УДК 05.13.15

Оптимизация программ для графических ускорителей с помощью преобразования вложенных циклов

Булавинцев В.Г. (ФГАОУ ВО Санкт-Петербургский Национальный Исследовательский Университет Информационных Технологий, Механики и Оптики, г. Санкт-Петербург)

Научный руководитель: Жданов Д.Д (ФГАОУ ВО Санкт-Петербургский Национальный Исследовательский Университет Информационных Технологий, Механики и Оптики, г. Санкт-Петербург)

Современные графические ускорители (англ. «graphics processing unit», GPU) демонстрируют высокую производительность в задачах моделирования физических процессов, обучении нейронных сетей, решении задач из области криптографии. В то же время, во многих комбинаторных и поисковых задачах GPU уступают в эффективности вычислительным устройствам традиционной архитектуры. Одной из причин этого является то, что GPU построены на архитектуре «одна команда, много данных» (ОКМД), которая плохо подходит для исполнения программ, включающих сложные конструкции из условных переходов. В архитектуре ОКМД одно устройство управления (УУ) обслуживает несколько процессорных элементов (ПЭ). Когда группа вычислительных потоков, обслуживаемая одним УУ, встречает условный переход, в результате которого разные потоки должны выполнить разные ветви программы, УУ вынуждено разбить эту группу на подгруппы, блоки команд в которых будут выполняться последовательно. Конструкция же из нескольких вложенных условных переходов может фактически перевести ОКМД устройство в режим последовательного исполнения, снизив производительность кратно размеру ОКМД-группы (т.е. в 32-64 раза для современных GPU). Частным случаем такой конструкции является пара циклов, один из которых вложен в другой, и при этом число повторений внутреннего цикла неизвестно заранее. Подобные конструкции встречаются, к примеру, в решателях систем логических уравнений и алгоритмах визуализации, основанных на методе трассировки лучей.

Целью настоящего исследования является разработка методов повышения эффективности исполнения ветвлений на ОКМД-архитектурах.

Проблема условных переходов может быть отчасти решена при помощи преобразования графа потока управления программы, которое приводит конструкцию из пары вложенных циклов может к одному циклу. Полученная программа будет семантически эквивалентна оригинальной, но при этом может стать более пригодной для исполнения на GPU и других ОКМД-устройствах. Описанное преобразование подробно исследовано в настоящей работе.

Преобразование может автоматически применяться компилятором на этапе оптимизации кода программы. В системе построения компиляторов LLVM исходный код программы транслируется в промежуточное представление (ПП) для виртуальной машины LLVM. К коду в ПП применяются различные оптимизирующие преобразования, и лишь затем он компилируется в машинный код для целевой платформы. Для автоматического применения описанного преобразования нами разработан соответствующий программный модуль LLVM (т.н. transformation pass). Поскольку модуль работает с ПП, возможность его применения не зависит от того, на каком языке программирования была написана оригинальная программа.

Эффективность предложенной оптимизации продемонстрирована на специально разработанном приложении для GPU, моделирующем различные паттерны условных переходов.

Булавинцев В.Г Жданов Д.Д.