605410080

2017 Computer Architecture – Midterm Exam

ISA.

1. (50%) True or false questions

F a. The RISC-based machines focused the attention of designers on instruction-level parallelism and the use of caches.

F b. Real-time performance is not highly application dependent.

T. c. Graphic Processor Units (GPUs) exploit data-level parallelism by applying a single instruction to a collection of data in parallel. SIMD

F d. All recent instruction set architectures (ISAs) of RISC processors are load-store. ?

T. e. The peak memory bandwidth does not grow as the number of cores grows.

V T f. For a cache, higher associativity can reduce conflict misses at the cost of increased hit time.

F g. An alternative to hardware prefetching is for the compiler to insert instructions to request data before the processor needs it. software

F h. Victim caches are used to reduce miss rates. penal

T i. To virtualize the processor, the VMM must control access to privileged mode and haldle interrupts..

F j. Simulating enough instructions can obtain accurate performance measures of the memory hierarchy.

multi core O

multi process.

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.

(a) (4%) What is the Amdahl's law?

(b) (8%) What is the speedup we have obtained from fast mode?

(c) (8%) In the textbook, Amdahl's law is suitable for uniprocessor. Justify if ~~it can be~~ applied to parallel computing. No.

MTTF

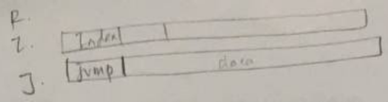
MTTF + MTTR

[illegible]

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 1 to state 2) or restorations (from state 2 to state 1). Give two main measures of dependability in detail.

Mean Time to Fa

4. (14%) 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 structure (configurations).. Direct mapping.

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

4[illegible] write through, write back, write allocation
write around.

✓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. loop Interchange

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

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. local miss rate

(a) (6%) Give the equation of average memory access time for a two-level cache. 40 ×

(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 to memory is 100 clock cycles, the hit time of L2 cache is 10 clock cycles, the hit time of L1 is 1 clock cycles, 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.

$0.04 \times 12 = 0.48$

$0.04 \times 12 = 0.48$

1 0.04 10. 0.02 100

d = Hit time + Miss Rate×(Hit time + Miss Rate×
Miss Palnaley.)

L2 miss penalty. — 100

= 1 + 100. +

$$CPI = \frac{\text{Clock cycle for a program}}{\text{Instruction Count}}$$

local miss rate

L1: $\frac{40}{1000}$ = 0.04. global miss rate.

L2: 0.02

$0.04 \times 0.02 = 0.0008$

hit ~~miss~~ time
miss rate
miss penalty.