Q1

Illustrate the architecture of proposed by Flynn. How they are different in from a programmer's point-of-view?

Consider a processor operating at 1 GHz (1 ns clock) connected to a DRAM with a latency of 100 ns (no caches). Assume that the processor has two multiply-add units and is capable of executing two instructions in each cycle of 1 ns. Show working to calculate peak processor rating in GFLOPs.

Q2

Show working to calculate peak speed of computing on the above machine for a dot product of two vectors with one multiply-add on a single pair of vector elements, i.e., each floating-point operation requires one data fetch.

How pre-fetching data from memory and multi-threading hide memory latency? Show two illustrations and hints to answer this question.