BAHIR DARE UNVERSITY IOT

SHOOL OF COMPUTING AND ELECTRICAL ENGINEERING

DEPARTMENT OF COMPUTER SCIENCE AND ELECTRICAL ENGINEERING

ASSIGNMENT_ONE ON CACHE MEMORY

4/11/2011

PREPARED BY:-

BEKELE HAILE

ID NO. ENG/R/069EC

ANSWER FOR REVIEW QUESTIONS

ANS.1

- Sequential access
  - In case of sequential access head search or reach at the memory address from start address of our storage device.
  - If you want to access the nth byte of a file, you must read all bytes 0 through n this is not work in random access.
  - Start at the beginning and read through in order
  - Access time depends on location of data and previous location
- Direct Access:-
  - Access data store on our storage media or RAM directly from that memory address.
  - Direct access means going straight to the record you want
  - Individual blocks have unique address
  - Access is by jumping to vicinity plus sequential search
  - Access time depends on location of data and previous location
- Random access
  - Random access means pick data randomly and then find that data which you required.
  - Individual addresses identify location exactly
  - Access time is independent of data location and previous location

    Advantages:

    (1) it provides rapid access to the desired information. In a decision-making environment where information is needed quickly, random access is a requisite to rapid retrieval;

    (2) it is efficient for retrieving a relatively few records at a time; and

(3) it provides a method of keeping files up to date as transactions or events occur

ANS.2

There is a trade-off among the three key characteristics of memory: cost, capacity, and access time. That means:-

- Faster access time – greater cost per bit
- Greater capacity – smaller cost per bit
- Greater capacity – slower access time

❖ As one goes down the memory hierarchy: (a) decreasing cost per bit; (b) increasing capacity; (c) increasing access time; (d) decreasing frequency of access of the memory by the processor.

ANS.3

The locality principle is the phenomenon that the collection of the data locations referenced in a short period of time in a running computer.

Locality of Reference principle

- Memory references by the processor, for both data and instructions, cluster
- Programs contain iterative loops and subroutines - once a loop or subroutine is entered, there are repeated references to a small set of instructions
- Operations on tables and arrays involve access to a clustered set of data word

ANS.4

Direct mapping:-

- Each block of main memory maps to only one cache line
- Direct mapping cache treats a main memory address as 3 distinct fields

These are Tag identifier, Line number identifier, Word identifier.

- No two blocks in the same line have the same Tag field

Pros of Direct Mapping

- Simple
- The tag memory is much smaller than in associative mapped cache.
- No need for an associative search, since the slot field is used to direct the comparison to a single field.

Cons of Direct Mapping

• Inexpensive

• Fixed location for a given block. If a program accesses two blocks that map to the same line repeatedly, then cache misses are very high.

- Consider what happens when a program references locations that are 2^{19} words apart, which is the size of the cache. Every memory reference will result in a miss, which will cause an entire block to be read into the cache even though only a single word is used.

Associative mapping:-

- A main memory block can be loaded into any line of the cache
- A memory address is interpreted as a tag and a word field
- The tag field uniquely identifies a block of main memory
- Each cache line's tag is examined simultaneously to determine if a block is in cache
- Main memory addresses are viewed as two fields

  These are tags(s) and words (w).

  - Pros of Associative Mapping

• Flexibility as to which block to replace when a new block is read into cache

- Replacement algorithms designed to maximize cache hit ratio
- Any main memory block can be placed into any cache slot.
- Regardless of how irregular the data and program references are, if a slot is available for the block, it can be stored in the cache.

  - Cons of Associative Mapping

• Complex circuitry required to examine the tags of all cache lines in parallel

- Considerable hardware overhead needed for cache bookkeeping.
- There must be a mechanism for searching the tag memory in parallel

Set Associative mapping:-

Compromise between direct and associative mapping. That takes the advantage of both direct and associative mapping.

• Cache divided into v sets

• Each set contains k lines

• A given block maps into any line in a given set

- Pros
- In our example the tag memory increases only slightly from the direct mapping and only two tags need to be searched for each memory reference
- Cons
- Expensive

ANS.5

Direct mapping cache treats a main memory address as 3 distinct fields

These are Tag identifier(s-r, 8bits), Line number identifier(r, 14bits), and Word (w, 2bits) identifier.

- Word identifier specifies the specific word (or addressable unit) in a cache line that is to be read
- Line identifier specifies the physical line in cache that will hold the referenced address
- The tag is stored in the cache along with the data words of the line

ANS.6

A memory address is interpreted as a tag(s, 22 bits) and a word (w, 2 bits) field. Line numbers (ids) have no meaning in the cache.

- The tag field uniquely identifies a block of main memory. Every line's tag must be examined for a match
- Least significant w bits = word position within block

ANS.7

- Compromise between direct and associative mapping. Divide cache into a number of sets (v), each set holding a number of lines (k).Therefore, a memory address is interpreted as a tag(s-d, 9 bits), set (d, 13 bits) and a word (w, 2 bits) field.
- Most significant s bits = tag used to identify which block is stored in a particular line of cache.
- Word field: - identifies the element (word) within the block that is requested by the processor. Word service as offset for words to be accessed or word position in block.
- Set field: - is used to uniquely identify the specific cache set that ideally should hold the targeted block.

ANS.8

- **Temporal Locality**: Concept that a resource will be referenced at one point in time will be referenced again. Cache miss traffic decreases fast when cache size increases and temporal locality determines sensitivity to cache size. This means the reuse of specific data and/or resources within relatively small time durations.
- **Spatial Locality**: Concept that likelihood of referencing a resource is higher if a resource near it was referenced. Cache miss traffic does not increase much when line size increases. Spatial locality determines sensivity to line size. And others like equidistance, branch. This refers to the use of data elements within relatively close storage locations.

ANS.9

Exploiting the locality of references is achieved usually on the hardware side. The temporal and special locality can be capitalized by hierarchical storage hard wares. In addition to this:-

- By using a data cache system
- Maximizing the effective cache memory space for any given cache size.
- Modern cache designs exploit spatial locality by fetching large blocks of data called cache lines on a cache miss. Subsequent references to words within the same cache line result in cache hits.

ANS.11

Least Recently Used (LRU)

- **Least Recently Used** (LRU): discards the least recently used items first. This algorithm requires keeping track of what was used when, which is expensive if one wants to make sure the algorithm always discards the least recently used item. General implementations of this technique require keeping "age bits" for cache-lines and track the "Least Recently Used" cache-line based on age-bits. In such implementation, every time a cache-line is used, the age of all other cache-lines changes. So, here are the simple techniques for implementing an LUR replacement algorithm in four ways set associative cache:-

- Associate a 2-bit counter with each of the four blocks in a set.
- Initially, arbitrarily set the four values to 0, 1, 2, and 3 respectively.
- When a hit occurs, the counter of the block that is referenced is set to 0. The other counters in the set with values originally lower than the referenced counter are incremented by 1; the remaining counters are unchanged.
- When a miss occurs, the block in the set whose counter value is 3 is replaced and its counter set to 0. All other counters in the set are incremented by 1.

ANS. 12

a. A reference to the first instruction is immediately followed by a reference to the second.

b. The ten accesses to a[i] within the inner “for loop” which occur within a short interval of time.