NTNU – Trondheim
Norwegian University of
Science and Technology

Department of Computer and Information Science

Examination paper for TDT4260 Computer Architecture

Academic contact during examination: Magnus Själander
Phone: +47 928 72 583

Examination date: 2016-06-01
Examination time (from-to): 09:00 – 13:00
Permitted examination support material: Simple calculator

Other information: The exam consists of 40 multiple choice questions. A correct answer to a question gives 2 points and a wrong answer deducts 1 point. Questions that are not answered gives 0 points. There are also five open questions that each give a maximum of 4 points. The maximum score is thus 100 points (80 from multiple choice and 20 from the five open questions).

Only questions answered on the answer sheet will be corrected.

Language: English
Number of pages (front page excluded): 9
Number of pages: 10

Checked by:

_____

Date Signature

This page is intentionally left empty.

Answers sheet. Only answers on this sheet will be corrected. Only those multiple choice answers circled here will be graded.

CANDIDATE NUMBER: _____

1. Fundamentals

1.1.	A	B	C	D
1.2.	A	B	C	D
1.3.	A	B	C	D
1.4.	A	B	C	D

2. Memory Hierarchy

2.1.	A	B	C	D
2.2.	A	B	C	D
2.3.	A	B	C	D
2.4.	A	B	C	D

3. Caches

3.1.	A	B	C	D
3.2.	A	B	C	D
3.3.	A	B	C	D
3.4.	A	B	C	D

4. Instruction-Level Parallelism

4.1.	A	B	C	D
4.2.	A	B	C	D
4.3.	A	B	C	D
4.4.	A	B	C	D

5. Data-Level Parallelism

5.1.	A	B	C	D
5.2.	A	B	C	D
5.3.	A	B	C	D
5.4.	A	B	C	D

6. Thread-Level Parallelism

6.1.	A	B	C	D
6.2.	A	B	C	D
6.3.	A	B	C	D
6.4.	A	B	C	D

7. Warehouse-Scale Computers

7.1.	A	B	C	D
7.2.	A	B	C	D
7.3.	A	B	C	D
7.4.	A	B	C	D

8. Networks

8.1.	A	B	C	D
8.2.	A	B	C	D
8.3.	A	B	C	D
8.4.	A	B	C	D

9. Prefetching

9.1.	A	B	C	D
9.2.	A	B	C	D
9.3.	A	B	C	D
9.4.	A	B	C	D

10. Articles

10.1.	A	B	C	D
10.2.	A	B	C	D
10.3.	A	B	C	D
10.4.	A	B	C	D

Answer each of the five open questions as concisely as possible. Only answers on this sheet will be graded so use the space wisely.

Answer 1:_____

Answer 2:

Answer 3:

Answer 4:

Answer 5:

1. Fundamentals

1. Which of Flynn's terms best describe the MMX and SSE x86 instruction set extensions?
    a. SISD
    b. SIMD
    c. MISD
    d. MIMD
2. What is the main cause for computer performance stopped improving around 2005?
    a. The memory wall
    b. The complexity wall
    c. The power wall
    d. The ILP wall
3. Dennard scaling states that
    a. the number of transistors double ever 18 months.
    b. the core complexity doubles with every architecture generation.
    c. the power is constant.
    d. the energy density remains the same across technology nodes.
4. The yield depends on
    a. the die area.
    b. the average number of defects per unit area.
    c. the size of the wafer.
    d. both **a** and **b**.

2. Memory Hierarchy

1. Which program will benefit most from a cache?
    a. A program that repeatedly processes a small chunk of data before moving on.
    b. A program that repeatedly processes a huge chunk of data all at once.
    c. A program that randomly processes its data.
    d. A program that never uses the same data twice.
2. A virtually indexed physically tagged cache improves
    a. Access latency
    b. Hit rate
    c. Energy efficiency
    d. Throughput
3. Which of the following has a lower average memory access time?
   #1: 32kB L1 with a hit time of 1 cycle and a miss time of 5 cycles to access a 256kB L2 (70% hit in the L1, 100% hit in L2)
   #2: 64kB L1 with a hit time of 2 cycles and a miss time of 4 cycles to access a 128kB L2 (90% hit in the L1, 100% hit in L2)
    a. #1
    b. #2
    c. Same
    d. Need to know something about the program
4. The purpose of a translation lookaside buffer (TLB) is
    a. To translate virtual pages to physical pages
    b. To translate virtual addresses to physical addresses
    c. To translate architectural registers to physical registers
    d. To translate physical addresses to disk accesses

3. Caches

1. What does banked caches improve?
    a. Access time
    b. Throughput
    c. Hit rate
    d. Miss penalty
2. Early restart is a technique where
    a. the pipeline is restarted as soon as the accessed data word reaches the cache on a miss.
    b. the pipeline is restarted as soon as a branch has been evaluated to reduce the branch penalty.
    c. load operations are restarted early when a memory ambiguity has been detected.
    d. None of the above
3. Blocking is a technique where
    a. the cache is divided into blocks to improve throughput.
    b. the critical word is fetched first to avoid stalling.
    c. the compiler organizes data accesses into smaller blocks to fit in the data cache.
    d. the compiler divides a loop into suitably sized blocks to fit in the instruction cache.
4. Way prediction is a technique to
    a. predict the way a branch takes.
    b. predict which way the data resides in the cache.
    c. predict the way in a superscalar pipeline.
    d. None of the above.

4. Instruction-Level Parallelism

1. Tomasulo's algorithm
    a. enables instructions to be dynamically reordered during execution.
    b. enables all memory accesses to be dynamically reordered during execution.
    c. enables precise exceptions during execution.
    d. Neither of the above.
2. A write after read (WAR) hazard
    a. is a true dependency.
    b. cannot happen in a fixed length in-order pipeline.
    c. is an anti-dependency.
    d. Both **b** and **c**.
3. What is **NOT** true about loop-unrolling?
    a. Increases register pressure.
    b. Is a compiler optimization.
    c. Reduces the code size.
    d. Can help hide load latencies.
4. A VLIW machine
    a. relies on the compiler to schedule all instructions.
    b. dynamically detects hazards to reduce their delay penalty.
    c. improves binary compatibility between generations of the architecture.
    d. improves code size.

5. Data-Level Parallelism

1. SIMD represents an organization that
    a. refers to a computer system capable of processing several programs at the same time.
    b. consist of a single computer containing a control unit, processor unit and memory unit.
    c. includes many processing units under the supervision of a common control unit.
    d. None of the above.
2. What best describes the gather-scatter technique?
    a. Is a technique for scattering the data to perform computations and later gather the results. This technique has become very popular for large data sets and warehouse scale computing.
    b. Is an addressing technique to collect sparse data to dense representations for efficient computations and then distribute the results.
    c. Is a technique to gather instructions into bundles and execute them in parallel on multiple data before scattering the results.
    d. Is a technique to gather data of the same type to efficiently perform computations on them before scattering the results.
3. What is true about vector machines?
    a. Can amortize the cost of fetching data over multiple operations.
    b. Can have multiple lanes to work on data in parallel since each element in the vector is independent.
    c. Perform chaining to reduce the latency between vector operations.
    d. All the above.
4. Conditional code causes problems for GPUs because
    a. of branch divergence, which complicates the hardware that need to keep track of threads executing down different paths of the application.
    b. it is costly to implement branch prediction for all the cores. Code with branches can, therefore, only be allocated to specialized cores that support branches.
    c. of branch divergence, which cause some threads to be idle, causing cores to stall, while other threads are executed.
    d. Both **a** and **c**.

6. Thread-Level Parallelism

1. What is the maximum speed up for any computer system with 100 cores executing a parallel application with a sequential section of 2%?
    a. 50x
    b. 100x
    c. 33.6x
    d. Need to know the locality of the application
2. Which statement about cache coherency protocols is **NOT** correct?
    a. Write invalidate protocols are more common than write update protocols.
    b. Write update protocols consume more bandwidth than write invalidate protocols.
    c. Snooping cache coherency protocols use a broadcast interconnect as the point of serialization.
    d. It is impossible to implement a directory-based cache coherency protocol in a system with a bus interconnect.
3. Which of the following statements is **NOT** true about locks?
    a. Locks do not prevent any other processor from writing the data while it is locked.
    b. Locks are a signaling mechanism that lets programs keep track of whether any other processors are currently accessing the data.
    c. One lock per piece of shared data is needed to synchronize processes properly.
    d. Locks are special variables administrated by the operating system.

4. Why are distributed shared memory systems often called NUMA machines?
    a. The access latency to the memory is not uniform.
    b. The bandwidth to memory is not uniform.
    c. The address space is not uniform.
    d. None of the above.

7. Warehouse-Scale Computers

1. What is the main form of parallelism that warehouse-scale computers benefit from?
    a. Instruction level parallelism
    b. Data level parallelism
    c. Thread level parallelism
    d. Request level parallelism
2. Why is a low PUE good?
    a. A low PUE means that there are small losses in the power utility elements.
    b. A low PUE means that the response time of a power utility emergency is low.
    c. A low PUE means that the warehouse-scale computer is energy efficient.
    d. Both **a** and **c**.
3. What is **NOT** true for warehouse-scale computers.
    a. Labor costs are significant part of the total cost.
    b. Warehouse-scale computers have to deal with both interactive and batch workloads.
    c. Typical server loads are less than 50%.
    d. They often rely on relaxed consistency models.
4. What is true for a warehouse-scale computer with a total cost of $120M for the facilities (amortization period of 10 years), $90M for servers (amortization period of 3 years), $24M for networking gear (amortization period of 4 years) and a total monthly cost of $0.75M?
    a. The total CAPEX is $0.75M
    b. The total OPEX is $4.75M.
    c. The total OPEX is 234M.
    d. Both **a** and **c**.

8. Networks

1. What is **NOT** true about crossbar switches?
    a. Crossbar switches are nonblocking.
    b. Crossbar switches scale easily.
    c. Crossbar switches are mostly used as on-chip networks.
    d. Crossbar switches have a high bisection width.
2. What does the bisection bandwidth describe?
    a. The maximum bandwidth through the network.
    b. The minimum bandwidth between any two nodes in the network.
    c. The maximum bandwidth in a bisected network.
    d. The maximum bandwidth between two equal halves of the network.
3. Which one of the following is NOT a network type?
    a. SAN
    b. WAN
    c. OCN
    d. TAN
4. Deadlocks can be avoided by
    a. using direction ordered routing.
    b. reducing the latency of network links.
    c. implementing a back-off mechanism.
    d. increasing the bandwidth of the network links.

9. Prefetching

1. Prefetching is the act of
    a. fetching data from a lower level cache into a higher level cache after a miss.
    b. predicting future memory references and fetching the data before it is referenced by the CPU.
    c. clearing the cache before new instructions are fetched from memory.
    d. fetching data into registers before the execution stage in the pipeline.
2. $\frac{\text{number of useful prefetches}}{\text{total number of prefetches}}$ is best described as
    a. coverage.
    b. accuracy.
    c. timeliness.
    d. correctness.
3. The prefetch degree determines
    a. the distance between the currently accessed cache block and the next fetched cache block.
    b. the number of cache blocks that are fetched.
    c. the number of entries in the prefetcher's history table.
    d. The number of good prefetches divided by the number of cache misses without prefetching.
4. Software prefetching is accomplished by
    a. the operating system when a program is executed.
    b. the compiler by altering the control flow of the program.
    c. the operating system before a program is loaded into memory for execution.
    d. the compiler by inserting special instructions into the program.

10. Articles

1. 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.
    d. None of the above.
2. Why has a multi-core architecture the potential to improve energy efficiency?
    a. Power scales quadratic with supply voltage so running two cores at half the frequency and half the supply voltage maintains the performance while reducing the power.
    b. Power scales quadratic with frequency so running two cores at half the frequency maintains the performance while reducing the power.
    c. Power scales quadratic with capacitance so running two cores with twice the capacity and half the frequency maintains the performance while reducing the power.
    d. Power scales quadratic with the activity factor so running two cores reduces the activity factor by half with maintained performance while reduces the power.
3. What will likely characterize future computer systems?
    a. They will consist of cores with various performance capability.
    b. They will employ accelerators to a large extent.
    c. The general purpose cores will have to deal with more irregular code.
    d. All the above.

4. Pollack's rule states that
    a. core performance is relative to the square root of the area of the core.
    b. core performance is relative to the square of the area of the core.
    c. core complexity is relative to the square root of the area of the core.
    d. core power is relative to the square of the area of the core.

Open questions

1. Concisely describe what is meant by energy proportionality and why it is important.

2. Concisely describe how a web search can be performed in a warehouse-scale computer.

3. Concisely describe what false sharing is and how it can occur in multi-core systems.

4. Concisely describe what each of the letters in MESI stands for (it is not enough to answer with a single word per letter).

5. Concisely describe how a conventional access to a direct mapped virtually indexed physically tagged cache is performed.