

QUESTION FOR CORE COURSE

CS665 - COMPUTER ARCHITECTURE

Question 1. A new processor with a 1.25 ns cycle time has both an on chip L_1 cache and an off chip L_2 cache. Measurements of a number of typical programs showed that 25% of all instructions had a data access. The L_2 cache has been chosen to be the largest cache available with a reasonable cost. You are to decide how the L_1 cache is to be configured.

The L_2 is a unified (instructions and data), 8-way associative mapped, 1MB cache with a block size of 256 bytes. If there is a miss in the L_2 cache, 40 CPU cycles are needed to access the main memory and then data/instructions are transferred from main memory to the L_2 cache at a rate of 4 bytes per CPU cycle. The missrate for this cache is 0.0004.

Several different designs for the L_1 cache are possible:

(a) A unified (instructions and data) 64KB, 4-way associative cache. The block size can be 32, 64 or 128 bytes. The corresponding miss rates for this cache are 0.0089, 0.0059, and 0.0047, respectively.

(b) Split caches, *i.e.* separate instruction and data caches, each of which is 32KB and 2-way associative. The block sizes of the instruction and data caches can be the same but do not have to be the same. The possible block sizes are 32, 64 or 128 bytes. The table gives the miss rates for each block size for an instruction cache and a data cache, each of which is 32KB and 2-way associative.

Block size (bytes)	Miss rate for an instruction cache	Miss rate for a data cache
32	0.0041	0.0380
64	0.0022	0.0289
128	0.0012	0.0245

Independent of the block size, if there is a miss in the L_1 cache, 10 CPU cycles are needed to access the L_2 cache and then data or instructions are transferred from the L_2 cache to the L_1 cache at the rate of 8 bytes per CPU cycle.

What is the optimal design for the L_1 cache?

Show ALL work; no credit will be given unless the work is shown.

Question 2. Consider a memory system with two level of cache, which has the following characteristics:

- L1 has a hit rate of 90
- L2 has a hit rate of 95
- Main memory has a latency of 50 cycles,
- The memory bus is two words wide, and can transfer two words per 8 cycles, read or write
- The caches use write-back, and assume that at any given time, a block has a 5% chance of being dirty.

- (a) How many cycles does a load instruction take, on average, on this memory system?
- (b) Assume we are able to clock the processor and cache at double the original rate, while leaving the main memory interface alone. Thus, main memory looks twice as slow to the processor as it did before. Compute the new average cycles for a load instruction.

Computer Architecture Paper Diagnostic – Spring 2002

[based on Culler and Singh, Chapter 5]

1. As the latency of cache misses increases, should an update protocol be more or less preferred relative to an invalidate protocol? Explain briefly the relevant trade-offs, and any additional key parameters that affect your answer. You are not required to quantify your answer.

2. The Firefly cache update protocol is obtained from the Xerox PARC Dragon update protocol by eliminating the Sm (shared-modified) state present in the latter. The Sm state is eliminated by ensuring that the main memory is also updated when caches are updated.
 - (a) Draw a state-transition diagram for the Firefly update protocol similar to that for the Dragon protocol shown in Figure 5.16.

 - (b) Show the actions taken by the processor and the bus, the states in processors, and who supplies the data for the Firefly protocol, on the following reference stream. Your answer should be a Table similar to Figure. 5.17 in the text book.
R1, R2, R3, W1, W2, W3, R1, R2, R3, W3, W2, W1.