1. GPU programming model

- The running program will have source code to run on CPU and code to run on GPU
- CPU and GPU have separated memone
- The data transferred from CPU to GPU to be computed.
- The data output from GPU computation is copied back to CPU memory

GPU execution model

GPU executes functions using a 2-level hierarchy of threads. A function's threads are grouped into equal-sized thread blocks, and a set of thread blocks are launched to execute the function. GPUs hide dependent instruction latency by switching to the execution of other threads.

2. The GPUs form logical groups of parallel threads belonging to the same instruction pack, named warps (wavefront in AMD terminology), and schedule a number of them for interleaved execution on an single instruction multiple-thread core, leading to a higher memory performance, reducing branch divergence.