Re-Exam

DAT278/DIT055: Sustainable Computing

Friday 09 April 2021, 08:30-12:30 (+10min tolerance)

Examiner and contact person during exam:
Pedro Petersen Moura Trancoso (contact through canvas email)

Grading intervals:
The maximum grade for this is exam is 100 points
DAT278 (Chalmers): U < 40p ≤ Grade 3 < 60p ≤ Grade 4 < 80p ≤ Grade 5
DIT055 (GU): U < 40p ≤ Grade G < 75p ≤ Grade VG

Examination solution and review:
Exam solution will be posted in canvas within 24h after the exam has finished.
A canvas announcement regarding the exam review.

Instructions:

- **ANSWERS**: All answers need to be typed in the space provided. Different questions need to start in a new page. Additional material may be added to your typed answers as pictures inserted in the same space with the typed text. (pictures can be used only for diagrams not for text)
- **LANGUAGE**: All answers should be written in English.
- **SUBMISSION**: Submit your answers in a SINGLE PDF file. Name your file "<course code>-<first name>-<last name>.pdf" (e.g. for John Smith taking DAT278 would be "DAT278-John-Smith.pdf"). You may submit many times during the exam time but only your LAST submission will be considered.
- **DEADLINE**: The deadline is 12:30, there is a 10 min tolerance but then the assignment closes at 12:40, no more submissions will be accepted after that. Submit before the deadline to avoid problems.
- **RESOURCES**: All available material is allowed. For any of your answers, if you use information from the internet, include the source (the web address).
- **PLAGIARISM**: Copying of material from any source (unless requested in the question) is FORBIDDEN (you need to write the answers in your own words). Communication and exchange of information with other people during the exam period is FORBIDDEN. Exam answers will be tested for plagiarism.
- **COMMUNICATION WITH EXAMINER:** Before communicating with the examiner test the Frequently Asked Questions (FAQ) page which will contain the answers given to other students. This page will be continuously updated. If you still need to communicate with the examiner, do this through canvas email.
- **CONTINGENCY PLAN**: In case there is a failure in canvas during submission you will have to submit your file with the answers as an attachment to an email to the examiner ppedro@chalmers.se. Use the same email in case the canvas site is down during the exam period and you need to communicate with the examiner.

Question 1 (10 points)
In the article with title ”Computational sustainability: computing for a better world and a sustainable future” the authors talk about using “computing for sustainability”. Start your answer by explaining the term “computing for sustainability” and how it compares to “sustainable computing”. Then:

- Select and describe a real example of where “computing for sustainability” is used (add an internet link to the source where you found the example)
- Explain the problem and how computing can help solve that problem.
- Briefly describe which computer systems are used to this problem
- Discuss if computing is just used or if there is any benefit to computing in this case of “computing for sustainability”

You answer should be brief but complete and statements should be clearly justified using references to articles and/or internet page links. Your answer should be around 200-300 words.

Computing for sustainability is the use of computer system to explore solutions for sustainability issues. This is opposed to sustainable computing which is applying sustainability concepts to computing, i.e. making computing more efficient.
One example is to process satellite images as to for example detect fires in a forest without easy access.
In this case computation with large systems able to process many images using AI algorithms can be used to determine if a fire is happening in a certain location.
In this case we could use large server systems in a data center since the images collected are quite large and the processing power needed is quite significant.
There may be a benefit to computing since maybe new architectures or algorithms need to be developed to solve the particular problem.

Question 2 (10 points)
In the article with title “A New Golden Age for Computer Architecture” the authors present the term “Domain Specific Architecture” (DSA). Regarding DSAs give the answers to the following questions:

- What is a DSA? Give an example of what a domain is and of an application within that domain.
- Why is it that it is so relevant now?
- How do DSAs compare to more traditional architectures like the multicore CPU in your own computer and the ASIC devices in many of our phones?

You answer should be brief but complete and statements should be clearly justified using references to articles and/or internet page links. Your answer should be around 200-300 words.

A DSA is an architecture that is designed to solve very efficiently a certain set of applications within a common domain. An example of a domain may be Machine Learning. An application within that domains could be object detection.
DSAs are relevant today because there are many new applications that put high demands on the memory and/or computation. These demands cannot be met with traditional CPUs neither these CPUs can scale effectively to provide the power required (end of Moore's Law). Then we require special dedicated hardware to solve these problems.
DSAs are placed between the generic CPUs and the very specialized ASICs – they are targeted to efficiently solve several different applications within the same domain. They are not generic as the CPUs which can solve all applications but have a penalty in the efficiency and ASIC which are the most efficient but can only solve a single application.

Question 3 (10 points)
Traditionally dynamic power has been the dominant factor in power consumption but in the latest years we have seen the focus shift to static power. Regarding static power, give the answers to the following questions:

- Explain what static power is and why is it that it is such a big problem these days?
- Describe a technique that has been presented in this course that is aimed at reducing the static power. Justify your answer with an article that you may find on the Internet which either describes the technique or an example of its application on a product or research work.

Justify your answer using references to articles and/or internet page links containing relevant information to your answer (e.g. an article where you can find information about the technique you decided to describe).

Static power is the power that is dissipated even if no circuit switching is happening. This is due to the leakage current. Static power is a serious problem these days because of the shrinking of the technology and reduction in the operating voltage which result in an increase in the leakage current and thus increase in the static power.
One example of a static power reduction technique is decay cache which turns off memory cells from the cache that are deemed “dead”, i.e. that contain data that is not going to be used again or that its reuse distance is very long. The description of the technique can be found in the following article:
Stefanos Kaxiras, Zhigang Hu, and Margaret Martonosi. 2001. Cache decay: exploiting generational behavior to reduce cache leakage power. SIGARCH Comput. Archit. News 29, 2 (May 2001), 240–251.

Question 4 (10 points)
In our course we have presented Single-Instruction Multiple-Data (SIMD) as a way to improve efficiency. Regarding this technique answer the following questions:

- What is SIMD?
- How can SIMD be used to improve the efficiency?
- Which systems support SIMD? (give a concrete example)
- Describe an example on how you can use SIMD in a real code and system (you can find this example online, just describe it in your own words and add the source where you found it).

SIMD instructions are instructions that apply the same operation to different data elements in the same clock cycle. In particular, these instructions improve efficiency by identifying that certain data values do not require the whole width of the registers (for example 32b in a 32b machine) and can be represented using fewer bits. If for example we could represent certain data elements using only 4 bits then we could have 8 such elements in a single 32b register and a SIMD operation which would take those 8 4b elements and perform the selected operation within the same clock cycle. For this particular operation we could potentially improve the execution by a factor of 8x and thus reduce the energy consumption for those operations by a factor of 8.
The AVX set of instructions is an example of a set of SIMD instructions for the Intel processors. Other CPUs also support their own SIMD instructions.

```
unsigned char A[MAX], B[MAX], C[MAX];
…
for( int i = 0; i < MAX; i++)
  C[i] = A[i] + B[i];
…
```

The for loop above can be simplified in a 32b machine with something like the following (using pseudo-instructions)

```
for( int i = 0; i < MAX/8; i++) {
  AX = simd-pack-elements(A[])
  BX = simd-pack-elements(B[])
  CX = simd-sum(AX, BX);
  C[] = simd-unpack-elements(CX)
}
```

Question 5 (10 points)
In our course we presented Approximate Computing as a technique to improve the efficiency. Regarding this technique answer the following questions:

- Explain what Approximate Computing is using your own words and give a small example of its use.
- How can Approximate Computing improve efficiency?
- What are the limitations of using Approximate Computing?
- We have discussed in class an article with title "Managing Performance vs. Accuracy Trade-offs With Loop Perforation" which presented an approximate computing based technique. Explain briefly the contents of this article and make your own code example where you show the use of this technique.

Approximate computing is a technique that recognizes that not all operations need to be performed with full precision and accuracy. As such, we may have more efficient executions if the hardware system is built to perform approximate operations or if the software is modified as to avoid all unnecessary operations in order to obtain a reasonably good result. For example, we may have in hardware an ALU which only operates on the higher bits leaving the lower bits 0 thus reducing the power consumption and making the operations more efficient (but introducing errors in the results!). The efficiency can be obtained by reducing the hardware or software used to produce the approximate results.
The main limitation of approximate computing is the fact that it is difficult to give guarantees regarding the bounding of the error introduced by the technique.
In the paper mentioned they introduce the technique of loop perforation where some iterations of a loop are skipped as to reduce the number of instructions executed.
One code example could be a code to calculate an average of a large array of values where we skip every second element:

```
sum = 0
for( i = 0; i < MAX; i += 2)
  sum += A[i];
avg = sum/(MAX/2);
```

Question 6 (10 points)
In our course we have discussed the Microsoft underwater datacenter – Project Natick (https://youtu.be/lBeepqQBpvU). Regarding this datacenter answer the following:

- Briefly describe the datacenter, its benefits and its limitations.
- Could this be the datacenter of the future? Justify your answer.

You answer should be brief but complete and statements should be clearly justified using references to articles and/or internet page links. Your answer should be around 200-300 words.

This datacenter is built to work on a closed environment which is airtight. Thus, the system needs to be designed in a way that it does not require any sophisticated cooling as the heat exchange needs to be done within the container – possibly through the container walls since the container will be immersed in cooler water. This is actually one advantage of this system! The system faces a serious challenge that the components cannot be replaced or repaired for the time that the container is submerged. Also, the system is not that large and thus in order to satisfy the requirements of current data centers there is a need to use several of these containers. The advantage is that this distributed organization will be placed closer to the users and thus the latency may be reduced as opposed to distance centralized data centers.
Given the above this is certainly a solution for future data centers.

Question 7 (20 points)
Consider two computer systems that you would like to evaluate in terms of their efficiency. Design and plan a simple laboratory assignment which goal is to learn how to perform this task. For this assignment you should define:

- Define the goals for this assignment.
- Define what efficiency is and how it can be measured.
- Determine and present the specifications for two systems that make sense to do this experiment.
- Determine and present the applications that should be used to execute and test the systems.
- Define (and justify) the metrics that should be used for this task.
- Define the methodology to measure the values for those metrics for the systems that you have selected.
- Summarize the assignment by providing clear steps on what students should do in order to reach the expected conclusions.

You may find inspiration for this assignment on the web and if so, you should mention the sources you found.

This is an open-ended question and thus the answer provided here just touches on the basic issues that need to be covered.
The clear goal is to be able to compare the efficiency of two systems using a simple experimental assignment.
For this assignment we could consider efficiency to be expressed by the energy-delay product or EDP. This metric can be obtained through the measurement of the execution time (T) and the average power consumption (P) and then EDP=P x T x T.
We could use a high-power system e.g. an Intel i9 based system (e.g. a powerful laptop or desktop) and a low-power system such as an ARM based system (e.g. a raspberry Pi). Alternatively, we could just use a simulator tool and two different configurations representing the systems mentioned above.
The applications used for this assignment could be any two SPEC applications used already in the Computer Architecture course or any simple small "kernel" applications such as a matric multiply code.
If a simulator is used like Sniper, then you can measure power using McPad as done in the Computer Architecture course. If a real system is used then you can get the power from the measurement of the performance counters or by applying a plug measurement tool to the power supply.

Question 8 (20 points)
In our course we have presented how to resize a cache in order to improve the efficiency.
(a) (10 points) Answer the following questions:

- Briefly describe what is cache resizing and how can we improve efficiency by applying it.
- Describe the steps involved in resizing the cache from a large cache to a small cache. Do not forget to describe how you determine that a cache resize should be done at runtime.
- Give a rough estimate of the improvement in efficiency when you resize a cache to its half size. Do not forget to mention what are the fixed overheads.

(b) (10 points) Plan a simple laboratory assignment where you design an experiment that illustrates the benefits of using cache resizing. Clearly state the goals for the assignment and mention the software and hardware you should use, the setup that should be used, the steps in the experiment that should be taken, and the analysis that should be made. Make it in a way that another student should be able to read your description and perform the experiment and draw the expected conclusions. You may find inspiration for this assignment on the web and if so, you should mention the sources you found.

(a) Cache resizing is a technique that can be used to reduce the cache size whenever it is not fully used. By "shutting down" the unused space (using a technique like what is used for the decay cache) we can reduce both the static and dynamic power of the cache. For the cache resizing you first need to determine if the resizing should happen, so we may trigger the resizing if for example we observe a very small miss rate – it could be an indication that the whole working set fits in the cache and it may be even smaller than the whole cache. Then if we have a n-way set associative cache we may "turn off" one or more ways of that cache. Before doing so you should flush the contents of that way to memory. Then you can restart the execution with the smaller cache configuration. If there is a request to a data element that was in one of those turned off ways, then that access will result in a miss and the data will be brought from the memory. In this resize we turn off all the cells in the way(s) as well as the tags used to identify the elements in those ways. The decoding logic though is the same and there is even an extra "masking" hardware as to not check the turned off ways. Overall, the overhead is not that large other than the fact that some data is evicted from the cache that may need to be loaded back again in the new configuration. So, if a cache is reduced by half size, its power could potential be reduced by almost a factor of 2.

(b) This is an open-ended question. In order to do this assignment in an effective way you should have two applications, one with a very small working set and one with a large working set. Then execute the applications on a simulator using two different system configurations. One with a cache of size C and another with a cache of size C/2 where the difference is just a reduction in the number of ways by half. So if the original cache was 16-way, the reduced cache is only 8-way. The difference in the execution time for the two configurations and two applications should be interesting to analyze as to draw any conclusions for this approach.