Computer Architecture

Quiz 2 → Guide to Reading Appendix B and Chapter 2

April 3, 2021

Abstract

The underlying ideas of Appendix B and Chapter 2 are Caching, Virtual Memory and Memory Hierarchy Design.

1 Vocabulary

We gain power over ideas which we can associate with a word. Instead of being fleeting collections of activations of neurons, alone, named ideas can be recalled, communicated, built upon.

Thus, we need vocabulary. Provide a definition for:

- virtual memory – moves pages between the two levels of the memory hierarchy. Virtual memory provides separation between processes that share one physical memory but having differing virtual address spaces.

- memory hierarchy – represents a separated version of computer storage based off of performance and response time.

- cache, including level 1 cache, level 2 cache – Cache as memory acts as a buffer between RAM and CPU. Level 1 cache is typically built into the CPU and is the fastest form of memory. Level 2 cache is located on a separate chip that feeds the level 1 cache and is slower than level 1 cache.

- inclusion property – says that the goal is to provide the system with a cost per byte that is almost as low as the cheapest level of memory with a speed that is closest to the fastest speed. The data contained in a lower level is a superset of the next higher level.

- memory bandwidth – the amount of information that can be transferred to/from memory.

- miss rate – a measure of the benefits of different cache organizations. It is the fraction of cache accesses that result in a miss – the number of accesses that miss / number of accesses

- tag (in the context of caching) – identifies which memory address a cache block corresponds to

- conflict (in the context of caching) – occurs when the block placement strategy is not fully associative. Conflict misses occur because a block has been discarded and later retrieved if several blocks map to its set and accesses to the different blocks are mixed.

- coherency (in the context of caching) – refers to uniformity of shared data that ultimately ends up in several local caches.

- page fault – an exception that is raised when a program attempts to access data that is in address space but not currently located in memory (RAM).

2 Relationships

Describe some relationship between the two terms, or explain why they are not related:

- Virtual address, Physical address – not related. Virtual addresses refer to the virtual store seen by the process, whereas physical addresses refer to hardware address of physical memory.

- write through, write back – Related in the sense that they both write information. Write through writes information to both the block in cache and the block in lower-level memory, whereas write back only writes to the block in cache.

- miss rate, miss penalty – Not related. Miss rate measures the benefits of different cache organizations whereas miss penalty measures the difference lower-level access time and cache access time.

- cache hit, cache miss – Related in the sense that they are both operations on data. Cache hits occur when the cache is able to retrieve data and display it. Cache miss is when the cache is not able to retrieve data and display it.

3 Calculations

Computer Architecture is quantitative, so we perform some computations.

1. An algorithm has poor data locality, causing multiple page faults. Calculate the CPU execution time of this algorithm, when the processor speed is 5GHz, instructions that commit per clock is 0.5, the algorithm contains (with iteration unrolled) 10,000 instructions, 10 cache misses occur, the miss penalty is 1000 clock cycles (see Figure 2.2, page 80). Assume that during a miss handling interval, no other progress is made on the computation.

   Time = (10,000 instructions/0.5 instructions per clock) -> 20,000 + (10 cache misses *1000 clock cycles) -> 30,000
   CPU execution time = 30,000/5,000,000,000 clock cycles per second -> **0.000006 seconds**

2. What would the CPU execution time be, if data locality were sufficient to remove all the page faults? Does attention to data locality significantly affect the performance of this algorithm?

   The CPU Execution time would be 20,000/5,000,000,000, yielding **0.000004** seconds. Attention to data locality does affect the performance of this algorithm but not significantly as the difference is the matter of 0.000002 seconds.