УДК 621.397.01

Ю.А. Затолокин, Э.И. Ватутин, В.С. Титов
e-mail: yuzato@list.ru
Юго-Западный государственный университет, Курск

ОЦЕНКА РЕАЛЬНОЙ ПРОИЗВОДИТЕЛЬНОСТИ ВЫЧИСЛЕНИЙ НА ГРАФИЧЕСКИХ ПРОЦЕССОРАХ С ПОДДЕРЖКОЙ ТЕХНОЛОГИИ OPENCL В ЗАДАЧЕ УМНОЖЕНИЯ МАТРИЦ

Приведено описание подходов к выполнению операции умножения матриц на видеокартах с поддержкой технологии OpenCL. Сделан сравнительный анализ производительности выполняемых действий без характерных для GPU оптимизаций и с оптимизациями.

Задача нахождения произведения матриц встречается в ряде научно-технических направлений (геометрическое моделирование, современная ускорительная техника, проектирование роботизированных средств, классификация бинарных отношений [1] и др.). При обработке больших объемов данных повышаются требования к времени решения задачи умножения матриц. Зачастую время ожидания выполнения расчетов с матрицами является определяющим фактором, от которого зависит общее время ожидания для получения результатов поставленной задачи в целом. Оптимизация выполняемых действий для получения произведения матриц включает различные подходы и использует распараллеливание выполняемых операций.

Анализ эффективности различных подходов выполнен с помощью решения задачи общего вида (умножения квадратных матриц размера N×N) с использованием графических процессоров, поддерживающих технологию OpenCL [4] (в рамках концепции GPGPU).

Выполнение операций умножения матриц на GPU имеет свою специфику и подразделяется на три этапа [2]:

- исходные матрицы A и B передаются из оперативной памяти в глобальную память GPU;
- выполняются операции с матрицами на GPU;
- для получения результирующей матрицы C данные передаются из глобальной памяти GPU в оперативную память.

Производительность обработки будет оцениваться по формуле $P = \frac{V}{t}$, где $V=2N^3$ – объем выполняемых вычислений, t – суммарное время, включающее выполнение операций с матрицами и обмен данными.

Для оптимизации времени выполнения операций применяются методики блочного умножения, кеширования данных, увеличения параллелизма на уровне инструкций (ILP).

Выполненная оценка производительности позволяет сделать вывод о том, что для оптимального использования вычислительных мощностей GPU необходимо соблюдать баланс между пропускной способностью глобальной памяти GPU, числом операций, выполняемых одной вычислительной единицей GPU и числом SIMD\PE-вычислителей.

Независимость OpenCL от аппаратной составляющей позволяет выполнять вычисления на любом вычислительном устройстве, с помощью драйвера [4], поставляемого разработчиком устройства. В выполняемом анализе вычислений с квадратными матрицами используются графические процессоры NVIDIA и AMD.

Выполнен сравнительный анализ архитектуры параллельных вычислений (CUDA) от NVIDIA с OpenCL с применением идентичных методов оптимизации. Результат обработки на машине с CPU Intel i7 4770 (Haswell) + GPU NVidia GeForce 970 GTX (Maxwell) в задаче параллельного умножения матриц одинарной точности размером 2048×2048 без алгоритмических оптимизаций с использованием трех циклов показал производительность GPU 13,2 GFLOP/s для CUDA [5] и 26 GFLOP/s для OpenCL. Оптимизация с помощью блочного умножения позволяет увеличить производительность GPU до 291 GFLOP/s для OpenCL. Максимальная производительность в выполняемых вычислениях была достигнута с оптимизацией блочным умножением с раскруткой внутреннего цикла с целью повышения параллелизма на уровне инструкций (ILP). При этом реальная производительность с использованием CUDA GPU 571 GFLOP/s и соответственно для OpenCL 365 GFLOP/s. Меньшая максимальная производительность с использованием OpenCL по-видимому объясняется реализацией дополнительной трансляции кода OpenCL под архитектуру CUDA, т.к. OpenCL драйвер для видеокарт NVIDIA использует вызов CUDA API.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. Ватутин Э.И., Зотов И.В. Построение матрицы отношений в задаче оптимального разбиения параллельных управляющих алгоритмов // Известия Курского государственного технического университета. Курск, 2004. № 2. С. 85–89.

2. Ватутин Э.И., Мартынов И.А., Титов В.С. Оценка реальной производительности современных видеокарт с поддержкой технологии CUDA в задаче умножения матриц // Известия Юго-Западного государственного университета. Серия: Управление, вычислительная техника, информатика. Медицинское приборостроение. 2014. № 2. С. 8–17.

3. OpenCL, 1 декабря 2015. Википедия: сайт. Режим доступа: https://ru.wikipedia.org/wiki/OpenCL (дата обращения 01.02.2017).

4. APP SDK – A Complete Development Platform // AMD: website. Available: http://developer.amd.com/tools-and-sdks/opencl-zone/amd-accelerated-parallel-processing-app-sdk/, accessed 01.02.2017.

5. Ватутин Э.И., Мартынов И.А., Титов В.С. Оценка реальной производительности современных процессоров и видеокарт с поддержкой технологии CUDA в задаче умножения матриц // CUDA альманах (май 2015). 2015. С. 9–10.