Библиотека статистического анализа данных в гетерогенной распределенной среде *

А.К. Богушов, А.А. Морозов

ФГБОУ ВПО "Южно-уральский государственный университет"(НИУ)

Открытые библиотеки научных инструментов для языка программирования Python (Numpy, Scipy) содержат большое количество статистических функций, в то же время, они имеют ряд недостатков: невысокая скорость обработки больших объемов данных, отсутствие поддержки длинной арифметики. Использование Numpy на многопроцессорных, многоядерных или распределенных системах приводит к усложнению программы [1], а работа в гетерогенной среде невозможна.

Применение графических ускорителей в библиотеках статистической обработки "больших данных"позволяет получить значительные преимущества. Например, библиотека gpustats [2] реализует эти преимущества с помощью технологии CUDA и техники метапрограммирования.

В данной работе предложен другой подход к реализации библиотеки статистических функций. В основе библиотеки лежит технология OpenCL, которая позволяет работать приложению в гетерогенной среде с большими объемами данных. Для хранения данных предлагается новая абстракция Resilient Distributed Arrays (RDA), основанная на Resilient Distributed Datasets (RDD) [3], и позволяющая хранить в оперативной памяти распределенные данные, осуществлять контроль разбиения и размещения данных по вычислительным узлам. Создать или изменить RDA можно с помощью крупномодульных операций. Примером таких операций служат функции высших порядков (map, reduce, filter, join, ...), которые применяют одни и те же команды ко всем элементам массива. Многие статистические функции могут быть выражены через эти функции. Данное ограничение дает возможность обеспечить отказоустойчивость через сохранение всей истории применения операций для каждого RDA, которая может быть потом использована для восстановления актуального состояния данных из исходных данных, вместо постоянного сохранения промежуточных состояний системы на жесткий диск в контрольных точках после каждой операции. Это позволяет хранить массивы данных непосредственно в оперативной памяти и обрабатывать данные в интерактивном режиме. RDA в отличие от RDD имеют ряд преимуществ, связанных с акцентом на хранение массивов данных: многомерность, доступ к срезам данных, только числовые типы данные, совместимые с типами Numpy, поддержка длинной арифметики.

Литература

1. Parallel Programming with numpy and scipy:
URL: http://www.scipy.org/ParallelProgramming (дата обращения: 01.12.2012).

2. Andrew Cron, Wes McKinney gpustats: GPU Library for Statistical Computing in Python
URL: http://ftp.stat.duke.edu/WorkingPapers/11-17.pdf (дата обращения: 01.12.2012).

3. M. Zaharia, M. Chowdhury, T. Das, A. Dave, J. Ma, M. McCauley, M.J. Franklin, S. Shenker, I. Stoica. Resilient Distributed Datasets: A Fault-Tolerant Abstraction for In-Memory Cluster Computing, NSDI 2012, April 2012.

*Работа выполнена при поддержке гранта РФФИ №12-07-31013