Lecture 13 Cache and Memory Review Exercises

Q1:
Consider two computers A and B:

Parameters	A	B
Base CPI	1.0	2.0
Percentage of load/store instructions	50%	50%
L1 caches hit time	1 cycle	2 cycles
L1 instruction cache miss rate	0%	0%
L1 data cache miss rate	4%	5%
L2 cache hit time	15 cycles	10 cycles
L2 cache local miss rate	75%	80%
Main memory access time	250 cycles	100 cycles

What is the CPI for each computer? What is the Average Memory Access Time (AMAT) for each computer?

Solution:
Given L1 instr and data caches hit time for A and B => all L1 caches hit time fits within the base CPI

So CPI = CPI_base + (%cache_access * %miss * Cycles_miss), where Cycles_miss = Cycles_hit + %miss * Cycles_miss at the next level

CPI of A = 1.0 + (50%*4%*(15+75%*250)) = 5.05
CPI of B = 2.0 + (50%*5%*(10+80%*100)) = 4.25

AMAT = t_avg(L1) = CPI_base + (%cache_access * %miss * Cycles_miss)
AMAT of A = 1.0 + (50%*4%*(15+75%*250)) = 5.05 cycles
AMAT of B = 2.0 + (50%*5%*(10+80%*100)) = 4.25 cycles

Q2:
Given a system with:

- 2 memory channels
- 2 DRAM DIMMS (1 DIMMS per channel)

Each DIMM has:

- 1rank
- 8 chips per rank
- 8-bit column size
- 4 banks per chip
- 32,768 rows per bank
- 2,048 columns per bank
- 8-byte bus

Assume a minimum number of bits needed to cover the physical address space are used for physical addresses. Physical addresses are assigned to use the row interleaving scheme. Also assume that the upper bit(s) of the physical address are used to select the channel. Please determine the address mapping of the physical address, i.e., which portions of the physical address bits are used for: column, channel, bank, row, byte in bus offset and in what order?

Solution:

Memory per chip = banks per chip * rows per bank * columns per bank * bits per column = 4 * 32768 * 2048 * 8 = 2Gbits
Memory per DIMM = memory per chip * ranks per DIMM * chips per rank = 2Gbit * 1 * 8 = 16Gbit
Total physical memory = memory per DIMM * DIMMs = 16Gbit x 2 = 32Gbits = 4GBytes

Minimum of physical address bits needed = log2(4G) = log2(4*1024*1024*1024) = 32 bits

Row interleaving is used, So the order of the IDs is as follows:

| Channel ID | Row ID | Bank ID | Column ID | Byte in Bus Offset |

2 channels → Channel ID is 1 bit
4 banks per chip → 4 banks per rank, because all the chips in a rank has the same number of banks → Bank ID is 2 bits
32,768 rows per bank → Row ID is log2(32768) = 15 bits
2048 columns per bank → Column ID is log2(2048) = 11 bits
8B bus → Byte in Bus offset is log2(8) = 3 bits

So, address mapping is as follows:
| Channel (31) | Row (30:16) | Bank (15:14) | Column (13:3) | Byte in Bus Offset (2:0) |

Q3 (Challenging problem, not required):

Consider two computers A and B:

Parameters	A	B
ISA	MIPS	x86

Clock rate	2 GHz	3 GHz
Base CPI	1	2
Number of pipeline stages	5	10
Percentage of branch instructions	10%	10%
Branch miss predictions	10%	5%
Branch misprediction penalty	1 cycle	3 cycles
Percentage of load/store instructions	30%	30%
L1 instruction cache hit time	1 cycle	1 cycle
L1 instruction cache miss rate	2%	2%
L1 data cache hit time	1 cycle	2 cycles
L1 data cache miss rate	8%	5%
L2 cache hit time	15 cycles	12 cycles
L2 cache global miss rate	2% for instruction fetch, 3% for data access	1% for instruction fetch, 4% for data access
Main memory access time	125 ns	100 ns

What is the CPI for each computer? What is the Average Memory Access Time (AMAT) for each computer? Which computer is faster and by how much, on average, if programs execute 1.25 times as many MIPS instructions as x86 instructions (hint: please compare based on execution time)?

Solution:

Given the clock rates of A and B,
Main memory access time for A = 125 ns = 250 cycles
Main memory access time for B = 100 ns = 300 cycles
Programs execute 1.25 times as many MIPS instructions as x86 instructions =>
Assume instr count of B is N, then instr count of A is 1.25N

Given L1 instr and data caches hit time for A and B => all L1 caches hit time fits within the base CPI

So CPI = CPI_base + (%cache_access * %miss * Cycles_miss) +
(%branch * %miss_prediction * miss_penalty)
= AMAT + (%branch * %miss_prediction * miss_penalty)

For computer A:

First, let's calculate AMAT_A:
AMAT = CPI_base + (%cache_access * %miss * Cycles_miss)
= CPI_base + (%instr_cache_access * %instr_miss * Cycles_instr_miss) +
(%data_cache_access * %data_miss * Cycles_data_miss)

Note that miss rates in the above equation (i.e., %instr_miss and %data_miss) are "local" miss rates, where
Local miss rate of L2 = number of misses in L2 / number of accesses to L2

Global miss rate of L2 = number of misses in L2 / total number of load and stores
= number of misses in L2 / number of accesses to L1

So,
L2 instr local miss rate for A = 2%/2% = 100%
L2 data local miss rate for A = 3%/8% = 37.5%
AMAT_A = 1 + 100% * 2% * (15 + 100% * 250) + 30% * 8% * (15 + 37.5% * 250) = 8.91 cycles

Calculate CPI_A:
CPI_A = AMAT_A + (%branch * %miss_prediction * miss_penalty)
= 8.91 + (10% * 10% * 1) = 8.92

Execution time for A = CPI_A * Instr_A * Cycle_time_A = 8.92 * 1.25N * 0.5ns = 5.575N ns

For computer B:

L2 instr local miss rate for B = 1%/2% = 50%
L2 data local miss rate for B = 4%/5% = 80%
AMAT_B = 2 + 100% * 2% * (12 + 50% * 300) + 30% * 5% * (12 + 80% * 300) = 9.02 cycles

CPI_B = AMAT_B + (10% * 5% * 3) = 9.02 + (10% * 5% * 3) = 9.035
Execution time for B = CPI_B * Instr_B * Cycle_time_B = 9.035 * N * 0.33ns = 2.98N ns

Therefore, Speedup of B over A = 5.575N / 2.98N = 1.87 => B is faster