COMPUTER SCIENCE TRIPOS Part II – 2021 – Paper 9

4 Comparative Architectures (rdm34)

(*a*) A superscalar processor may speculatively execute loads even when one or more earlier stores have not yet computed their memory addresses. In practice, we would need to restart execution from the speculative load if a memory-carried dependency is subsequently detected.

(*i*) With the help of some additional hardware it is possible to record which loads cause such ordering violations. Briefly outline how this could be done and how such a record could be used to help improve performance. [3 marks]

(*ii*) Describe why such a scheme may unnecessarily delay the issuing of a load even when the mechanism correctly recalls that the load has led to an order violation between a store and load in the past? [4 marks]

(*b*) Why might it also be advantageous for a superscalar processor to predict whether a particular load will hit or miss in the processor's L1 data cache? [3 marks]

(*c*) You are asked to design hardware to run artificial neural network applications in a high-performance and energy-efficient manner. Such workloads can typically make good use of many multiply-accumulate (MAC) units operating in parallel and narrow datatypes. Your system is required to support a range of different neural networks that vary considerably in the type of computations they perform. You consider three approaches: (1) to use a multicore processor; (2) to design a single domain-specific accelerator; (3) to compose your design from two or more domain-specific accelerators where each is specialised for different types of neural network.

(*i*) What are the advantages and disadvantages of each approach? [6 marks]

(*ii*) Describe one possible way of organising the multicore processor and a possible choice for the architecture(s) of its individual cores. Briefly justify your design decisions. [4 marks]