COMPUTER SCIENCE TRIPOS Part II – 2022 – Paper 8

1 Advanced Computer Architecture (rdm34)

(*a*) A simple snoopy coherence protocol requires that bus transactions are broadcast and observed by all processors in the same order. How could a ring interconnect be used to support this protocol? [4 marks]

(*b*) Imagine a multicore processor with private L2 caches and an inclusive L3 or Last-Level Cache (LLC). We wish to reduce the miss rate of the L2 caches so replace the LLC with a smaller non-inclusive cache allowing for cache chip area to be redistributed to create larger L2 caches. We then discover that the miss rate of the smaller LLC changes very little, why might this be? [3 marks]

(*c*) In the case of a directory-based cache coherence protocol, why might we want to retain an inclusive directory even if our LLC is non-inclusive? [3 marks]

(*d*) It is suggested that directory-based cache coherence can scale with the aid of a hierarchy of on-chip caches. For example, we could group 64 cores into 8 clusters of 8 cores each. Each processor has its own private cache and each cluster has its own shared inclusive "cluster" cache. The chip also contains a shared inclusive Last-Level Cache (LLC). We assume sharers are tracked precisely. Describe how the cache hierarchy can be exploited to reduce the storage cost of tracking sharers. [4 marks]

(*e*) Imagine a System-on-a-Chip (SoC) that consists of multiple cores and a Domain-Specific Accelerator (DSA). The DSA could be given its own cache and be kept fully coherent with the other on-chip caches. Alternatively, the accelerator may be non-coherent and instead DMA data directly from main memory into a local scratchpad. When might each approach be preferable? [6 marks]