System-On-Chip Design And Modelling Exercises kg1-kg3 Feb 2018

1 Exercise Sheet ONE - Jan/Feb 2018.

QP1. *a)* Why might a subsystem on a SoC never be turned on in that SoCs lifetime?

b) What low-level and high-level mechanisms might control such permanent power gating?

c) A laptop CPU has a thermal inertia that means its fastest rate of change of temperature is 5 degrees per second. Given the CPU can easily generate more average power than can be dissipated by the fan on average, how might the clock frequency and supply voltage be controlled in a laptop and for what reason?

d) When viewing a DVD (including moving video and audio) on a laptop, what is the best clock frequency policy?

QP2. (Mainly for discussion in supervisions)

a) Data can be compressed before being sent by a hand-held unit over a radio link or later on at a central server. What are the advantages and disadvantages? [4 Marks]

b) How might you estimate the battery life of a sensor device that has a 3 watt radio transmitter, no LEDs or displays and needs to run about five minutes worth of CPU cycles per day? Would solar cells serve it well? [4 Marks]

c) In the last decade or so, digital sensors with radio (wireless) backlinks have been installed in all rooms in all public buildings. What have been the main forward steps in technology that have facilitated this? Does the price of copper make a difference?

QP3. *a)* What are the principal differences between an FPGA and a masked ASIC for implementation of a SoC? [5 Marks]

b) How can a SoC design team use FPGAs to prototype their product before SoC fabrication? [5 Marks]

c) When would it be sensible to ship an FPGA instead of a masked ASIC in production runs? [5 Marks]

QP4. A hardware accelerator can perform a block processing operation in 20 clock cycles that on the main CPU requires 1500 instructions. What would the main considerations be regarding whether the accelerator is of benefit in performance terms? State any dependencies data size, cache affinity, pipelining, CPU instructions per clock, accelerator relative clock frequency and so on.

QP5. Consider now the data size per operation is 16 bytes of argument and 16 bytes of result. The accelerator has dimension of 1 sq mm, it is 2 mm distant from the CPU and tracking capacitance is 1 fF/mm. It is made of logic gates with a density of 800,000 per sq mm. Will using the accelerator save electricity? State further assumptions needed to form your answer.

QP6. A convolutional neural network is implemented on an FPGA. The main computational burden in CNNs is matrix multiplication.

a) Why might a custom arithmetic precision be helpful compared with the 1+8+24 single-precision floating point typically used?

b) Give a formula or a specific estimate based on example concrete values for the throughput of such a CNN. You will need to case split on whether there are sufficient RAM blocks in the programmable logic to hold all of the coefficients or whether these need to be streamed from outside somehow. (If you are not intrested in CNNs just answer for matrix multiplication with fixed coefficients and streaming data).

c) Given that the DSP blocks of an FPGA are uniformly distributed over the programmable logic area, is there a problem with the 'shell and role' approach to virtualisation?

QP7. *a)* Why is it that system performance can be enhanced if a DRAM controller is able to respond to operations out-of-order? If the DRAM is cached, is it still helpful for the font side of the cache to still be able to respond out-of-order?

b) This question requires you have a little familarity with MPEG. Feel free to skip it. Consider the MPEG compression problem mentioned in lectures and make . Why is compression more power hungry than decompressing? You may assume a standard resolution 320x240 pixels using square tiles of 256 pixels each

and a search radius in units of one pixel for a radius of +/- 7 pixels in each direction. Estimate the number of ALU operations needed for a motion prediction operation following the diamond search technique of 'A New Diamond Search Algorithm for Fast Block-Matching Motion Estimation' Shan Zhu and Kai-Kuang Ma, IEEE Tans Imag Proc 2000. As it says in the caption to Figure 4, there are 24 search points in totaltaking nine, five, three, three, and four search points at each step.

Also, make initial notes as to why might this operation might better implemented on custom hardware compared with a CPU?

We will discuss this more on the accelerator exercise sheet.

c) ... still missing in this draft of this sheet.... Repeat for the Huffman Encocding step of MPEG.

QP8. *a)* What is the penalty for putting more and more transistors on a chip? (Or is there none?).

b) How can silicon yield problems be mitigated for regular structures?

c) What would be the potential uses of having a small PROM using a ten thousandth of the die area of a chip? A PROM is a write-once programmable ROM. Bits are all initially clear when the die is finished but are programmed with a high-power laser or ion gun or electrically with a masked circuit structure that uses hefty transistors to irreversibly melt targeted bit cells.

d) What is the Power Wall?

 END OF DOCUMENT.