Гибридные системы с прямым доступом к общей памяти на примере процессора Tegra K1: тесты Roofline и оптимизация МД алгоритма

В.С.Вечер, В.В. Стегайлов
Объединенный институт высоких температур РАН
Московский физико-технический институт (государственный университет)

Применение графических ускорителей для молекулярно-динамических расчетов уже давно является весьма обыденным делом [1]. Увеличение размеров моделируемых систем до миллионов атомов, а также применение все более точных моделей приводит к соответствующему росту требуемой вычислительной мощности. В то же время развитие графических ускорителей трансформировало их в компактные высокопроизводительные устройства, ориентированные на интенсивную обработку больших массивов данных.

Однако, следствием переноса вычислений на отдельное устройство стало возникновение ограничения, связанного с задержками по передаче данных по системной шине [2]. Для начала вычислений, данные должны быть скопированы на графический акселератор, и, чем больше объем требуемых данных – тем больше задержка на их передачу. Вышеизложенного недостатка лишены гибридные системы, поддерживающие технологию прямого доступа в память (DMA), – например, системы, где у процессора и графического ядра имеется физический доступ к одной и той же памяти [3][4]. В таких системах исчезает необходимость копировать данные через медленную шину, что снижает время доступа к ним и увеличивает вычислительную эффективность системы для решения требовательных к памяти задач.

Оценка выигрыша гибридных систем с общей памятью может быть произведена с использованием различных моделей и подходов. Одной из таких моделей оценки вычислительной эффективности является оценка по модели Roofline [5], связывающей производительность вычислений над числами с плавающей точкой с пиковой производительностью, пропускной способностью оборудования, а также интенсивностью арифметических расчетов. Модель также учитывает эффекты кэширования, параллелизма инструкций и данных, а также неодноуровневость памяти. Результатом разработки этой модели стало создание пакета программного обеспечения Empirical Roofline Tool, реализующего практический анализ вычислительной эффективности имеющегося оборудования.

В докладе будут представлены результаты применения данной модели с целью анализа эффективности DMA-подобной архитектуры процессора Nvidia Tegra K1, а также соответствующей оптимизации МД алгоритма для GPU для этого процессора.

Литература

1. *Anderson J. A., Lorenz C. D., Travesset A.* General purpose molecular dynamics simulations fully implemented on graphics processing units // Journal of Computational Physics. – 2008. – Т. 227. – №. 10. – С. 5342-5359.
2. *Vuduc R. et al.* On the limits of GPU acceleration // Proceedings of the 2nd USENIX conference on Hot topics in parallelism. – USENIX Association, 2010. – С. 13-13.
3. *Doerksen M.* An in-depth performance analysis of irregular workloads on VLIW APU // hgpu.org – 2014.
4. *Negrut D. et al.* Unified Memory in CUDA 6.0. A Brief Overview of Related Data Access and Transfer Issues. – Tech. Report TR-2014-09 // Simulation-Based Engineering Laboratory, University of Wisconsin-Madison, 2014.
5. *Lo Y. J. et al.* Roofline Model Toolkit: A practical tool for architectural and program analysis //High Performance Computing Systems. Performance Modeling, Benchmarking, and Simulation. – Springer International Publishing, 2014. – С. 129-148.