CHALMERS UNIVERSITY OF TECHNOLOGY | UNIVERSITY OF GOTHENBURG

Exam

DAT278 / DIT055: Sustainable Computing

Monday 11 April 2022, 08:30-12:30

Examiner:
Pedro Petersen Moura Trancoso

Contact person during exam:
Pedro Petersen Moura Trancoso, tel. 7726319

Supporting Materials/tools:
Chalmers approved calculator.

Instructions:
All answers should be written in English.

Grading intervals:
The maximum grade for this is exam is 100 points
DAT278 (Chalmers):

Fail < 40p ≤ Grade 3 < 60p ≤ Grade 4 < 80p ≤ Grade 5

DIT055 (University of Gothenburg):

Fail < 40p ≤ Grade G < 75p ≤ Grade VG

Examination solution:
Will be posted in canvas within 24h after the exam has finished.

Examination review session:
A canvas announcement with the date and location will be sent to all students.

Question 1 (4+4+8 points)

(a) Define what "sustainable computing" is in your own words.
(b) In the course we have also discussed what "computing for sustainability" is. Define it in your own words and give a practical example.
(c) One of the factors that affects the efficiency of large data centers is the cooling. Explain why cooling is a problem that affects the efficiency and present one technique that we have discussed in class that can reduce the cooling problem. Justify your answer.

(a) Sustainable computing can have many different definitions. One could be that computers are designed to consume as little energy as possible to perform their tasks and that they are design to be easily recycled at end of life.
(b) Computing for sustainability is usually the term used when computational power is used to solve sustainability problems. For example, if we want to monitor the fishing of a certain species then we can use computers to process satellite imaging and help in identifying any possible issues and follow up on the evolution.
(c) Cooling is needed for the correct operation of the high-performance devices but it requires energy that is not really "useful" so it then affects the efficiency. One technique that was discussed in class is placing the data centers in cool locations and then use the cold air outside to cool the systems ("open windows" technique). A good flow of air is required for this method to be effective.

Question 2 (4+8+4 points)

In our course we have discussed a technique known as "subthreshold voltage".
(a) Explain in your own words what "subthreshold voltage" is.
(b) Explain how "subthreshold voltage" can help in improving the efficiency. Do not forget to mention which are the issues associated with operating at "subthreshold voltages".
(c) Is this technique available currently in general-purpose systems, for example in the CPU of your laptop computer?

(a) Threshold voltage for traditional semiconductors is 0.7V so anything below that voltage is regarded as "sub-threshold".
(b) The reduction of the voltage results in a square reduction in the power so this implies a great benefit to the efficiency. Nevertheless, lower voltage also means lower frequency and thus slower switching and slower operation thus somehow leading to increased energy. Then operating at sub-threshold means that the results are not necessarily easy to identify given the possible noise and thus we may get incorrect results.
(c) Not on general-purpose systems, it is available for ultra-low power system that need to operate with minimal energy maybe even with energy harvesting techniques.

Question 3 (6+12 points)

One technique to improve efficiency is "cache resizing". Let's assume a system with a CPU that contains two caches: (i) a Level-1 (L1) cache of 8KB capacity, 64-byte cache line and 2-way set associative; and (ii) a Level-2 (L2) cache of 512KB, 64-byte line and 16-way set associative.

(a) Which of the two caches (L1 or L2) do you think is a better candidate for cache resizing. Justify your answer.

(b) Let's apply cache resizing to the cache you have selected in (a). Describe the different steps in resizing the cache to half of its original size.

(a) L2 for different reasons. L1 is a critical cache and thus not good to be making changes dynamically – would introduce too much overhead. Also, L1 has only 2-ways and thus if we halved it, it would become direct-mapped which could affect the performance.

(b) When applying cache re-sizing, the easiest is to resize the number of ways. So, in the case of L2 and resizing into half, we can downsize the 16-way to a 8-way cache. The steps are the following:

(1) Block and new incoming accesses to the cache (have a busy wait) and wait for all in-flight accesses to complete.

(2) Decide which ways to disable.

(3) If the cache followed the write-back policy then flush to memory all dirty data in the ways to disable.

(4) Turn off the ways to disable.

(5) Allow for requests to go through.

Question 4 (4+4+4+4 points)

Dynamic Voltage-Frequency Scaling (DVFS) is a technique that can be applied at different levels of the computer system. For this question we are interested in exploring DVFS at the level of the Operating System (OS). In this case the DVFS is controlled at runtime by a program running in the background also known as "governor".

(a) Can you give examples of different DVFS governors implemented on the Linux system?

(b) Decisions on the DVFS are taken at certain time steps. Discuss the tradeoffs in the size of these timesteps. Should they be small and often or large and less often? Why? Justify your answer.

(c) When the governor decides to change the frequency from the lowest setting to the highest setting, how is this done? In one or several steps? Why? Justify your answer.

(d) To take decisions on the voltage and frequency settings, do you think that using AI would help improve the performance? Justify your answer.

(a) Some of the Linux governors are: Performance (always high frequency), Powersave (always low frequency), on-demand and conservative which take into account the current load of the processor.

(b) Small time steps mean a more reactive system, so it can react fast to changes in the load. At the same time, small time steps mean more decisions, more interrupts, more overhead to the performance. So, the tradeoff is between reaction and overhead in measuring and switching.

(c) From lowest-to-highest it is usually in one step since this is just a power penalty not performance and in case of a needed reaction it is better to have a penalty in power than to miss the opportunity to give the performance that was required. A reason to go in slower steps would be to avoid quick peaks in the load which cannot be captured, and which would just lead to increased unnecessary power consumption.

(d) Decision making is a critical step in the DVFS setup process. So, using AI to improve the quality of the decisions would be a good strategy. Nevertheless, AI may require considerable resources in terms of storage and execution to perform the prediction. Also, it may require several training phases and dynamically during execution to capture changes in the workloads. The training could again be leading to a considerable overhead. Basically, as for all techniques, there is a need to weigh the benefit of the new technique versus the overheads that it may incur. If the benefits overcome the overheads, then the technique should be adopted. Otherwise, it is not that interesting.

Question 5 (6+6+6 points)

Consider a technique discussed in the course known as "data compression" and assume that this technique is applied at the DRAM memory level. Assume that the original DRAM of the system is 16GB and that the data compression rate is in average 2x.

(a) For this scenario, where do you think the memory compression and decompression operations are executed? Justify your answer.

(b) Describe all the steps that are involved in an access to memory, from the CPU decoding the Load instruction. Assume that the requested data is not available on any of the CPU internal caches.

(c) How is the dynamic and static energy affected by the implementation of memory compression at the memory level as presented in this question? Justify your answer.

(a) Since we are considering here memory compression, the compression-decompression could be done at the exit of the CPU before the bus or at the memory side. To save power in the communication to memory by reducing traffic from the compression, it is preferable to have it at the exit/entry of the CPU.

(b) The steps are the following: (1) decode the load instruction and get the address; (2) with the address check if the data is in the L1 or L2 caches. If so then return the address to the CPU otherwise issue the request to memory; (3) Wait for the memory request to arrive; (4) Since the memory is compressed, the data will arrive compressed, decompress the data; (5) Pass the data to the caches and to the CPU for its use in the next operations

(c) In terms of static energy we could consider that if the memory is compressed then we could disable part of the memory chips and reduce static energy, if we are not in need of the extra memory. In terms of dynamic energy, since the activity in the storage is reduced, the memory energy will be reduced, also since that the traffic will be compressed the communication energy will also be reduced. As the downside, the compressing and decompressing at the exit/entry side of the CPU will require an extra dynamic energy and if this is performed in hardware then an extra static energy too.

Question 6 (8+8 points)

We have presented in this course a technique known as “heterogenenous computing”.

(a) In your own words describe why we have moved away from homogeneous to heterogenenous systems? How can these heterogenenous systems help achieve better efficiency?

(b) Give an example of an heterogenenous system. Show how an example application would benefit from this system as opposed to a homogeneous one.

(a) Homogeneous systems are usually designed in a way that they satisfy all the demands in the best average way. This is not the most efficient way since in some cases the homogeneous system may be under- or over-provisioned. The best match of a task to the hardware is with heterogenenous systems which have different hardware components that better match the needs to the workloads. Heterogenenous systems also started to be more popular to overcome the problem with “dark silicon” when it is not possible to activate all transistors on the chip at the same time.

(b) There are same-ISA and different-ISA heterogenenous sytems. One common system currently is a System-on-chip configuration where different processing units dedicated to different tasks co-exist. The simplest implementation of this is a combined CPU-GPU chip. In this case, the graphics tasks are sent to the GPU while the general computation tasks are handled by the CPU. This is much more efficient as the graphics tasks are quite “computational heavy” and would incur in a large performance overhead if executed in software on the CPU, reducing significantly the efficiency of the overall system.