COMPUTER SCIENCE TRIPOS Part II – 2023 – Paper 9

1 Advanced Computer Architecture (rdm34)

(*a*) Describe why it is often important to try to limit off-chip memory accesses when designing a high-performance domain-specific accelerator that is part of a System-on-a-Chip (SoC)? [4 marks]

(*b*) What techniques or strategies might be employed to reduce off-chip memory accesses when designing a domain-specific accelerator? [8 marks]

(*c*) Imagine a domain-specific accelerator with numerous individual compute units. Each compute unit is programmable, has specialised functional units and produces requests to access memory. The compute units share a L1 cache.

In order to help improve the effective bandwidth to the shared L1 cache, memory requests that access the same address may be coalesced (or merged). Experiments confirm that there is scope to coalesce requests both from a single unit and from different units.

You are asked to design a hardware scheme for coalescing load requests only, store requests are not coalesced. The compute units generate loads and stores in program order. For all requests from a single compute unit, the coalescing unit must ensure that stores are not reordered with respect to loads or other stores from the same compute unit. The accelerator provides no guarantees about the way in which requests from different compute units may be reordered.

The design should be efficient and scalable, i.e. make good use of any on-chip memory required and make some attempt to minimise the number of address comparisons required. It should be possible to coalesce requests that are made from different compute units. There is no need to discuss how data is returned from the L1 cache to the compute units.

Describe your design, list any assumptions and justify your decision decisions. [8 marks]