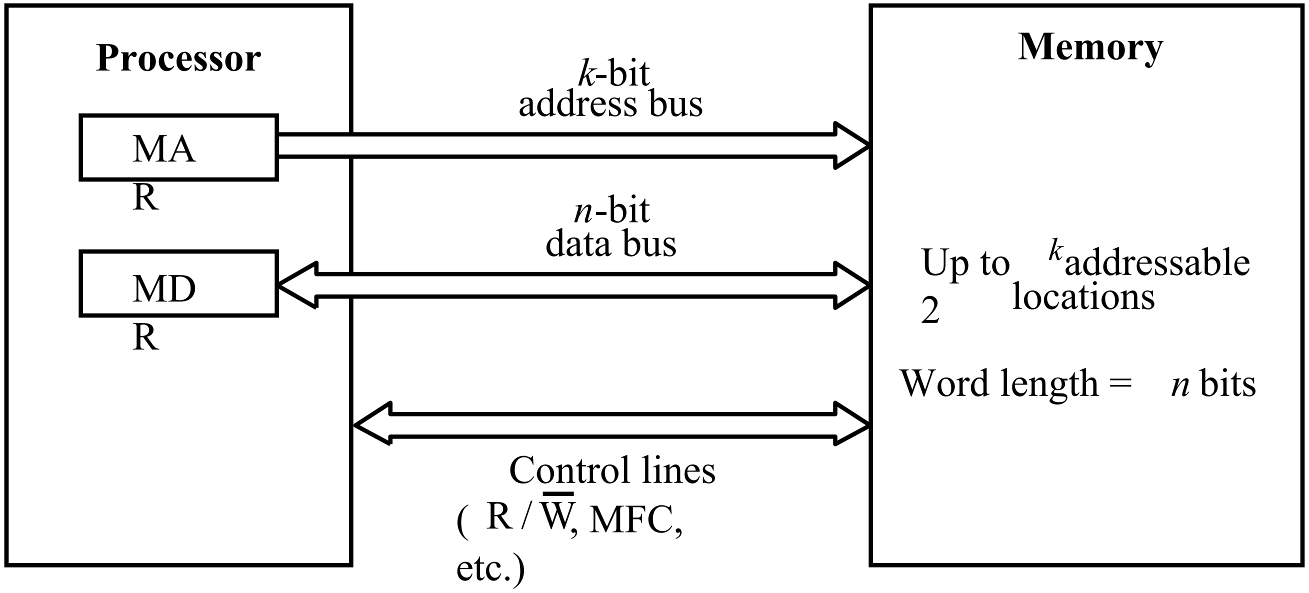
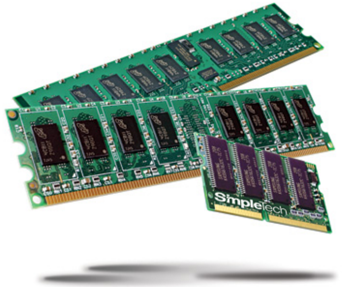
Fundamental Concepts

The Memory System

Some basic concepts

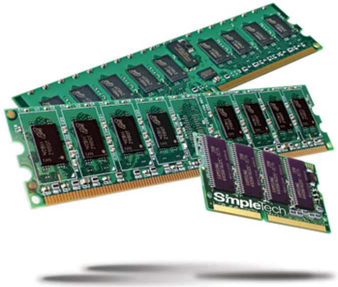
- Maximum size of the Main Memory
- byte-addressable
- CPU-Main Memory Connection





Some basic concepts(Contd.,)

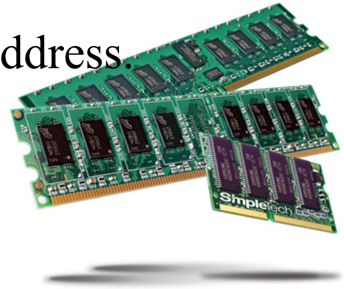
- Measures for the speed of a memory:
  - memory access time : Time that elapses between the initiation of an operation and the completion of that operation
  - memory cycle time: Time delay between initiation of two successive memory operation. Ex: Delay between two successive read operation
- An important design issue is to provide a computer system with as large and fast a memory as possible, within a given cost target.



Some basic concepts(Contd.,)

- Several techniques to increase the effective size and speed of the memory:
  - Cache memory (to increase the effective speed).
  - Cache is small, fast memory that is inserted between Main memory and processor
  - Holds currently active program and the data.

  - Virtual memory (to increase the effective size).
  - Increase the apparent size of the physical memory
  - Data are addressed in a virtual address space that can be as large as the addressing capability of the processor
  - Address generated by the processor is referred as a virtual or logical address.
  - MMU unit translates virtual address into physical address

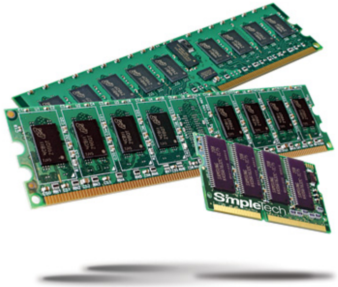


Semiconductor RAM memories

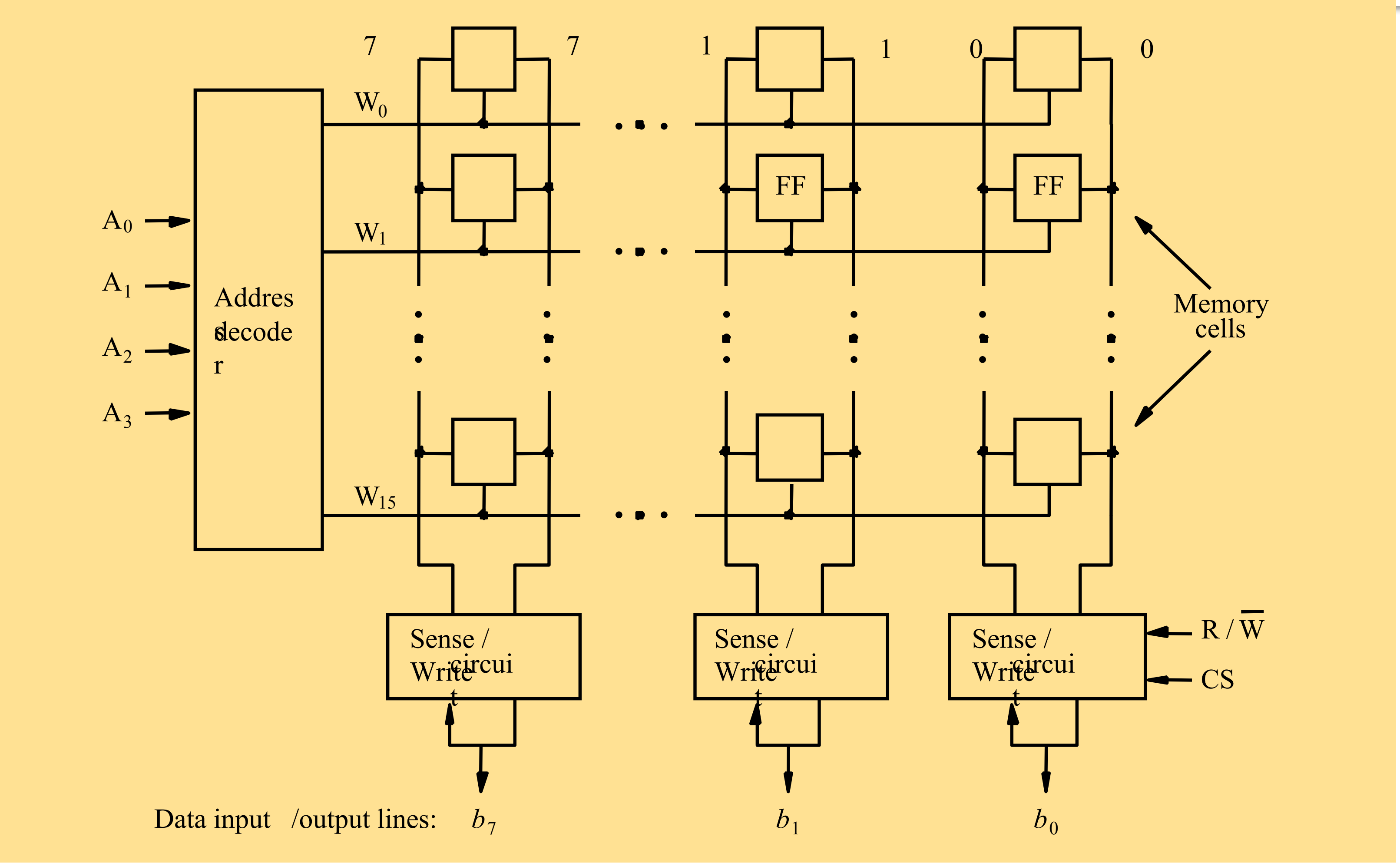
The Memory System

Internal organization of memory chips

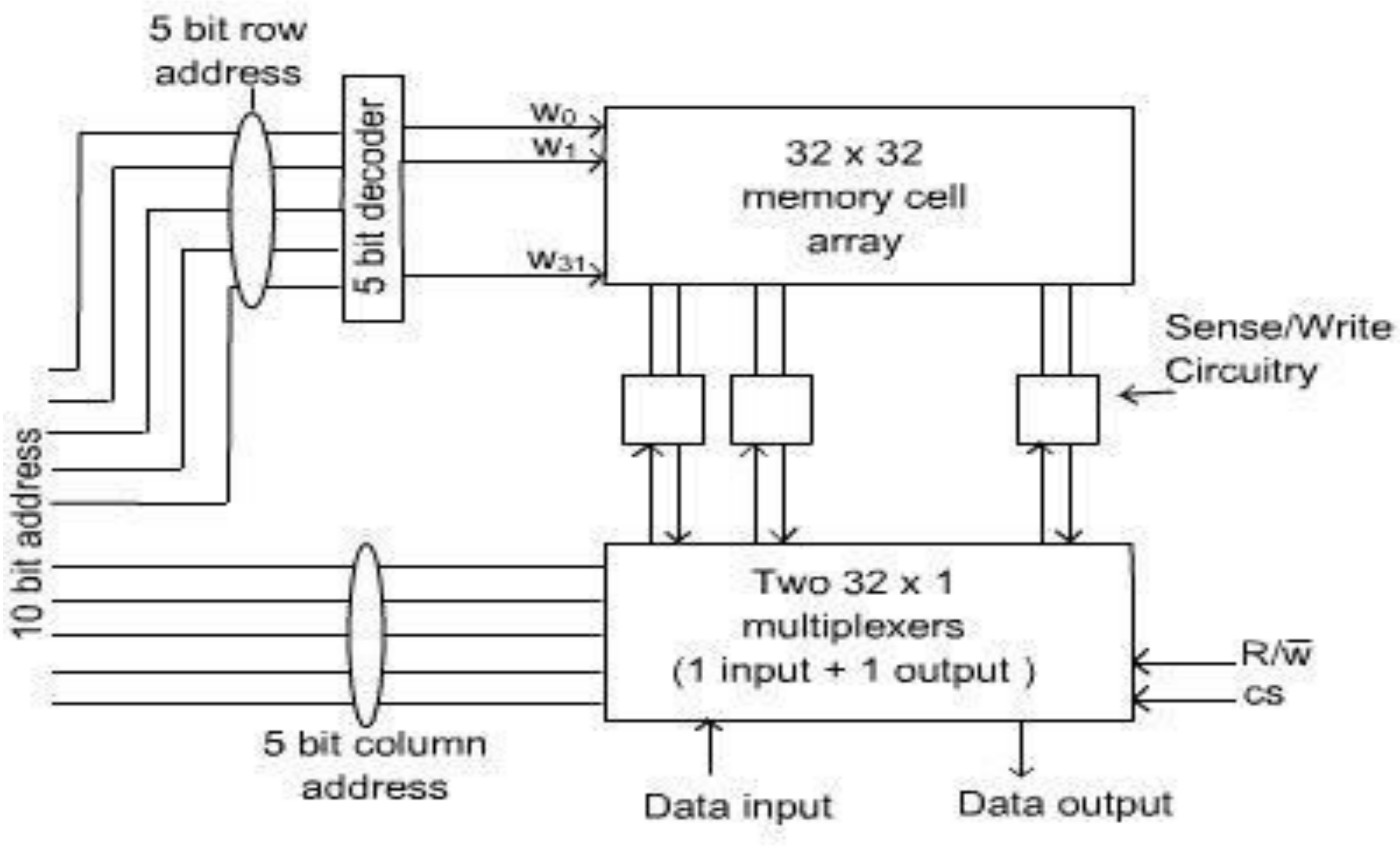
- Each memory cell can hold one bit of information.
- Memory cells are organized in the form of an array.
- One row is one memory word.
- All cells of a row are connected to a common line, known as the "word line".
- Word line is connected to the address decoder.
- Sense/write circuits are connected to the data input/output lines of the memory chip.



Internal organization of memory chips (Contd.,)



Organization of 1k x 1 memory

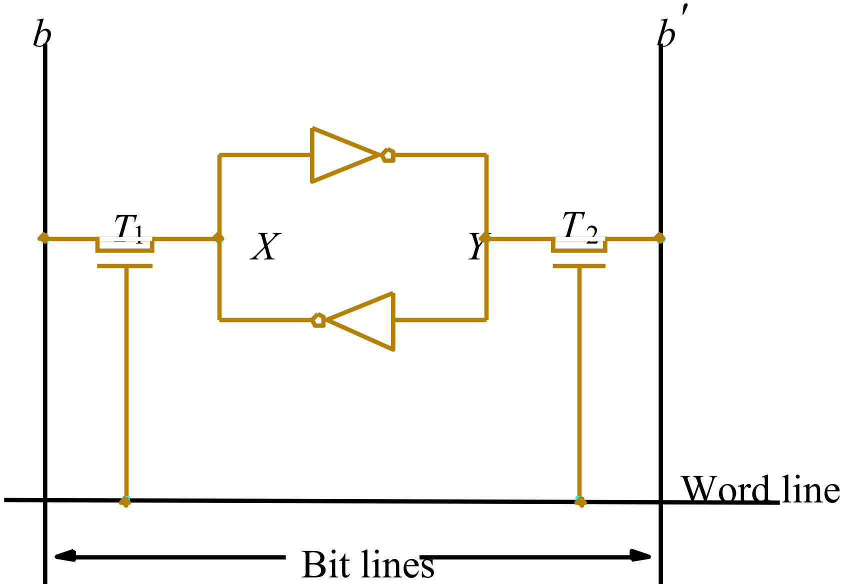


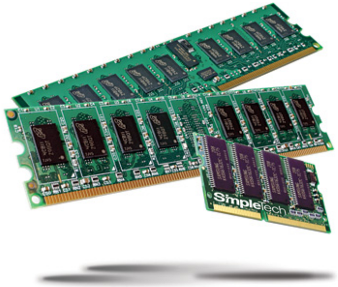
STATIC MEMORIES

- Memories that consist of circuits capable of retaining their state as long as power is applied.
- Memory can be accessed very quickly
- Access time is few nanoseconds
- SRAMs are used in applications where speed is of critical concern.

SRAM Cell

- Two transistor inverters are cross connected to implement a basic flip-flop.
- The cell is connected to one word line and two bits lines by transistors T1 and T2
- When word line is at ground level, the transistors are turned off and the latch retains its state



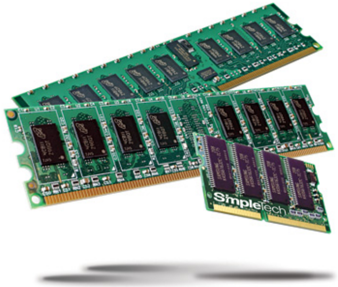


SRAM CELL continued…..

Read operation: In order to read state of SRAM cell, the word line is activated to close switches T1 and T2. Sense/Write circuits at the bottom monitor the state of b and b’ and set the output accordingly.

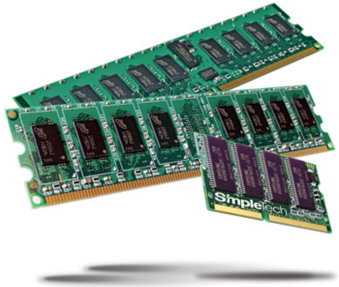
write Operation:

Set appropriate value on line b and b’, and then activate the word line. The required signals on the bit line are generated by sense/write circuit. The value on line b is retained in the cell.



Asynchronous DRAMs

- Static RAMs (SRAMs):
  - Consist of circuits that are capable of retaining their state as long as the power is applied.
  - Volatile memories, because their contents are lost when power is interrupted.
  - Access times of static RAMs are in the range of few nanoseconds.
  - However, the cost is usually high.
- Dynamic RAMs (DRAMs):
  - Do not retain their state indefinitely.
  - Contents must be periodically refreshed.
  - Contents may be refreshed while accessing them for reading.



Asynchronous DRAMS

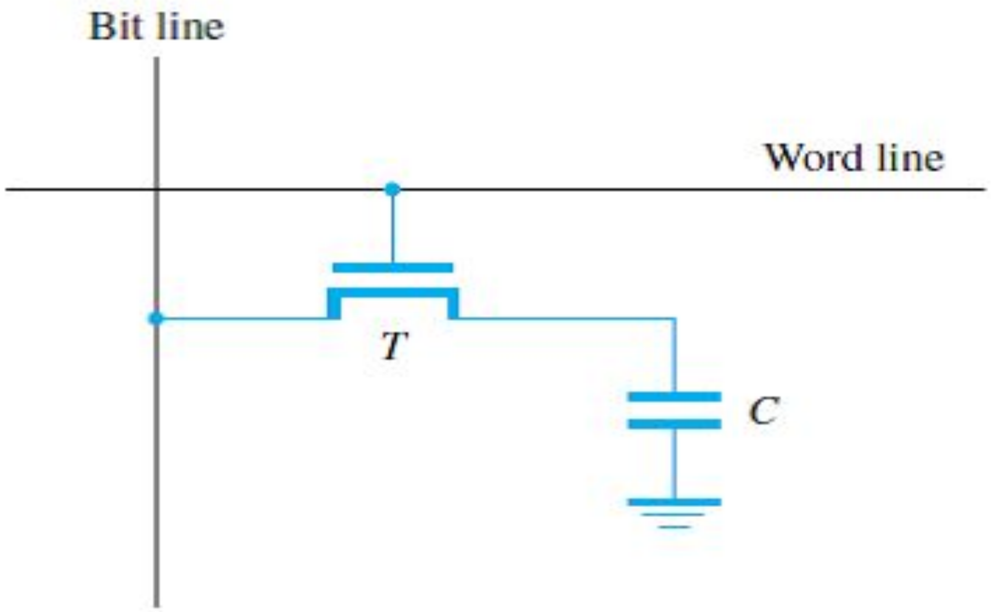
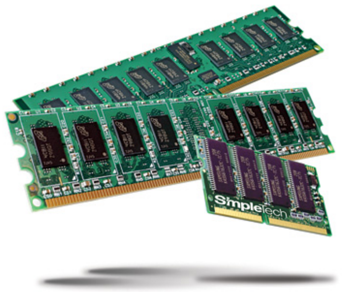
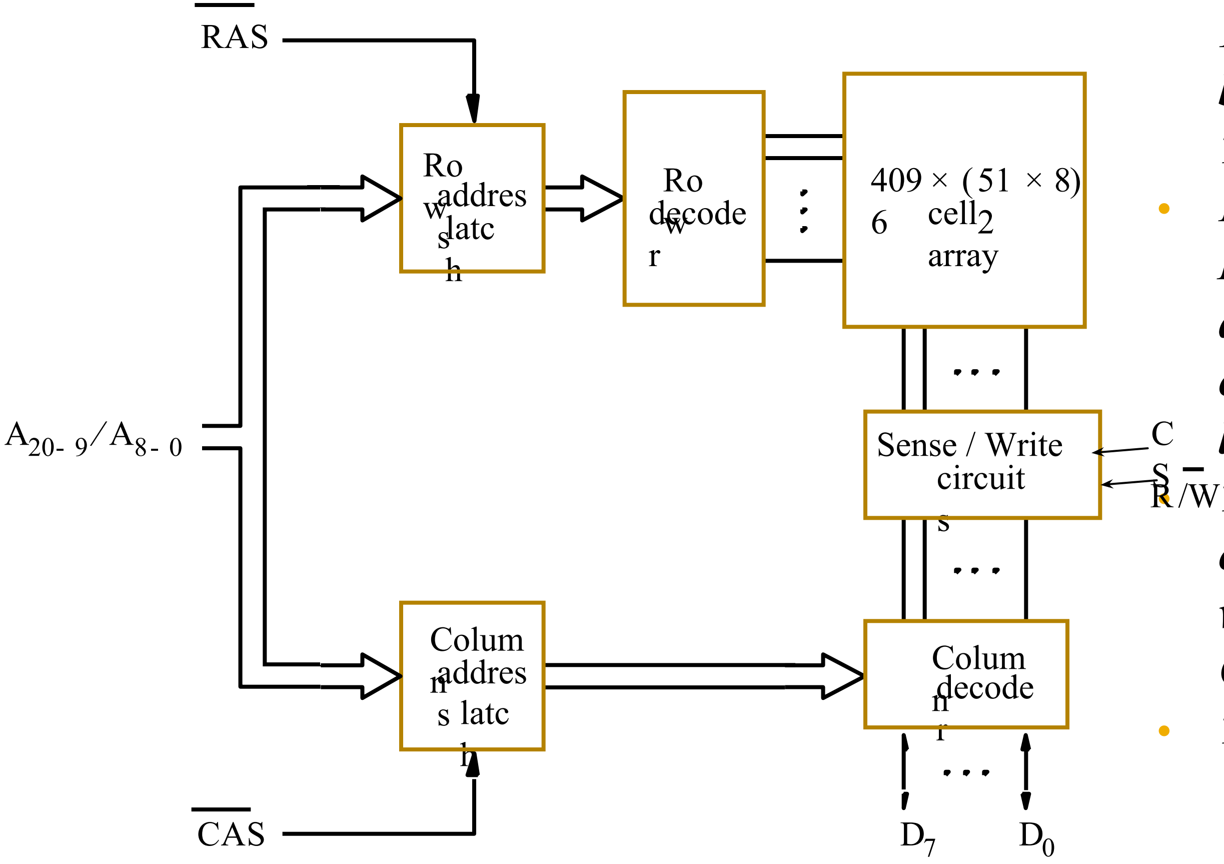


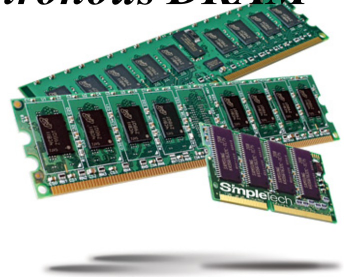
Figure 8.6 A single-transistor dynamic memory cell.



Asynchronous DRAMs

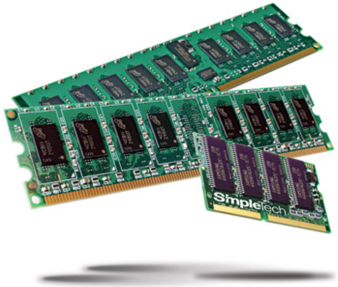


- *Each row can store 512 bytes. 12 bits to select a row, and 9 bits to select a group in a row. Total of 21 bits.*
- *First apply the row address, RAS signal latches the row address. Then apply the column address, CAS signal latches the address.*
- *Timing of the memory unit is controlled by a specialized unit which generates RAS and CAS.*
- *This is asynchronous DRAM*

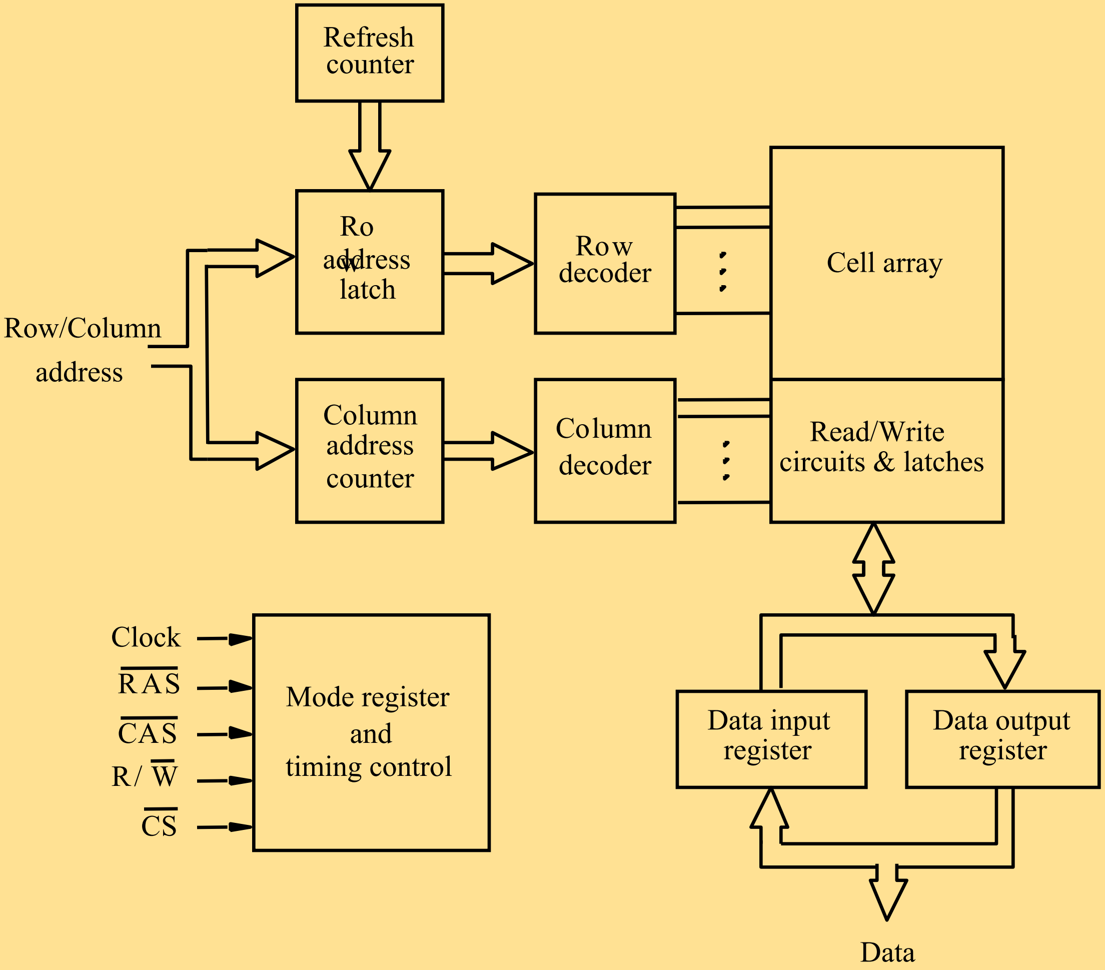


Fast Page Mode

- Suppose if we want to access the consecutive bytes in the selected row.
- This can be done without having to reselect the row.
  - Add a latch at the output of the sense circuits in each row.
  - All the latches are loaded when the row is selected.
  - Different column addresses can be applied to select and place different bytes on the data lines.
- Consecutive sequence of column addresses can be applied under the control signal CAS, without reselecting the row.
  - Allows a block of data to be transferred at a much faster rate than random accesses.
  - A small collection/group of bytes is usually referred to as a block.
- This transfer capability is referred to as the fast page mode feature.

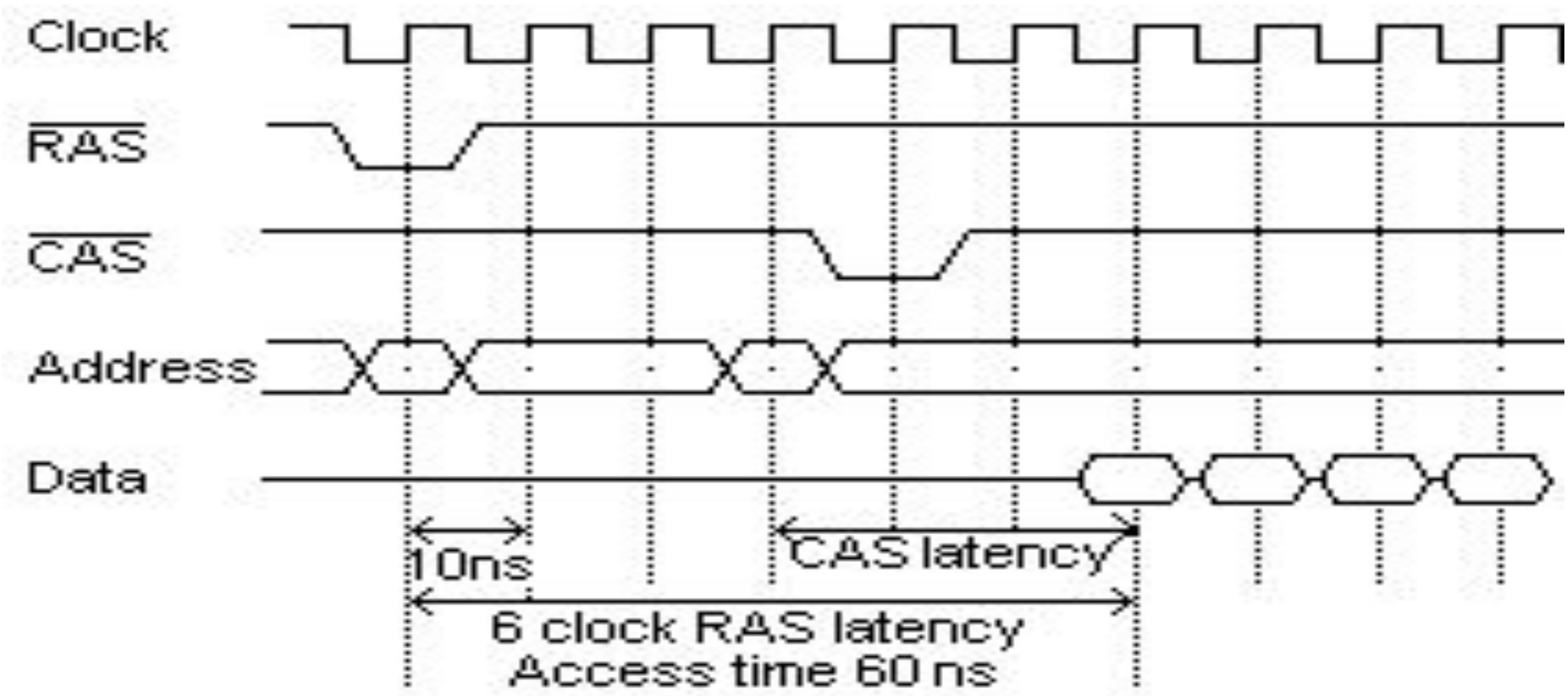


Synchronous DRAMs

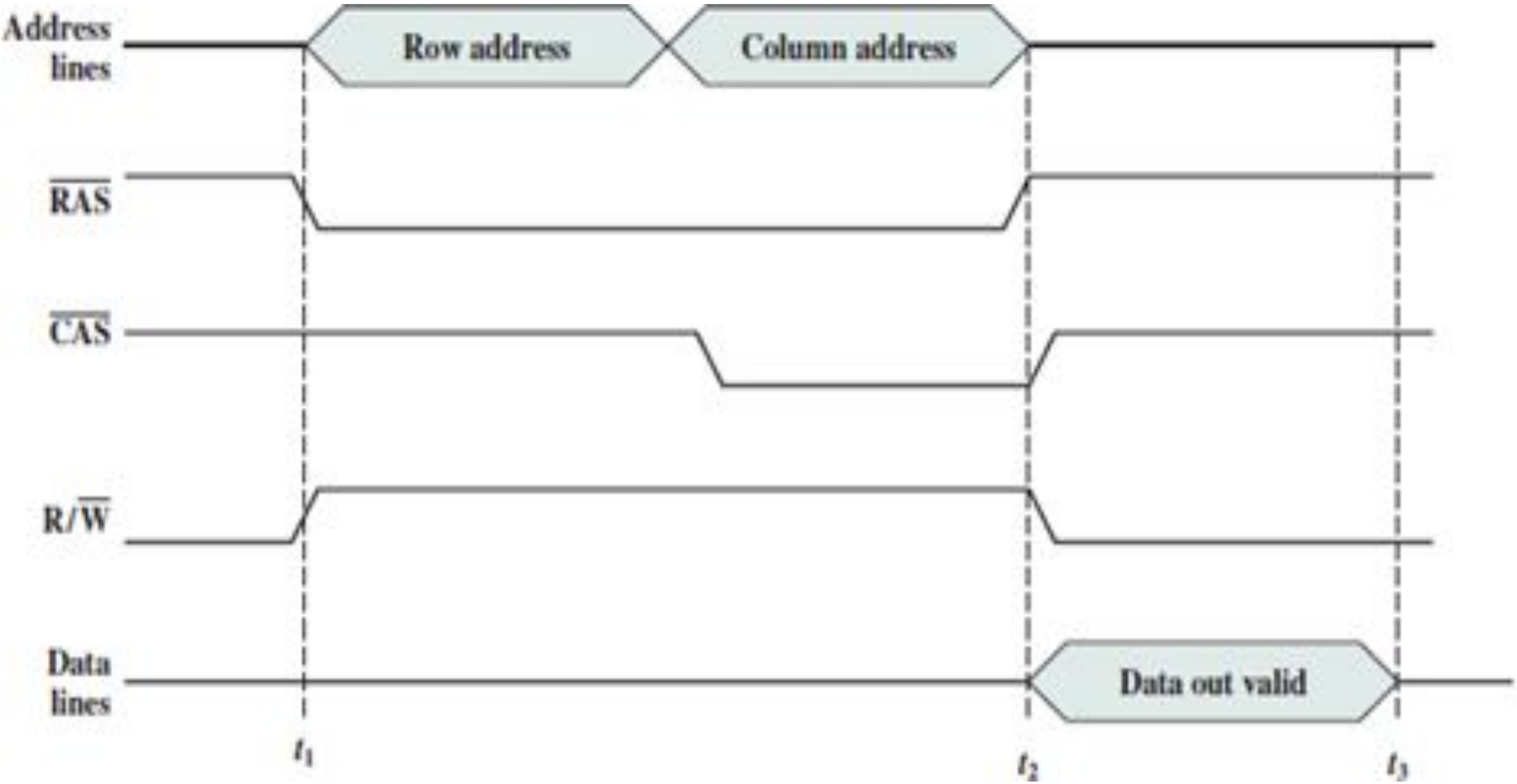


- *Operation is directly synchronized with processor clock signal.*
- *The outputs of the sense circuits are connected to a latch.*
- *During a Read operation, the contents of the cells in a row are loaded onto the latches.*
- *During a refresh operation, the contents of the cells are refreshed without changing the contents of the latches.*
- *Data held in the latches correspond to the selected columns are transferred to the output.*
- *For a burst mode of operation, successive columns are selected using column address counter and clock. CAS signal need not be generated externally. A new data is placed during raising edge of the clock*

Synchronous DRAM Timing Diagram

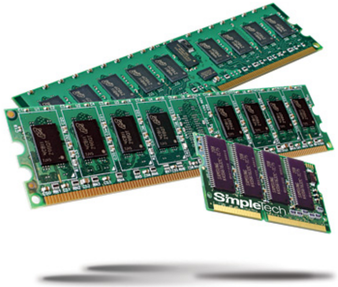


Asynchronous DRAM Timing Diagram continued ….

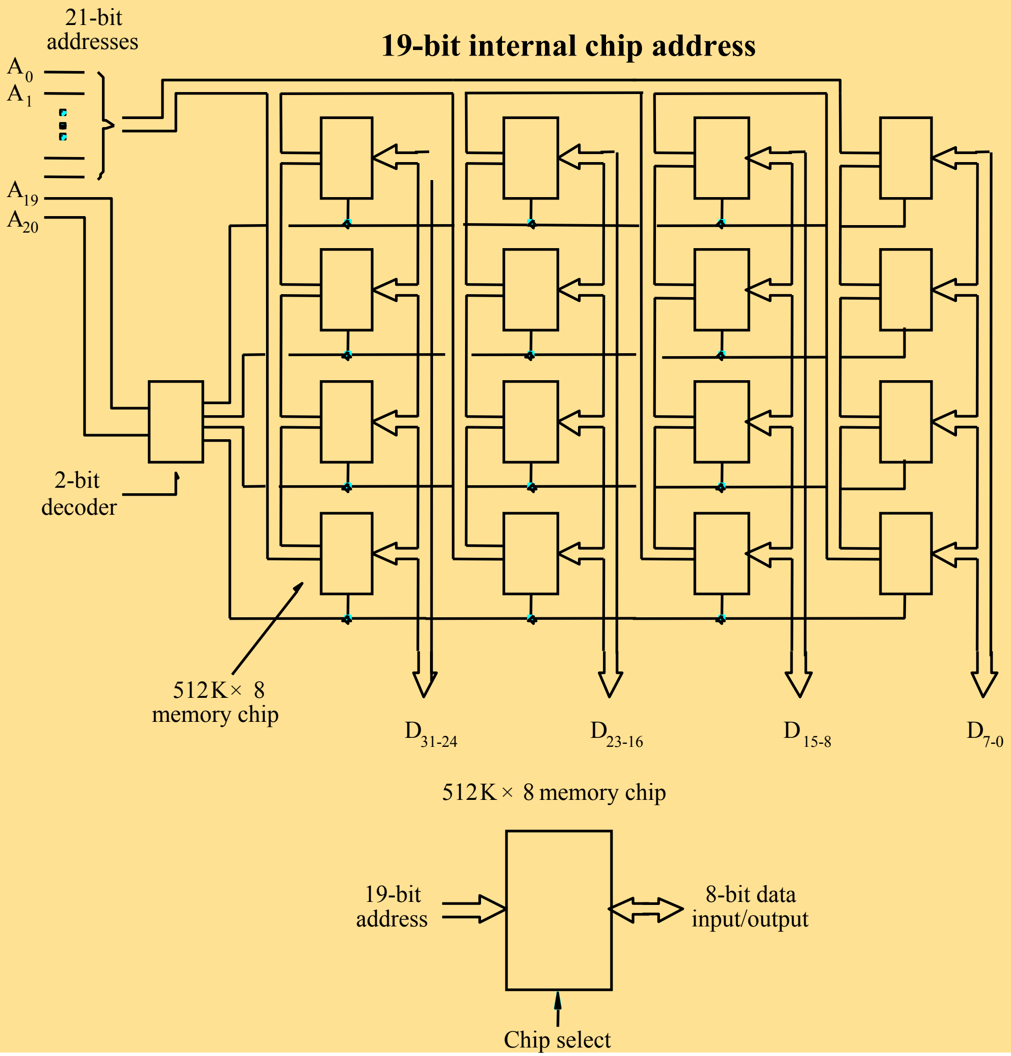


Latency, Bandwidth, and DDRSDRAMs

- Memory latency is the time it takes to transfer a word of data to or from memory
- Memory bandwidth is the number of bits or bytes that can be transferred in one second.
- DDRSDRAMs
  - Cell array is organized in two banks



Static memories



Implement a memory unit of 2M words of 32 bits each.

Use 512x8 static memory chips.

Each column consists of 4 chips.

Each chip implements one byte position.

A chip is selected by setting its chip select control line to 1.

Selected chip places its data on the data output line, outputs of other chips are in high impedance state.

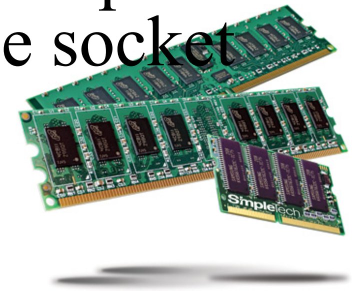
21 bits to address a 32-bit word.

High order 2 bits are needed to select the row, by activating the four Chip Select signals.

19 bits are used to access specific byte locations inside the selected chip.

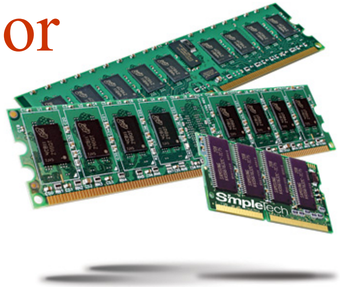
Dynamic memories

- Large dynamic memory systems can be implemented using DRAM chips in a similar way to static memory systems.
- Placing large memory systems directly on the motherboard will occupy a large amount of space.
  - Also, this arrangement is inflexible since the memory system cannot be expanded easily.
- Packaging considerations have led to the development of larger memory units known as SIMMs (Single In-line Memory Modules) and DIMMs (Dual In-line Memory Modules).
- Memory modules are an assembly of memory chips on a small board that plugs vertically onto a single socket on the motherboard.
  - Occupy less space on the motherboard.
  - Allows for easy expansion by replacement.

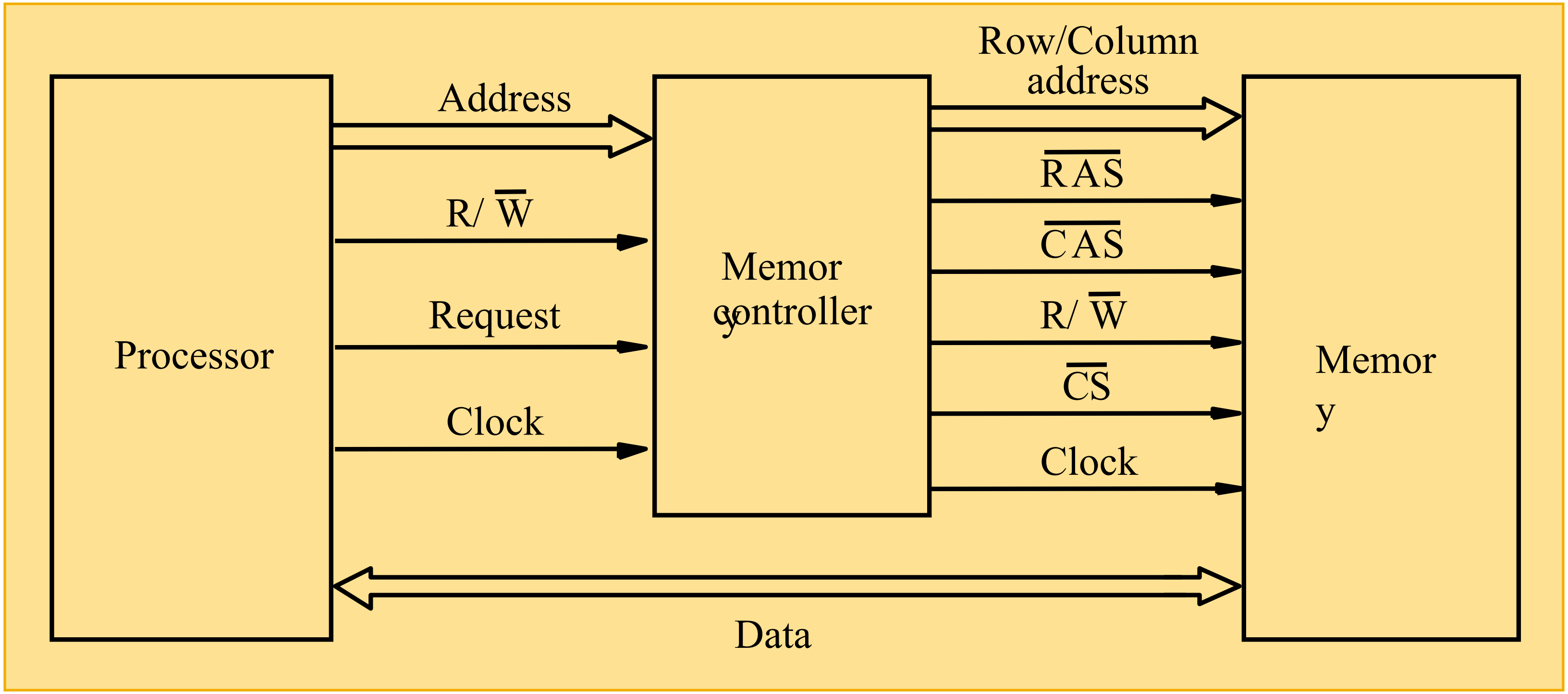


Memory controller

- Recall that in a dynamic memory chip, to reduce the number of pins, multiplexed addresses are used.
- Address is divided into two parts:
  - High-order address bits select a row in the array.
  - They are provided first, and latched using RAS signal.
  - Low-order address bits select a column in the row.
  - They are provided later, and latched using CAS signal.
- However, a processor issues all address bits at the same time.
- In order to achieve the multiplexing, memory controller circuit is inserted between the processor and memory.



Memory controller (contd..)

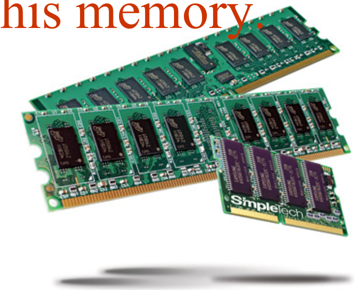


Read-Only Memories (ROMs)

The Memory System

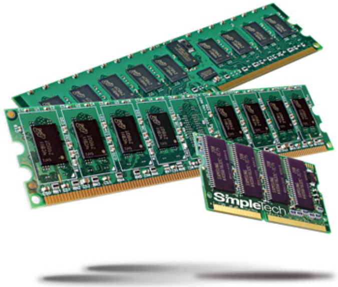
Read-Only Memories (ROMs)

- SRAM and SDRAM chips are volatile:
  - Lose the contents when the power is turned off.
- Many applications need memory devices to retain contents after the power is turned off.
  - For example, computer is turned on, the operating system must be loaded from the disk into the memory.
  - Store instructions which would load the OS from the disk.
  - Need to store these instructions so that they will not be lost after the power is turned off.
  - We need to store the instructions into a non-volatile memory.
- Non-volatile memory is read in the same manner as volatile memory.
  - Separate writing process is needed to place information in this memory.
  - Normal operation involves only reading of data, this type of memory is called Read-Only memory (ROM).



Read-Only Memories (Contd.,)

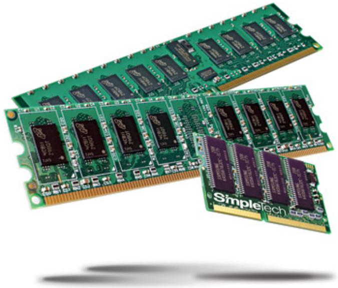
- Read-Only Memory:
  - Data are written into a ROM when it is manufactured.
- Programmable Read-Only Memory (PROM):
  - Allow the data to be loaded by a user.
  - Process of inserting the data is irreversible.
  - Storing information specific to a user in a ROM is expensive.
  - Providing programming capability to a user may be better.
- Erasable Programmable Read-Only Memory (EPROM):
  - Stored data to be erased and new data to be loaded.
  - Flexibility, useful during the development phase of digital systems.
  - Erasable, reprogrammable ROM.
  - Erasure requires exposing the ROM to UV light.



Read-Only Memories (Contd.,)

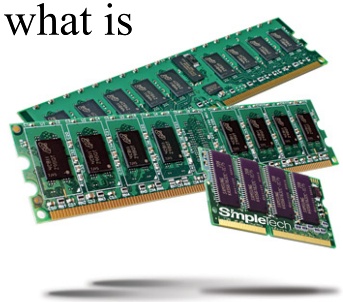
- Electrically Erasable Programmable Read-Only Memory (EEPROM):
  - To erase the contents of EPROMs, they have to be exposed to ultraviolet light.
  - Physically removed from the circuit.
  - EEPROMs the contents can be stored and erased electrically.
- Flash memory:
  - Has similar approach to EEPROM.
  - Read the contents of a single cell, but write the contents of an entire block of cells.
  - Flash devices have greater density.
    - Higher capacity and low storage cost per bit.
  - Power consumption of flash memory is very low, making it attractive for use in equipment that is battery-driven.
  - Single flash chips are not sufficiently large, so

  larger memory modules are implemented using flash cards and flash drives.

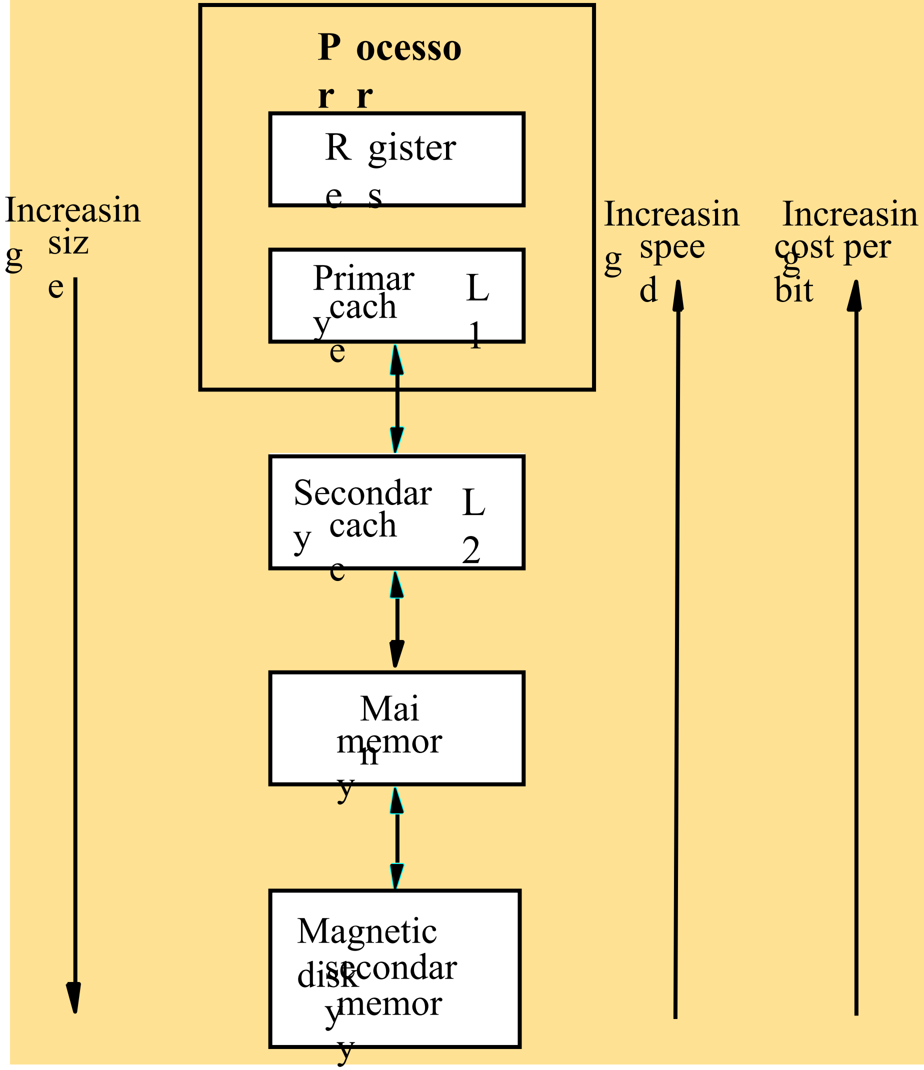


Speed, Size, and Cost

- A big challenge in the design of a computer system is to provide a sufficiently large memory, with a reasonable speed at an affordable cost.
- Static RAM:
  - Very fast, but expensive, because a basic SRAM cell has a complex circuit making it impossible to pack a large number of cells onto a single chip.
- Dynamic RAM:
  - Simpler basic cell circuit, hence are much less expensive, but significantly slower than SRAMs.
- Magnetic disks:
  - Storage provided by DRAMs is higher than SRAMs, but is still less than what is necessary.
  - Secondary storage such as magnetic disks provide a large amount of storage, but is much slower than DRAMs.



Memory Hierarchy



- *Fastest access is to the data held in processor registers. Registers are at the top of the memory hierarchy.*
- *Relatively small amount of memory that can be implemented on the processor chip. This is processor cache.*
- *Two levels of cache. Level 1 (L1) cache is on the processor chip. Level 2 (L2) cache is in between main memory and processor.*
- *Next level is main memory, implemented as SIMMs. Much larger, but much slower than cache memory.*
- *Next level is magnetic disks. Huge amount of inexepensive storage.*
- *Speed of memory access is critical, the idea is to bring instructions and data that will be used in the near future as close to the processor as possible.*

Cache Memories

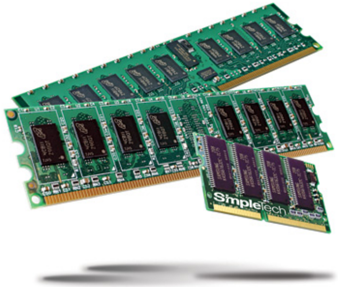
The Memory System

Cache Memories

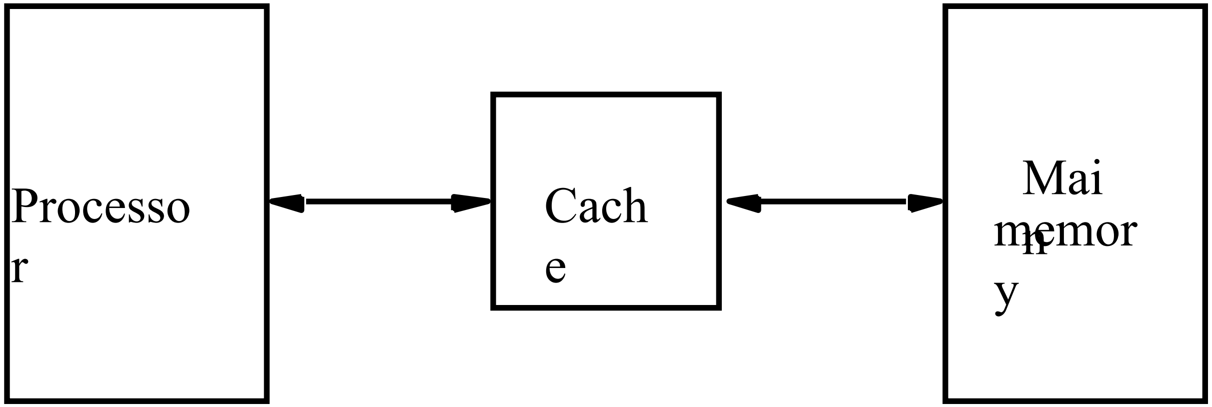
- Processor is much faster than the main memory.
  - As a result, the processor has to spend much of its time waiting while instructions and data are being fetched from the main memory.
  - Major obstacle towards achieving good performance.
- Speed of the main memory cannot be increased beyond a certain point.
- Cache memory is an architectural arrangement which makes the main memory appear faster to the processor than it really is.
- Cache memory is based on the property of computer programs known as “locality of reference”.

Locality of Reference

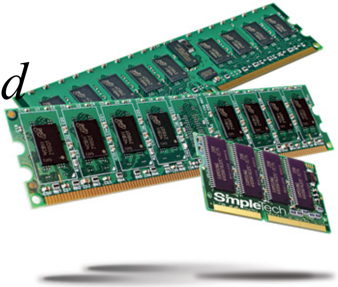
- Analysis of programs indicates that many instructions in localized areas of a program are executed repeatedly during some period of time, while the others are accessed relatively less frequently.
  - These instructions may be the ones in a loop, nested loop or few procedures calling each other repeatedly.
  - This is called "locality of reference".
- Temporal locality of reference:
  - Recently executed instruction is likely to be executed again very soon.
- Spatial locality of reference:
  - Instructions with addresses close to a recently instruction are likely to be executed soon.



Cache memories



- *Processor issues a Read request, a block of words is transferred from the main memory to the cache, one word at a time.*
- *Subsequent references to the data in this block of words are found in the cache.*
- *At any given time, only some blocks in the main memory are held in the cache. Which blocks in the main memory are in the cache is determined by a "mapping function".*
- *When the cache is full, and a block of words needs to be transferred from the main memory, some block of words in the cache must be replaced. This is determined by a "replacement algorithm".*



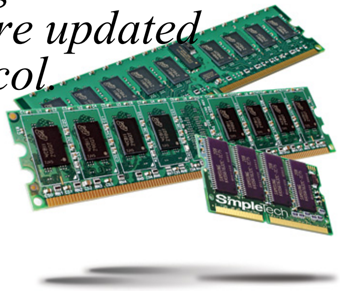
Cache hit

- *Existence of a cache is transparent to the processor. The processor issues Read and Write requests in the same manner.*

- *If the data is in the cache it is called a Read or Write hit.*

- *Read hit:*
  - *The data is obtained from the cache.*

- *Write hit:*
  - *Cache has a replica of the contents of the main memory.*
  - *Contents of the cache and the main memory may be updated simultaneously. This is the write-through protocol.*
  - *Update the contents of the cache, and mark it as updated by setting a bit known as the dirty bit or modified bit. The contents of the main memory are updated when this block is replaced. This is write-back or copy-back protocol.*

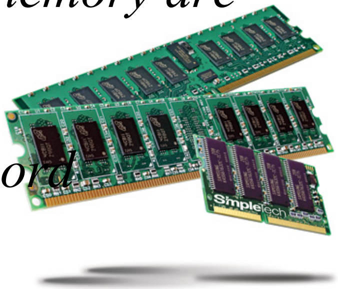


Cache miss

- *If the data is not present in the cache, then a Read miss or Write miss occurs.*

- *Read miss:*
  - *Block of words containing this requested word is transferred from the memory.*
  - *After the block is transferred, the desired word is forwarded to the processor.*
  - *The desired word may also be forwarded to the processor as soon as it is transferred without waiting for the entire block to be transferred. This is called load-through or early-restart.*

- *Write-miss:*
  - *Write-through protocol is used, then the contents of the main memory are updated directly.*
  - *If write-back protocol is used, the block containing the addressed word is first brought into the cache. The desired word is overwritten with new information.*

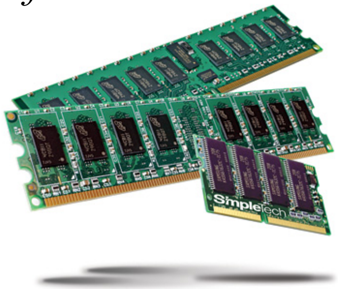


Cache Coherence Problem

- *A bit called as "valid bit" is provided for each block.*
- *If the block contains valid data, then the bit is set to 1, else it is 0.*
- *Valid bits are set to 0, when the power is just turned on.*
- *When a block is loaded into the cache for the first time, the valid bit is set to 1.*

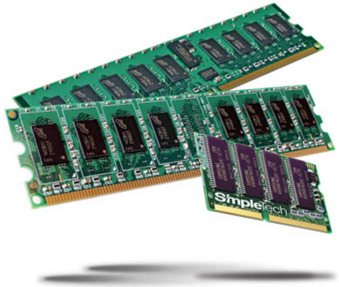
- *Data transfers between main memory and disk occur directly bypassing the cache.*
- *When the data on a disk changes, the main memory block is also updated.*
- *However, if the data is also resident in the cache, then the valid bit is set to 0.*

- *What happens if the data in the disk and main memory changes and the write-back protocol is being used?*
- *In this case, the data in the cache may also have changed and is indicated by the dirty bit.*
- *The copies of the data in the cache, and the main memory are different. This is called the cache coherence problem.*
- *One option is to force a write-back before the main memory is updated from the disk.*

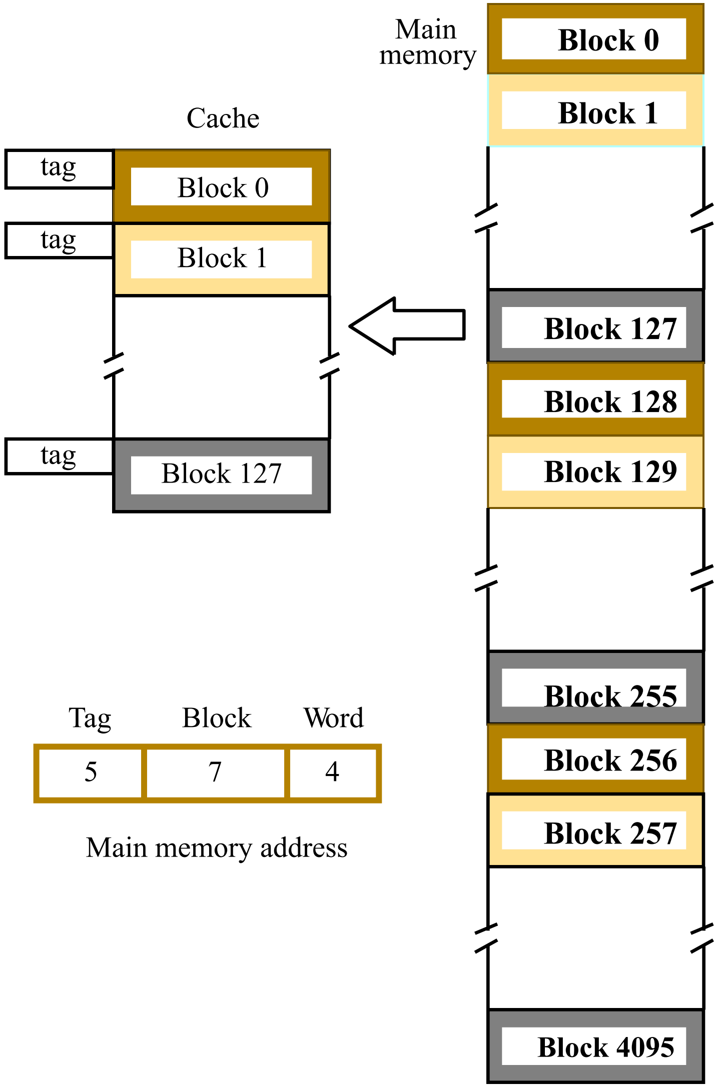


Mapping functions

- Mapping functions determine how memory blocks are placed in the cache.
- A simple processor example:
  - Cache consisting of 128 blocks of 16 words each.
  - Total size of cache is 2048 (2K) words.
  - Main memory is addressable by a 16-bit address.
  - Main memory has 64K words.
  - Main memory has 4K blocks of 16 words each.
- Three mapping functions:
  - **Direct mapping**
  - **Associative mapping**
  - **Set-associative mapping.**



Direct mapping



- *Block j of the main memory maps to j modulo 128 of the cache. 0 maps to 0, 129 maps to 1.*
- *More than one memory block is mapped onto the same position in the cache.*
- *May lead to contention for cache blocks even if the cache is not full.*
- *Resolve the contention by allowing new block to replace the old block, leading to a trivial replacement algorithm.*
- *Memory address is divided into three fields:*
  - *Low order 4 bits determine one of the 16 words in a block.*
  - *When a new block is brought into the cache, the the next 7 bits determine which cache block is placed in.*
  - *High order 5 bits determine which of the possible 32 blocks is currently present in the cache. These are tag bits.*
- *Simple to implement but not very flexible.*

BLOCK 0:

0	128	256	384	512	640	768	896	1024	1152
1280	1408	1536	1164	1792	1920	2048	2176	2304	2432
2560	2688	2816	2944	3072	3200	3328	3456	3584	3712
3840	3968								

BLOCK 1:

1	129	257	385	513	641	769	897	1025	1153
1281	1409	1537	1165	1793	1921	2049	2177	2305	2433
2561	2689	2817	2945	3071	3201	3329	3457	3585	3713
3841	3969								

Examples of 12-bit address generated by CPU

Block 0

Address	Blocks from MM
00000	Block 0
00001	Block 128
00010	Block 256
00011	Block 384
00100	Block 512
00101	Block 640
00110	Block 768
00111	Block 896
01000	Block 1024
01001	Block 1152

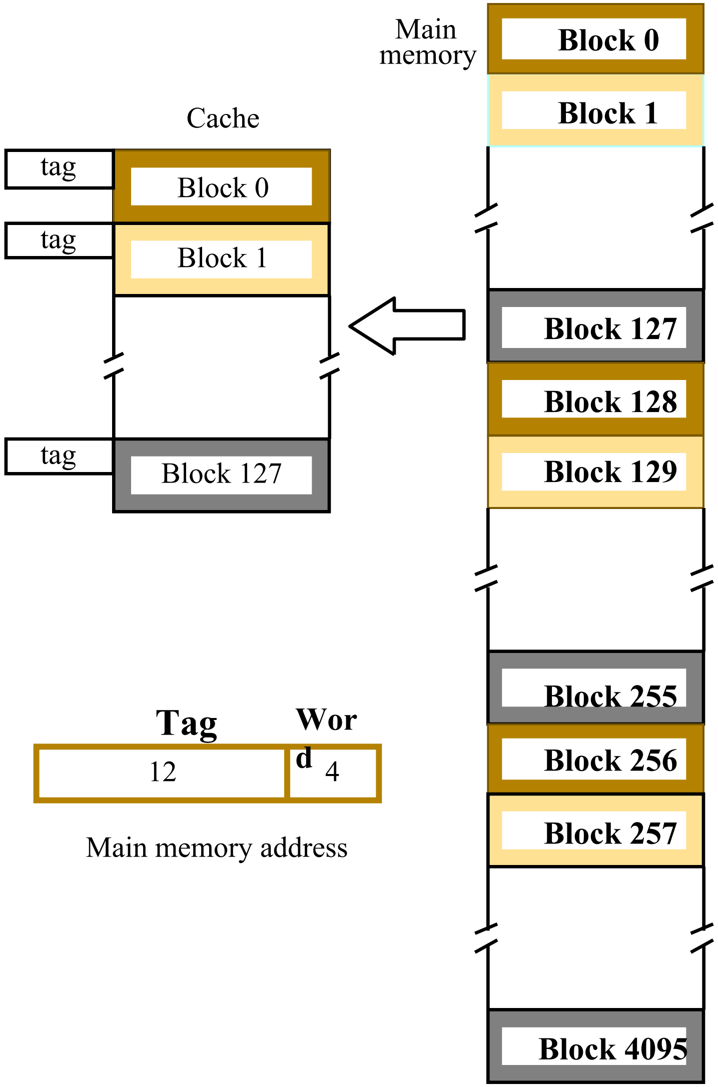
Address	Blocks from MM
01010	Block 1280
01011	Block 1408
01100	Block 1536
01101	Block 1164
01110	Block 1792
01111	Block 1920
10000	Block2048
10001	Block 2176
10010	Block 2304
10011	Block 2432

Address	Blocks from MM
10100	Block 2560
10101	Block 2688
10110	Block 2816
10111	Block 2944
11000	Block 3072
11001	Block3200
11010	Block 3328
11011	Block 3456
11100	Block 3584
11101	Block 3712
11110	Block3840
11111	Block 3968

Examples of 12-bit address generated by CPU

TAG BITS – 5Bits	BLOCK -7Bits	WORD -4 Bits
00000	0000000	0010
11111	1000001	1100
00011	1111111	0011

Associative mapping

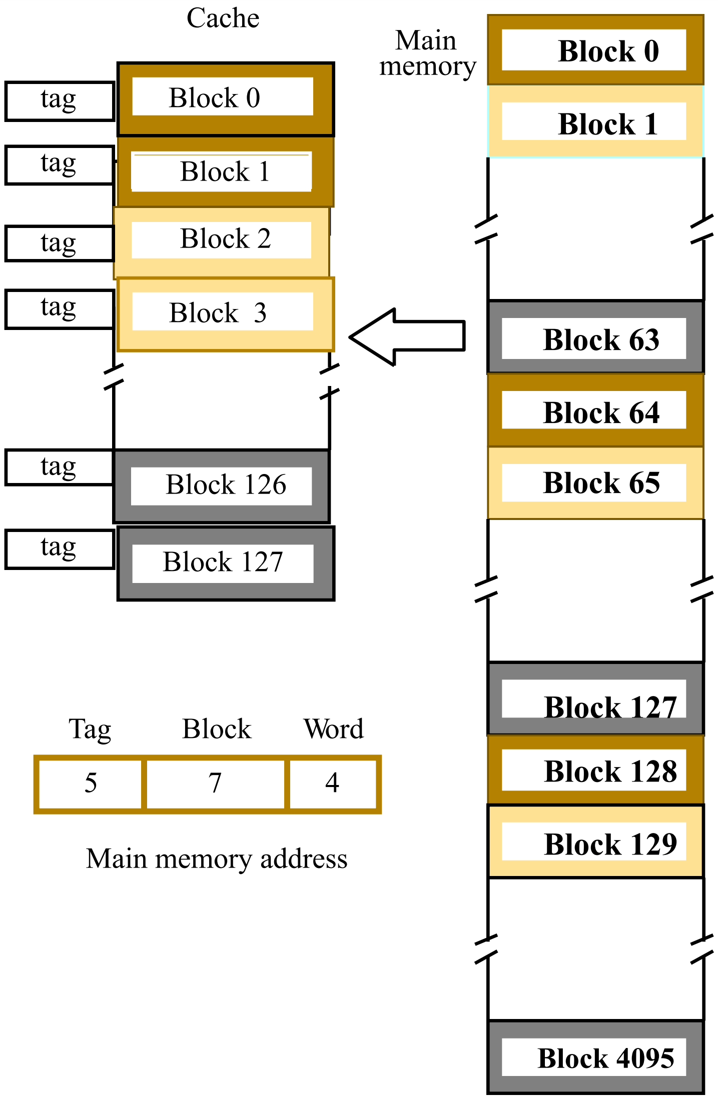


- *Main memory block can be placed into any cache position.*
- *Memory address is divided into two fields:*

  - Low order 4 bits identify the word within a block.

  - High order 12 bits or tag bits identify a memory block when it is resident in the cache.
- *Flexible, and uses cache space efficiently.*
- *Replacement algorithms can be used to replace an existing block in the cache when the cache is full.*
- *Cost is higher than direct-mapped cache because of the need to search all 128 patterns to determine whether a given block is in the cache.*

Set-Associative mapping



•Blocks of cache are grouped into sets.

•Mapping function allows a block of the main memory to reside in any block of a specific set.

•Divide the cache into 64 sets, with two blocks per set.

•Memory block 0, 64, 128 etc. map to block 0, and they can occupy either of the two positions.

•Memory address is divided into three fields:

- 6 bit field determines the set number.

- High order 6 bit fields are compared to the tag fields of the two blocks in a set.

•Set-associative mapping combination of direct and associative mapping.

•Number of blocks per set is a design parameter.

- One extreme is to have all the blocks in one set, requiring no set bits (fully associative mapping).

- Other extreme is to have one block per set, is the same as direct mapping.