Computer Organization and Architecture (EET 2211)

Chapter 4
CACHE MEMORY

Replacement Algorithms (1) Direct mapping

- No choice
- Each block only maps to one line
- Replace that line

Replacement Algorithms (2) Associative & Set Associative

- Hardware implemented algorithm (speed)
- Least Recently used (LRU)
- e.g. in 2 way set associative
  - Which of the 2 block is LRU?
- First in first out (FIFO)
  - replace block that has been in cache longest
- Least frequently used
  - replace block which has had fewest hits
- Random

Write Policy

- The old block is not altered , then over written with a new block without first writing out the old block
- Must not overwrite a cache block unless main memory is up to date

Problems:

- More than one device may have access to main memory

e.g. An I/O may able to read-write main memory directly . If a word has been altered only in cache , then corresponding memory word is invalid .If the I/O device has altered main memory ,the cache word is invalid.

Write through

- All writes go to main memory as well as cache
- Multiple CPUs can monitor main memory traffic to keep local (to CPU) cache up to date
- Lots of traffic
- Slows down writes

Write back

- Updates initially made in cache only
- Update bit for cache slot is set when update occurs
- If block is to be replaced, write to main memory only if update bit is set
- Other caches get out of sync
- I/O must access main memory through cache
- N.B. 15% of memory references are writes

- In bus organization a cache and main memory is shared , a new problem is occurred .If data in one cache are altered, corresponding word in main memory is invalids , also same word in other caches.
- A system that prevents this problem is said to maintain cache coherency .Possible approaches to cache coherency include followings:
  - Bus watching with write through
  - Hardware transparency
  - Noncacheable memory

Line Size

- Retrieve not only desired word but a number of adjacent words as well

- Increased block size will increase hit ratio at first
  - The principle of locality

- Hit ratio will decreases as block becomes even bigger
  - Probability of using newly fetched information becomes less than probability of reusing replaced . Two specific effects come:

1. Larger blocks -
   - Reduce number of blocks that fit in cache
   - Data overwritten shortly after being fetched.
2. Each additional word is less local, so less likely to be needed.

- No definitive optimum value has been found
- 8 to 64 bytes seems reasonable
- For HPC systems, 64- and 128-byte most common

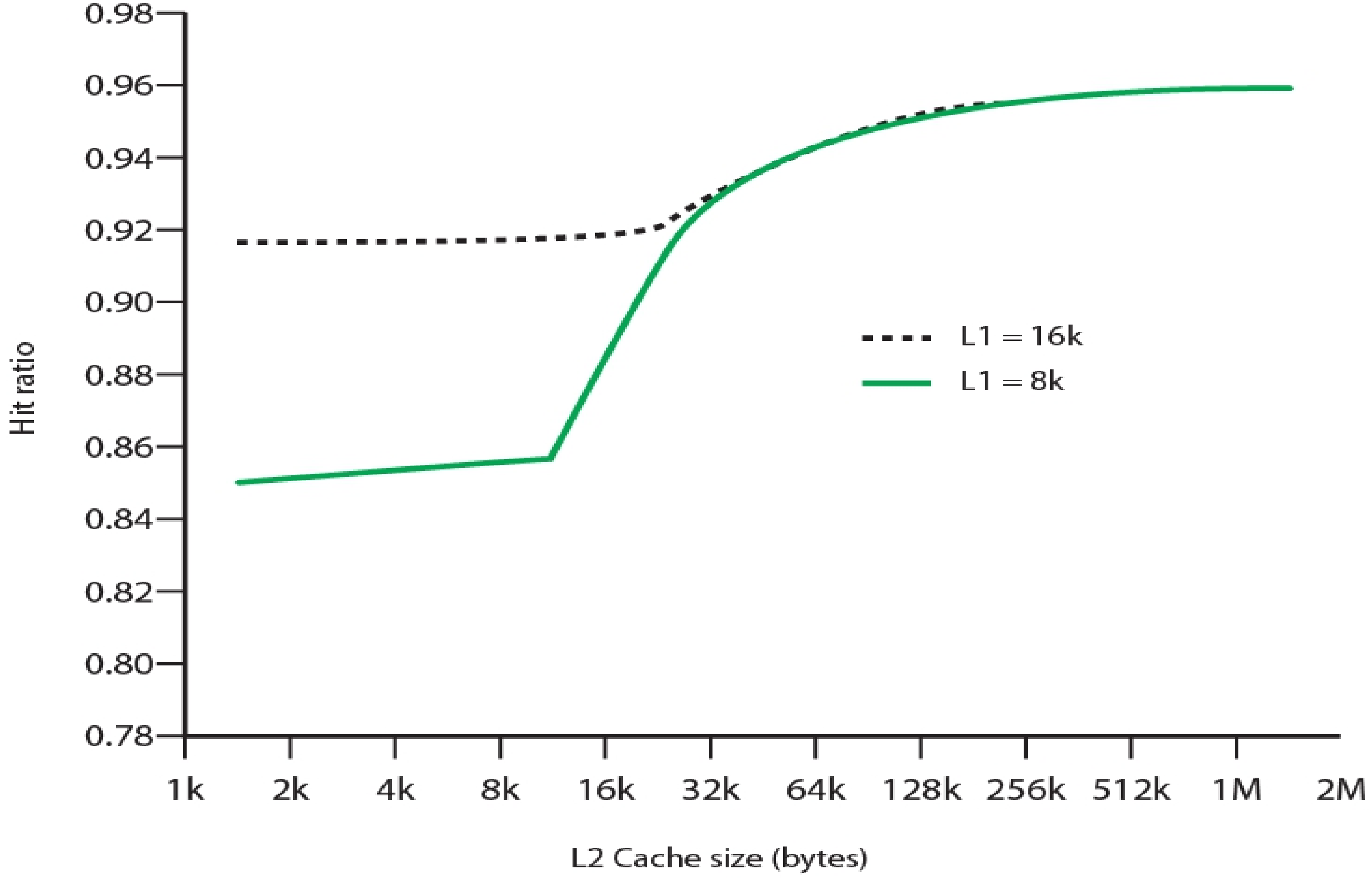
Number of Caches

- MULTILEVEL CACHES
- UNIFIED VERSUS SPLIT CACHES

Multilevel caches

- High logic density enables caches on chip
  - Faster than bus access
  - Frees bus for other transfers
- Common to use both on and off chip cache
  - L1 on chip, L2 off chip in static RAM
  - L2 access much faster than DRAM or ROM
  - L2 often uses separate data path
  - L2 may now be on chip
  - Resulting in L3 cache
    - Bus access or now on chip…

Hit Ratio (L1 & L2) For 8kbytes and 16 kbytes L1



Unified Versus Split Caches

- Split the cache into two: one for instructions and one for data
- Both exist at same level(two L1 caches)
- Processor attempts to fetch an instruction from main memory – the instruction L1 cache
- Processor attempts to fetch an data from main memory – the data L1 cache
- Advantages of unified cache
  - Higher hit rate
    - Balances load of instruction and data fetch
    - Only one cache to design & implement
- Advantages of split cache
  - Eliminates cache contention between instruction fetch/decode unit and execution unit
    - Important in pipelining

Review Questions

1 .What are the differences among sequential access, direct access, and random access?

2.What is the access time for a random-access memory and a non-random access memory?

3 . What is the general relationship among access time, memory cost, and capacity?

4 . What are the differences among direct mapping, associative mapping, and set-Associative mapping?

5 . For a direct-mapped cache, a main memory address is viewed as consisting of three fields. List and define the three fields.

cond..

6 . For an associative cache, a main memory address is viewed as consisting of two fields . List and define the two fields.

7. For a set-associative cache, a main memory address is viewed as consisting of three fields. List and define the three fields.

8 . What are the advantages of using a unified cache?

Thank You !