CHALMERS UNIVERSITY OF TECHNOLOGY | UNIVERSITY OF GOTHENBURG

Exam

DAT278 / DIT054: Sustainable Computing

Thursday 16 january 2020, 14:00-18:00

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

DIT054 (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.

Q1 (12 points) Sustainability

Q1.1 (6 points) Briefly explain what the following concepts are and for each one of them give an example that fits the concept:
(a) Sustainability
(b) Sustainable Computing
(c) Computing for Sustainability

Sustainability: There are many definitions for Sustainability. It can be seen as "…meeting the needs of the present without compromising the ability of future generations to meet their needs…". In principle is to continue the development but without compromising the future. So, we should for example develop new materials that are not consuming so many new resources as for example when using recycled materials to produce new materials (e.g. clothes made of fibers produced out of recycled plastic bottles).

Sustainable Computing: This concept refers to making the field of computing sustainable, i.e. reducing the impact on the environment while computing. This can be achieved with more efficient compute systems (less energy consumption) and using renewable energies to power the computer systems, for example.

Computing for Sustainability: This concept refers to using computer systems to help solve sustainability issues, for example to use computer systems to develop new materials that are more environmental friendly, or in order to use artificial intelligence to better manage the execution in large data centers as to save overall energy.

Q1.2 (6 points) In the paper by C. Gomes et al. titled "*Computational sustainability: computing for a better world and a sustainable future*", among other things the authors describe how focusing on sustainability issues can help the development of computer system. Further elaborate on this and justify it giving a real example.

As sustainability problems become more and more challenging, they are not easily solved in traditional computer systems. Consequently, these complex problems drive the development of new algorithms to solve them as well as the development of new systems able to cope with the demands. In order to efficiently manage large data centers, there are many factors to consider and tune and thus new Artificial Intelligence models have been developed in order to come up with better suggestions for a more efficient management. The parallel systems to execute these complex algorithms also needed further development (e.g. the development of the Google Tensor Processing Unit – TPU).

Q2 (12 points) Basics, Metrics & Models

Q2.1 (2 points) The Total Power in an integrated circuit component (e.g. processor or memory) is composed of two parts: Dynamic and Static Power. Describe what Dynamic Power is and which are the factors it depends on.

Dynamic Power is the part of the Total Power that is consumed as a consequence of the switching activity. Dynamic power depends on the frequency and voltage (square) of the circuit.

Q2.2 (4 points) When evaluating a system, different metrics are used for different purposes. Briefly describe for each metric bellow which is the main goal for the system evaluation and give an example of a type of system for which that metric is very important:

(a) Energy (E)
(b) Energy-Delay-Squared (ED2)

For Energy (E) the main goal is the battery life for the system. This is a very relevant metric for mobile devices.

For Energy-Delay-Squared (ED2) the main goal is performance (the execution time) with a secondary goal the energy consumption. This is an efficiency metric for High-Performance systems.

Q2.3 (6 points) Knowing the values for the Energy-Per-Instruction for the instruction types listed below, determine the impact on the total energy of the compiler optimization that transforms the code on the left to the code on the right presented in Figure below.

Instruction type	ALU	Load	Store	Control
EPI [nJ]	50	80	60	40

Original	Optimized
`ADDI R4,R0,#100 ; R4 <- 100`	`ADDI R4,R0,#100 ; R4 <- 100`
`L1: LD R2,0(R10)`	`L1: LD R2,0(R10)`
`LD R3,0(R11)`	`LD R3,0(R11)`
`ADD R1,R2,R3`	`ADD R1,R2,R3`
`SD 0(R12),R1`	`ADDI R1,R1,#16`
`SUBI R4,R4,#1`	`SD 0(R12),R1`
`ADDI R10,R10,#4`	`SUBI R4,R4,#1`
`ADDI R11,R11,#4`	`ADDI R10,R10,#4`
`ADDI R12,R12,#4`	`ADDI R11,R11,#4`
`BNEZ R4,L1`	`ADDI R12,R12,#4`
`ADDI R4,R0,#100 ; R4 <- 100`	`BNEZ R4, L1`
`L2: LD R1,0(R13)`	
`ADDI R1,R1,#16`	
`SD 0(R13),R1`	
`SUBI R4,R4,#1`	
`ADDI R13,R13,#4`	
`BNEZ R4,L2`	

Original code:
1xALU + 100x(5xALU + 2xLoad + 1xStore + 1xControl) + 1xALU + 100x(3xALU + 1xLoad + 1xStore + 1xControl) = 802xALU + 300xLoad + 200xStore + 200xControl = 40100 + 24000 + 12000 + 8000 = 84100nJ = 84.10uJ

Optimized code:
1xALU + 100x(6xALU + 2xLoad + 1xStore + 1xControl) = 601xALU + 200xLoad + 100xStore + 100xControl = 30050 + 16000 + 6000 + 4000 = 56050nJ = 56.05uJ

The optimized code reduces the energy by 33% ((84.10-56.05)/84.10)

Q3 (12 points) Technology and Circuits

Q3.1 (6 points) Each processor is classified by its manufacturer with a Thermal Design Power (TDP) value. For example, the latest 12-core AMD Rizen 9 3900X has a TDP rating of 105 Watt. Explain what TPD is in practical terms, how it is determined, and if a processor can consume power higher than its TPD.

TDP is the maximum amount of heat a processor can stand for the execution of regular applications. TDP is determined by running a set of different benchmarks and observing the maximum power dissipated when running those relevant workloads. A processor can consume in extreme cases higher power than TDP. Power virus programs are synthetic programs that can force a processor to consume more power than TDP.

Q3.2 (6 points) Miniaturization of the circuits was a major driver for the processor development that we have observed in the latest years. In your own words explain what the impact is of reducing the technology point to half such as from 14nm to 7nm.

By reducing the technology point, the size of the transistors is being reduced as well as the wires. Consequently, in the same wafer you are able to fit more transistors. More chips in the same wafer means a reduction in the cost and also an increase in the "yield" since relatively fewer chips will be damaged during the production. By reducing the technology point to half and maintaining the same chip design (same transistors per chip) you are actually increasing the number of chips by a factor of 4x thus achieving a reducing in the cost by the same 4x factor.

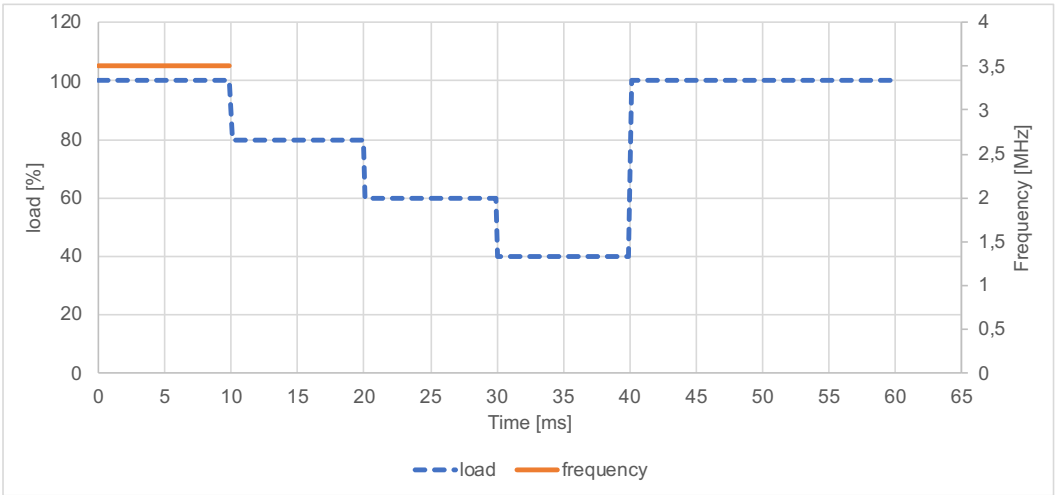
Q4 (14 points) Dynamic Power DVFS

Q4.1 (8 points) Assume a system with a processor with an operating frequency of 3.5GHz. For efficiency, the processor can execute in one of the 5 different frequency states: P0 3.5GHz, P1 3.0GHz, P2 2.5GHz, P3 2.0GHz, and P4 1.5GHz. Assume also that the Operating System (OS) has a Governor that monitors the load of the system every 10ms (at 5ms, 15ms, etc.) and decides on the frequency state for the processor every 10ms (at 10ms, 20ms, etc.) in the following way:

(a) If the load is greater than what was in the last time it was read (t-10ms) then set the frequency to the next higher frequency state (e.g. if currently in P3 then set the frequency to P2). If frequency is already at P0 then no change is made.

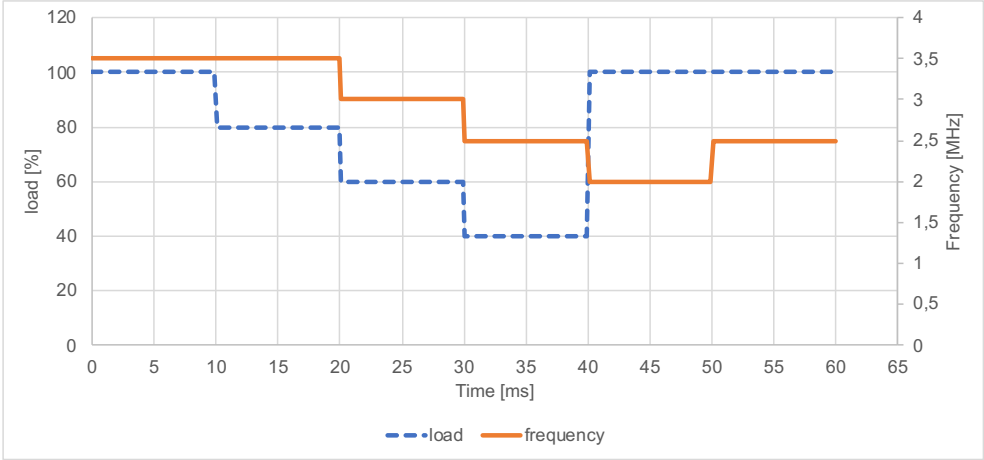
(b) If the load is lower than what was in the last time it was read (t-10ms) then set the frequency to the next lower frequency state (e.g. if currently in P3 then set the frequency to P4). If frequency is already at P4 then no change is made.

Using the load diagram below draw the frequency over time that is set by the above Governor.



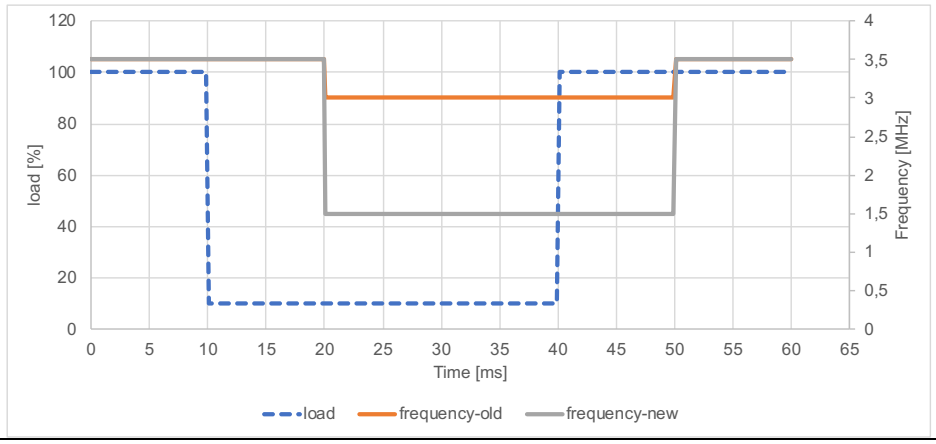
Load(t=5ms) = 100%
Load(t=15ms) = 80%

Since Load(t=15ms) < Load(t=5ms) the frequency will be changed from P0 to P1 in t=20ms
Load(t=25ms) = 60%
Since Load(t=25ms) < Load(t=15ms) the frequency will be changed from P1 to P2 in t=30ms
Load(t=35ms) = 40%
Since Load(t=35ms) < Load(t=25ms) the frequency will be changed from P2 to P3 in t=40ms
Load(t=45ms) = 100%
Since Load(t=45ms) > Load(t=35ms) the frequency will be changed from P3 to P2 in t=50ms
Load(t=55ms) = 100%
Since Load(t=55ms) = Load(t=45ms) the frequency will not be changed in t=69ms



Q4.2 (6 points) Comment on what situations are not well handled by the Governor described above. Justify your answer by giving an example as a load diagram that shows those badly handled cases. Suggest changes to the Governor to better handle these cases.

This Governor works well if the load changes over time but when it is constant, no change is made, so if the change in load is abrupt from one end to another this will only result in a single change of a single step in the frequency. So, in addition to the decision points a and b maybe we could add two more: (c) If the load is higher than 80% then set the frequency to P0; and (d) If the load is lower than 20% then set the frequency to P4.



Q5 (20 points) Dynamic & Static Power Techniques

Q5.1 (4 points) Single Instruction Multiple Data (SIMD) instructions are very popular now. Explain in your own words what a SIMD instruction is, how it works and what are their

benefits as compared with regular instructions. Give an example of an application (or type of application) that can benefit from the SIMD instructions.

SIMD stands for Single-Instruction Multiple Data, so SIMD instructions perform the same operation on multiple data at the same time. For example, a regular add instruction takes 2 inputs and produces 1 output, the sum of the inputs. A SIMD add instruction takes 2n inputs and produces n outputs, each one being the sum of the corresponding pair of inputs. Usually SIMD instructions operate on the same registers as the regular operations, for example a 32bit register can be used to store a single value of 32bits or 4 values of 8bits each. In this case with minimal overhead, the SIMD instruction is being much more efficient for the cases that the values are small and fit a small width (e.g. 8bits instead of 32bits) and thus 4 8bit operations can be done with the same energy as a single 32bit operation.
Multimedia applications as well as Machine Learning applications can benefit from SIMD as their variables can have values limited to 8 or 16bits instead of the regular 32 or 64bits.

Q5.2 (10 points) Assume a processor with a cache with the following characteristics: 2MB capacity, 64B cache line, 16-way set associativity, and 2048 sets. The execution conditions have determined that the power of the cache should be reduced to approximately to 25% of its current power.

(a) Use the cache resizing technique that was described in class to achieve this goal. Describe what are the changes that need to be applied to the cache as well as the resized cache characteristics. Is there also a need to change the address translation to access the data in the cache?

(b) What needs to be done to assure the correctness of the execution, i.e. what needs to be done so that an application executing before the cache resize continues its execution correctly during and after the cache resizing.

(c) After the resizing the system monitor detects a high miss rate indicating that the working set may be larger than the resized cache. What can be done in order to address this issue, without changing the cache size?

In order to reduce the power to 25% of the original we need to resize the cache to ¼ of its original size. As learned in class, the easiest way to resize dynamically a cache is by disabling a number of the ways. In this particular case, the cache has 16 ways. Since we want to have ¼ of the original size, we can disable 12 ways and just keep 4 ways active. Since we just disable the ways, there is no need to make any changes to the address translation. The only thing that needs to change is that with 4 ways we need only to check 4 tags per access to check if a variable is in the cache or not. The resized cache will be 512KB capacity, 64B cache line, 4-way set associative, and 2048 sets.

In order to assure correctness, the data in the ways that are to be disabled needs to be flushed (as a side note, if the cache is write through, then nothing needs to be done, the data can just be discarded as the latest data will be in the memory already, if the cache is write-back then the dirty data needs to be updated in memory). Then the resizing needs to take action – basically just disabling the ways and re-configuring the tag check comparators to compare only the active ways. Then after this the execution can continue as normal.

If the data needed to be stored is larger than the capacity and the capacity cannot be changed, the technique to be used would be data compression. In compression is successful, it will add an overhead to the access latency but the compressed working set would be able to fit in the cache and thus large gains would come from avoiding memory accesses.

Q5.3 (6 points) In class we presented two techniques to reduce the static power for caches: Cache Decay and Drowsy Cache. Explain in your own words what Cache Decay is,

how it works and why it can be effective in reducing the static power. Justify your answer.

Cache Decay is a technique that disables (switches off) parts of the cache that contain data that will not be used again or that its reuse distance is very long. This technique exploits the fact that in the cache there is a large portion of data that has very long reuse distance or that it will not be reused again. Instead of spending energy in maintaining this data "alive" its storage is switched off and thus energy can be saved from that.

Q6 (16 points) Communication & Memory / Heterogenenity & Specialization

Q6.1 (8 points) A Scratchpad memory is a piece of memory used to store temporary data such as a cache, but its management is done in software instead of hardware. Describe how a Scratchpad can be used to improve the efficiency of a system and how it can be effectively used (i.e. what is needed to be done in order to use it effectively).

A Scratchpad can be used to improve the efficiency of a system because the software (the programmer) has a better knowledge of the data use and reuse than what the hardware can guess. Thus, the available space will be used better and also the evictions can be controlled in order to effectively keep the data that needs to stay for reuse in the Scratchpad space. In order to use it though the programmer or the compiler need to issue specific instructions to add required data to the Scratchpad and also to remove unwanted data from the Scratchpad. This can be painful for the programmer and difficult to be done automatically.

Q6.2 (8 points) Approximate Computing is a concept that trades-off accuracy for efficiency. How and under which conditions can it be exploited? Give examples of applications that can benefit from it. Give an example of a software technique that can be used to exploit approximate computing.

Approximate Computing can be used when "good enough" solutions are accepted, i.e. when the software is tolerant to some sort of error in the result. This approximation should in turn offer a significant gain in efficiency. Certain applications like image processing where small errors in some of the pixels may not show in the overall picture or in applications where ranking is more relevant than an absolute number such as in ranking which pages are the top search results in a search engine. Other applications like machine learning are also tolerant to some error by construction of their models and algorithms.
One simple software technique that can be used is to randomly eliminate some iterations in a long running loop. This is also known as loop perforation.

Q7 (14 points) Reconfigurable Computing / Data Centers & Papers

Q7.1 (6 points) The Power Usage Effectiveness (PUE) is a metric used to classify the efficiency of Data Centers and is defined as the ratio between Total Facility Energy and the IT Equipment Energy. Describe two techniques that are used by Google and/or Facebook in their Data Centers in order to reduce the PUE

The reason for the PUE to be larger than 1 is because of the extra power needed to support the IT system in cooling and other. In order to reduce the power used in cooling you need to start by having systems that require less cooling and then use simpler cooling solutions like placing the data center in a cooler location and then just using the outside air to cool it down. Facebook has chosen this approach for their Data Center in Luleå. Another way to reduce the PUE is by having a better power distribution and management of the overall system.

Google has introduced an Artificial Intelligence approach to better manage the system and reduce it energy cost.

Q7.2 (8 points) In the paper by K. Ganesan et al. with title "*System-level Max Power (SYMPO) - A Systematic Approach for Escalating System-level Power Consumption using Synthetic Benchmarks*" the authors describe their approach to produce a "power virus". Explain in your own words what a "power virus" is and what its intended use is.

A power virus is a synthetic program (i.e. artificially created, not representing the code in a regular application or workload) that is designed to make the system use as much power as possible. This is done as a way to determine the maximum power that can be drawn by a system. This maximum power is usually higher than the rated TDP (Thermal Design Power) as this is achieved with artificial programs while TDP is achieved by regular programs. Power viruses are important though to dimension correctly the necessary cooling for the system even in a case of peak that is not usually achieved from regular application execution. Without a power virus the cooling would be done in a very conservative way (i.e. overestimated and thus not as efficient) as to guarantee the correctness and reliability of the components.