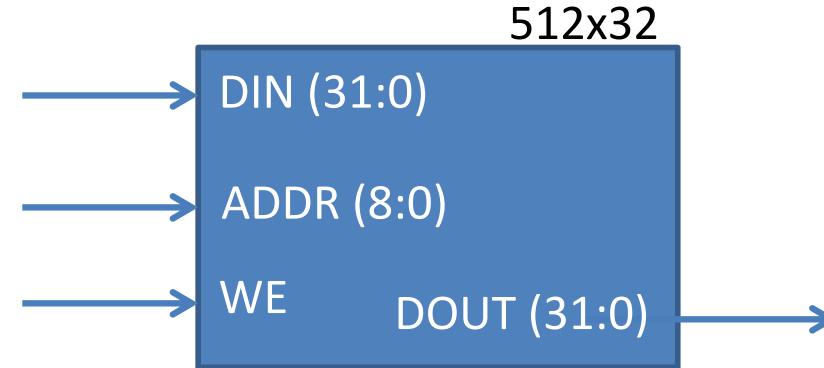
Configuraciones

- Por defecto la BRAM tiene palabras de 32 bits, sin embargo, tiene las siguientes posibles organizaciones:
 - 16Kx1
 - 8Kx2
 - 4Kx4
 - 2Kx8
 - 1Kx16
 - 512x32



Configuraciones

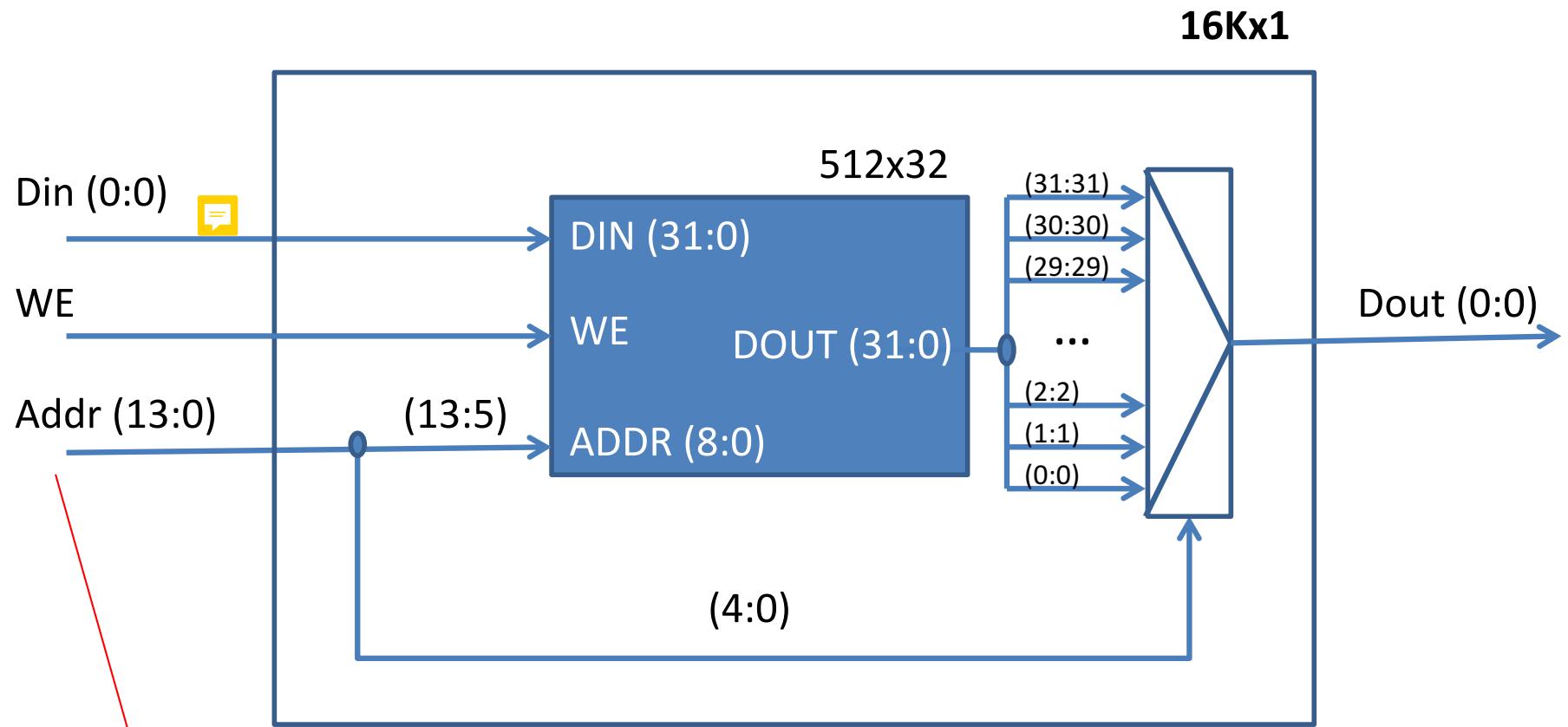
- Configuración *original* simplificada:





Configuraciones

- Configuración 16Kx1:



Longitud de puerto de direcciones: 16K posiciones = $2^4 \times 2^{10}$ posiciones = 2^{14} posiciones



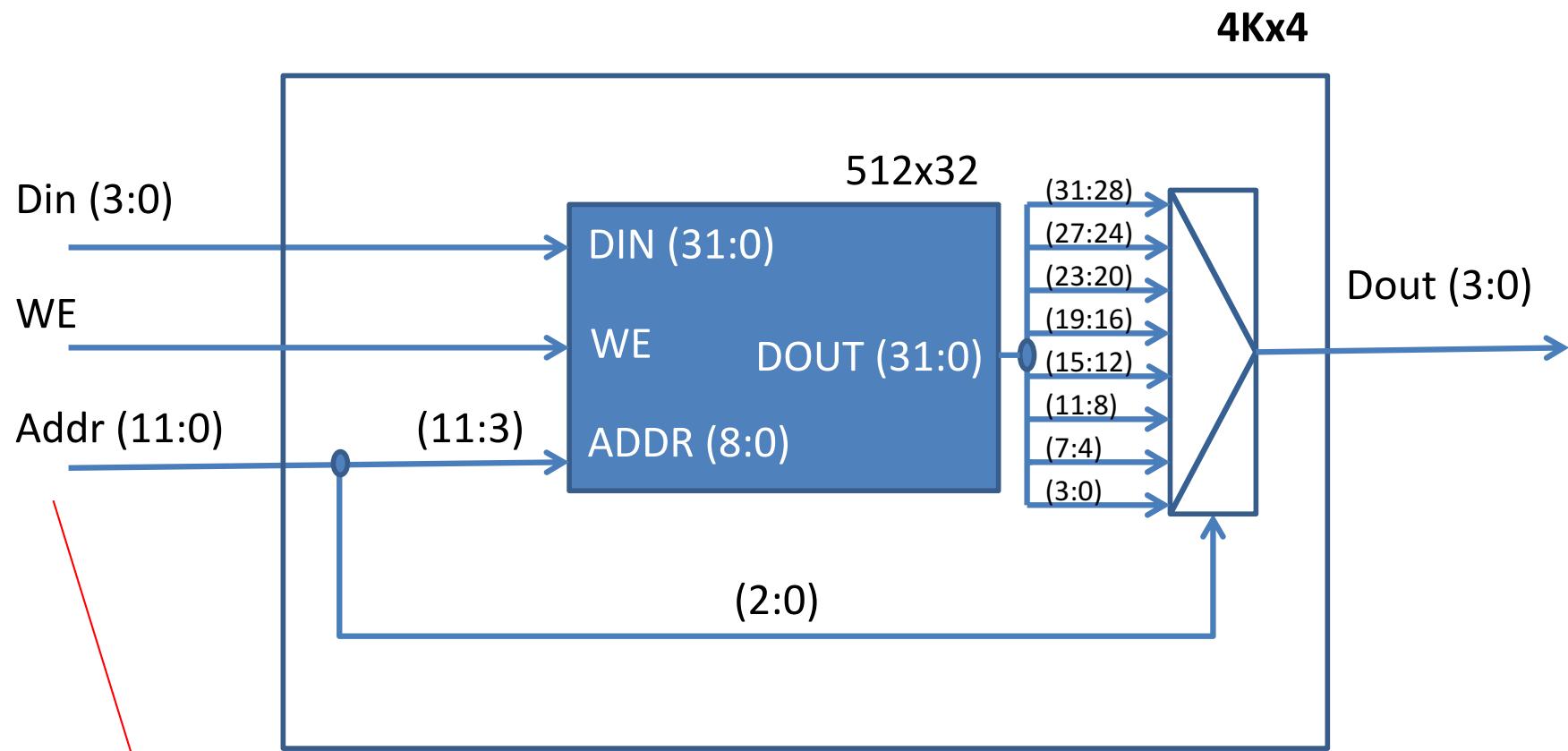
Configuraciones

- Ejemplos:
 1. ¿Cómo sería el HW necesario para hacer la lectura en 4Kx4?
 2. ¿Cómo sería el HW necesario para hacer la lectura en 1Kx16?
 3. ¿Cómo sería el HW necesario para hacer la lectura en 256x64?
 4. ¿Cómo sería el HW necesario para hacer la escritura en 16Kx1?
 5. ¿Cómo sería el HW necesario para hacer la escritura en 8Kx2?



Configuraciones

- Lectura para configuración 4Kx4:

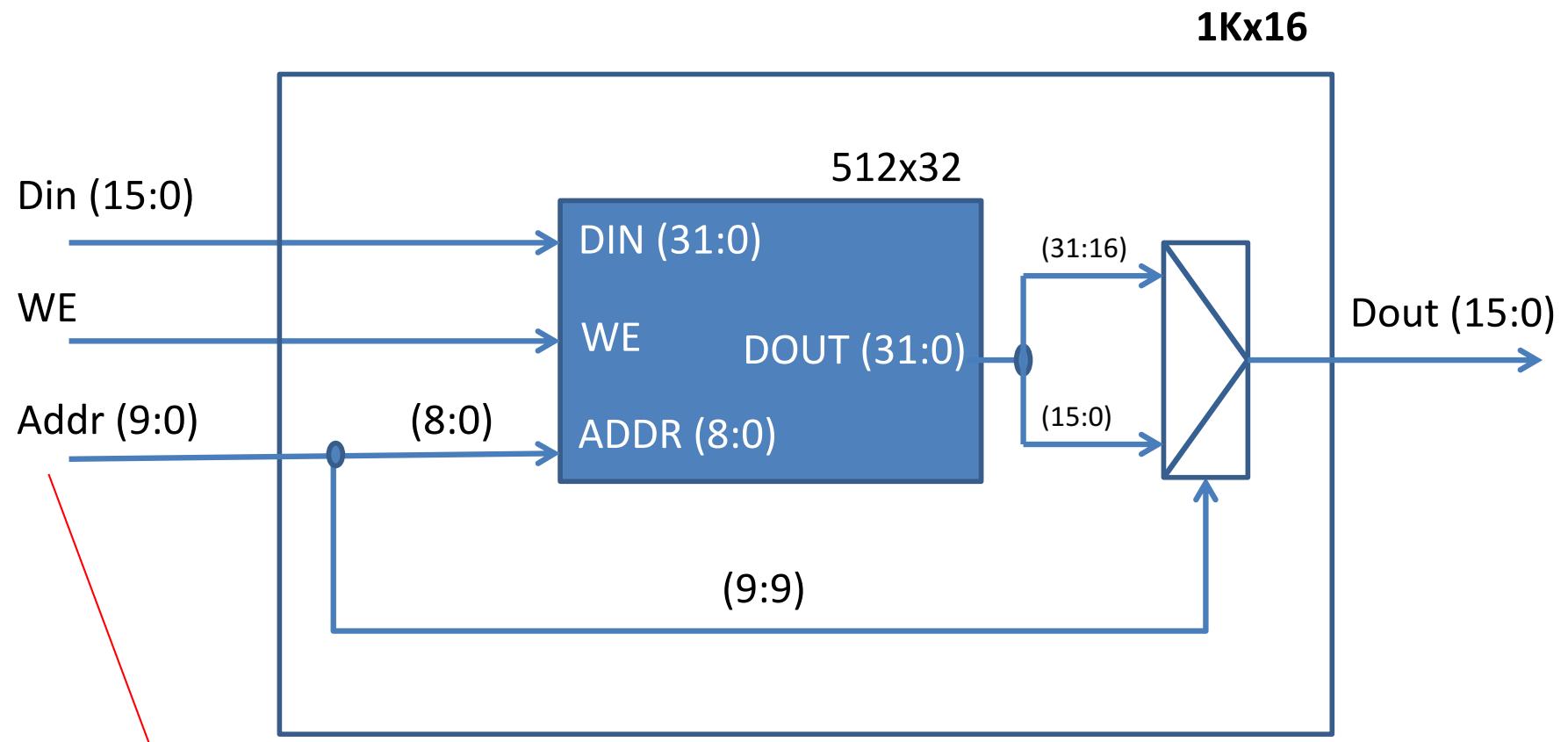


Longitud de puerto de direcciones: 4K posiciones = $2^2 \times 2^{10}$ posiciones = 2^{12} posiciones



Configuraciones

- Lectura para configuración 1Kx16:

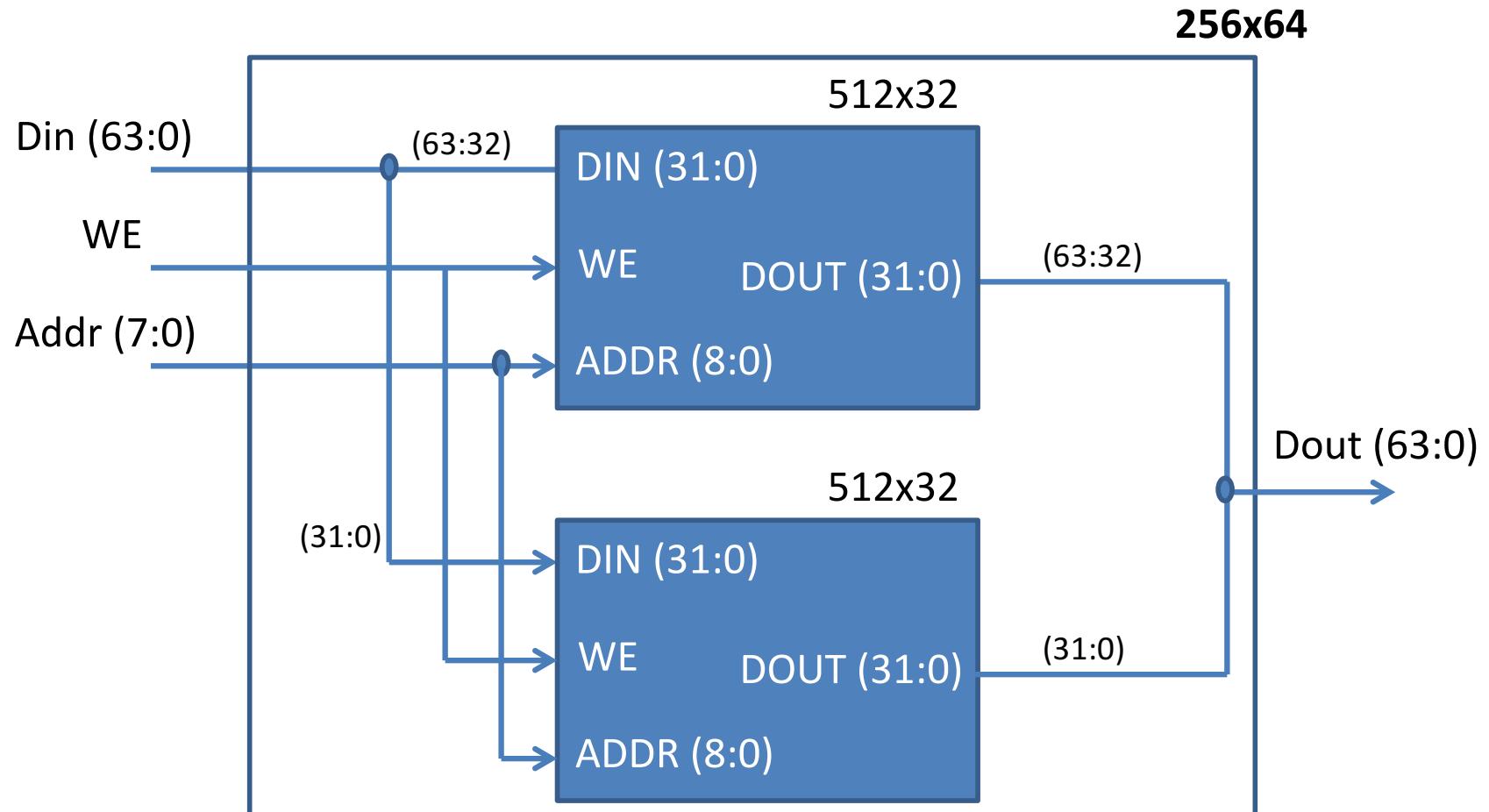


Longitud de puerto de direcciones: 1K posiciones = 2^{10} posiciones



Configuraciones

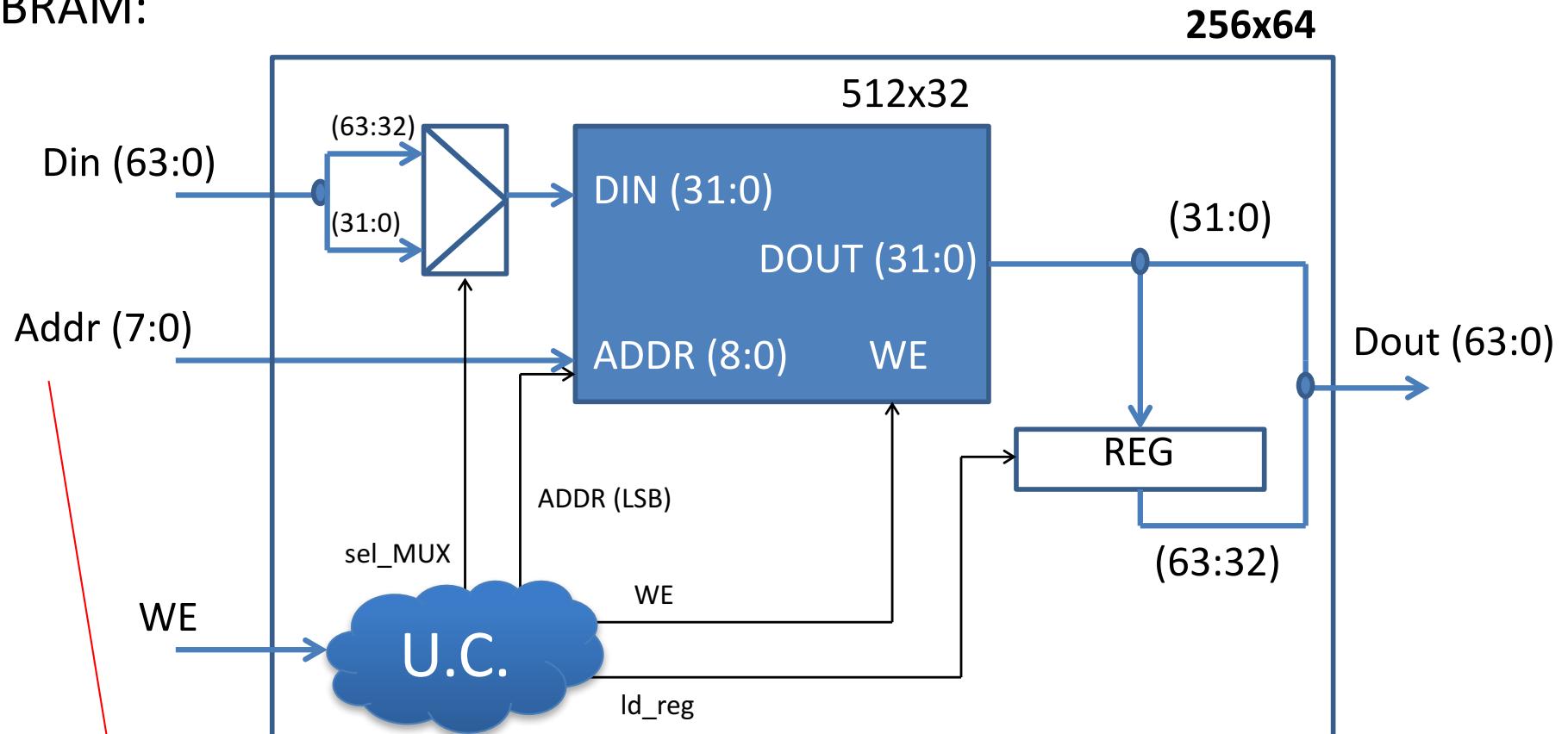
- Lectura para configuración 256x64:
 - Hacen falta 2 BRAMs para realizar las lecturas en un solo ciclo





Configuraciones

- Lectura para configuración 256x64, diseño alternativo con una sola BRAM:



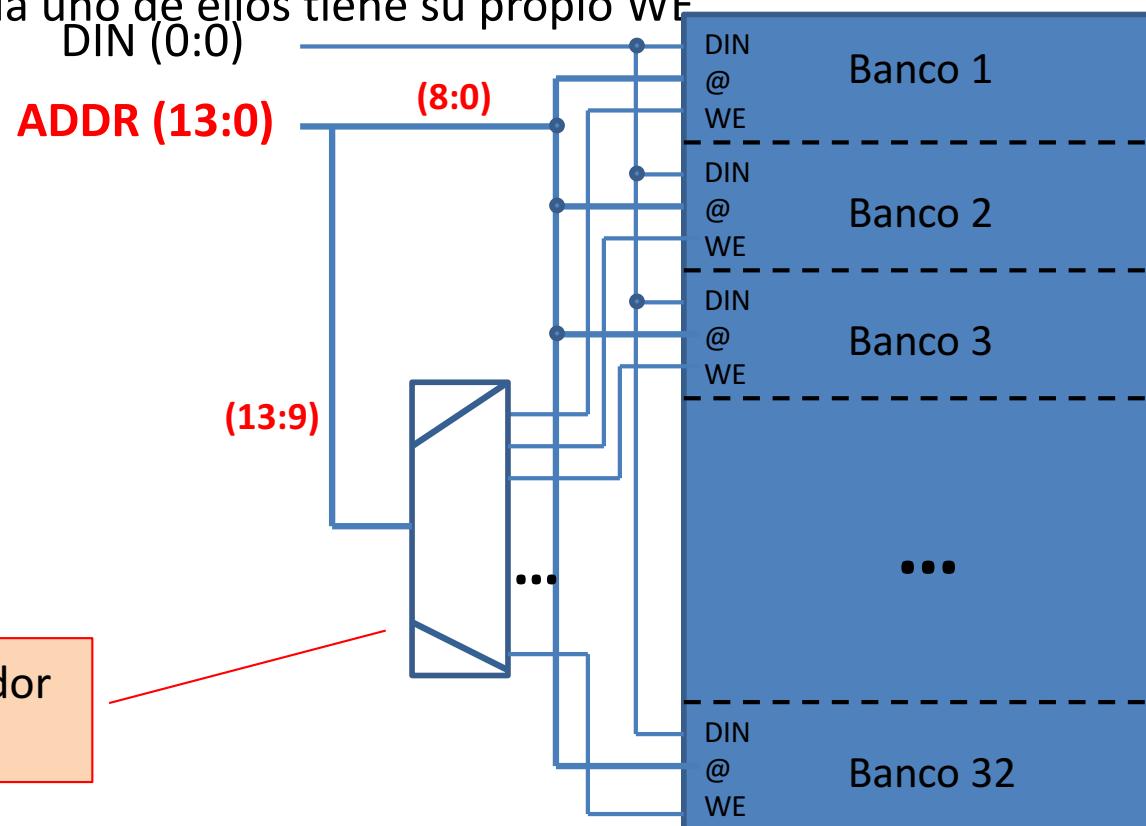
Longitud de puerto de direcciones: $256 = 2^8$ posiciones

Las lecturas se efectúan en 2 ciclos, controlados por la U.C.



Configuraciones

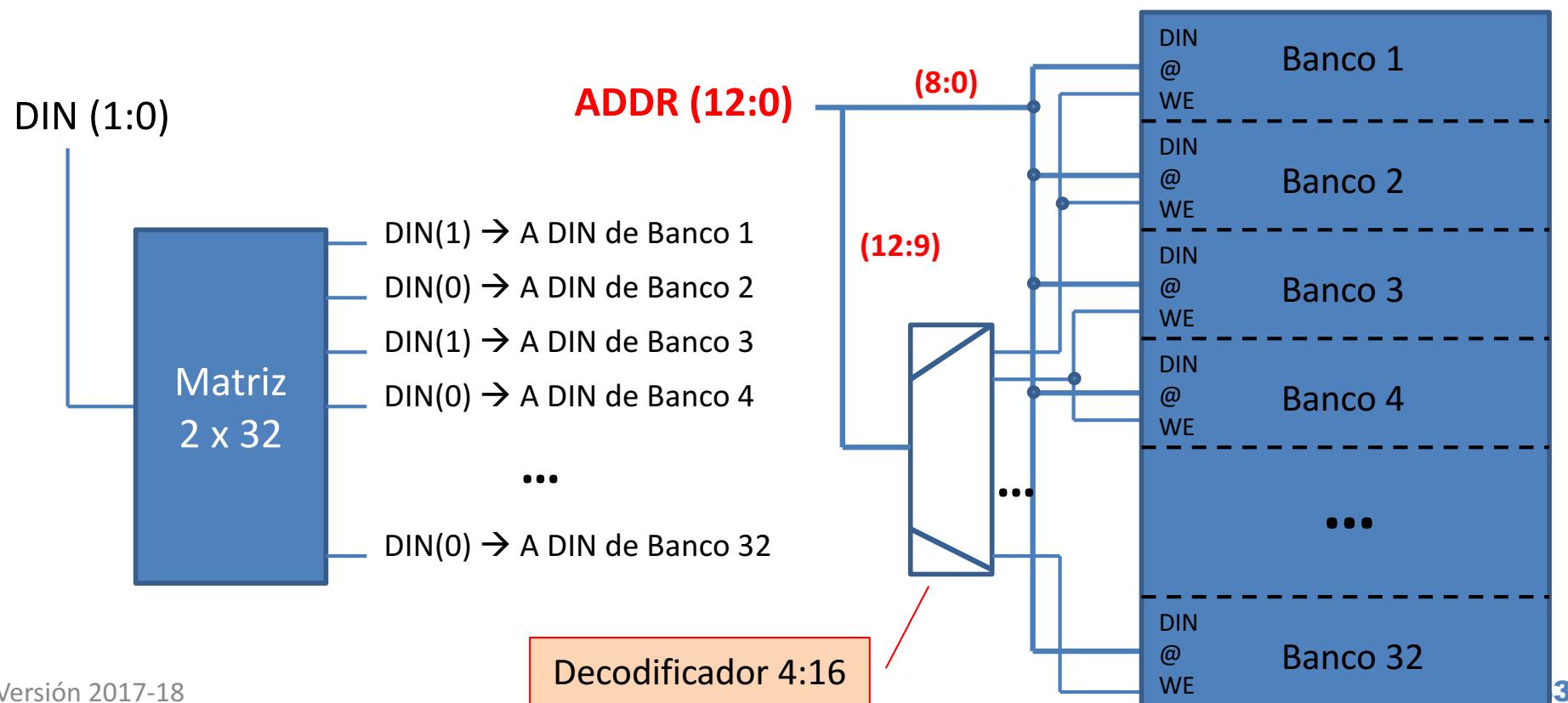
- Escritura para configuración 16Kx1:
 - **Realmente las BRAMs de Xilinx son direccionables por bit.** Es decir, realmente, el puerto ADDR tiene 14 bits en vez de 9.
 - Internamente, están organizadas en 32 bancos de memoria idénticos de 512x1 bits. Cada uno de ellos tiene su propio WE





Configuraciones

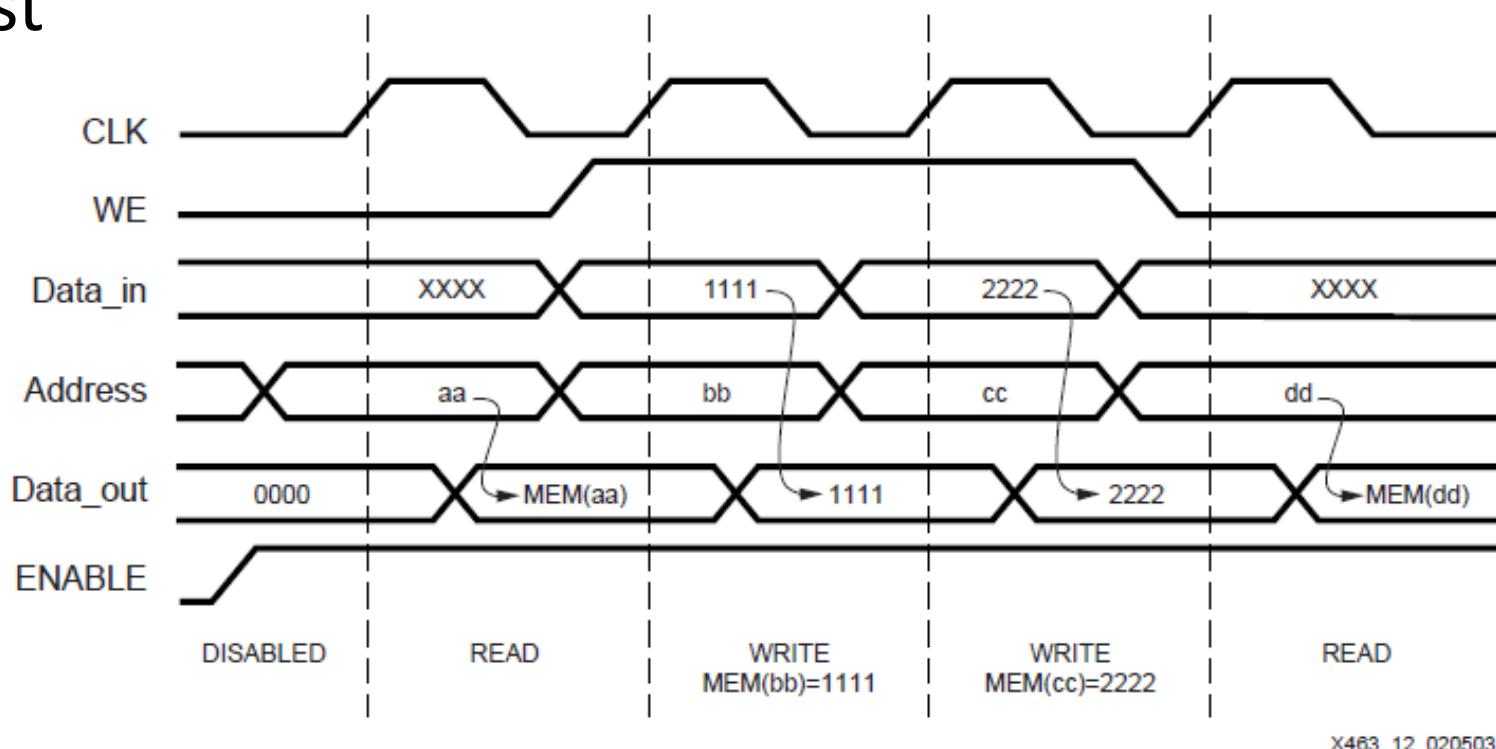
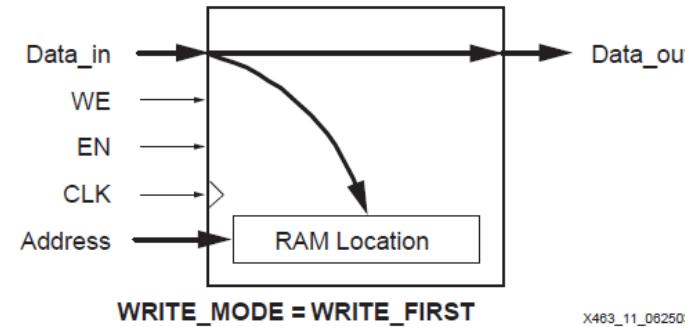
- Escritura para configuración 8Kx2:
 - La idea es la misma que antes, añadiendo un decodificador de 4 a 16.
 - Esta vez es necesario añadir también una matriz de interconexión para dirigir los datos DIN
 - **Conclusión:** para soportar diferentes tamaños de palabra, al bloque de BRAM de Xilinx se le añade un decodificador al puerto ADDR, un MUX antes de la salida DOUT (visto en transparencias anteriores), y una matriz de interconexión en DIN; todos ellos de tamaño variable.





Modo **WRITE_FIRST**

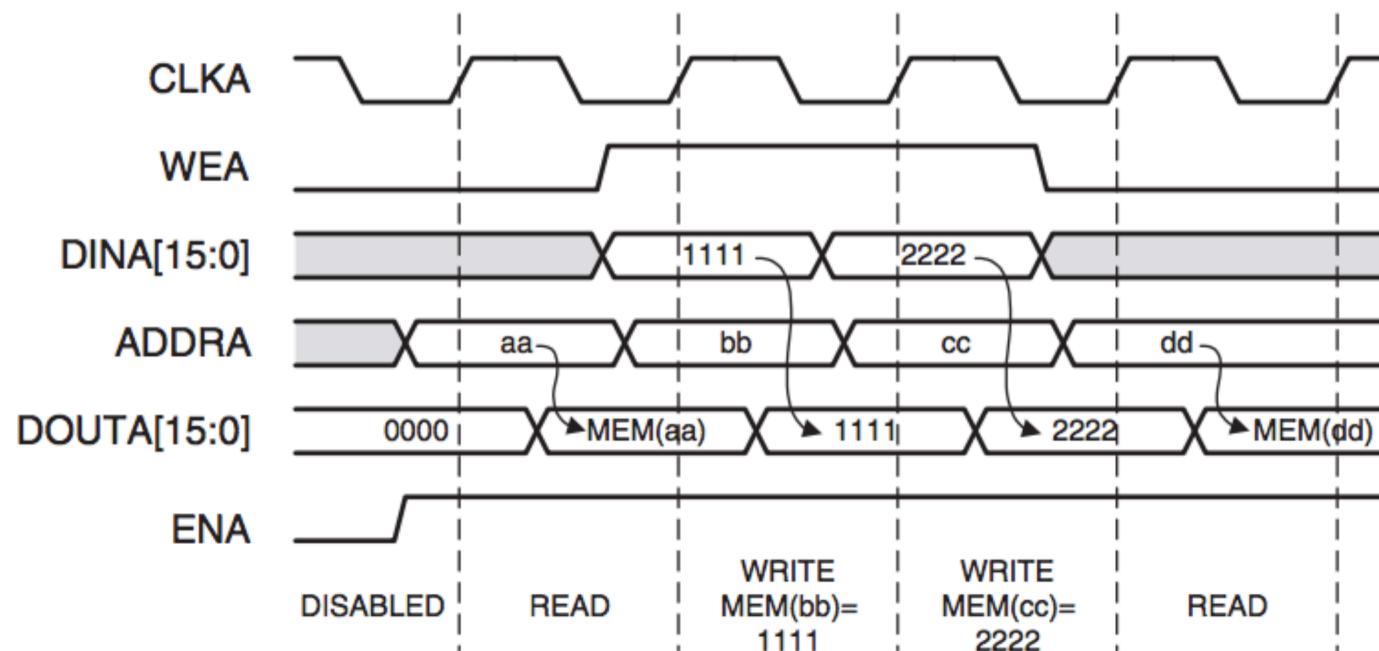
- Por defecto las memorias de las FPGAs trabajan en el modo write-first





Modo **WRITE_FIRST**

- Los datos de entrada se escriben en memoria y, simultáneamente, se visualizan por el puerto de salida. Este modo ofrece la flexibilidad de utilizar el puerto de datos de salida durante una operación de escritura de esos mismos datos





Modo **WRITE_FIRST**

- Cómo infiere Xilinx ISE una memoria BRAM de tipo **WRITE_FIRST**:

Synthesizing Unit <blockram2>.

Related source file is "C:/blockram2/blockram2.vhd".

Found 32x4-bit single-port block RAM for signal <ram>.

mode	write-first			
aspect ratio	32-word x 4-bit			
clock	connected to signal <clk>			
enable connected to signal <en>		low	rise	
write enable	connected to signal <we>		low	
address	connected to signal <addr>			
data in	connected to signal <di>			
data out	connected to signal <do>			
ram_style	Auto			

Summary:

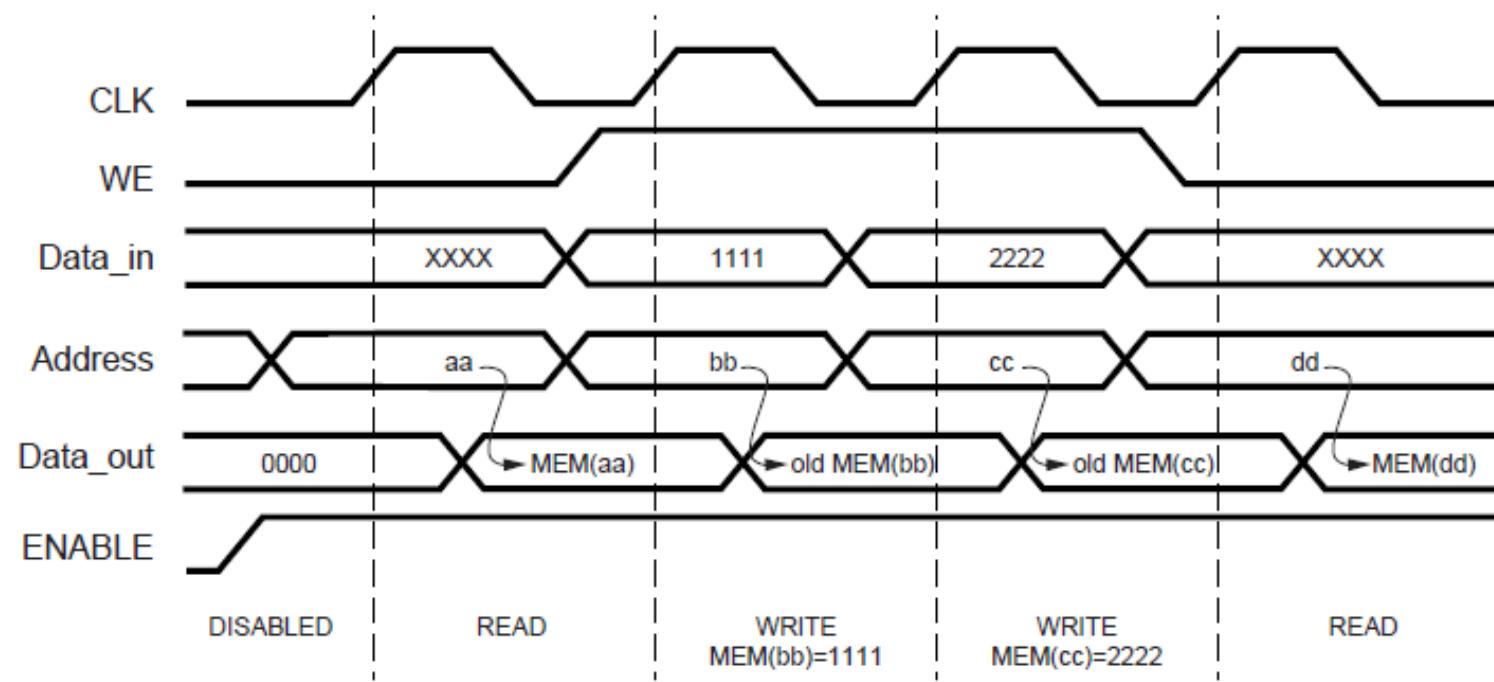
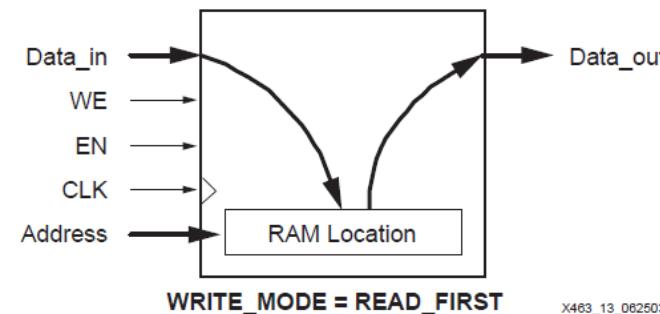
inferred 1 RAM(s).

Unit <blockram2> synthesized.



Modo READ_FIRST

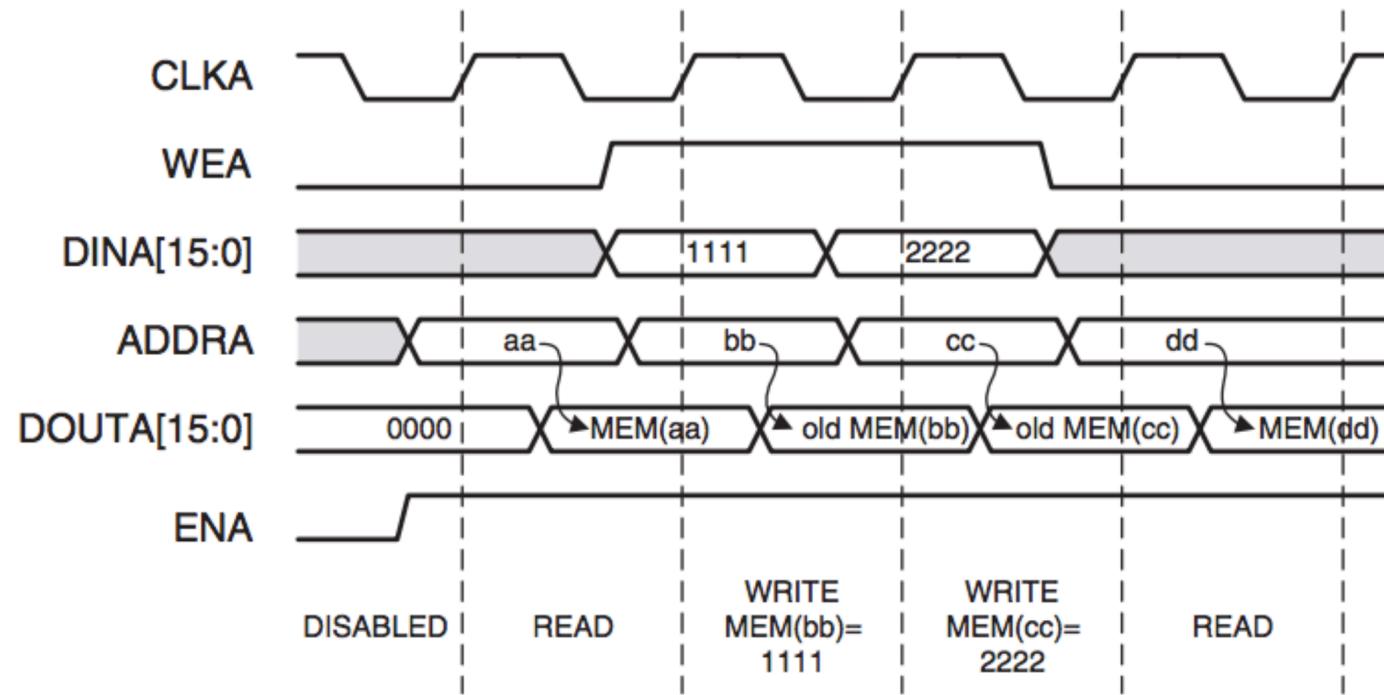
- Otro modo de trabajo de las memorias de las FPGAs es read_first





Modo READ_FIRST

- El dato almacenado en la dirección dada por Address aparece en el puerto de salida, a la vez que el dato de entrada se escribe en esa misma posición de memoria





Modo READ_FIRST

- Cómo infiere Xilinx ISE una memoria BRAM de tipo READ_FIRST:

```
Synthesizing Unit <blockram3>.  
Related source file is "C:/blockram3/blockram3.vhd".  
Found 32x4-bit single-port block RAM for signal <ram>.
```

mode	read-first			
aspect ratio	32-word x 4-bit			
clock	connected to signal <clk>		rise	
enable	connected to signal <en>	high		
write enable	connected to signal <we>		high	
address	connected to signal <addr>			
data in	connected to signal <di>			
data out	connected to signal <do>			
ram_style	Auto			

Summary:

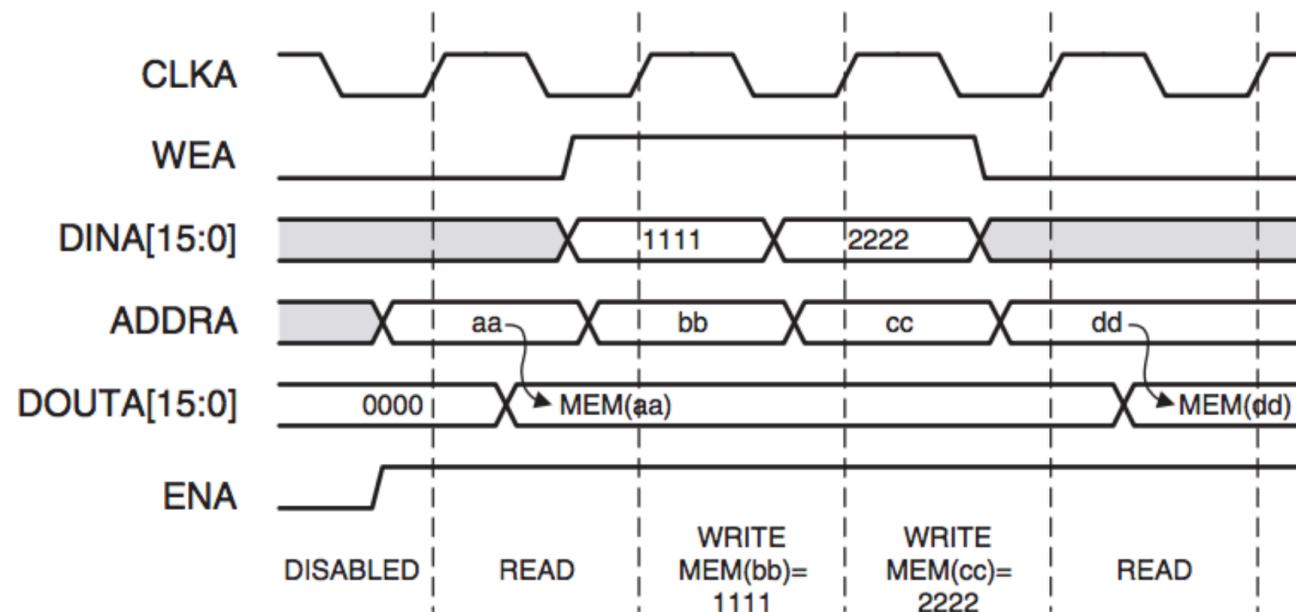
inferred 1 RAM(s).

Unit <blockram3> synthesized.



Modo NO_CHANGE

- Las memorias SRAM que hemos utilizado hasta ahora trabajan en modo *NO_CHANGE*
 - El puerto de salida **no se actualiza** cuando se produce una operación de escritura





Modo NO_CHANGE

- Cómo infiere Xilinx ISE una memoria BRAM de tipo NO_CHANGE:

```
Synthesizing Unit <blockram1>.  
Related source file is "C:/blockram1/blockram1.vhd".  
Found 32x4-bit single-port block RAM for signal <ram>.
```

mode	no-change			
aspect ratio	32-word x 4-bit			
clock	connected to signal <clk>		rise	
enable	connected to signal <en>	high		
write enable	connected to signal <we>		high	
address	connected to signal <addr>			
data in	connected to signal <di>			
data out	connected to signal <do>			
ram_style	Auto			

Summary:

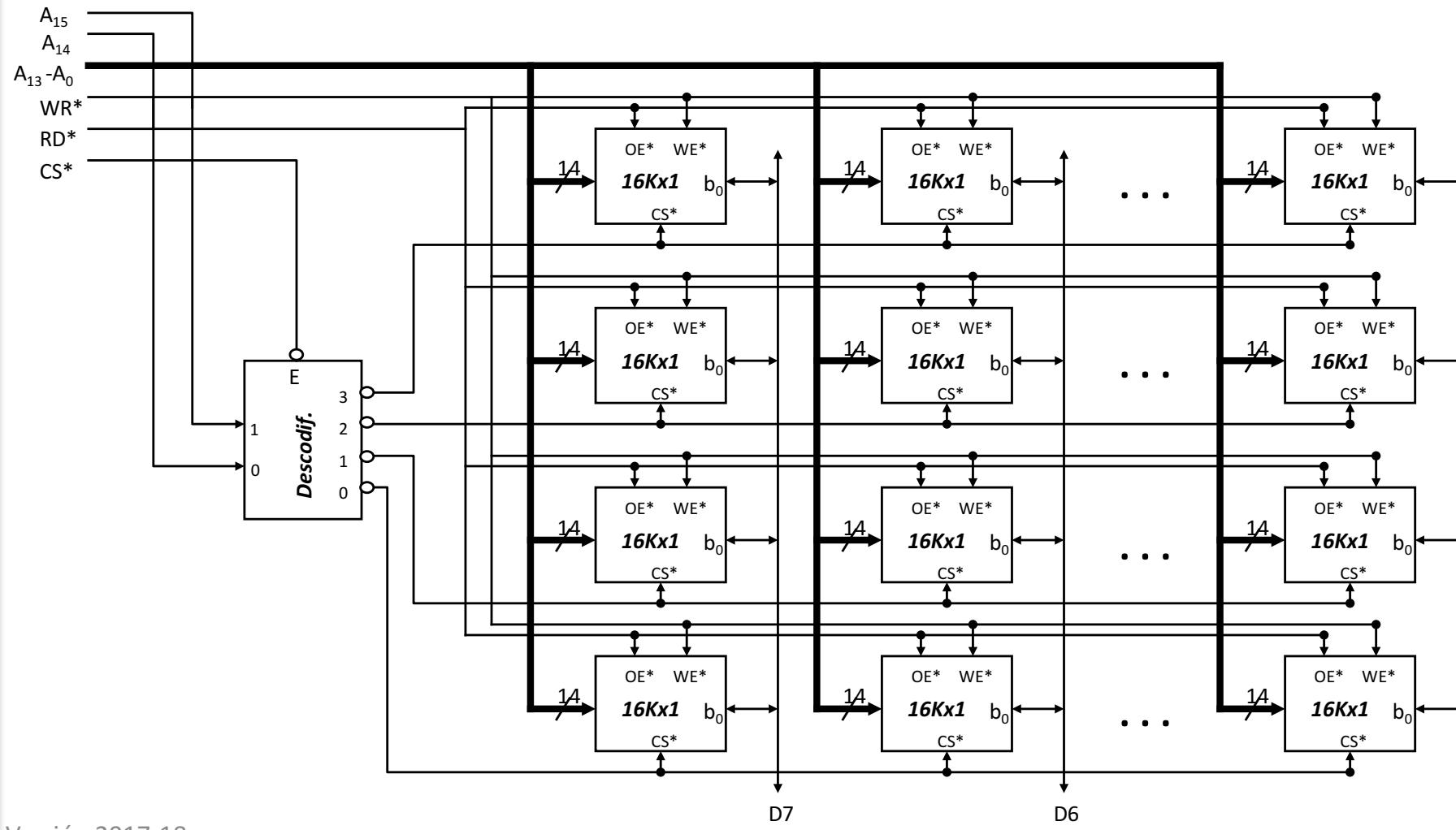
inferred 1 RAM(s).

Unit <blockram1> synthesized.



Diseño SRAM

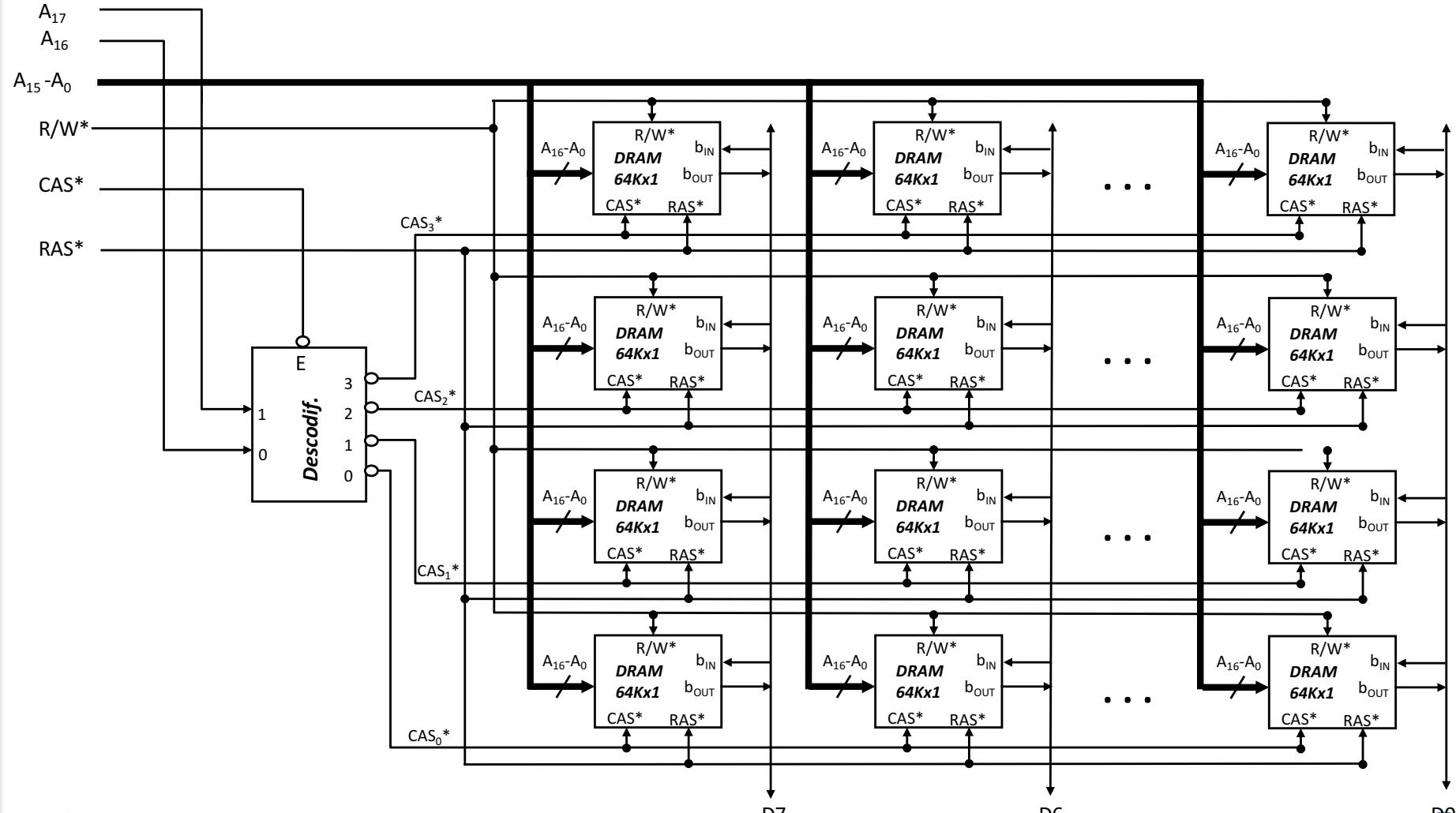
- Diseñar memoria 64Kx8 usando memorias de 16Kx1





Ejemplo diseño DRAM

- Diseñar memoria 256Kx8 usando memoria de 64Kx1





Más ejemplos

- Cualquier memoria puede transformarse prácticamente en cualquier estructura de datos de almacenamiento:
 - FIFO
 - PILA
 - Buffer circular
 - ...



Diseño buffer circular

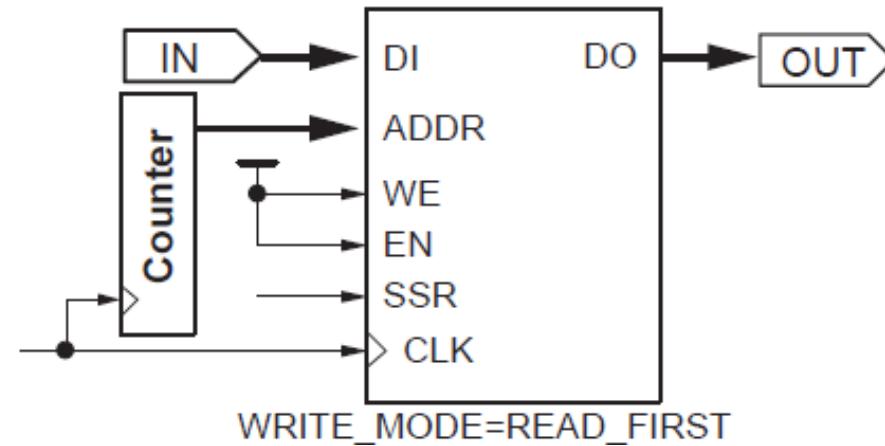
■ Ejercicio:

- Los buffers circulares se utilizan en una variedad de aplicaciones de procesamiento de señales digitales, tales como filtrado de múltiples canales, correlación, FIR correlación cruzada.
- El funcionamiento es el siguiente: un dato se escriben en la memoria y después de n ciclos de reloj (n es igual al número de datos que almacena el buffer circular), el dato almacenado sale fuera del buffer mientras que un nuevo dato se escribe en esa ubicación.



Diseño buffer circular

- Solución:





Disco

- La tecnología de los discos ha avanzado enormemente en los últimos años. Proporciona prácticamente ilimitada capacidad de almacenamiento en línea a muy bajo coste, permite crear software de alta complejidad sin preocuparse por limitación de tamaño, así como libera a los usuarios de tener que ser conscientes de qué y cuántos datos mantener
- Mientras que el almacenamiento en disco no puede ser el motor central de un computador, su dinámica de crecimiento y desarrollo sin duda jugó un papel fundamental en el tremendo avance de los sistemas informáticos desde sus primeros días hasta donde están hoy
- Imagínese lo que el PCsería como si su almacenamiento secundario tenía una capacidad de sólo decenas de megabytes en lugar de las decenas y hasta centenas de gigabytes que ahora damos por sentado



Disco

- Características comunes:
 - El disco de almacenamiento:
 - magnético guarda la información en la polarización del material ferro-magnético
 - DVD o CD guarda la información en distintas reflexiones de la luz
 - Ese material se moldea en forma de disco que puede almacenar información por una o por las dos caras
 - Cabeza lectora: transductor encargado de convertir la señal magnética (o luminosa) en señal eléctrica y viceversa
 - El datos se localiza indicando el radio (moviendo la cabeza) y el ángulo (moviendo el disco)

Disco



■ Rendimiento

- Tiempo de respuesta: desde que se produce una petición hasta que la transferencia se ha completado. Aunque el valor que se suele dar es el tiempo medio de respuesta.
- Tiempo de servicio: es el tiempo de respuesta cuando se hace una petición al disco una vez
- Throughput:
 - Número de operaciones de entrada/salida por segundo IOPS
 - Tamaño de datos transferidos por segundo MB/s
 - IOPS x Tamaño de bloque = MB/s
- Tiempo de respuesta vs throughput



Disco

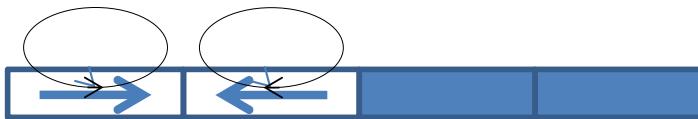
■ Escritura



vacio



Escrito primer compartimento



Escrito segundo compartimento

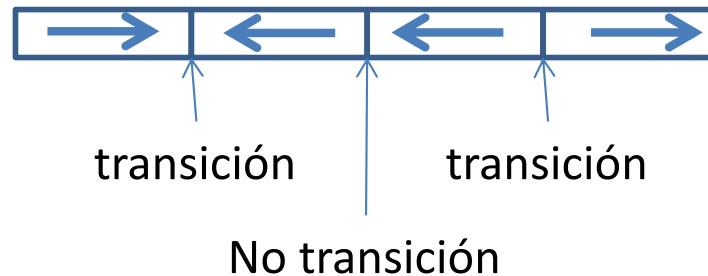


Escritura completa



Disco

Lectura



- Si conseguimos un cabezal capaz de leer con precisión la orientación del material ferro-magnético tendríamos almacenado por ejemplo 0110
- Sin embargo lo normal es que sólo seamos capaces de leer si existe una transición en ese caso tendríamos almacenado 101



Disco

- Driver ATA: www.T13.org
 - Parallel ATA (desde 1986)
 - Serial ATA (desde 2003)
 - Presenta conexión punto a punto
 - Cada ciclo ofrece 1 bit
 - Se utilizan 10 bits para codificar 8 bits

	SATA I	SATA II	SATA III
Frecuencia	1500 MHz	3000 MHz	6000MHz
Velocidad real	150 MB/s	300 MB/s	600 MB/s