

АНАЛИЗ И РАЗРАБОТКА ВЫСОКОПРОИЗВОДИТЕЛЬНЫХ АЛГОРИТМОВ ПРИ АНАЛИЗЕ БОЛЬШИХ ОБЪЕМОВ РАЗНОРОДНЫХ ДАННЫХ С ИСПОЛЬЗОВАНИЕМ ГИБРИДНЫХ АРХИТЕКТУР

Мохаммед С. (Университет ИТМО)

Научный руководитель – доцент ФЦТ, к.т.н. Болгова Е. В.
(Университет ИТМО)

Аннотация. Статья посвящена разработке, реализации и анализу высокопроизводительных алгоритмов с использованием программно-аппаратной архитектуры параллельных вычислений на базе графических процессоров для ускорения расчетов гетерогенных графов, описывающих сложные процессы оказания медицинской помощи пациентам.

Введение. Вычисление матриц расстояний подобия является сложной вычислительной задачей, на которую приходится значительная часть времени обработки, особенно при работе с большими и многомерными наборами данных, и многие реальные приложения используют меры сходства, чтобы увидеть, как два объекта связаны друг с другом.

Основная часть. В данной статье в качестве примера вычислительной задачи была выбрана задача поиска сходства среди пациентов с закодированными событиями.

В качестве входных данных использовалась выборка, описывающая процесс и траекторию лечения пациентов из медицинского центра им. Алмазова, записанная в файл объемом 1.05 GB. Выборка представляет собой набор векторов, где каждый вектор - это отдельный пациент, проходивший лечение в центре, а элементы каждого вектора - закодированные события (назначения, операции, анализы, перевод пациента между отделениями и т.д.).

Результатом работы должен быть мета алгоритм, позволяющий обрабатывать файлы любого размера, при этом использовать вычислительные и архитектурные особенности вычислительной системы в зависимости от параметров системы, наличие или отсутствие технологических особенностей и объема входного файла. Для построения такого алгоритма был проведен ряд экспериментов для оценки производительности алгоритма при реализации на различных технологиях параллельного программирования.

Экспериментальные исследования показали:

- 1- Меры сходства могут быть рассчитаны для файлов размером менее 200 мегабайт (но это занимает много времени, около 4500 секунд)
 - 2- Файл размером более 200 мегабайт невозможно обрабатывать последовательным способом, так как при выполнении обработки возникает ошибка нехватки памяти
 - 3- Для файлов большего размера необходимо использовать параллельные вычисления в целом и графические карты в частности с использованием технологии GPGPU.
- В данной работе алгоритм был реализован с использованием технологии CUDA. Эксперименты показали, что для определения меры сходства с использованием глобальной памяти в графическом процессоре и размера блока (32,32) требуется около 4200 секунд по сравнению с примерно 1290 секундами для реализации с использованием разделяемой памяти (для такого же размера блока).
- 4- Лучшие результаты были получены с Apache Spark (технология быстрых кластерных вычислений) и время, затраченное на поиск меры сходства, около 120 секунд.

Выводы. В данной статье сравнивая результаты измерения сходства между последовательной и многопроцессорной обработкой обнаружили, что параллельные вычисления могут решить эту проблему, поэтому было проведено несколько тестов на GPU (с использованием глобальной и разделяемой памяти). Также часть работы посвящена исследованию

зависимости времени выполнения алгоритма от выбранной технологии и размера обрабатываемого файла и, как результат, выписана соответствующая формула. Данную формулу планируется использовать при дальнейшей работе, заключающейся в проектировании и реализации мета-алгоритма, который позволит обрабатывать файлы любого размера максимально быстро в зависимости от вычислительных возможностей системы.

Мохаммед С. (автор)

Подпись

Болгова Е.В. (научный руководитель)

Подпись