

國立中山大學 112 學年度

碩士班暨碩士在職專班招生考試試題

科目名稱：計算機結構【資工系碩士班甲組、乙組】

— 作答注意事項 —

考試時間：100 分鐘

- 考試開始鈴響前不得翻閱試題，並不得書寫、劃記、作答。請先檢查答案卷（卡）之應考證號碼、桌角號碼、應試科目是否正確，如有不同立即請監試人員處理。
- 答案卷限用藍、黑色筆(含鉛筆)書寫、繪圖或標示，可攜帶橡皮擦、無色透明無文字墊板、尺規、修正液（帶）、手錶(未附計算器者)。每人每節限使用一份答案卷，請衡酌作答(不得另攜帶紙張，亦不得使用應考證空白處作為計算紙使用)。
- 答案卡請以 2B 鉛筆劃記，不可使用修正液（帶）塗改，未使用 2B 鉛筆、劃記太輕或污損致光學閱讀機無法辨識答案者，後果由考生自負。
- 答案卷（卡）應保持清潔完整，不得折疊、破壞或塗改應考證號碼及條碼，亦不得書寫考生姓名、應考證號碼或與答案無關之任何文字或符號。
- 可否使用計算機請依試題資訊內標註為準，如「可以」使用，廠牌、功能不拘，唯不得攜帶具有通訊、記憶或收發等功能或其他有礙試場安寧、考試公平之各類器材、物品（如鬧鈴、行動電話、電子字典等）入場。
- 試題及答案卷（卡）請務必繳回，未繳回者該科成績以零分計算。
- 試題採雙面列印，考生應注意試題頁數確實作答。
- 違規者依本校招生考試試場規則及違規處理辦法處理。

國立中山大學 112 學年度碩士班暨碩士在職專班招生考試試題

科目名稱：計算機結構【資工系碩士班甲組、乙組】

題號：434001

※本科目依簡章規定「不可以」使用計算機(問答申論題)

共 3 頁第 1 頁

1. (20% total) Performance Calculations
 - 1.1 (5%) The miss latency is 10 cycles. If the direct-mapped cache has a two-cycle hit latency, what is the average memory latency with a 15% miss rate?
 - 1.2 (5%) The miss latency is 10 cycles. If the set-associative cache has a three-cycle hit latency, what the miss rate must the set-associative cache obtain to have a lower average latency than the direct-mapped cache with 15% miss rate?
 - 1.3 (5%) Consider a workload with 30% branches and a 66.66% branch prediction accuracy (33.33% misprediction rate). You may ignore memory operations. For a simple five-stage pipeline, what is the CPI of this workload assuming a three-cycle misprediction penalty and a single-cycle latency for all other instructions?
 - 1.4 (5%) Consider a workload with 30% branches and a 66.66% branch prediction accuracy (33.33% misprediction rate). You may ignore memory operations. Calculate under what condition a much deeper pipeline with double the core clock frequency will outperform the shallower pipeline?
2. (15% total) A memory system has four channels, and each channel has two ranks of DRAM chips. Each memory channel is controlled by a separate memory controller. Each rank of DRAM contains eight banks. A bank contains 32K rows. Each row in one bank is 8KB. The minimum retention time among all DRAM rows in the system is 64 ms. In order to ensure that no data is lost, every DRAM row is refreshed once per 64 ms. Every DRAM row refresh is initiated by a command from the memory controller which occupies the command bus on the associated memory channel for 5 ns and the associated bank for 40 ns. Let us consider a 1.024 second span of time. We define utilization (of a resource such as a bus or a memory bank) as the fraction of total time for which a resource is occupied by a refresh command. (Note: For each calculation in this question, you may leave your answer in simplified form in terms of powers of 2 and powers of 10.)
 - 2.1 (5%) How many refreshes are performed by the memory controllers during the 1.024 second period in total across all four memory channels?
 - 2.2 (5%) What command bus utilization, across all memory channels, is directly caused by DRAM refreshes?
 - 2.3 (5%) What bank utilization (on average across all banks) is directly caused by DRAM refreshes?
3. (15% total) A benchmark searches for an entry in a linked list build from the following structure, which contains a key, a pointer to the next node in the linked list, and a pointer to the data entry.

```
struct node {
    int key;
    struct node *next;
    struct data *ptr;
}
```

The following RISC-V code shows the core of the benchmark, which traverses the linked list and finds an entry with a particular key.

```
loop: LW    x3, 0(x1)      # load a key
      LW    x4, 4(x1)      # load the next pointer key
      SEQ   x3, x3, x2     # set x3 if x3 == x2
      BNEZ  x3, end        # find the entry
      ADD   x1, x0, x4
      BNEZ  x4, loop       # check the next code
end:
```

We run this benchmark on a single-issue in-order processor. The processor can fetch and issue (dispatch) one instruction per cycle. If an instruction cannot be issued due to a data dependency, the processor stalls. Integer instructions take one cycle to execute and the result can be used in the next

國立中山大學 112 學年度碩士班暨碩士在職專班招生考試試題

科目名稱：計算機結構【資工系碩士班甲組、乙組】

題號：434001

※本科目依簡章規定「不可以」使用計算機(問答申論題)

共 3 頁第 2 頁

cycle. For example, if SEQ is executed in cycle 1, BNEZ can be executed in cycle 2. We also assume that the processor has a perfect branch predictor with no penalty for both taken and not-taken branches.

3.1 (5%) Assume that the system does not have a cache. Each memory operation directly accesses main memory and takes 50 CPU cycles. The load/store unit is fully pipelined, and non-blocking. After the processor issues a memory operation, it can continue executing instructions until it reaches an instruction that is dependent on an outstanding memory operation. How many cycles does it take to execute one iteration of the loop in steady state?

3.2 (5%) Now we add zero-overhead multithreading to the pipeline. A processor executes multiple threads, each of which performs an independent search. Hardware mechanisms schedule a thread to execute each cycle. In the first implementation, the processor switches to a different thread every cycle using fixed round robin scheduling. Each of the N threads executes one instruction every N cycle. What is the minimum number of threads that we need to fully utilize the processor, i.e., execute one instruction per cycle?

3.3 (5%) We change the processor to only switch to a different thread when an instruction cannot execute due to data dependency (data-dependent switching). What is the minimum number of threads to fully utilize the processor now? Note that the processor issues instructions in order in each thread.
Hint: consider how many instructions each thread can execute in steady state.

4. (15% total) For a given compute kernel, we define a tensor's reuse interval (RI) as the number of different elements of that tensor that have been referenced between each re-reference of the same element. For example, consider the following:

for m in $[0, M)$

for n in $[0, N)$

$Z[m, n] = A[m] * B[n]$

Since A 's element is used at every iteration of the inner loop, its RI is 1. Each element of B is re-referenced after N references, so its RI is N . Z has "infinite" reuse interval (i.e., no data reuse) since no element is re-referenced throughout the computation:

RI of $A = 1$

RI of $B = N$

RI of $Z = \text{infinite} / \text{no reuse}$

4.1 (5%) Consider the following Matrix-Matrix multiply pseudocode, which multiplies two dense matrices A and B to produce Z :

;; Multiply two matrices A and B to produce Z

;; First matrix A is $M \times K$

;; Second matrix B is $K \times N$

;; Thus, resulting matrix Z is $M \times N$

for m in $[0, M)$

for n in $[0, N)$

for k in $[0, K)$

$Z[m, n] += A[m, k] * B[k, n]$

What are the reuse intervals for the three matrices? Provide your answers in terms of M , N , and K .

4.2 (5%) Now consider the following where we re-order the loop nest:

for k in $[0, K)$

for m in $[0, M)$

for n in $[0, N)$

$Z[m, n] += A[m, k] * B[k, n]$

What are the reuse intervals for the three matrices? Provide your answers in terms of M , N , and K .

國立中山大學 112 學年度碩士班暨碩士在職專班招生考試試題

科目名稱：計算機結構【資工系碩士班甲組、乙組】

題號：434001

※本科目依簡章規定「不可以」使用計算機(問答申論題)

共 3 頁第 3 頁

4.3 (5%) Now consider the following scenario:

- $M = 500$, $N = 1000$, and $K = 30$
- A , B , and Z are dense matrices.
- The matrices are resident in DRAM at the beginning of the computation kernel.

Your wish to design a matrix-matrix multiply accelerator given the above assumptions. Your goal is to maximize compute intensity, defined as the average number of computations you perform per DRAM access. You are given the following hardware budget:

- A multiply-accumulate (MAC) unit with a flip-flop on each input and the output.
- An SRAM array large enough to store 32 elements.

Which order of loop nest in Questions 4.1 or 4.2, would you choose, and given the order how would you partition the SRAM budget among the different elements to maximize compute intensity? You can reorder the loop nests from Questions 4.1 and 4.2, but do not add more loops (e.g., no tiling).

5. (15% total) Considering a system which applies the Sequential Consistency (SC). Suppose that three processes, P1, P2 and P3 are applied with different processors on a system (the values of RA, RB, RC were all zeros before the execution):

P1	P2	P3
P1.1: ST (A), 1	P2.1: ST (B), 1	P3.1: ST (C), 1
P1.2: LD RC, (C)	P2.2: LD RA, (A)	P3.2: LD RB, (B)

After all processes are executed, it is possible for the system to have multiple machine states. For example, $\{RA, RB, RC\} = \{1, 1, 1\}$ is possible if the execution sequence of instructions is $P1.1 \rightarrow P2.1 \rightarrow P3.1 \rightarrow P1.2 \rightarrow P2.2 \rightarrow P3.2$. Also, $\{RA, RB, RC\} = \{1, 1, 0\}$ is possible if the sequence is $P1.1 \rightarrow P1.2 \rightarrow P2.1 \rightarrow P3.1 \rightarrow P2.2 \rightarrow P3.2$. For each state of $\{RA, RB, RC\}$: $\{0, 0, 0\}$ $\{0, 1, 0\}$ $\{1, 0, 0\}$ $\{0, 0, 1\}$, specify the execution sequence of instructions that results in the corresponding state. If the state is NOT possible with SC, just put X.

6. (20% total) Assuming that you want to design a more efficient, machine-specific implementation of an algorithm by observing an existing cache system. The patterns that access various bytes in the system can be used to determine the cache characteristics. However, the only statistic that you can collect on this system is cache hit rate after performing the access patterns as listed below:

Sequence No.	Addresses Accessed (Oldest to Youngest)							Hit Rate
1	0	4	8	16	64	128		1/2
2	31	8192	63	16384	4096	8192	64	5/8
3	32768	0	129	1024	3072	8192		1/3

Assume that the cache is initially empty at the beginning of the first sequence, but not at the beginning of the second and third sequences. The sequences are executed back-to-back, i.e., no other accesses take place between the three sequences. Thus, at the beginning of the second (third) sequence, the contents are the same as at the end of the first (second) sequence. Based on what you observe, what are the following characteristics of the cache? Explain to get points!

- 6.1 (5%) Cache block size (8, 16, 32, 64, or 128 B)?
- 6.2 (5%) Cache associativity (1-, 2-, 4-, or 8-way)?
- 6.3 (5%) Cache size (4 or 8 KB)?
- 6.4 (5%) Cache replacement policy (LRU or FIFO)?