# Исследование эффективности хранения и обработки баз данных в графической памяти видеокарт с поддержкой CUDA*

А.И. Семенов, П.С. Костенецкий

Южно-Уральский государственный университет

Исследования, посвященные обработке баз данных в оперативной памяти, известны достаточно давно, нов связи с появлением технологии CUDA возникло новое направление исследований, посвященное обработке баз данных с использованием графических ускорителей [1]. В данном направлении известны работы, посвященные ускорению обработки запросов к базе данных [3] и оптимизации процесса интеллектуального анализа данных (data mining) [2]. Так как пропускная способность видеопамяти достигает 327 ГБ/с [5], что примерно в 3 раза превышает среднюю скорость работы с оперативной памятью [4], возникает вопрос эффективности не только обработки, но и размещения базы данных непосредственно в графической памяти. На сегодняшний день на рынке доступны гибридные вычислительные серверы с объемом видеопамяти, достигающим 16 Гб. Таких объемов памяти обычно достаточно для эффективного хранения и обработкибольшинства баз данных. Для увеличения объема хранимой в видеопамяти базы данных, отдельные серверы могут бытьобъединены в вычислительные кластеры. Например, объем графической памяти суперкомпьютера Tianhe-1A, находящегося на второй позиции ТОП 500, составляет 21.5 ТБ. Кроме того, при необходимости можно реализовать механизмы загрузки блоков данныхиз оперативной памятивычислительных узлов.

В даннойработе описывается незаконченное исследование, посвященноеоценкеэффективности хранения и обработки баз данных непосредственно в графической памяти видеокарт с поддержкой технологии программирования CUDA. На текущий момент авторами разработана система, моделирующая выполнение запроса JOINнепосредственно в памяти GPU ускорителя. Для хранения отношений базы данных, над которым выполняется запрос, используется глобальная памятьмодели программирования CUDA. Нитям необходимы очень частые обращения к их исполняемому программному коду и метаданным, поэтому для размещения этих данных используется константная память, имеющая значительно меньшую латентность, чем у глобальной памяти. Локальная память являетсяабстракцией модели программирования CUDA и означает блок памяти отдельной нити, расположенный в глобальной памяти. Локальная памятьиспользуется для хранения промежуточных данных во время выполнения запроса. Регистровая память используется для хранения локальных переменных нитей.Текстурная память на текущем этапе работы не используется.

Следующим этапом работы будет выполнение вычислительных экспериментов с целью исследования эффективности использования памяти GPU для хранения и обработки баз данных.

## Литература

1. Bakkum P., Skadron K. Accelerating SQL Database Operations on a GPU with CUDA // Proceedings of the 3rd Workshop on General-Purpose Computation on Graphics Processing Units (GPGPU'10), Pittsburgh, USA, March 14, 2010. ACM, 2010. P. 94-103.
2. Fang W., Lau K.K., Lu M., et al. Parallel Data Mining on Graphics Processors // Technical Report HKUST-CS08-07, 2008. P. 1-10.
3. Govindaraju N., Lloyd B., Wang W., et al. Fast computation of database operations using graphics processors. In SIGGRAPH ’05: ACM SIGGRAPH 2005 Courses, New York, NY, USA, 2005. ACM. P. 206.
4. Measuring Memory Bandwidth (White Paper). Intel Corporation. URL: http://www.intel.com/performance/resources/briefs/memband.pdf (датаобращения: 02.12.2011). 2010.
5. Спецификации NvidiaGeForceGTX 590. URL: http://www.nvidia.ru/object/product-geforce-gtx-590-ru.html (дата обращения 01.12.2011). 2011.

* Работа выполнена при финансовой поддержке Минобрнауки РФ (государственный контракт № 07.514.11.4036) и Российского фонда фундаментальных исследований (проект 12-07-00443-а).