NTNU – Trondheim
Norwegian University of
Science and Technology

Department of Computer and Information Science

Examination paper for TDT4260 Computer Architecture

Academic contact during examination: Lasse Natvig

Phone: 906 44 580

Examination date: 18. May 2015

Examination time (from-to): 09:00 – 13:00

Permitted examination support material: code D; No written or handwritten examination support materials are permitted. A specified, simple calculator is permitted.

Other information:

The exam accounts for 80% of the final grade, and the provided points show the maximal number of points that can be achieved on each assignment. Read the problem texts thoroughly. You can answer the questions in English or Norwegian.

For all multiple choice questions: Answer by writing the question-ID and one alternative, like this: "X1 b" where X1 is the question ID and b is your answer. You are awarded 3.0 points for a correct answer and 0 points if you do not answer. If your answer is wrong or you give more than one alternative, you will get -1.5 points.

Language: English

Number of pages: 4

Checked by:

……..

10/5-2015 *Antonio Garcia* (sign)

Date Signature

Problem A: Benchmarks and technology trends (Max 12 points)

A1: *SPEC2006* is a desktop benchmark focusing on processor performance. It is based on real programs that

a) typically use 1 CPU second when executing on an average CPU in year 2006.
b) are modified to be portable and to minimize the effect of I/O on performance.
c) have approximately the same amount of integer and floating point operations.

A2: *Kernels or microbenchmarks* are often better than applications for

a) gaining understanding of the performance of details in the computer architecture.
b) testing the overall usefulness of a computer.
c) measuring the throughput of multiprocessor servers.

A3: The term "*dark silicon*" is used to denote the fact that in the future processors

a) will be made in silicon with a feature size that is so small that impurities will lead to several parts of the chip to be malfunctioning, or dark. These can be tolerated using redundancy.
b) will be made in a new material composed of both silicon and dark (black) carbon nanotubes.
c) will have so many transistors that only a subset of them can be powered at the same time to stay within the power budget of the chip.

A4: *Pollack's rule* is based on historical evidence and says that

a) the performance of a processor scales linearly with the numer of transistors used to implement the processor.
b) the performance we get by microarchitectural techniques when the number of transistors grow with a factor *n* is roughly the square root of *n*.
c) 50% of the chip area should always be used to implement caches.

Problem B: Cache systems (Max 9 points)

B1: The difference in the development of CPU and memory performance has

a) motivated for using fully associative caches instead of set-associative caches.
b) enabled integration of cache on the same chip as the processor.
c) been a main reason for using extensive, often multi-level cache systems.

B2: The 2:1 *cache rule of thumb* states that

a) the instruction set cache should be twice the size of the data cache to achieve a balanced hit rate.
b) a direct-mapped cache of size N has about the same miss rate as a two way set-associative cache of size N/2.
c) the miss rate in a cache is roughly doubled when its size is reduced from N to N/2.

B3: The *principle of locality* is an important property of programs that is exploited to improve performance. Which statement regarding locality is most correct?

a) Temporal locality applies only to data and spatial locality only to instructions.
b) A sequence of memory references that has good (high) temporal locality will most probably have a bad (low) spatial locality.
c) A sequence of memory references may have both good temporal locality and good spatial locality.

Problem C: Processors (Max 9 points)

C1: A so-called write-after-read (WAR) hazard can occur in a processor pipeline when

a) some instruction writes to a dataword that has been loaded into cache before the next instruction is able to read this dataword from the cache.
b) a stage in the pipeline writes to one of the operands of the instruction that already has been fetched in an operand-fetch stage.
c) there is a write stage early in the processor pipeline and a read stage late in the pipeline.

C2: Load linked (LL) and store conditional (SC) are two instructions that

a) are easier to implement in a processor than an atomic swap, and can be used to implement an atomic swap (exchange) operation.
b) are executed after each other as a sequence of two instructions to implement a lock-operation, and the last instruction (SC) will not return before it has got exclusive access to the lock.
c) that is available in *all* RISC processors for implementing a barrier synchronization.

C3 : The use of multimedia SIMD extensions like SSE and AVX can give significant performance improvement for some applications. A concept called *partitioned adders* has been used to achieve efficient implementation of such instructions. The advantage of partitioned adders is that

a) the processor efficiently can operate on short vectors of operands of variable length (measured in bits).
b) the number of pipeline-stages for doing an addition is increased.
c) addition and subtraction is performed in different logic, so that more parallelism can be exploited.

Problem D: Parallel processing and energy efficiency (Max 12 points)

D1: The Thermal Design Power (TDP) is a metric that

a) describes the peak (maximum) power consumption of a processor.
b) gives the average power consumption used by a processor during a computation.
c) determines the cooling requirement since the cooling system is usually designed to match or exceed TDP.

D2: Gather-scatter operations are used to

a) support moving of matrices or vectors between a compressed representation and a normal representation (i.e. where zeros are included in a sparse matrix)
b) gather several data-words from a set of memory addresses before they are loaded as one block into the L1 data cache.
c) convert floating-point numbers into more energy efficient fixed point numbers.

D3: Directory based cache coherence protocols

a) implement cache coherence in the operating system by using the directories of the file system.
b) are often used in systems with scalable interconnection networks that do not offer a efficient way of broadcasting information to all processors.
c) implement broadcasting in a more efficient way than a bus.

D4: The purpose of the tag in a tagged token dynamic data flow computer is to

a) allow multiple instances of an arc in a data flow graph (operand in a data flow program) to exist concurrently in different contexts.
b) tag every operand with the name of the storage location storing that operand.
c) tag the input token so that the program input can be routed to their correct functional units.

Problem E: Multiprocessors and memory systems (Max 18 points)

E1: (Max 6 points) Explain the difference between fine grained and coarse grained multithreading, and describe the concept simultaneous multithreading (SMT).

E2: (Max 6 points) Three techniques to interconnect memory modules and processors in a multiprocessor are a) crossbar switch, b) multistage interconnection network and c) bus.
Assume that a multiprocessor has 4 processors (P) and 4 memory modules (M). Sketch each of these three interconnection-techniques by using this multiprocessor as an example. Do not draw caches and IO-modules, but expose the structure of the interconnection.

E3: (Max 6 points) One optimization of cache performance to reduce miss penalty is called *write merging* and is used in cache systems with write buffer. In systems where input/output device registers are mapped into the physical address space, write merging cannot be used for these I/O addresses. Explain why, and how this problem typically is solved in such a cache system.

Problem F: Research papers and multiprocessors (Max 20 points)

F1: (Max 5 points) Explain briefly the main principles that are used in the Matching Unit in the *Manchester Dataflow Machine* (MDM). Two keywords are hashing and an overflow unit.

F2: (Max 6 points) a) Explain briefly how you interconnect two hypercubes of size 8 nodes when you are building a 16 node hypercube. Hint: Which nodes are connected together? How are the nodes addressed? b) Describe also a clear disadvantage of the hypercube as interconnection topology when it comes to extending a parallel computer with more processors.

F3: (Max 5 points) A central concept used in the paper «*Exploring the Design Space of Future CMP's*» is Cache Byte Equivalents (CBE). Explain briefly how this concept was necessary and made it possible to evaluate and compare various designs that had different processors and different cache sizes.

F4: (Max 4 points) A programming technique used to process vectors of arbitrary length on a vector-processor with fixed length vector registers is called strip mining. Explain the technique briefly.

…---oooOOOooo---…