**1. (50%)** True or false questions

1. The RISC-based machines focused the attention of designers on instruction-level

parallelism and the use of caches.

1. Real-time performance is not highly application dependent.
2. Graphic Processor Units (GPUs) exploit data-level parallelism by applying a single

instructed a collection of data m parallel.

1. All recent instruction set architectures (JSAs) of RlSC processors are load-store.
2. The peak memory bandwidth does not grow as the number of cores grows.
3. For a cache, higher associativity can reduce conflict misses at the cost of increased hit time.
4. An alternative to hardware prefetching is for the compiler to insert instructions to request data before the processor needs 1t.
5. Victim caches are used to reduce miss rates.
6. To virtualize the processor, the VMM must control access to privileged mode and handle interrupts..
7. Simulating enough instructions can obtain accurate performance measures of the memory hierarchy.

ANS: a:F b:F c:T d:F e:T f:T g:F h:F i:T j:F

**2. (20%)** Assume that we make an enhancement to a computer that improves some mode of execution by a factor of 10. Enhanced mode is used 50% of the time, measured as a percentage of the execution time when the enhanced mode is in use. Recall that Amdahl’s law depends on the fraction of the original, unenhanced execution time that could make use of enhanced mode. Thus, we cannot directly use this 50% measurement to compute speedup with Amdahl’s law.

(5%) What is the Amdahl’s law?

(6%) What is the speedup we have obtained from fast mode?

(6%) In the textbook, Amdahl’s law is for uniprocessor. Justify if it can be applied to multicore and multiprocessor architectures.

ANS: same as 2016 mid

**3. (12%)** One difficult question is deciding when a system is operating properly. Infrastructure providers started offering service level agreements (SLAs) or service level objectives (SLOs). Systems alternate between Service accomplishment and Service interruption with respect to SLA. Transitions between these two states are caused by failures (from state l to state 2) or restorations (from state 2 to state 1). Give two main measures of dependability in detail.

ANS:

Module reliability = measure of continuous service accomplishment (or time to failure).

Module availability measures service as alternate between the 2 states of accomplishment and interruption

**4.**A cache is a hardware or software component that stores data so future requests for that

data can be served faster. A cache hit occurs when the requested data can be found in a cache, while a cache miss occurs when it cannot. Cache hits are served by reading data from the cache,which is faster than recomputing a result or reading from a slower data store; thus, the more requests can be served from the cache, the faster the system performs.

(a) (6%) Give three cache structures. (configurations).

(b)(8%) What are the four main questions of a cache?

ANS:

(a)Direct mapped、Fully associative、Set associative

(b)Block placement、Block identification、Block replacement、Write strategy

**5.(16%)** Compiler optimization is generally implemented using a sequence of optimizing

transformations, algorithms which take a program and transform it to produce a semantically

equivalent output program that uses fewer resources. In particular, it can be applied to improve performance by reducing cache misses.

(a) (6%) Give two compiler optimizations to reduce cache misses.

(b) (10%) Give examples to explain why your answers of(a) can work well.

ANS:

挑兩個寫 b的舉例 寫下code

Merging Arrays: improve spatial locality by single array of compound elements vs. 2 arrays

Loop Interchange: change nesting of loops to access data in order stored in memory

Loop Fusion: Combine 2 independent loops that have same looping and some variables overlap

Blocking: Improve temporal locality by accessing “blocks” of data repeatedly vs. going down whole columns or rows

**6.(20%)** The cache is a piece of hardware which reduces the access time to the data in the

memory by keeping some part of the frequently used data of the main memory in itself. It is smaller and faster than the main memory.

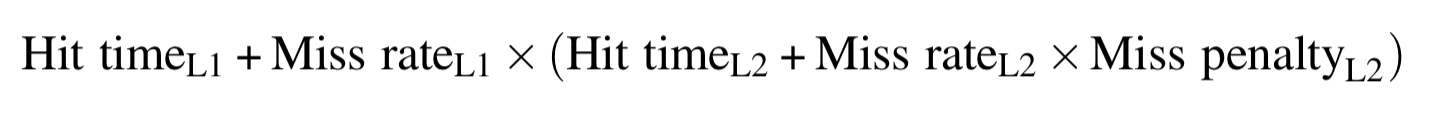
(a) (6%) Give the equation of average memory access time for a two-level cache.

(b) (8%) Suppose that in 1000 memory references there are 40 misses in the first-level cache and 20 misses in the second-level cache. What are the various miss rates?

(c) (8%) Assume the miss penalty from L2 cache lo memory is 100 clock cycles, the hit time

of L2 cache is 10 clock cycles, the hit time of L1 iis 1 clock cycle, and there are 1.5 memory references per instruction. What is the average memory access time and average stall cycles per instruction? Ignore the impact of writes.

ANS:

(a)

(b)

1st-level local and global miss rate = 40 / 1000 = 4%

2nd-level local miss rate = 20 / 40 = 50%

2nd-level global miss rate = 20 / 1000 = 2%

c)

Average memory access time

　= 1 + 4%(10 + 50% × 100 ) = 3.4 clock cycles.

(若無L2 => Average memory access time = 1 + 4% × 100 = 5 clock cycles.)

1.5 memory references per instruction => 1000 memory reference per 667 instructions.

所以 每千個指令的失誤率 Miss Rate L1 = 40\*1.5 = 60 , Miss Rate L2 = 20\*1.5=30

Average memory stalls per instruction

= Misses per instruction L1× Hit time L2 + Misses per instruction L2 × Miss penalty L2

= (60/1000) × 10 + (30/1000) × 100

= 0.060 × 10 + 0.030 × 100 = 3.6 clock cycles

另一種算法是 (Average memory access time - L1 Hit time ) × 平均Cache存取次數

= (3.4 – 1.0) \* 1.5 = 3.6 clock cycles.