PREPARATION FOR THE FINAL

# Information about the exam

Duration: 3 hours

One sheet of paper (cheat-sheet) is allowed.

Format:

* 1 theoretical problem with multiple parts
* 4 problems with multiple parts

Complaining:

* You can come to see the exam on Tues. Dec 21st 13:00-14:00
* To complain: you will fill the form and I will notify you about the decision.

# Course Material

**Introduction and caches**

Objectives:

* Review of processor architectures including RISC, vector, SIMD, superscalar
* Understanding performance, benchmarking, Amhdal’s and Gustafson laws
* Understanding performance challenges
* Understanding cache organization, hierarchy, performance. Understanding advanced topics including prefetching and non-blocking caches

Topics:

* Introduction: Chapter 1
* Cache memories: Chapter 4.3

 Problem to practice:

* Assignments 1 (with solutions from <http://www.site.uottawa.ca/~mbolic/ceg4131/index.shtml>).

**Multiprocessor systems**

Objectives:

* Understanding parallel programming for both shared and distributed memory systems. Being able to understand and modify parallel programs.
* Understanding snooping protocols, implementation of cache controller and design issues with multilevel caches, split-transaction buses and non-atomic cache states
* Understanding directory protocols

Topics:

* Parallel-programming model abstractions and message passing MP systems: Chapter 5.2 and 5.3
* Bus-based shared memory systems: Chapter 5.4
* Scalable and cache only memory systems: Chapter 5.5

Problem to practice:

* Assignments 3 (with solutions from <http://www.site.uottawa.ca/~mbolic/ceg4131/index.shtml>).
* Quizzes from last years
* Problems 5.2. 5.3, 5.14, 7.7

**Interconnection networks**

Objectives:

* Understanding different network topologies and being able to calculate their parameters
* Learning how to calculate latencies for different netwoeks and switching strategies
* Understanding routing, switching and arbitration
* Describe solutions for deadlock and understand virtual channels
* Being able to design a switch

Topics:

* Interconnection networks: design space, switching and topologies: Chapters 6.2, 6.3 and 6.4
* Routing and switching architectures: Chapters 6.5 and 6.6

The following slides have been used – they are based on this document [**Lecture scribing from 2006**](http://www.site.uottawa.ca/~mbolic/ceg4131/CEG4131-LectureScribing2006.pdf):

* [Slides for interconnection networks](https://uottawa.blackboard.com/bbcswebdav/pid-295632-dt-content-rid-998234_1/xid-998234_1) - starting from slide 51-56 and then from 126 to 138
* [**Buses**](http://www.site.uottawa.ca/~mbolic/ceg4131/ceg4131_buses.ppt)
* [**Dynamic interconnection networks**](http://www.site.uottawa.ca/~mbolic/ceg4131/ceg4131_dynamic_networks.ppt)
* [**Static networks**](http://www.site.uottawa.ca/~mbolic/ceg4131/ceg4131_static_networks.ppt)
* [**Deadlock**](http://www.site.uottawa.ca/~mbolic/ceg4131/CEG4131_Deadlock_4465359.ppt)
* [Router microarchitecture](http://www.eecg.toronto.edu/~enright/interconnects-microarch.pdf)

Problems to practice:

* Assignments 2 and 4 (with solutions from <http://www.site.uottawa.ca/~mbolic/ceg4131/index.shtml>). Assignment 3 problem 5 and 6.
* Problems 6.1-6.5 from the textbook

**Synchronization and consistency**

Objectives:

* To be able to implement synchronization primitives using basic command. To analyze traffic due to synchronization requirements. To understand synchronization concepts.

Topics:

* Synchronization: 7.5

Problem to practice:

* Problem 7.3

**Chip multiprocessors**

Objectives:

* Being able to apply what we studied in chapters 5, 6 and 7 in more complex systems
* Basic understanding of superscalar architecture enhanced with multithreading
* Understanding of multithreading approaches
* Understanding of multi core architectures and classifications
* Understanding of transactional memory

Topics from the book:

* Multithreading: 8.2
* Multicore architecture: 8.3
* Chip processor architectures and programming models: 8.4 - 8.5.3

Problem to practice: 8.3, 8.5 and 8.8,

### Additional topics

**Graphical processing units – architecture and programming**

Objectives:

* Describe the architecture and memory organization of a GPU
* Understand OpenCl execution model

**Embedded multiprocessors**

* **Understand terminology**
* **Reasons for having a hypervisor**

## Questions related to the lectures

Unfortunately, questions from a number of lectures are missing.

**Snooping caches - implementation**

Why do we need dual directories in snooping caches?

Why do we need to store the data in a FIFO buffer for write-back caches?

Draw and explain the architecture of a cache controller with multiple caches?

What is the purpose of wired-NOR bus line for the MESI protocol?

Why is there a shared signal line between caches? How is this line implemented?

What are advantages and disadvantages of MESI in comparison with MSI?

What are the major features of MOESI protocols?

List cache protocol optimizations.

What is migratory sharing?

What is producer-consumer sharing? Give an example.

Where is the snooping cache controller places in multilevel caches?

What does cache0inclusivity have with placement of the controller?

What are the information that have to be exchanged between two caches (L1 and L2) for the snooping protocol with multilevel caches? Assume that the caches are inclusive. Describe information exchange both in L1->L2 and L2->L1 directions.

What are the transient states that need to be added to MSI protocol? What is the reason for adding transient states? Give an example.

**Questions: Directory based protocols**

Give an example of coherence bandwidth requirements for snooping based systems

Why aren't snooping protocols scalable?

What are two main approaches in handling scalability:

* Replace  non-­‐scalable  bandwidth substrate  (bus) with  scalable bandwidth network (point-­‐to-­‐point  network,  mesh)
* Replace  non-­‐scalable  broadcast protocol (spam  everyone) with scalable directory protocol (only communicate with processors that have desired cache copies)

Show how central directory protocol works.

Describe directory protocol for a read-miss when there is another modified copy and for a write miss when there are other shared copies.

What are the sizes of full mapped (baseline directory) and limited (limited-pointer) directories?

Be able to do the example where there is a small program with loads and stores and you will need to show the content of directories and caches during and after the execution of the program for different types of directory protocols. You will be given the state-transition diagram of the protocol.

Describe actions that every cache coherence protocol has to perform

(0) Determine when to invoke coherence protocol

(a) Find source of info about state of line in other caches

whether need to communicate with other cached copies

(b) Find out where the other copies are

(c) Communicate with those copies (inval/update)

What are home directory, requester node, dirty node and shared node?

Why the architecture with local memories scale better that the one with the shared memory?

What is NUMA, cc-NUMA and COMA architectures? Draw organizations of these different architectures.

Classify full mapped, limited and chained directory protocols as memory centric or cache centric.

If MSI protocol is executed in each cache in directory protocol, what will happen when there is cache line replacement and that line is not dirty?

If MSI protocol is executed in each cache in directory protocol, what will happen when there is cache line replacement and that line is dirty?

 Why do we need a busy bit or a lock bit at the home node?

What is the number of network transactions in case of read miss for the baseline and for the Stanford Dash system? What has been done to reduce the number of transactions?

List advantages and disadvantages of chained directory protocols (cache centric directory protocols).

List properties of trhee types of intra-inter cluster coherence solutions.

What are the advantages of static and of dynamic page placement to cc-NUMA architectures? What are round-robin and first touch static placements?

When is it beneficial to perform page replication and when it is disadvantageous?

**Hypervisors**

Why is hypervisor type 2 less efficient than type 1?

Why would we like to have more Operating systems in one system?

What is the function of the hypervisor?

**Interconnection networks**

What is the width of the link?

What is the link bandwidth?

What is the advantage of asynchronous communication?

What is switch degree?

What is the maximum degree of the switches in 4x4 mesh network?

Distinguish between unicast, broadcast and multicast.

Describe the packet? Why is the routing information stored in the header?

Define and compare switching strategies.

What is routing, flow control and network topology?

What contributes to the end-to-end packet latency?

What are flit and phit?

Compute end-to-end delay for a given switching strategy.

Define network diameter and average routing distance.

What is the split-transaction bus? Why is it used? What changes in the system does one need to make in order to support split transaction buses?

What is the burst mode of the bus?

What is the difference between synchronies and Asynchronous buses? What is hand-shaking in buses?

Calculate the transaction time and latency when buses are used.

Understand the following topologies: crossbar switch, tree, multistage networks including omega and butterfly.

Compute network parameters/characteristics of all types of networks including both direct and indirect networks.

What are hypercubes and k-ary n-cubes?

How is routing performed in meshes, omega networks and hypercubes? Give examples of deterministic routing?

What is the difference between deterministic, oblivious and adaptive routing?

What are virtual channels? What are they used for?

How can deadlock be avoided for deterministic routing and how for adaptive?

Draw the architecture of the switch. Explain its components.

What is flow control and how is it implemented in the switch.

**Multithreading**

Will context switching take longer for the OoO superscalar processor then in-order in case of blocked multithreading implementation.

Is it common to have vacant slots in simultaneous multithreading?

Does SMT require large area to be implemented?

What needs to be replicated for SMT?

# Additional list of questions

**Caches and cache coherence**

Q. Consider a 32-bit microprocessor that has an on-chip 16Kbytes four-way set-associative cache. Assume that a cache line has a line size of four 32 bit words. Where in the cache is the word from memory location 0xABCDE8F8 mapped?

a. 15

b. 248

c.143

d.128

Q. A computer system contains a main memory of 32K bytes. It has also a 2-way set-associative cache. The size of the cache is 64Bytes and it has 4 words (16 bytes) per set. Assume LRU policy. The processor runs a program (that is a long sequence of load instructions) that fetches 18 words from locations 0, 4, 8,…, 68 in that order sequentially 10 times. Assume that the processor is a 5-stage pipelined and fully loaded with hazard detection unit, forwarding unit, stalling support, … Assume that the clock period of the processor is 10ns, that the access time for the cache is 10ns. Regarding memory timing, assume that it takes – 1 clock cycle to send the referenced address – 10 clock cycles for each DRAM access initiated – 1 clock cycle to send a word of data. How long will it take to execute this program if there is no cache and the processor is directly connected to the memory using 32-bit bus.

1. 2164
2. 2160
3. 180
4. 12

Q. Consider a shared memory system with two processors A and B. Each processor has a cache that is direct-mapped and has 64 blocks with block size of 16 bytes. The cache is initialized with all zeros. The following sequence is executed 1. Processor A writes a word 11 to the address 1200 2. Processor B reads a word from the address 1208 3. Processor A writes a word 33 to the address 1200 4. Processor B writes a word 44 to the address 1208 5. Processor A reads a word from the address 1208

Give the contents of the caches A after the last step step, if basic MSI write-back invalidation protocol is used. Show the whole cache line and use X for the words that you do not know.

Q. Consider a parallel multiprocessing system on chip in which processors have 2 levels of local caches (L1 and L2). Caches are inclusive. A snooping invalidate cache coherence protocol is used. Is L1 or L2 cache controller responsible for enforcing coherence within the chip.

a. L1

b. L2

**Interconnection networks**

Q. Select preferable routing choices for multiprocessor systems on a chip with 64 processors where reduction of power is one of main design goals

1. Deterministic
2. Oblivious
3. Adaptive

Q. Which of the following flow control strategies use buffer to store a packet being denied the channel and keep it stored until the channel is free?

A Blocking flow control

B. Virtual cut-through routing

C. Detour after being blocked

Q. Is the number of physical channels the same as the number of virtual channels?

a. yes

b. no

Q. (We did not cover it in 2013) In oblivious routing a packet arrives to a channel that is not available. In this case,

1. the packet will be rerouted to an available channel
2. the packet will wait until the channel is available or it will be dropped and retransmitted.

Q. If reducing the amount of memories and buffers in a system-on-a-chip is a main design goal, which switching technique would you use:

1. circuit switching
2. virtual cut through
3. wormhole

Q. What is the bisection width of a star network with 33 processors (1 in the middle and 32 leaf processors)?

Q. How many nodes are there is 4-ary tree (4-ary means that each intermediate node has 4 children) which height is 3 (3 levels of links and 4 levels of nodes)?

Q. Consider simple comparison between 16x16 Omega network and 16x16 crossbar network. While the crossbar uses cross points, the Omega network is using 2x2 switching elements (SE). Assume that the cost of the SE is four times that of a cross point. How many times is crossbar network more expensive than Omega network if we assume that the cost of the Omega network is determined only by its switching elements and the cost of the crossbar network is determined only by its cross points.

Q. Is it possible to use a crossbar switch for message passing networks. Please think about how you can connect different processors?

a. yes

b. no

Q. How many stages of 4x4 switches are needed for 16x16 multistage network?

Q. What bus allocation technique would end up in having empty slots if the master does not have data to send

1. TDMA
2. Round-robin
3. Unequal priority scheme

Q. Why isn't daisy chain arbitration fair:

a) It can starve the masters that are further from the arbiter in the chain if the devices that are closer have a lot of data to transfer.

b) When one device starts its sequence of transactions it does not stop until it is finished.

Q. What bus allocation technique would work the best for systems where one master is more important than the others.

1. TDMA
2. Round-robin
3. Unequal priority scheme

Q. What techniques/transactions can be used for the buses where fast and slow devices need to communicate? List at least 2.

Q. Is it possible to have concurrent transfers in hierarchical busses

1. Yes
2. No
3. Only if the transfers do not go through bridges.

**Additional topics**

Q. Is it possible to run multiple operating systems on different cores in a multicore embedded chip?

a. yes

b. no

Q. We see in the lecture on GPU architectures that SIMD processor in is actually a symmetric VLIW4 in AMD's GPU. Nvidia has an opposite approach and requires dynamic resolution of the dependencies. What GPU architecture would be in this case more power and area efficient?

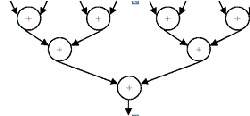
1. AMD
2. NVIDIA

Q. Which of the following architectures does not rely on data paralelism

1. Stream processors
2. Risk Processors
3. Multimedia SIMD
4. GPUs

**Programming and performance**

Q. How long does it take to add 64 numbers using 4 processors. Communication and computation take one unit time. Assume that processors are connected using a single bus.



1. 18
2. 20
3. 22
4. 24

Q. Scalability means that the speedup increases linearly when the number of processors increases and the problem size increases as well.

a. yes

b. no

Q. What is the maximum efficiency of a parallel processing system?

a. 1

b. >1

c. 0

d. n

Q. What is the reason for using Barrier in parallel programming?

1. To provide mutex mechanism
2. To provide a point for synchronization for programs running on different processors
3. To lock the shared variable

Q. Where is the mutex stored in general purpose processing systems:

a. In memory

b. In special hardware module next to each processor

Q. Consider a multiprocessor system with 2 processors (A and B) with their individual caches that are connected via a bus with a shared memory. Variable x is originally x=3. Given is the following sequence of operations: 1. A reads variable x 2. B reads variable x 3. B updates x so that x=x+2 4. A updates x so that x=x\*2 5. B reads variable x. What will processor B read in step 5 if no cache coherence mechanism is implemented.

1. 3
2. 5
3. 10

Q. Is it possible to implement message passing method on a system with shared memories such as COMA?

a. yes

b. no

Q. Are the assumption that there is no overhead in communication and synchronization the only assumptions made for derivation of Amhdals law?