Индексация больших объемов данных на вычислительном кластере, основанном на платформе Cloudera и в системе Enterprise Desktop Grid, основанной на платформе BOINC*

Никитина Н. Н.[1], Ивашко Е. Е.[1] , Мочалов В. А.[2], Мочалова А. В.[2]

[1]Институт прикладных математических исследований КарНЦ РАН, республика Карелия, г. Петрозаводск
[2]Институт космофизических исследований и распространения радиоволн ДВО РАН, Камчатский край, с. Паратунка

В течение нескольких последних десятилетий в различных областях человеческой деятельности наблюдается бурный рост объемов собираемой и анализируемой информации. В связи с этим активно разрабатываются новые методы обработки, анализа и хранения больших объемов данных с привлечением технологий высокопроизводительных вычислений. Одной из важных ресурсоемких задач обработки больших данных является индексация — создание структуры данных, хранящей ключевую информацию и обеспечивающей выполнение целого ряда задач по анализу данных, например, быстрого и точного поиска по ним. Свойства задачи индексации позволяют разрабатывать параллельные и распределенные алгоритмы ее решения. Так, в работе [1] рассматривается задача распределенной индексации и поиска по большим массивам данных на вычислительном кластере, построенном на базе связующего программного обеспечения Cloudera. Распределенная программная реализация индексации с применением библиотеки Lucene одного миллиона документов (в сумме 10 Гб текстов) на экспериментальном кластере позволила при ограниченных поисковых возможностях сократить время расчета индекса более чем в 55 раз по сравнению со встроенным в систему Polyanalyst [2] узлом индексации текстов.

Начиная с 90-х гг. XX века, все более широкое распространение получают вычислительные системы Desktop Grid, объединяющие территориально распределенные неспециализированные вычислители (например, персональные компьютеры), связанные с управляющим узлом сетью Интернет или локальной сетью передачи данных. Как правило, в таких системах вычислительные узлы не связаны между собой и предоставляют вычислительные ресурсы на нерегулярной основе. Desktop Grid могут быть построены на основе добровольно предоставляемых вычислительных ресурсов компьютеров частных лиц и организаций (добровольные вычисления) или на основе локальных ресурсов в масштабах организации или группы организаций (Enterprise Desktop Grid). Для организации и управления распределенными вычислениями в Desktop Grid создан ряд программных платформ, наиболее популярной из которых является BOINC (Berkeley Open Infrastructure for Desktop Computing). Данная платформа является свободным программным обеспечением с открытым исходным кодом, универсальна для решения вычислительных задач из различных областей науки и промышленности, поддерживает высокую производительность и масштабируемость Desktop Grid, предоставляет богатую функциональность.

Цель представленной работы — разработка и реализация алгоритмов распределенной индексации больших объемов данных в гетерогенной системе распределенных вычислений и выработка рекомендаций по применению в гетерогенной системе различных программно-аппаратных компонетов. Выработанные рекомендации позволят сформировать набор решений по используемому в гетерогенной системе множеству узлов, при этом в каждом из узлов будут определяться используемые аппаратные компоненты и программное обеспечение. Реализация и апробация разработанных алгоритмов будут производиться в Enterprise Desktop Grid Карельского научного центра РАН, Институте космофизических исследований и распространения радиоволн ДВО РАН и Петрозаводском государственном университете.

Список литературы:

1. Мочалов В.А., Мочалова А.В., Никитина Н.Н., Шутов А.А., Маряхина А.А. Некоторые вопросы применения открытых систем распределенных вычислений и обработки больших данных // Материалы V Всероссийского Симпозиума "Инфраструктура научных информационных ресурсов и систем", Санкт-Петербург, 5-8 октября 2015 г.
2. PolyAnalyst – Анализ данных. Анализ текста. Единый инструментарий. URL: http://megaputer.ru/polyanalyst.php

* Работа выполнена при поддержке РФФИ (проект 13-07-00008 А) и РГНФ (проект 15-04-12029).