The University of Alabama in Huntsville
ECE Department
CPE 431 01/01R, CPE 531 01/91
Fall 2022

Due November 1, 2022
1.0(15), 2.0.1(10), 2.0.2(10), 2.0.3(20), 3.0.1(5), 3.0.2(10), 4.0.1(5), 4.0.2(10)

1.0 **<5.3>** Caches are important to providing a high-performance memory hierarchy to processors. Below is a list of 32-bit hexadecimal memory addresses, given as byte addresses. 74, A0, 78, 38C, AC, 84, 88, 8C, 7C, 34, 38, 13C, 388, 18C

For each of these references, identify the index and the tag, given a three-way set associative cache with two word blocks and a total of 24 words. List if each reference is a hit or a miss, assuming the cache is initially empty and show every entry to the cache, including the tag value and the addresses of all data items stored. Use hexadecimal or binary, whichever is easier.

2.0 **<5.4.>** This exercise examines the impact of different cache designs, specifically comparing associative caches to the direct-mapped caches from Section 5.4. For these exercises, use the word address stream given in hexadecimal: 15, A6, C9, 8F, 3D, A6, 3E, 85, 6F, 8F, 90, 3D

2.0.1 Using the references given and a fully associative cache with two-word blocks and a total size of 16 words, identify the index bits, the tag bits, and if it is a hit or miss. Use LRU replacement. Also show all entries made in the cache.

Multilevel caching is an important technique to overcome the limited amount of space that a first level cache can provide while still maintaining its speed. Consider a processor with the following parameters.

Base CPI, no memory stalls	Processor speed	Main memory access time	First-level cache miss rate per instruction	Second-level cache, direct-mapped speed	Global miss rate with second-level cache, direct-mapped	Second-level cache, eight-way s et associative speed	Global miss rate with second-level cache, eight-way set associative
1.5	2.4 GHz	70 ns	4.5 %	22 cycles	2.5 %	28 cycles	1.5 %

2.0.2 Calculate the CPI for the processor in the table using: 1) only a first level cache, 2) a second level direct-mapped cache, and 3) a second level eight-way set associative cache. How do these numbers change if main memory access time is doubled? If it is cut in half?

2.0.3 In older processors such as the Intel Pentium or Alpha 21264, the second level of cache was external (located on a different chip) from the main processor and the first-level cache. While this allowed for large second-level caches, the latency to access the cache was much higher, and the bandwidth was typically lower because the second-level cache ran at a lower frequency. Assume a 512 KiB off-chip second-level cache has a global miss rate of 4 %. If each additional

512 KiB of cache lowered the global miss rate by 0.5 %, and the cache had a total access time of 50 cycles, how big would the cache have to be to match the performance of the second-level direct-mapped cache listed in the table? Of the eight-way set-associative cache?

3.0 **<5.5>** This exercise examines the single error correcting, double error detecting (SEC/DED) Hamming code.

3.0.1 What is the minimum number of parity bits required to protect a 256-bit word using the SEC/DED code?

3.0.2 Consider an SEC code that protects 8 bit words with 4 parity bits. If we read the value 0x876, is there an error? If so, correct the error.

4.0 Media applications that play audio or video files are part of a class of workloads called "streaming" workloads (i.e., they bring in larg e amounts of data but do not reuse much of it). Consider a video streaming workload that accesses a 2048 KiB working set sequentially with the following address stream:

0, 1, 2, 3, 4, 5, 6, 7, 8, 9, …

4.0.1 Assume a 128 KiB direct-mapped cache with a 16-byte block. What is the miss rate for the address stream above?

4.0.2 "Prefetching" is a technique that leverages predictable address patterns to speculatively bring in additional cache blocks when a particular cache block is accessed. One example of prefetchcing is a stream buffer that prefetches sequentially adjacent cache blocks in to a separate buffer when a particular cache block is brough in. If the data is found in the prefetch buffer, it is considered a a hit and moved into the cache and the next cache block is prefetched. Assume a two-entry stream buffer and assume that the cache latency is such that a cache block can be loaded before the computation on the previous cache block is completed. What is the miss rate for the address stream above?