國 立 交 通 大 學 試 題 紙

一百零八學年度第二學期
博士班資格考

科目：計算機架構 B　　日期： 109 年 7 月 14 日　第 1 頁 共 3 頁

請 "✓" 明　　✓不可看書　　可看書

* 請將答案依題號順序寫入答案卷

答題時字跡需工整，否則不予計分。Write your answers legibly; otherwise you will get zero score.

1. [10%] Nowadays, deep neural networks achieve outstanding performance on applications such as image classification and object detection. To handle huge amount of calculation efficiently, we should choose suitable architecture for performing these deep learning (DL) applications.
   (a) What is the bottleneck of traditional instruction sets and corresponding general-purpose processors (e.g., single-core processors) for executing the calculation of DL applications? (3%)
   (b) Some x86 architectures which support instruction set extensions for multimedia, e.g., AVX, are used to accelerate these DL applications. What are the advantage(s) and limitation(s) of the instruction set extensions? (3%)
   (c) Graphics processing units (GPUs) are usually the first choice for DL computation. What type(s) of parallelism does a GPU intend to exploit or take advantage of? How does a GPU exploit the type(s) of parallelism? (4%)

2. [10%] Average memory access time (AMAT) is an important metric for evaluating memory/cache performance and can be reduced by different cache optimization techniques.
   HINT: AMAT = (Hit Time) + (Miss Rate) * (Miss Penalty)
   (a) Please describe how "non-blocking cache" works, and how does it influence the cache performance. (4%)
   (b) Please describe how "critical word first and early restart" works, and how does it influence the cache performance. (4%)
   (c) Between "non-blocking cache" and "critical word first and early restart", which one introduces higher hardware cost? (2%)

◎請用深黑色鋼筆或原子筆出題

命題老師簽名：

3. [18%] Given a 32-byte cache (byte-addressable, initially empty) and a sequence of access addresses: $(6)_{10}$, $(24)_{10}$, $(34)_{10}$, $(0)_{10}$, $(28)_{10}$, $(4)_{10}$, please derive the corresponding hit/miss sequence for each of the following cache designs. For each design, draw the same table in your answer sheet and fill out the table you draw.

(a) 16 bytes per block, direct-mapped (6%)

Address	Cache index	Hit or miss?
6		
24		
34		
0		
28		
4		

(b) 8 bytes per block, 2-way set associative (6%)

Address	Cache index	Hit or miss?
6		
24		
34		
0		
28		
4		

(c) 8 bytes per block, fully associative (6%)

Address	Cache index	Hit or miss?
6		
24		
34		
0		
28		
4		

4. [12%] You are given two n-by-n matrices (A and B) and asked to multiply matrix A by matrix B (note: it's A×B, not B×A). The conceptual implementation is shown below.

```
// Matrix multiplication: C = A×B
n = SIZE_OF_MATRIX;
int A[n][n], B[n][n], C[n][n]; // assuming row-major order
// Set matrices A and B, and initialize matrix C here!
for (int i = 0; i < n; i++)
    for (int j = 0; j < n; j++)
        for (int k = 0; k < n; k++)
            C[i][j] += A[i][k] * B[k][j];
```

*** **DO THIS** ***

(a) From the perspective of cache vs. memory hierarchy and assuming that the cache has a block size significantly smaller than n^2, please predict the cache hit/miss behavior based on the implementation shown above, by discussing the reference locality of matrix A and the reference locality of matrix B during the multiplication. Calculating the values of cache hit/miss rate is NOT required. (6%)
HINT: There are two types of reference locality, spatial and temporal.

(b) Can we change the block size to improve the cache hit/miss behavior? Why? (3%)

(c) Can we change the replacement policy to improve the cache hit/miss behavior? Why? (3%)