CHALMERS UNIVERSITY OF TECHNOLOGY | UNIVERSITY OF GOTHENBURG

Exam

DAT278/DIT054: Sustainable Computing

Thursday 14 January 2021, 14:00-18:00 (+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

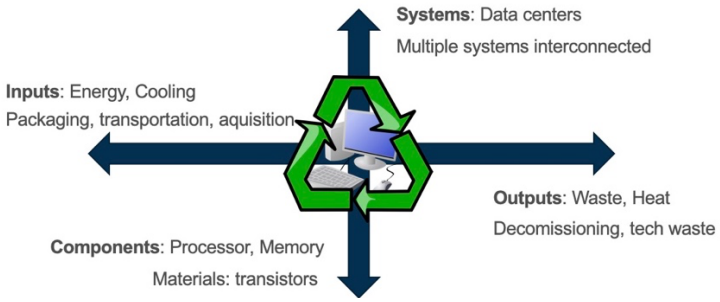
Examination solution and review:
Exam solution will be posted in canvas within 24h after the exam has finished. A canvas announcement regarding the exam review.

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). Solutions may also be provided as written in e-ink into the pdf document as long as the text is well readable.
- **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 18:00, there is a 10 min tolerance but then the assignment closes at 18:10, 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 text 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 check the Frequently Asked Questions (FAQ) page in canvas 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: Sustainability and Sustainable Computing (20 points)

(a) (10 points) Consider the sustainable computing "dimensions" as presented in class and the diagram shown below:



Explain what the horizontal axis represents and give three (3) examples that can be used to improve sustainability of computing according to that axis.

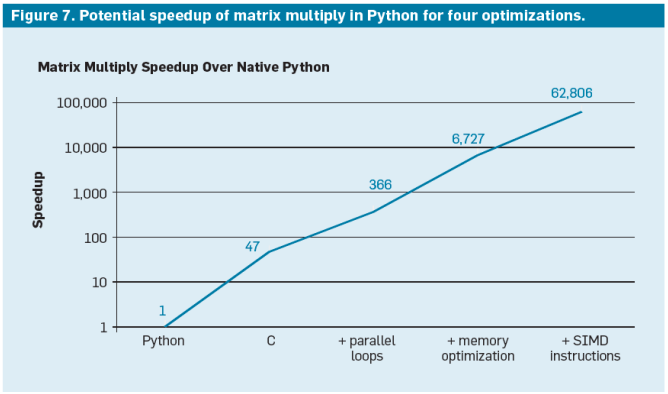
The horizontal axis represents the different factors affecting the sustainability of computing products from the time they are designed, built, transported, sold, operated, and later during the end-of-life, decommissioned.
Examples to improve sustainability: (a) make boxes for computer devices smaller so that you are able to pack more boxes (i.e. devices) per same unit of space and thus shipping will be more efficient; (b) have the computer systems consume energy that is produced from renewable energy sources (e.g. solar, wind, etc.); (c) Build the systems in such a way that they are easy to disassemble in order to separate their parts and sort them accordingly to the different recycling units.

(b) (10 points) In the article titled “Once is Never Enough” Google explains how they are achieving a circular economy model by applying four strategies: Maintain, Refurbish, Reuse, and Recycle. Explain in your own words what this is about and how these strategies are helping towards sustainability. (https://sustainability.google/progress/projects/circular-economy/)

The goal is to improve the sustainability by reducing the waste. So, first thing is to Maintain, i.e. fix components that fail as much as possible. If components cannot be fixed, then Refurbish them and bring them again to the systems after. If the component cannot be re-used within the Google institution, then it is made available for Reuse outside the corporation. Finally, if the component cannot be reused at all then it should be split apart into different sub-components as to be binned for recycling accordingly.

Question 2: Energy Efficiency (20 points)

(a) (10 points) The paper “A new golden age for computer architecture“ (John L. Hennessy and David A. Patterson. 2019) includes the figure below.



Explain why a large speedup (47x) is observed between Python and C? What causes this and what is the impact of this speedup in the energy efficiency of a system? Determine the factor by which the energy efficiency (energy-delay product or EDP) changes for this 47x speedup. Justify your answer. You may use references from the internet to justify your answer. Please include the websites where the information was found.

The main reason why a program written in C may run much faster than a program written in Python is because while C is a compiled language i.e. the programs in C will pass through a compiler that creates an executable, Python is an interpreted language and thus its source code is translated to processor instructions at runtime. The translation and execution, as well as the lack of global knowledge of the program, make interpreted programs slower than compiled programs. In terms of energy efficiency, if the speedup observed is 47x then the EDP will be improved by 2209x (47x47) since it is affected to the square of the time (for the energy and for the delay).

(b) (10 points) In one of the assigned papers in this course (Parthasarathy Ranganathan. 2010. Recipe for efficiency: principles of power-aware computing. Commun. ACM 53, 4 (April 2010), 60-67) the authors mention that a strategy to reduce waste is “Spend power to save power.” Explain what the authors mean with this, find two (2) concrete examples for this strategy on the internet and justify your answer. Present the website link(s) where you found the information.

“Spend power to save power” means that you should not be afraid of solutions that introduce some overhead (spend power) as long as the benefit is much larger (save power). Many of the techniques we have analyzed in class follow this advice. A concrete example may be the gcc compiler and the fact that it includes specific flags to analyze the code during compilation for better optimizations. This is a process that is more time-consuming and thus spends extra power. The optimizations may include the analysis of code that can use SIMD instructions. After optimizing the code, the optimized code will run much faster resulting in a significant improvement of the efficiency.
Another example may be to add extra hardware for the execution such as a GPU, which adds overheads in terms of static energy as well as overheads for coordination of the execution, but its benefits are very significant for the workloads that benefit from a GPU which can execute several orders of magnitude faster than a CPU (example game image rendering).

Question 3: Static and Dynamic Power (40 points)

(a) (10 points) Consider the “cache compression” optimization technique.

i. This technique is used to reduce the dynamic or static power? Justify your answer.

ii. Describe how cache compression works and in which cases can it be successfully applied?

iii. Describe how the memory access path is affected by cache compression.

You may use internet sources and/or any of the assigned reading assignment articles to help justify your answers. Please include all websites where the relevant information was found.

Cache compression is a technique that can be used to reduce both the dynamic and the static power. It can reduce the dynamic power if by compressing the data of the cache, the cache space is freed and whatever free space is just not utilized, so no switching activity. If the cache space is not utilized at all then it may be “turned off” (decay cache) and thus static power may be saved. If on the other hand the contents of the cache are compressed thus allowing for more data to populate the physical cache space, dynamic power can be reduced by the fact that fewer memory accesses will be done.

Cache compression reduces the space that is needed to store the original data. For example, we can use a dictionary mechanism if the number of unique data elements is limited. The encoding of the dictionary entries requires much smaller space than the original data. As mentioned before the condition is that the number of unique data values is limited.

The memory access path is affected by the decompression phase upon a memory read. So, when a memory read is issued, if it is directed to the compressed cache, then the compressed data is retrieved and needs to be expanded before being sent to the CPU to be used. In terms of stores and compression, the path is not that affected since compressing and storing the updated data is not in a critical path.

(b) (10 points) Consider the “DVFS” optimization technique.

i. This technique is used to reduce the dynamic or static power? Justify your answer.
ii. Describe how DVFS work and in which cases can it be successfully applied?
iii. At which level is DVFS implemented? User-level, system-level, or hardware-level? Justify your answer.
iv. Consider one system A which has 10 different operating VF levels (e.g. P0, P1, etc.) and a system B which has the ability to flexibly change V and F though a dynamic range of values for V and F respectively. Which system is better to achieve better energy efficiency? Justify your answer.

You may use internet sources and/or any of the assigned reading assignment articles to help justify your answers. Please include all websites where the relevant information was found.

DVFS is a technique that targets the reduction of dynamic power by dynamically changing the voltage and frequency according to the application/workload requirements. Reducing the frequency and voltage will reduce the activity and thus the dynamic power.
DVFS works by changing the voltage and the frequency of the CPU operation. The voltage and frequency can not be changed arbitrarily but instead they follow together, i.e. a voltage and frequency either reduce or increase together, not in separate.
DVFS can be implemented at all the levels mentioned. The higher the level the better and more timely decisions. For example, at the user-level, with all the application knowledge, the user can take much better decisions. But it can also be done at the HW level where circuits that are not. In the critical path may be slowed down to exploit the slack.
To achieve better energy efficiency, a system that allows for a continuous change of values over its range is much better than a system that only allows a set of discrete points in that range. This is because at each point in time we can offer a better VF state which better matches the current need of the application. On the downside is the fact that a more fine-grain change means that possibly changes are done more frequently and thus introducing more overheads.

(c) (10 points) In the article titled “Amdahl's Law in the Multicore Era” (Mark D. Hill and Michael R. Marty. 2008) the authors show a figure with three different types of multicore chip:

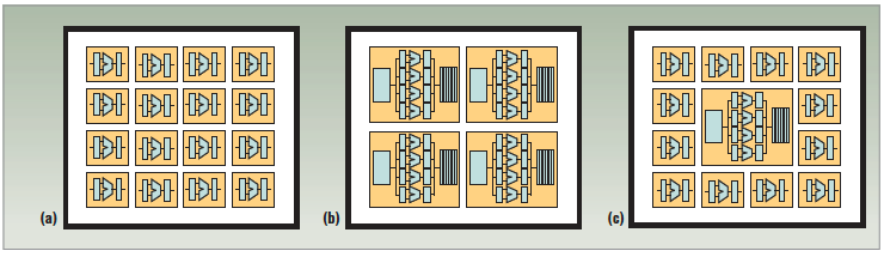


Figure 1. Varieties of multicore chips. (a) Symmetric multicore with 16 one-base core equivalent cores, (b) symmetric multicore with four four-BCE cores, and (c) asymmetric multicore with one four-BCE core and 12 one-BCE cores. These figures omit important structures such as memory interfaces, shared caches, and interconnects, and assume that area, not power, is a chip's limiting resource.

i. Which type is best? Justify your answer.
ii. If no one of the types is best, then for which workloads/application is better each type? Justify your answer.
iii. How do you relate this issue discussed in (i) and (ii) with the architecture proposed in the article “Core Fusion: Accommodating Software Diversity in Chip Multiprocessors” (Engin Ipek, Meyrem Kırman, Nevin Kırman, and Jose F. Martinez, ISCA 2007). Please elaborate and justify your answer.

You may use internet sources to help justify your answers. Please include all websites where the relevant information was found.

No one of these types is best, it depends! If only type needs to be selected, then the (c) is the most “flexible” solution since it includes both larger and smaller cores.
An architecture with many small cores (a) is more suited to embarrassingly parallel workloads. This architecture can be seen as a GPU and thus applications with high degree of parallelism will work well in this type. An architecture with a few large cores (b) can be used best for multiple programs running in simultaneously, i.e. for throughput. As for the (c) architecture, as mentioned before it is an architecture that includes both small and large cores and thus can satisfy the execution of an application with different phases such as the sequential phase in the large core and parallel phase in the small cores. Core fusion is an architecture that has the ability to “mutate” from one setup to another. This is very good as it can adapt dynamically to the different applications or different phases of the application and also depending on the demands of the current application.

(d) (10 points) Cache memories are increasing its size in the processor die and correspondingly consuming a larger share of the total processor power. Mention two (2) techniques that can be used to make these large caches more energy efficient. Give a complete answer to each technique by presenting the overview of the technique, how it works, which are its benefits, and if possible, giving a real-world example of its application. Justify your answer.

You may use internet sources and/or any of the assigned reading assignment articles to help justify your answers. Please include all websites where the relevant information was found.

We have studied different techniques. Given that the cache is not always fully used we could change dynamically is utilized space thus reducing the static power. We can use a technique like cache decay or drowsy cache to achieve this or even cache partitioning where one or more ways are turned off when the utilization is low. Another technique that could be used is way prediction for caches with large associativity. This technique allows for one a single way to be activated when a read operation is issued.

Question 4: Heterogeneity and Data Centers (20 points)

(a) (10 points) Describe what an FPGA is – what is its hardware like, what can you do it, how do you use it, etc. In terms of energy efficiency, how does it compare to a CPU or GPU? What is the frequency of operation of an FPGA compared to CPU and GPU? Justify your answer. Find a use-case for FPGAs on the Internet that justifies your point. Present the website link(s) where you found the information.

An FPGA – Field-Programmable Gate Array - is a reconfigurable hardware which can be programmed to perform different tasks. You need to usually use several tools to transform a piece of software code into a design that can be uploaded to the FPGA which can then execute that code directly in hardware. In terms of energy efficiency, it can be much better than general-purpose devices like CPUs and better even than dedicated but still general-purpose GPUs. Its ability to implement dedicated operations and its high degree of parallelism and throughput are the key factors for its efficiency. This is achieved even though the frequency of operation of FPGAs is much lower than CPUs and GPUs.
Microsoft uses FPGAs in their data centers to accelerate their bing search engine and also to accelerate AI applications.

(b) (10 points) According to Facebook (in https://sustainability.fb.com/innovation-for-our-world/sustainable-data-centers/), their datacenters are "…are 80% more water efficient than average data centers". Describe what is meant by that claim, how is it exactly that the data center achieves water efficiency? What is one of the most common strategies used by Facebook to achieve Data Center cooling? Justify your answer and give an example of a data center where the techniques discussed are implemented.

You may use internet sources to help justify your answers. Please include all websites where the relevant information was found.

One of the points made by Facebook is that the water used for the cooling of data centers is somehow wasted. Consequently, Facebook focuses on designing their systems in ways that it benefits from air cooling and thus reduces the need for water cooling, and consequently results in an improvement of the water efficiency. So, Facebook mainly tries to place their data centers in locations which can benefit from air cooling even though the climate conditions may not be favorable. The benefits in water savings though are significant. The Facebook data center in Luleå is an example of such an approach.