Секция «Вычислительная математика и кибернетика»

Построение суффиксного массива на GPU

Пироженко Илья Сергеевич
Студент
Филиал МГУ имени М.В.Ломоносова в г. Севастополе, Факультет компьютерной математики, Севастополь, Украина
E-mail: ilya.pir@gmail.com

Секвенирование генома - биохимический процесс определения порядка нуклеотидов в молекуле ДНК. Кроме того это необходимый процесс получения важной информации для биологических и химических исследований. Длина генома колеблется от нескольких миллионов нуклеотидов (для бактерии) до нескольких миллиардов нуклеотидов (для человека), таким образом, задача секвенирования генома требует обработки и анализа огромных объемов данных (генерируемых с частотой около терабайта в день), что требует высокопроизводительных вычислений. [1, с. 2]

Современные GPU(Graphics Processor Unit) представляют собой массивно-параллельные вычислительные устройства, с высоким быстродействием и большим объемом памяти. Многие ресурсоемкие вычислительные задачи хорошо ложатся на архитектуру GPU, позволяя заметно ускорить их численное решение. [3,с. 11]

Так как большая часть времени секвенирования генома приходится на поиск подстроки в строке, а строка с течением времени не изменяется, выгодно провести препроцессинг строки, т. е. обработать е таким образом, чтобы в дальнейшем поиск одного образца был как можно быстрее. [4] Для препроцессинга необходимо представить строку в виде эффективного набора данных, под этот критерий подходят: суффиксное дерево и суффиксный массив.

В работе используется суффиксный массив, из-за несравнимо больших требований к объему памяти для суффиксного дерева (что пока критично для GPU). Два из трех основных способов построения суффиксного массива реализованы на GPU:

1. Дублирование префикса (Prefix-Doubling)
2. Индуцированное копирование (Inducing Copy)

Сопоставлены скорости работы и требования к памяти. Построение суффиксного массива с помощью рекурсии (KA и KS алгоритмы) не было осуществлено, из-за архитектурных особенностей технологии CUDA.

Литература

1. Abdullah Gharaibeh, Matei Ripeanu, «Size matters: Space/Time tradeoffs to improve GPGPU application performance». – University of British Columbia, November 2010.- 20 с.

2. Donald Adjeroh, Tim Bell, Amar Mukherjee, «The barrows-wheeler transform: data compression, suffix arrays, and pattern matching ». - MIS:Press, NY, USA 2009 . - 360 с.

3. Боресков А.В., Харламов А.А., «Основы работы с технологией CUDA». – М.: ДМК Пресс, 2010. – 232 с.: ил

4. Гасфилд Д., «Строки, деревья и последовательности в алгоритмах: Информатика и вычислительная биология». – СПб.: Невский Диалект, 2003 – 150 с.

Слова благодарности

Огромная благодарность Александру Александровичу Дрозду, за помощь и поддержку.