ENGR 325

HOMEWORK #8

ANSWER KEY

1. (5) Recall that we have two write policies and write allocation policies, and their combination can be implemented either in L1 or L2 cache. Assume the following choice for L1 and L2 caches:

L1	L2
Write-through, non-write allocate	Write-back, write allocate

Describe the procedure of handling an L1 write-miss, considering the component involved and the possibility of replacing a dirty block. (P&H 5.4, §5.3, 5.8)

***SOLUTION**:*

When an L1 write miss occurs, a cache block is not allocated in L1 for the block in question. L2 is written directly. Since L1 is a write-through cache, there are no issues with dirty blocks in L1.

If there is an L2 cache hit, L2 is simply updated and the dirty bit set.

If there is an L2 cache miss, a block must be allocated in L2 and the evicted block must be written back to main memory if it is dirty.

2. (10) Media applications that play audio or video files are part of a class of workloads called "streaming" workloads; i.e., they bring in large amounts of data but do not reuse much of it. Consider a video streaming workload that accesses a 512 kiB working set sequentially with the following address stream (P&H 5.5, §5.1, 5.4, 5.8, 5.13):

   0, 2, 4, 6, 8, 10, 12, 14, 16, …

   a. Assume a 64 kiB direct-mapped cache with a 32-byte block. What is the miss rate for the address stream above? How is this miss rate sensitive to the size of the cache or the working set? How would you categorize the misses the workload is experiencing, based on the 3C model?

   b. Re-compute the miss rate when the cache block size is 16 bytes, 64 bytes, and 128 bytes. What kind of locality is this workload exploiting?

   c. "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 prefetching is a stream buffer that prefetches sequentially adjacent cache blocks into a separate buffer when a particular cache block is brought in. If the data is found in the prefetch buffer, it is considered as 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?

   d. What is the optimal block size for a miss latency of 20xB cycles?

   e. What is the optimal block size for a miss latency of 24+B cycles?

   f. For a constant miss latency, what is the optimal block size?

***SOLUTION**:*

a. Given a 64 kiB cache, a 32-byte block implies 2kiB blocks. The access pattern is:

Address 0: miss

2: hit

4: hit

|

30: hit, after which the pattern repeats.

*So there are 15/16 hits, for a **miss rate of 1/16 or 6.25%**. Misses of this type are **compulsory**. This continues until the whole 64 kiB cache is filled. Then address 65,536 has a **conflict miss** at cache line 0. The hit/miss pattern continues (15/16). All further misses are conflict misses (even though the cache is full as well). **The miss rate for this pattern is not sensitive to the size of the cache or the size of the working set, but only the block size**.*

*b. **If the cache size is 16 bytes, the miss rate is 1/8, or 12.5%. If 64 bytes, the miss rate is 1/32, or 3.125%. If 128 bytes, the miss rate is 1/64, or 1.5625%**. The workload is exploiting **spatial locality**.*

*c. For this pattern, every access is a hit after the first one. Using a 32-byte block, and a 512 kiB (=2^{19} bytes) data set, **the miss rate is $1/2^{19} = 2^{-19}$, effectively 0**.*

Cache block size (B) can affect both miss rate and miss latency. Assuming a 1-CPI machine with an average of 1.35 references (both instruction and data) per instruction, find the optimal block size given the following miss rates for various block sizes.

8: 4%	16: 3%	32: 2%	64: 1.5%	128: 1%

d. Average Memory Access Time (AMAT) = (Time for a Hit) + (Miss Rate) x (Latency). Since CPI is given as one, the time for a hit in this case is one cycle.

Block Size (B)	*Miss Rate*	*Latency*	*AMAT*
8 bytes	*4%*	*160 cycles*	*7.4 cycles*
16 bytes	*3%*	*320 cycles*	*10.6 cycles*
32 bytes	*2%*	*640 cycles*	*13.8 cycles*
64 bytes	*1.5%*	*1280 cycles*	*20.2 cycles*
128 bytes	*1%*	*2560 cycles*	*26.6 cycles*

A block size of 8 has the lowest AMAT.

e. See the table below.

Block Size (B)	*Miss Rate*	*Latency*	*AMAT*
8 bytes	*4%*	*32 cycles*	*2.28 cycles*
16 bytes	*3%*	*40 cycles*	*2.20 cycles*
32 bytes	*2%*	*56 cycles*	*2.12 cycles*
64 bytes	*1.5%*	*88 cycles*	*2.32 cycles*
128 bytes	*1%*	*152 cycles*	*2.52 cycles*

A block size of 32 has the lowest AMAT.

f. For a constant miss latency, the block size with the lowest miss rate will be optimal – ***128 bytes****.*

3. (10) In this exercise, we will look at the different ways capacity affects overall performance. In general, cache access time is proportional to capacity. Assume that main memory accesses take 70 ns and that memory accesses are 36% of all instructions. The following table shows data for L1 caches attached to each of two processors, P1 and P2. (P&H 5.6, §5.4)

	L1 Size	L1 Miss Rate	L1 Hit Time
P1	2 kiB	8.0%	0.66 ns
P2	4 kiB	6.0%	0.90 ns

a. Assuming that the L1 hit time determines the cycle times for P1 and P2, what are their respective clock rates?

b. What is the Average Memory Access Time for P1 and P2?

c. Assuming a base CPI of 1.0 without any memory stalls, what is the total CPI for P1 and P2? Which processor is faster?

For the next three problems, we will consider the addition of an L2 cache to P1 to presumably make up for its limited L1 cache capacity. Use the L1 cache capacities and hit times from the previous table when solving these problems. The L2 miss rate indicated is its local miss rate.

L2 Size	L2 Miss Rate	L2 Hit Time
1 MiB	95%	5.62 ns

d. What is the AMAT for P1 with the addition of an L2 cache? Is the AMAT better or worse with the L2 cache?

e. Assuming a base CPI of 1.0 without any memory stalls, what is the total CPI for P1 with the addition of an L2 cache?

f. Which processor is faster, now that P1 has an L2 cache? If P1 is faster, what miss rate would P2 need in its L1 cache to match P1's performance? If P2 is faster, what miss rate would P1 need in its L1 cache to match P2's performance?

***SOLUTION**:*

a. P1 Cycle Time = 0.66 ns. $f_1 = 1/T = 1/(0.66 \times 10^{-9})$ = ***1.52 GHz***

P2 Cycle Time = 0.90 ns. $f_2 = 1/T = 1/(0.90 \times 10^{-9})$ = ***1.11 GHz***

b. AMAT = Hit Time + (Miss Rate) x (Miss Latency)

For P1: AMAT = 0.66 + (0.08)(70) = ***6.26 ns***

For P2: AMAT = 0.90 + (0.06)(70) = ***5.10 ns***

c. $CPI = Base\ CPI + (Miss\ Rate)\ x\ \left(\frac{Memory\ Accesses}{Instruction}\right) x\ (Miss\ Penalty)$

For P1: CPI = 1 + (0.08)(1.36)(70/0.66) = ***12.54 cycles/instruction***

For P2: CPI = 1 + (0.06)(1.36)(70/0.90) = ***7.34 cycles/instruction***

d. Find AMAT for L2 cache: AMAT = Hit Time + (Miss Rate) x (Miss Latency)

AMAT = 5.62 + (0.95)(70) = 72.12 ns

Find AMAT for processor core: AMAT = Hit Time + (Miss Rate) x (Miss Latency)

AMAT = 0.66 + (0.08)(72.12) = ***6.43 ns, worse than the AMAT for P1 found in part b****.*

e. $CPI = Base\ CPI + (L1\ Miss\ Rate)\ x\ \left(\frac{Memory\ Accesses}{Instruction}\right) x\ (L1\ Miss\ Penalty) + (L2\ Miss\ Rate)\left(\frac{Memory\ Accesses}{Instruction}\right) x\ (L2\ Miss\ Penalty)$

*CPI = 1 + (0.08)(1.36)(5.62/0.66) + (0.95*0.08)(1.36)(70/0.66) =* ***12.89 cycles/instruction***

f. P1's CPI is 12.89 cycles/instruction, from 5.6.5. P2's CPI is 7.34 cycles/instruction (5.6.3). In terms of AMAT, P1's AMAT is 6.43 ns with an L2 cache (5.6.4). P2's AMAT is 5.10 ns (5.6.2). ***P2 is faster.***

Since P2 is faster, set P1's AMAT to equal P2's and solve for the miss rate:

AMAT = Hit Time + (Miss Rate) x (Miss Latency) = 5.10

Miss Rate = (AMAT - Hit Time) / Miss Latency = (5.10 - 0.66)/70 = ***6.3%****.*

4. (5) Chip multiprocessors (CMPs) have multiple cores and their caches on a single chip. CMP on-chip L2 cache design has interesting trade-offs. The following table shows the miss rates and hit latencies for two benchmarks with private vs. shared L2 cache designs. Assume L1 cache misses once every 32 instructions. (P&H 5.18, §5.13)

	Private	Shared
Benchmark A misses-per-instruction	0.30%	0.12%
Benchmark B misses-per-instruction	0.06%	0.03%

Assume the following hit latencies:

Private Cache	Shared Cache	Memory
5	20	180

a. Which cache design is better for each of these benchmarks? Use data to support your conclusion.

b. Shared cache latency increases with the CMP size. Choose the best design if the shared cache latency doubles. Off-chip bandwidth becomes the bottleneck as the number of CMP cores increases. Choose the best design if off-chip memory latency doubles.

***SOLUTION**:*

a. AMAT = Hit Time + (Miss Rate) x (Miss Latency) = 5.10

Benchmark A, Private Cache: AMAT = (1/32 x 5) + (0.003)(180) = 0.70 cycles

Benchmark A, Shared Cache: AMAT = (1/32 x 20) + (0.0012)(180) = 0.84 cycles

Benchmark B, Private Cache: AMAT = (1/32 x 5) + (0.0006)(180) = 0.26 cycles

Benchmark B, Shared Cache: AMAT = (1/32 x 20) + (0.0003)(180) = 0.68 cycles

***Private cache is better** for each of these benchmarks.*

b. If latency doubles for the shared cache and memory latency doubles for the private cache:

Benchmark A, Private Cache: AMAT = (1/32 x 5) + (0.003)(360) = 1.24 cycles

Benchmark A, Shared Cache: AMAT = (1/32 x 40) + (0.0012)(180) = 1.47 cycles

Benchmark B, Private Cache: AMAT = (1/32 x 5) + (0.0006)(360) = 0.37 cycles

Benchmark B, Shared Cache: AMAT = (1/32 x 40) + (0.0003)(180) = 1.30 cycles

***Private cache is better** for each of these benchmarks.*