ECE4700J Homework 4

Yiwen Yang

Q1

1. All instructions cause I-Cache miss. Since CPI is 2, then on average 0.5 instruction is accessed per cycle. This yields $0.5 \times 0.3\% \times 64 = 0.096B/cycle$ to read bandwidth.

   Data read instructions contribute $0.5 \times 0.25 \times 2\% \times 64 = 0.16B/cycle$ to read bandwidth.

   Data write instructions uses write-through and write-allocate, so an 8-byte block is written directly to RAM regardless of cache hit or miss, and a 64-byte clock is fetched from RAM on cache miss. So write instructions contribute $0.5 \times 0.1 \times 8 = 0.4B/cycle$ to write bandwidth, and $0.5 \times 0.1 \times 2\% \times 64 = 0.064B/cycle$ to read bandwidth.

   In total, read bandwidth should be $0.096+0.16+0.064 = 0.32B/cycle$, write bandwidth should be $0.4B/cycle$.

2. With a write-back, write-allocate cache, the read bandwidth is the same in (1) as $0.32B/cycle$.

   Write-back occurs both in read and write instructions on cache miss, so write bandwidth should be $0.5 \times (0.25 + 0.1) \times 2\% \times 30\% \times 64 = 0.0672B/cycle$.

Q2

1. Cache is a real memory-unit that physically stores data, while virtual memory is not a memory-unit but a technique to map virtual address to physical address.

2. Cache is completely handled by the hardware, but virtual memory can be controlled by the operating system.

3. Cache size is much smaller than other memory units, yet virtual memory assumes a much larger size than the physical memory.

Q3

1. $$AMAT_1 = 0.22 + 100m_1$$
   $$AMAT_2 = 0.52 + 100m_2$$

   When $AMAT_1 < AMAT_2$, $m_1 - m_2 < 0.3\%$, smaller cache is more advantageous.

2. (a) Miss penalty is 10ns.

$$AMAT_1 = 0.22 + 10m_1$$

$$AMAT_2 = 0.52 + 10m_2$$

When $AMAT_1 < AMAT_2$, $m_1 - m_2 < 3\%$, smaller cache is more advantageous.

(b) Miss penalty is 1000ns.

$$AMAT_1 = 0.22 + 1000m_1$$

$$AMAT_2 = 0.52 + 1000m_2$$

When $AMAT_1 < AMAT_2$, $m_1 - m_2 < 0.03\%$, smaller cache is more advantageous.

From above, we can conclude that when miss penalty is smaller, cache of smaller size is more likely to be advantageous, otherwise not because smaller cache tends to have higher miss rates.

Q4

The access time is 0.86ns, with 0.5ns one cycle, so it takes 2 cycles to access the cache if prediction correct and 3 cycles if wrong. Then $AMAT = ((80\% \times 2 + 20\% \times 3) + 0.33\% \times 20) \times 0.5 = 1.133$ns.

Q5

Harvard architecture stores instruction cache and data cache separately, while unified stores all in one. The advantage is that instruction and data memory can be accessed simultaneously, so the speed can be greatly scaled in a superscalar pipeline. In addition, separated L1 cache can be optimized, no write ports are needed for I-Cache and fewer read ports for D-Cache specifically. The disadvantage is that two sets of cache need to be built on the physical processor on different dies, also separate paths between cache and memory are needed, which means that the cost will be higher than one die holding all L1 cache.

Q6

1. When there is L1 cache miss, it will first access L2 cache, then if hit, the block is fetched into L1 cache, if miss the block is fetched from main memory to both L1 and L2 cache. In both hit and miss, if storing an L1 evicted block in L2 causes a block to be evicted from L2, then L1 cache should also check the corresponding block and evict it if exists.

2. When there is L1 cache miss, it will first access L2 cache, then if hit, the block is fetched into L1 cache, if miss the block is fetched from main memory to only L1 cache. In both hit and miss, if storing an L1 evicted block in L2 causes a block to be evicted from L2, then only L2 cache needs to evict the block.