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: 25. May 2013

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-identificator and one alternative, like this: "X1 b" where X1 is the question identificator 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: 7

Checked by:

……..

10/5/2013 *Magnus Jahre* (sign)

Date Signature

Problem A: Performance, multiple choice (Max 15 points)

A1: An important measure of computer performance is clock *cycles per instruction* (CPI). Which one of the following three claims is most correct?

a) The CPI measure depends on the computer architecture.
b) The CPI measure depends on the application running on the computer.
c) The CPI measure depends on both the application and the computer architecture.

A2: A common expression for execution time is

CPU time = y × Clock cycle time × cycles per instruction

where y is

a) the number of instructions in the program.
b) the number of instructions executed.
c) the average time for an instruction including cache misses.

A3: The TPC benchmark measures performance as transactions per second. In this context

a) the focus is on throughput but the response time must be within a certain limit.
b) the response time is proportional to the throughput.
c) response time and throughput are equally important.

A4: The *decreasing feature size* in integrated circuits affects the performance of both the transistors and the wires.

a) Transistors performance and wire delay improve at roughly the same rate resulting in a steady improvement of well-balanced circuits.
b) Global wire delay scales poorly compared to transistor performance.
c) The decrease in feature size has made the wires shorter so that the wire delay is no longer important for the performance of large integrated circuits.

A5: *Superlinear speedup* is

a) a popular term in many press releases, but can never be achieved in practice.
b) may occur as a result of cache or memory effects.
c) is achievable on superscalar processors using out-of-order execution.

Problem B: Memory and cache (Max 24 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 *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.

B3: Where may a new block of data be placed in a 4-way set associative cache?
a) Anywhere
b) (Address MOD (number of sets)) to select set, direct mapping inside the set.
c) (Address MOD (number of sets)) to select set, anywhere inside the set.

B4: Assume a two-level cache with 200 cycles miss penalty from L2 cache to memory and the hit time in the L2 cache is 10 clock cycles. The hit time of L1 is 2 clock cycles, the miss rate to the L1 cache is 10% and the miss rate for accesses to the L2 cache is 25%. What is the average memory access time?
a) 8 clock cycles.
b) 14 clock cycles.
c) 140 clock cycles.

B5: Assume a bus-based multiprocessor with a bus snooping cache coherence protocol. Five processors share the same data variable, and one of the processors updates the variable very frequently. In this situation a write-invalidate cache coherence policy
a) will cause less bus-traffic than a write-update protocol.
b) will cause more bus-traffic than a write-update protocol.
c) will invalidate its own copy to avoid an inconsistent memory.

B6: Directory based cache coherence protocols
a) implements 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 an efficient way of broadcasting information to all processors.
c) implements broadcasting in a more efficient way than a bus.

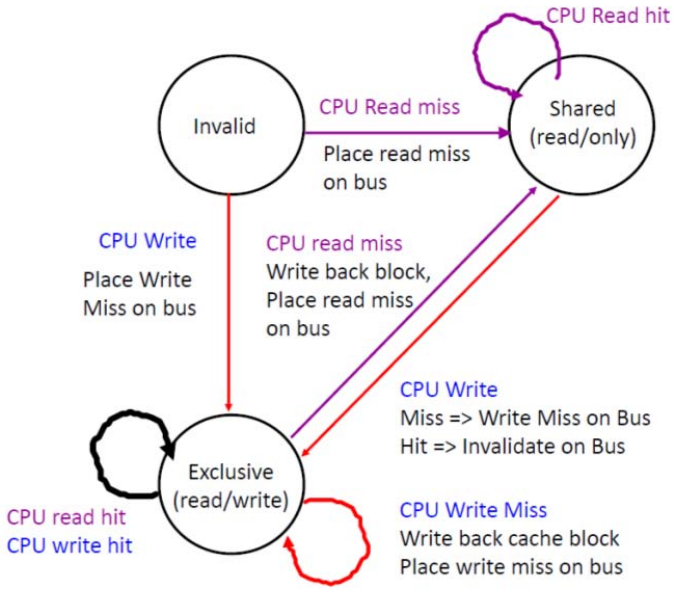

Figure 1: Example cache coherence protocol.

B7: Figure 1 shows a cache coherence protocol for a multiprocessor. In this protocol each cache block can be in one of three states: Invalid, Shared and Exclusive. In the figure we have removed the transition for the case when the processor (CPU) causes a read miss for a block that is in state Shared. This should have been shown as

a) a transition from state Invalid to state Shared marked with the text “Place read miss on bus”
b) a transition from state Shared to state Shared marked with the text “Place read miss on bus”
c) a transition from state Shared to state Exclusive marked with the text “Place read miss on bus and let processor get exclusive access for read operations”

B8: Figure 1 shows the transitions initiated from the CPU. There are also transitions that the cache controller performs upon requests from the bus side (bus requests). If the addressed cache block is in state exclusive and the cache controller observes (by snooping) a write miss for that block, the correct action will be

a) a transition from state Exclusive to state Shared marked with the text “Write back block”
b) a transition from state Exclusive to state Shared marked with the text “Upgrade state of block to a shared stated with combined read/write access”
c) a transition from state Exclusive to state Invalid marked with the text “Write back block”

Problem C: Multiprocessors and HW/SW interface (Max 21 points)

C1: *False sharing* occurs in a multiprocessor

a) when at least one private cache has a copy of a memory block that is inconsistent (unequal) with the contents in the memory.
b) that has a bus-based cache coherence scheme based on broadcasting such that all processors will see all memory updates.
c) when a cache block is shared, but no word in the cache is actually shared.

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: A *distributed shared memory* (DSM) multiprocessor

a) usually has non-uniform memory access.
b) has multiple address spaces on a set of distributed nodes.
c) cannot offer message-passing as a communication mechanism for parallel programs.

C4: Instruction scheduling may be done by compiler (static) or processor (dynamic). Which of the following statements are <u>not</u> true?

a) The compiler has more time to do complex scheduling algorithms than the processor.
b) The compiler must be conservative and assume more conflicts than what actually will occur.
c) Profiling gives the compiler more accurate information about conflicts than what is available to the dynamic scheduler in the processor.

C5: The *Omega interconnection* network

a) was invented by the famous student Carl G. Armfeldt in Trondheim during studies of the Illiac IV research project. The event has given name to the EE-student alumni organization Omega.
b) is less expensive than a crossbar network.
c) can avoid blocking by use of adaptive routing.

C6: The purpose of the tag in a tagged token 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 receiving instruction.
c) tag the input token so that the program input can be routed to their correct functional units.

C7: The advantage of virtual cut through routing compared to wormhole routing is that it

a) more easily finds the shortest route to the destination by using short cuts.
b) spools messages into buffers on its way to the destination to reduce blocking.
c) implements routing in software in a virtual machine through all its software layers.

Problem D: Really Big Computers (Max 8 points)

D1: (Max 4 points) The course textbook describes that architects of a warehouse-scale computer (WSC) share many goals and requirements with server architects. Describe briefly four such goals or requirements. A WSC will process both interactive and batch processing workloads. Describe this using the main service provided by google as an example.

D2: (Max 2 points) Two characteristics of a WSC are *ample parallelism* and *opportunities/problems associated with scale*. Describe these two briefly.

D3: (Max 2 points) Describe briefly the main hardware characteristics of the new Vilje supercomputer at NTNU. (Keywords: processor type and architecture, node architecture, memory system, interconnection network and total size in number of cores)

Problem E: Research papers (Max 12 points)

E1: (Max 4 points) In the paper *Exploring the Design Space of Future CMP's* the authors presents a graph of results shown in Figure 2. Explain briefly how two main architectural parameters were varied to give 8 values in this graph (Eg. see 8 filled squares for the application sphinx in the middle of the figure). Explain briefly also how varying one of these parameters can make an application to "move" from one to another of the three applications classes indicated in the figure.

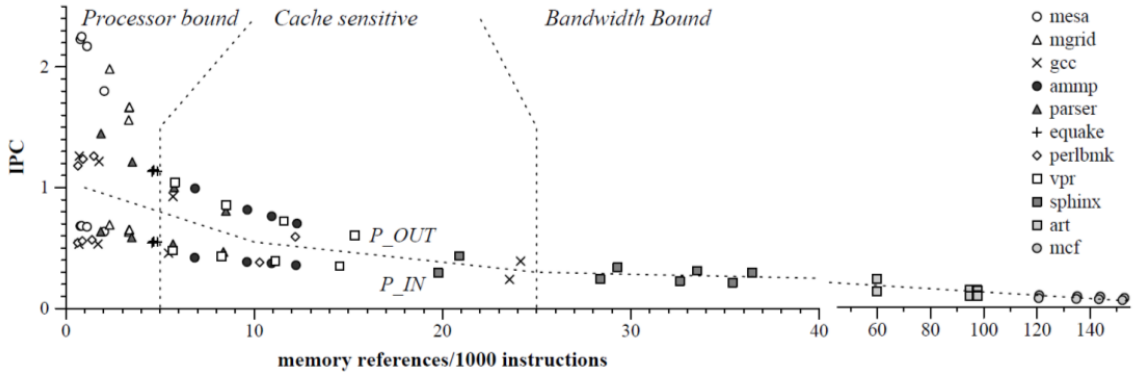

Figure 2: Design Space Exploration

E2: (Max 4 points) In the paper by Borkar and Chien, the authors state that the rapid growth in microprocessor performance for the past 20 years has been enabled by three key technology drivers. Explain briefly these three. The family name of Robert Dennard has often be used to denote one of these trends, and the "classical Dennard scaling" has provided three major benefits that made possible rapid growth in computer performance. Explain these three benefits briefly (exact numbers are not needed)

E3: (Max 4 points) Figure 3 shows a heterogeneous many-core system with variation. Describe briefly how this kind of architecture can be energy efficient. (Hints: workload scheduling and dynamic system configuration)

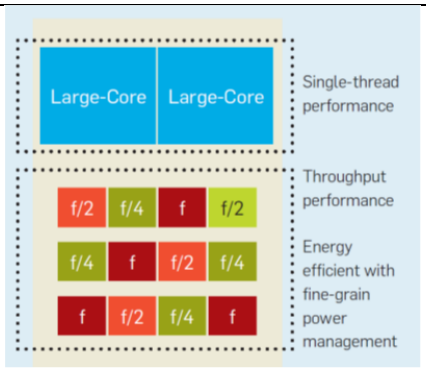

Figure 3: A heterogeneous many-core system with variation.

…---oooOOOooo---…