Name: Awais Ghaffar
Dep: BSCS
Semester 15269 4th
ID 4
Assig. vYo Computer Architecture
Subject
Submitted to. Amin Sir.

(i)

What is the general relationship among access time, memory Cost and Capacity?

As access time becomes faster the Cost per bit increases, As memory Size increases the Cost per bit is smaller, Also with greater capacity, the access time becomes slower.

ii) Discuss different memory access methods in detail?

Another distinction among memory types is the method of accessing units of data. These include the following.

* Sequential access:

Memory is organized into units of data called records access must be made in a specific linear sequence stored adressing information is used to separate records and assist in the retrieval process. A shared read. Write mechanism is used and this must be moved from its current location to the desired location, Passing and refecting each intermediate record.

Thus the time to access an arbitrary record is highly variable.

Direct access.

As with sequential access direct access involves a shared a shared read, write mechanism. However individuals blocks or records have an unique address based on physical location. Access is accomplished by direct access to reach a general vicinity plus sequential searching counting or waiting to reach the final

the final location. Again access time is variable.

Random access :;

Each adressable location in memory has a unique, physically wired addressing mechanism. the time to access a given location is independent of the sequence of prior accesses and is constant. Thus, any location can be selected at random and directly addressed and accessed main memory and some cache system are random access.

Associative :;

This is random access type of memory that enables one to make a comparison of desired bit location with in a word of specified match, and to do this for all words simultaneously Thus a word for a specified portion of its contents rather then its address.

As with ordinary random-access memory each location has its own addressing mechanism and retrieval time is constant independent of location or prior access patterns. Cache memories may apply associative access.

iii) Discuss the importance of memory hierarchy.

iv) How does the Principal of locality relate to the use of multiple memory levels?

Ans) slower and less expensive memory is used in higher stages with the most expensive being the register in the processor as well as cache. Main memory is slower and less expensive and is outside of the processor.

v) How many main memory address is interpreted in direct, associative and set associative?

Mapping ?

Direct mapping ;;?

The simplest technique maps each block of main memory

into only one possible cache line.

Associative line mapping:

Permits each main memory into block to be loaded into any line of cache.

The cache control logic interprets a memory adress simply as a tag and a word field.

To determine wether a block is in the cache The cache control logic must simultaneously examine every lines Tag for a match.

Set Associative mapping;

A compromize that exhibits the strengths of both and direct and associative approuches while reducing Their disadvantage.

Q: NO: 2

Note each of following:

i) Memory unit transfer:?

It is the maximum number of bits that can be read or written into the memory.

at a time. In case of main memory it is mostly equal to word size. In case external memory unit of transfer is not limited to the word size it is often larger is refered to as blocks.

Memory of Performance Parameter:
The two most important characteristic of memory are capacity and performance Three performance Parameter are used

Access time:
For random access memory, this is the time it takes to perform a read or write operation that is the time from the instant that an adress is presented to the memory to the instant that data have been stored or made available for use. For non-random access memory access time is the time it takes to position the read write mechanism and the desired location.

Memory cycle time:
This concept is primarily applied to random access memory and consist of the access time plus any additional time required before a sound access can commence. This additional time may be required for transients to die on signal lines or to regenerate data if they are read destructively. Note that memory cycle is concerned with the system bus not the processor.

iii) Disk Cache.
A portion of main memory can be used as a buffer to hold data temporarily that is to be read out to disk.
A few large transfer of data can be used instead of many small transfer of data
Data can be retrieved rapidly from the software cache rather than slowly from the disk.

iv) Principal of locality ::

The principal of locality states that data in the vicinity of a referenced word are likely to be referenced in the near future.

v) Logical cache and Physical cache.

A logical cache, also known as a virtual cache, stores data using adresses. The processor accesses the cache directly without going through the MMU. A physical cache stores data using main memory physical adresses.

One obvious advantage of the logical cache is that cache access speed is faster then for a physical cache. disadvantage has to do with the fact most virtual memory system supply each aplication with the same virtual memory adress space

Replacement Algorithams:

Once the cache has been filled when a new block is brought into the cache. One of the existing blocks must be replaced.

For direct mapping there is only one possible line for any particular block. and no choice is possible. For the associative and set associative techniques a replacement algorithm is needed. To achieve hight speed. Such an algorithm must be implemented in hardware.

---

vii) Possible approaches to cache coherancy.

Possible approaches to cache coherancy include the following

Bus watching with write Through

Each cache controller monitors the adress lines to detect write operation to memory by other bus master.

if another master writes to a location in shared memory that also resides in the cache memory. The cache controller invalidates that cache entry. This strategy depends on the use of a write through policy by all cache controllers.

Hardware transparency:
Additional hardware is used to ensure that all updates to main memory via cache are reflected in all caches. Thus, if one processor modifies a word in its cache this update is written to main memory.

Non-cacheable memory:
only a portion of main memory is shared by more than one processor and this is designated as non-cacheable. In such a system all access to shared memory are cache misses. because the shared

The shared memory is never copied into the cache. The non-cacheable memory can be identified using chip select logic or high address bits.

Q.NO: 3.

Differentiate each of the following:

i) Sequential:

Sequential access memory is organized into units of data called records. Access must be made in a specific linear sequence. Stored addressing information is used to separate records and assist in the retrieval process. A shared read-write mechanism is used and this must be moved from its current location to the desired location. Passing and rejecting, Searching, counting or waiting to reach the final location. Again time is variable.

iii) Direct :

As with sequential access direct access involve a shared read-write mechanism. However individual blocks or records have an unique address based on Physical location. Access is accomplished by direct access to reach the final location. Again access time is variable.

Random access :

Each addressable location in memory has a unique Physically wired in addressing mechanism. The time to access a given is independent of the sequence of prior accesses and is constant. Thus any location can be selected at random and directly addressed and accessed. main memory and some cache system or random

ii) Direct associative and setassociative mapping.?

Direct

The direct mapping technique is simple and inexpensive to implement its main disadvantage is that there is fixed cache location for any given block. Thus. if a program happens to reference words repeatedly from two different blocks that map into the same line then the blocks will be continually swapped in the cache. and the hit ratio will be low.

Associative

With associative mapping. there is flexibility as to which block to replace when a new block read into the cache. Replacement algorithms discussed later in this section. are designed to maximize the hit hit ratio.

The principal disadvantage of associative mapping is the complex circuitry required to tags of

of all cache lines in parallel.

Set associative

Set associative mapping is a compromise that exhibits the strengths of both the direct and associative approaches while reducing their disadvantages. In this case the cache consist of a number sets each of which consists of a number of lines. The relationship

$$m = v \times k$$

$$i = j \bmod v$$

where

i = Cache set number

j = main memory block number

m = number of lines in the cache

k = number of lines in each set

---

iii) Split Cache and unified Cache.

- Has become common to split Cache
  - One dedicated to instruction
  - One dedicated to data
- Both exist at the same level typically as two L1 Caches.

Advantages of unified Cache:-

- Higher hit rate.
- Balance load of instruction and dlta fetches automatically.
only one cache needs to be designed and implemented.
Trend is toward split caches at the L1 and unified caches for higher levels.

Advantages of split cache.
Eliminates cache contention b/w instruction fetch/decode unit and execution unit
- important in pipelining.

iv) write through write back.
Write Through
- Simplest technique.
- all write operation are made to main memory as well as to the cache memory.
The main disadvantage of this technique is that it generates substantial memory traffic and may create a bottleneck.

Write back
- minimize memory writes.
- updates are made only in cache.
- Portion of main memory are invalid and hence accesses by I/O modules can be allowed only through the cache.

(Q: No 4.)

Direct Solve.

i) In example. Suppose 95% of memory accesses are found in level 1. The average time to access a word can be expressed as

$(0.95)(0.01\text{ms}) + (0.05)(0.01\text{ms} + 0.1\text{ms})$

$= 0.0095 + 0.0055 = 0.015\text{ms}$

The average access time is much closer to 0.01ms than to 0.1ms as desired.

(ii)

There are total of 8 Kbytes/16 bytes. 512 lines in the cache. Thus the cache consist of 256 sets of 2 lines each. Therefore 8 bits are needed to identify

Set number. For the 64-Mbyte main memory a 26 bit address is needed. main memory consists of 64 M byte. / 16 bytes = 222 blocks. Therefore the set plus Tag length must be 22 bits so the Tag length is 14 bits and the word field length is 4 bits of main memory address.

Tag	set	word
14	8	4

END...