4 –way module interleaving

0000	0001	0010	0011	0100	0101	0110									1111
0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15

	M0(00)	M1(01)	M2(10)	M3(11)
00	0	1	2	3
01	4	5	6	7
10	8	9	10	11
11	12	13	14	15

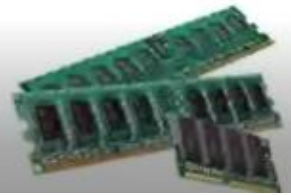
low order memory

M0(00)	M1(01)	M2(10)	M3(11)
0	4	8	12
1	5	9	13
2	6	10	14
3	7	11	15

high order memory

LSB-address of module
MSB-address of words

MSB-address of module
LSB-address of words



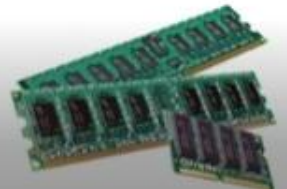
Hit Rate and Miss Penalty

- The number of hits stated as a fraction of all attempted accesses is called the **Hit Rate**.
- The extra time needed to bring the desired information into the cache is called the **Miss Penalty**.
- High hit rates well over 0.9 are essential for high-performance computers.
- Performance is adversely affected by the actions that need to be taken when a miss occurs.
- A performance penalty is incurred because
- of the extra time needed to bring a block of data from a slower unit to a faster unit.
- During that period, the processor is stalled waiting for instructions or data.
- We refer to the total access time seen by the processor when a miss occurs as the miss penalty.
- Let h be the hit rate, M the miss penalty, and C the time to access information in the cache. Thus, the average access time experienced by the processor is
- $t_{avg} = hC + (1 - h)M$
- .

Caches on the processor chip

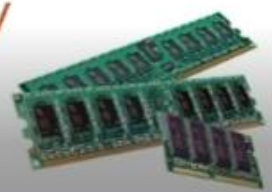
- In high performance processors 2 levels of caches are normally used.
- Avg access time in a system with 2 levels of caches is

$$T_{ave} = h1c1+(1-h1)h2c2+(1-h1)(1-h2)M$$



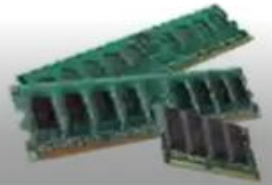
Virtual memories

- Recall that an important challenge in the design of a computer system is to provide a large, fast memory system at an affordable cost.
- Architectural solutions to increase the effective speed and size of the memory system.
- Cache memories were developed to increase the effective speed of the memory system.
- Virtual memory is an architectural solution to increase the effective size of the memory system.



Virtual memories (contd..)

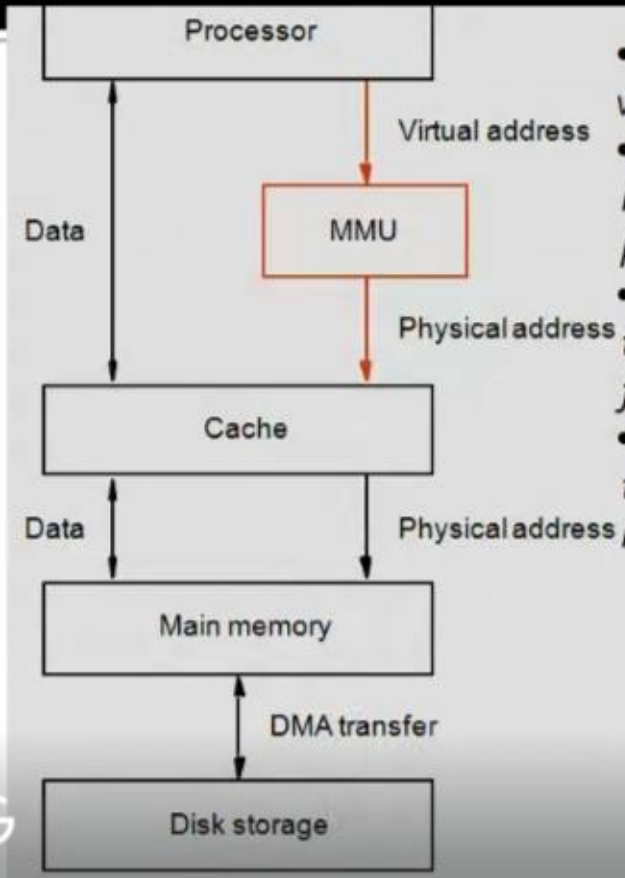
- When a new piece of a program is to be transferred to the main memory, and the main memory is full, then some other piece in the main memory must be replaced.
  - Recall this is very similar to what we studied in case of cache memories.
- Operating system automatically transfers data between the main memory and secondary storage.



Virtual memories (contd..)

- Techniques that automatically move program and data between main memory and secondary storage when they are required for execution are called virtual-memory techniques.
- Processor issues binary addresses for instructions and data.
  - These binary addresses are called logical or virtual addresses
- Virtual addresses are translated into physical addresses by a combination of hardware and software subsystems.
  - If virtual address refers to a part of the program that is currently in the main memory, it is accessed immediately.
  - If the address refers to a part of the program that is not currently in the main memory, it is first transferred to the main memory before it can be used.

Virtual memory organization



- *Memory management unit (MMU) translates virtual addresses into physical addresses.*
- *If the desired data or instructions are in the main memory they are fetched as described previously.*
- *If the desired data or instructions are not in the main memory, they must be transferred from secondary storage to the main memory.*
- *MMU causes the operating system to bring the data from the secondary storage into the main memory.*

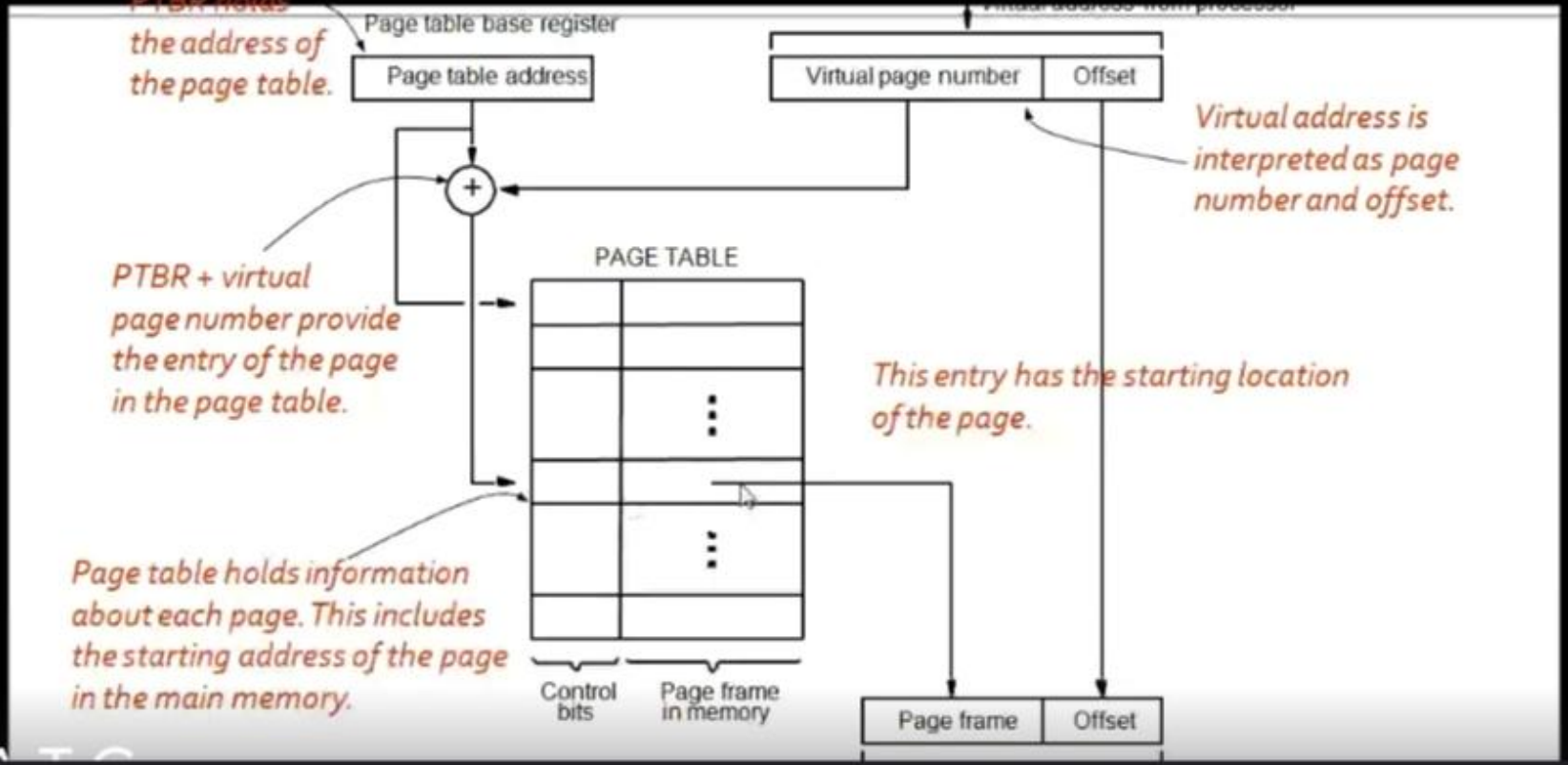
VIRTUAL MEMORY ADDRESS TRANSLATION

- All programs and data are composed of fixed length units called **Pages** (Figure)
  - The Page consists of a block-of-words. The words occupy contiguous locations in the memory. The pages are commonly range from 2K to 16K bytes in length.
- **Like Cache Bridge** speed-up the gap between main-memory and secondary-storage.
- **Each virtual-address contains**

1) Virtual Page number (Low order bit) and
2) Offset (High order bit).

- Virtual Page number + Offset → specifies the location of a particular word within a page.

Address translation (contd..)



Making VM work: translation

How does a program access memory?

1. Program executes a load specifying a virtual address (VA)