GraphicEra
(Deemed to be University)
Accredited by NAAC with Grade A

MCA I

Computer Organization and Architecture

TMC 102

By

Jaishankar Bhatt

Assistant Professor

Graphic Era Deemed to be University

Dehradun

Unit 4

Topic Name : Memory Organization

Table of Contents

- Memory Organization
- Memory Hierarchy
- Cache Memory
- Mapping
- References.

Memory Organization: -Memory in a computer system is required for storage and subsequent retrieval of the instructions and data. A computer system uses variety of devices for storing these instructions and data which are required for its operations. A memory system is a very simple system yet it exhibits a wide range of technology and types. The basic objective of a computer system is to increase the speed of computation. Likewise the basic objective of a memory system is to provide fast, uninterrupted access by the processor to the memory such that the processor can operate at the speed it is expected to work.

Classification of memories: Memory in a computer system is required for storage and subsequent retrieval of the instructions and data. A computer system uses variety of devices for storing these instructions and data which are required for its operations.

The basic objective of a memory system is to provide fast, uninterrupted access by the processor to the memory such that the processor can operate at the speed it is expected to work.

A memory system can be considered to consist of three groups of memories. These are:

a. **Internal Processor Memories:** These consist of the small set of high speed registers which are internal to a processor and are used as temporary locations where actual processing is done.

b. Primary Memory or Main Memory: It is a large memory which is fast but not as fast as internal processor memory. This memory is accessed directly by the processor. It is mainly based on integrated circuits.

c. Secondary Memory/Auxiliary Memory/Backing Store: Auxiliary memory in fact is much larger in size than main memory but is slower than main memory. It normally stores system programs (programs which are used by system to perform various operational functions), other instructions, programs and data files. Secondary memories cannot be accessed directly by a processor. First the information of these memories is transferred to the main memory and then the information can be accessed as the information of main memory.

Characteristics Terms for Various memory Devices : - The following terms are most commonly used for identifying comparative behavior of various memory devices and technologies.

1. **Storage Capacity : -** It is the amount of data a storage device can hold. Storage capacity is measured in kilobytes (KB), megabytes (MB), gigabytes (GB) and terabytes (TB). The capacity of internal memory and main memory can be expressed in terms of number of words or bytes.
2. **Unit of transfer : -** Unit of transfer is defined as the number of bits read in or out of the memory in a single read or write operation, For main memory and internal memory, the normal unit of transfer of information is equal to the word length of a processor.

3. Access Modes: - Once we have defined the unit of transfer next important characteristics is the access mode in which the information is accessed from the memory. A memory is considered to consist of various memory locations. The information from memory devices can be accessed in the following ways:

a) **Random Access:** It is the mode in which any memory location can be accessed in any order in the same amount of time.

b) Sequential access: On the other hand we have memories which can be accessed in pre-defined sequences for example, the songs stored on a cassette can be accessed only one by one. The example of sequential access memory is Magnetic Tape.

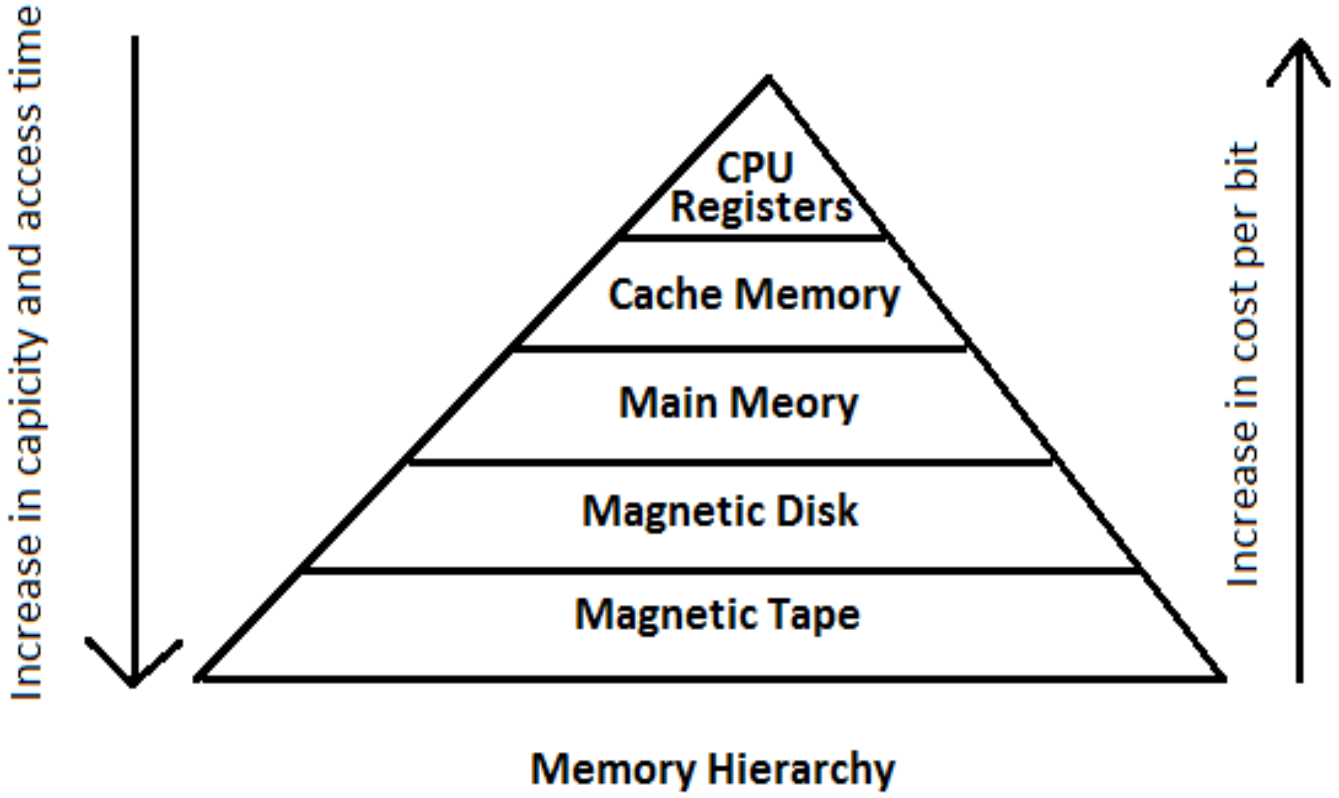
4. Access Time: - The access time is the time required between the request made for a read or write operation till the time the data is made available or written at the requested location. Normally it is measured for read operation. The access time depends on the physical characteristics and access mode used for that device.

5. Cycle Time: - It is defined as the minimum time elapsed between two consecutive read requests. Is it equal to access time? Yes, for most of the memories except the ones in which destructive readout is encountered. Cycle time for such memories is the access time (time elapsed when a read request is made available) plus writing time as after the data has been made available the information has to be written back in the same memory location as the previous value has been destroyed by reading. But for most of the commonly used semiconductor memories cycle time is equal to the access time.

6. Data Transfer Rate: -The amount of information that can be transferred in or out of the memory in a second is termed as data transfer rate or bandwidth. It is measured in bits per second. Maximum number of bits that can be transferred in a second depends on how many bits can be transferred in or out of the memory simultaneously and thus the data bus width become one of the controlling factors.

Memory Hierarchy: - The total memory capacity of a computer can be visualized as being a hierarchy of components. The memory hierarchy system consists of all storage devices employed in a computer system from the slow but high-capacity auxiliary memory to a relatively faster main memory, to an even smaller and faster cache memory accessible to the high-speed processing logic.

In practice, a memory system is a hierarchy of storage devices with different capacities, costs, and access times. CPU registers hold the most frequently used data. Small, fast cache memories nearby the CPU act as staging areas for a subset of the data and instructions stored in the relatively slow main memory. The main memory stages data stored on large, slow disks, which in turn often serve as staging areas for data stored on the disks or tapes of other machines connected by networks.


Increase in capicity and access time
CPU Registers
Cache Memory
Main Meory
Magnetic Disk
Magnetic Tape
Increase in cost per bit
Memory Hierarchy

Level-0 (CPU Registers): -

- At level-0, registers are present which are contained inside the CPU.
- Since they are present inside the CPU, they have least access time.
- They are most expensive and therefore smallest in size (in KB).
- Registers are implemented using flip flops.

Level-1 (Cache Memory): -

- At level-1, cache memory is present.
- It stores the segments of program that are frequently accessed by the processor.
- It is expensive and therefore smaller in size (in MB).
- Cache memory is implemented using static RAM.

Level-2 (Main Memory): -

- At level-2, main memory is present.
- It can communicate directly with the CPU and with auxiliary memory devices through an I/O processor.
- It is less expensive than cache memory and therefore larger in size (in few GB).
- Main memory is implemented using dynamic RAM.

Level-3 (Magnetic Disk): -

- At level-3, secondary storage devices like Magnetic Disk are present.
- They are used as back up storage.
- They are cheaper than main memory and therefore much larger in size (in few TB).

Level-4 (Magnetic Tape): -

- At level-4, tertiary storage devices like magnetic tape are present.
- They are used to store removable files.
- They are cheapest and largest in size (1-20 TB).

Observations-

The following observations can be made when going down in the memory hierarchy-

- Cost / bit decreases
- Frequency of access decreases
- Capacity increases
- Access time increases

Goals of Memory Hierarchy-

The goals of memory hierarchy are-

- To obtain the highest possible average access speed
- To minimize the total cost of the entire memory system

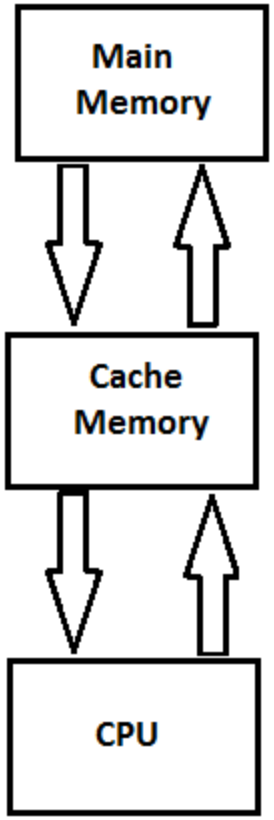
Cache Memory : - Cache memory is a high speed small memory. It is placed between the CPU and main memory. The cache memory access time is less than the access time of main memory by a factor of 5 to 10. The cache is the fastest component in the memory hierarchy and approaches the speed of CPU components.

The idea behind using a cache is to keep the information expected to be used more frequently by the CPU in the cache (a small high-speed memory that is near the CPU). The end result is that at any given time some active portion of the main memory is duplicated in the cache. Therefore, when the processor makes a request for a memory reference, the request is first sought in the cache.

If the data required by the CPU is not found in the cache memory, it is then accessed from the main memory and copied into the cache memory at the same time. Next time whenever the same piece of data is required by the CPU, it is accessed from the cache memory.

Cache memory is used to reduce the average time to access data from the Main memory. The cache is a smaller and faster memory which stores copies of the data from frequently used main memory locations. There are various different independent caches in a CPU, which store instructions and data.

Cache memory is an extremely fast memory type that acts as a buffer between RAM and the CPU. It holds frequently requested data and instructions so that they are immediately available to the CPU when needed.



Cache Memory Organization

Terms related to cache memory: - Following terms are related to cache memory.

Hit: - If the data required by the CPU is found in cache memory, it is known as hit.

Miss: - If the data required notby the CPU is found in cache memory, it is known as miss.

Hit ratio: - The performance of cache memory is frequently measured in terms of a quantity called hit ratio. Which is calculated as

Hit ratio = Hit/Hit + Miss

Locality of reference: - Analysis of a large number of typical programs has shown that the references to memory at any given interval of time tend to be confined within a few localized areas in memory. This phenomenon is known as the property of locality of reference locality of reference. This is of two types

1- Temporal locality reference 2- Spatial locality reference

Temporal locality reference: -Temporal locality reference states that the instructions which are accessed recently, likely to accessed in near future. For example the instructions which are under the loop accessed first time and likely to accessed in near future.

Spatial Locality reference: - Spatial Locality Reference states that the instructions which are very close to each other are likely to be accessed in near future. For example if one location of array is accessed, then the chance to access second location is very high.

Writing into Cache : -- An important aspect of cache organization is concerned with memory write requests. When the CPU finds a word in cache during a read operation, the main memory is not involved in the transfer. However, if the operation is a write, there are two ways that the system can proceed - write-through and write-back

Write through: - In this method when CPU update a data item in cache memory, the data item is also updated in the main memory simultaneously by the CPU. This method has the advantage that main memory always contains the same data as the cache.

Write back: - In this method only the cache memory is updated every time and when the updated data item is removed from the cache memory, its original value is updated in main memory.

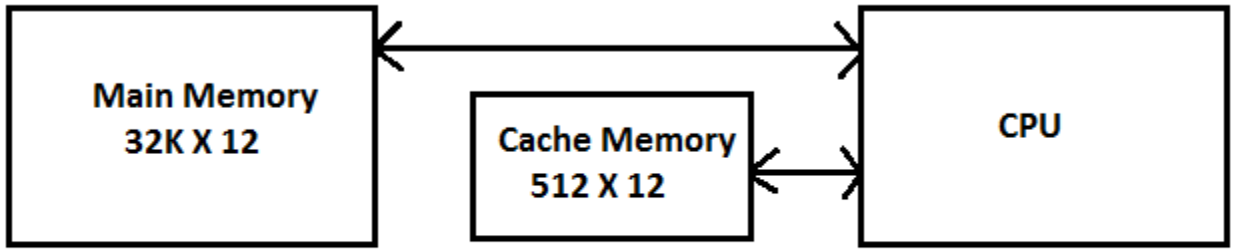
Cache Basics : -

• The cache memory is a high speed memory that keeps a copy of the frequently used data

• When the CPU wants a data value from memory, it first looks in the cache

• If the data is in the cache, it uses that data

• If the data is not in the cache, it copies a line of data from RAM to the cache and gives the CPU what it wants

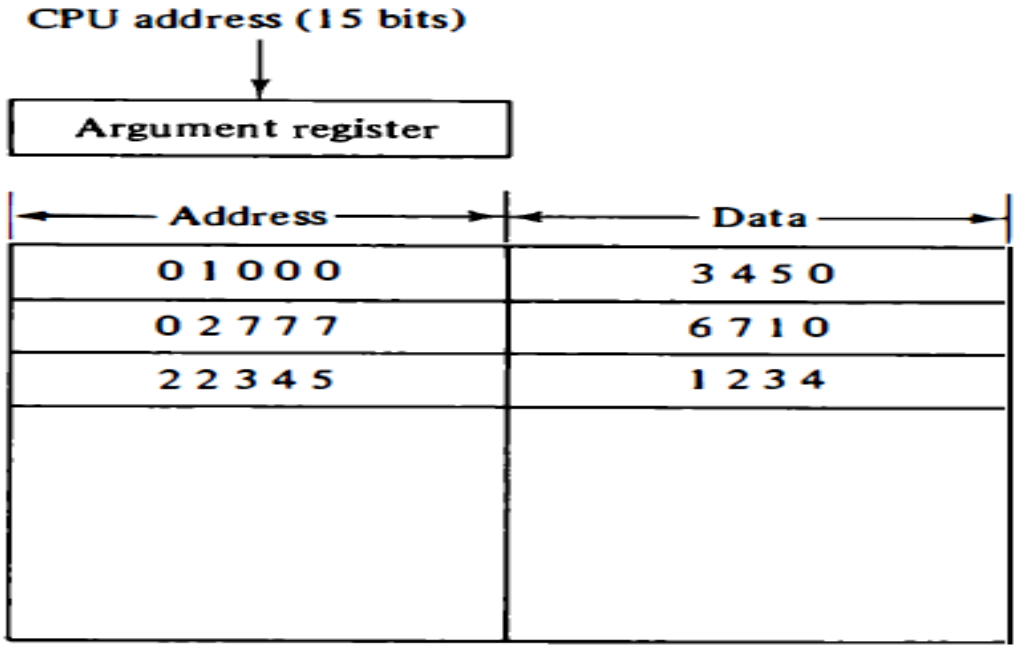


Cache Memory Model

Mapping Process: - The transformation of data from main memory to cache memory is referred to as a mapping process. Three types of mapping procedures are of practical interest when considering the organization of cache memory:

1. Associative Mapping
2. Direct Mapping
3. Set Associative Mapping

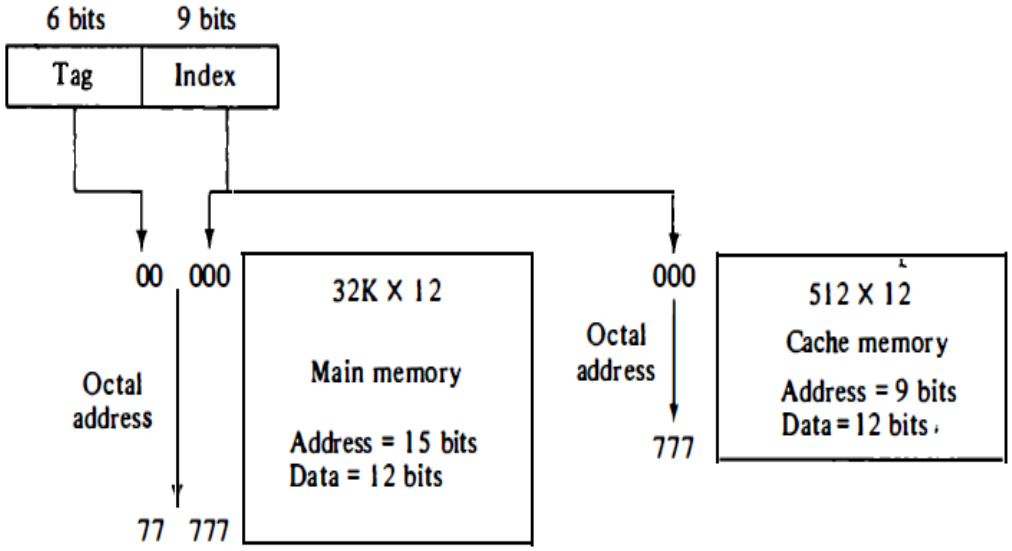
Associative Mapping: - The fastest and most flexible cache organization uses an associative memory. The associative memory stores both the address and content (data) of the memory word. This permits any location in cache to store any word from main memory.



A CPU address of 15 bits is placed in the argument register and the associative memory is searched for a matching address. If the address is found, the corresponding 12-bit data is read and sent to the CPU. If no match occurs, the main memory is accessed for the word. The address data pair is then transferred to the associative cache memory.

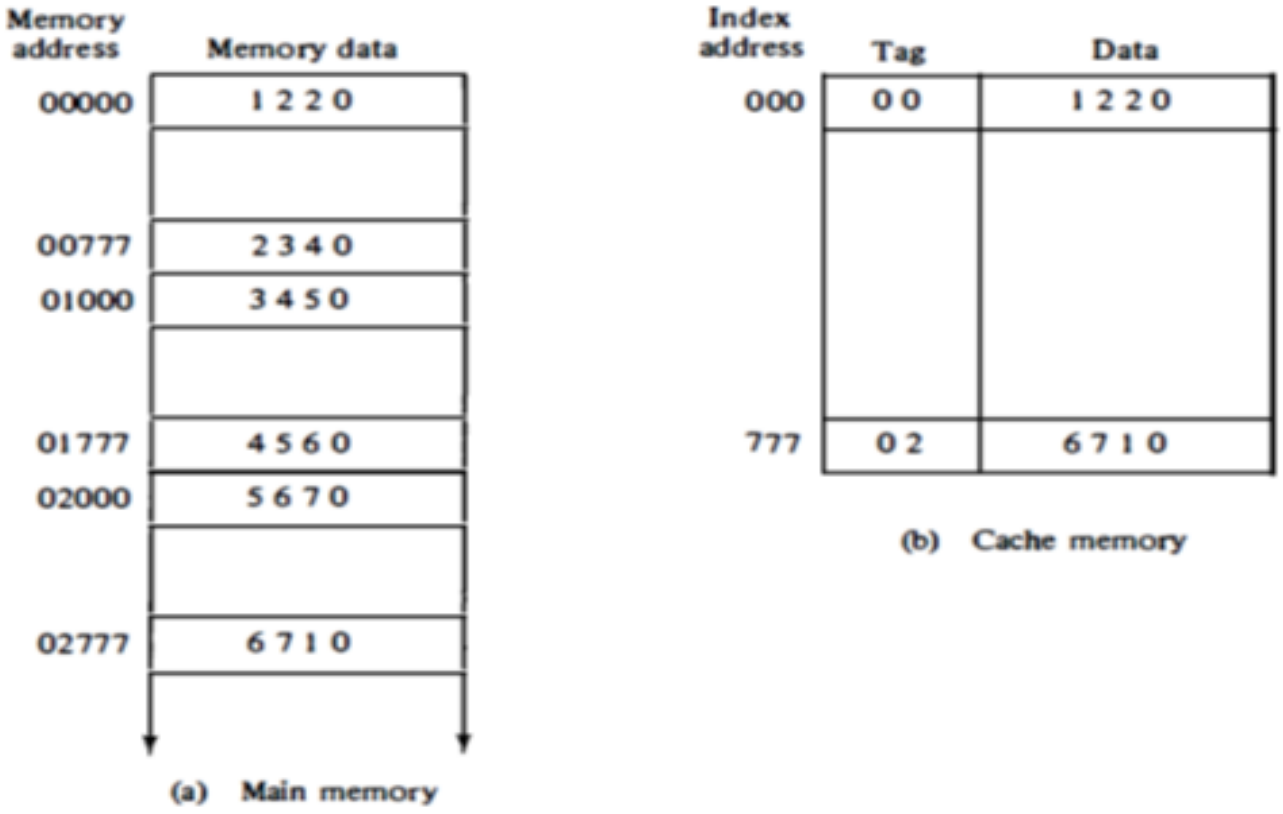
If the cache is full, an address data pair must be displaced to make room for a pair that is needed and not presently in the cache. The decision as to what pair is replaced is determined from the replacement algorithm that the designer chooses for the cache. A simple procedure is to replace cells of the cache in round-robin order whenever a new word is requested from main memory. This constitutes a first-in first-out (FIFO) replacement policy.

Direct Mapping: - Associative memories are expensive compared to random-access memories because of the added logic associated with each cell. The CPU address of 15 bits is divided into two fields. The nine least significant bits constitute the index field and the remaining six bits form the tag field. Each word in cache consists of the data word and its associated tag.



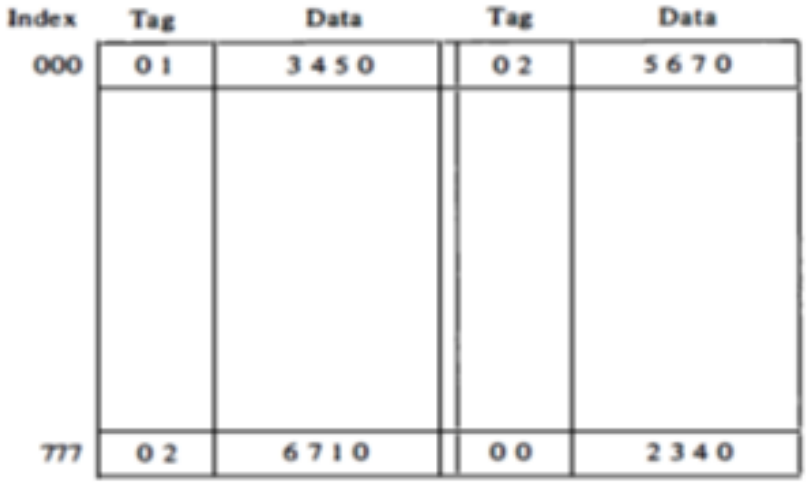
When a new word is first brought into the cache, the tag bits are stored alongside the data bits. When the CPU generates a memory request, the index field is used for the address to access the cache. The tag field of the CPU address is compared with the tag in the word read from the cache. If the two tags match, there is a hit and the desired data word is in cache.

If there is no match, there is a miss and the required word is read from main memory. It is then stored in the cache together with the new tag, replacing the previous value. The disadvantage of direct mapping is that the hit ratio can drop considerably if two or more words whose addresses have the same index but different tags are accessed repeatedly.



Direct mapping cache organization.

Set associative Mapping: - A third type of cache organization, called set-associative mapping, is an improvement over the direct mapping organization in that each word of cache can store two or more words of memory under the same index address. Each data word is stored together with its tag and the number of tag-data items in one word of cache is said to form a set.



Index	Tag	Data	Tag	Data
000	01	3450	02	5670
777	02	6710	00	2340

Two-way set-associative mapping cache.

References

- Computer System Architecture, Morris Mano, PHI .
- Computer Organization & Architecture: Designing for Performance, Stalling, PHI.
- Computer Organization and Architecture, Stalling, Pearson Education.
- Computer Architecture and Organization, J.P. Hayes McGraw Hill Company, New Delhi.