Одной из реализаций методологии GPGPU является технология CUDA фирмы NVIDIA, используемая в видеокартах семейства GeForce [3]. Её программная модель показана на рис.2:



Рис.2. Программная модель технологии CUDA

Согласно данной модели все вычисления делятся на осуществляемые на центральном процессоре (хост или host) и на осуществляемые на видеокарте (устройство или device). Во время выполнения вычислений хост обращается к устройству, передавая на исполнение программное ядро (kernel) – скомпилированный шейдер графического процессора. Перед вызовом ядра происходит конфигурация потоков устройства. Предполагается, что при работе ядра будет создано большое количество потоков, которые объединяются в блоки. Все потоки одного блока исполняются на одном мультипроцессоре устройства, причём физически потоки в блоке объединяются в пулы (warp), в которых все команды исполняются одновременно. Множество блоков образует сетку (grid). В настоящее время число потоков в пуле составляет 32, размерность блока и сетки может составлять до 3 измерений включительно, однако максимальное число потоков в блоке не должно превышать 1024. На одном мультипроцессоре может одновременно исполняться до 16 блоков с суммарным числом потоков, не превышающим 2048.

Таким образом, технология CUDA позволяет разрабатывать алгоритмы под мощный потоковый процессор с возможностью одновременной обработки данных в тысячах потоков. Разумеется, у этой технологии есть ограничения, например:

1. Размер блока надо подбирать с учётом потребления регистровой памяти в каждом потоке, так как её количество ограничено в каждом мультипроцессоре.

2. Обращения к глобальной памяти рекомендуется оптимизировать таким образом, чтобы, упрощённо, соседние по индексу потоки обращались к соседним четырёхбайтовым словам в глобальной памяти видеокарты. Это приведёт к считыванию данных для нескольких потоков в одно обращение, а не в несколько независимых.

3. Невозможно использовать рекурсивные функции.

4. Ветвление в разных потоках блока на разные пути выполняется медленнее, чем в случае, когда все потоки ветвятся на один и тот же путь.

Таким образом, перевод на GPGPU некоторых алгоритмов может быть затруднён.