Министерство образования Республики Беларусь
Учреждение образования
Белорусский государственный университет
информатики и радиоэлектроники

УДК 004.42

Казимирчик
Дмитрий Васильевич

Алгоритмы и программные средства распределённой обработки экспериментальных данных

АВТОРЕФЕРАТ
на соискание степени магистра технических наук

по специальности 1-40 80 05 – Математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей

Научный руководитель
Бранцевич П.Ю.
к.т.н., доцент

Минск 2014

КРАТКОЕ ВВЕДЕНИЕ

Сегодня стоимость хранения информации настолько низка, что зачастую представляется целесообразным постоянное накопление “сырых” экспериментальных данных получаемых от датчиков, систем мониторинга событий и т.д. в реальном времени. В будущем над собранными в процессе наблюдения данными можно проводить различные виды анализа для получения наиболее полной картины происходящего, выявления закономерностей и аномалий. В результате этого появляется необходимость в организации эффективной и быстрой обработки таких данных. Так как ресурс повышения производительности отдельно взятых процессоров давно исчерпан, постоянно ведётся поиск путей ускорения вычислительных процессов с использованием методов распараллеливания и распределения обработки данных между множеством машин.

MapReduce – это распространённая модель организации вычислений предназначенная для использования при обработке и генерации больших объёмов данных. Преимущество MapReduce заключается в том, что он позволяет распределенно производить операции предварительной обработки и свертки. Операции предварительной обработки работают независимо друг от друга и могут производиться параллельно (хотя на практике это ограничено источником входных данных и/или количеством используемых процессоров).

На основании вышеизложенного можно выделить актуальную проблему использования вычислительных ресурсов множества машин, работающих вместе для решения ресурсоёмких задач обработки большого количества данных, накопленных в процессе экспериментальных наблюдений.

Диссертационная работа посвящена разработке алгоритмов и ПО для систем, позволяющих решать задачи распределённой обработке сигнальных данных на вычислительных кластерах, построенных на базе ЭВМ общего назначения, что позволит создать гибкие, универсальные недорогие системы для обработки больших объёмов экспериментальных данных.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Цель и задачи исследования

Целью диссертационной работы является анализ и разработка алгоритмов и программного обеспечения для решения задач по распределённой обработке экспериментальных данных на базе персональных компьютеров общего назначения.

Для достижения поставленной цели необходимо решить следующие задачи:

1. Определить возможности применения существующих моделей организации распределённых вычисления для обработки экспериментальных сигнальных данных.
2. Проанализировать возможность адаптации распространённых алгоритмов и методов цифровой обработки сигналов для работы в распределённых системах.
3. Реализовать ПО для организации распределённой обработки экспериментальных данных.
4. Провести экспериментальные исследования разработанной системы.

Объектом исследования являются системы распределённой обработки данных.

Предметом исследования является математическое и программное обеспечение компьютерных систем для решения задач распределённой обработки сигнальных данных.

Основной *гипотезой*, положенной в основу диссертационной работы, является возможность использования вычислительных кластеров, состоящих из компьютеров общего назначения для эффективной обработки больших объёмов сигнальных данных. В частности рассматривается возможность применения модели распределённых вычислений MapReduce для организации работы широкого спектра алгоритмов цифровой обработки сигналов.

Связь работы с приоритетными направлениями научных исследований и запросами реального сектора экономики

Работа выполнялась в соответствии научно-техническим заданием и планом работ кафедры «Программное обеспечение информационных технологий» по теме:

«Разработать модели, методы, алгоритмы для оценки качества и состояния, обеспечения отказоустойчивости, защищенности и диагностируемости аппаратно-программных средств сложных систем и внедрения в современные обучающие комплексы» (ГБ № 11-2004, № ГР 20111065, научный руководитель НИР – В. В. Бахтизин).

Личный вклад соискателя

Результаты, приведенные в диссертации, получены соискателем лично. Вклад научного руководителя П.Ю. Бранцевича, заключается в формулировке целей и задач исследования.

Апробация результатов диссертации

Основные положения диссертационной работы докладывались и обсуждались на 50-й научной конференции студентов и аспирантов БГУИР (Минск, Беларусь, 2014)

Опубликованность результатов диссертации

По теме диссертации опубликована 1 печатная работа в сборнике материалов научно-технической конференции.

Структура и объем диссертации

Диссертация состоит из введения, общей характеристики работы, четырех глав, заключения, списка использованных источников, списка публикаций автора и приложений. В первой главе представлен анализ предметной области, выявлены основные существующие проблемы в рамках тематики исследования, показаны направления их решения. Вторая глава посвящена обзору алгоритмов анализа и обработки цифровых сигнальных данных. В третьей главе предложены методы и подходы к реализации вышеперечисленных алгоритмов в рамках распределённых систем. В четвертой главе предложена практическая реализация ПО для применения некоторых алгоритмов цифровой обработки сигналов в распределённых системах, проведён экспериментальный анализ полученных реализаций алгоритмов и сделаны выводы о возможности и целесообразности применения такого подхода на практике.

Общий объем работы составляет 61 страниц, из которых основного текста – 42 страниц, 12 рисунков на 8 страницах, 1 таблицы на 1 странице, список использованных источников из 30 наименований на 2 страницах и 1 приложения на 9 страницах.

ОСНОВНОЕ СОДЕРЖАНИЕ

Во **введении** определена область и указаны основные направления исследования, показана актуальность темы диссертационной работы, дана краткая характеристика исследуемых вопросов, обозначена практическая ценность работы.

В **первой главе** проведен анализ применяемых архитектурных решений систем, используемых при решении задач распределённой обработки информации. Сформулированы основные принципы построения распределённых си-

стем. Проанализировано применяемое для этих целей аппаратное и программное обеспечение. Выполнен анализ применяемых методов и алгоритмов распределённой обработки данных, выявлены их достоинства и недостатки.

Модель MapReduce была специально разработана для организации обработки очень больших массивов данных на большом кластере, состоящем из недорогих общедоступных компьютеров, не требующим надёжного сетевого соединения и подразумевает, что дисковое пространство стоит дёшево, а сетевые соединения являются дорогостоящей операцией.

MapReduce не является единственной моделью, развивающей идею распределённой обработки больших объёмов данных. Таким образом, большинство проблем, которые помогает решить использование MapReduce, уже имеют альтернативные решения в других моделях распределённой обработки данных. Однако, большинство из них не получило такого широкого распространения как она. MapReduce является более гибкой моделью, по сравнению с грид, так как может работать на кластерах состоящих общедоступных компьютеров, легко подвергается горизонтальному масштабированию, а также предоставляет чётко определённый и относительно простой интерфейс программирования. Что делает её основным кандидатом для исследования в данной работе.

Вторая глава посвящена рассмотрению основных методов и алгоритмов применяемых для анализа цифровых сигналов. Подавляющее большинство сигналов, обрабатываемых современными техническими системами, так или иначе имеет цифровое представление поэтому является целесообразным рассмотреть подходы к их анализу.

Базовым и простейшим подходом к анализу цифровых сигналов является вычисление их числовых характеристик. Исходно анализируемый сигнал представляется в цифровом виде (дискретный и квантованный) как массив данных $x(i), i = 0,1,2.....$

Для количественной оценки сигналов наиболее часто применяются следующие параметры. Абсолютные значения максимума и минимума сигнала на рассматриваемом отрезке времени T=[0, T], называемые пиковыми значениями, разница между пиковыми значениями называется размахом колебаний. Также к числовым характеристикам относят среднее значение (постоянная составляющая), мощность сигнала, среднее квадратическое значение и пик-фактор (характеризующий наличие амплитудных выбросов в сигнале).

Многие сигналы удобно анализировать, раскладывая их на синусоиды (гармоники). Для этого применяется преобразование Фурье. Дискретное преобразование Фурье, по возможности вычисляемое быстрыми методами, лежит в основе различных технологий спектрального анализа, предназначенных для исследования случайных процессов. ДПФ находит многочисленные применения в области анализа сигналов.

Одним из применений ДПФ в области анализа сигналов является спектральный анализ. Спектральный анализ – разложение сложного сигнала на некоторое множество простых сигналов (колебаний) с целью определения интенсивности каждого колебания в этом сложном сигнале. Также ДПФ позволяет

легко восстанавливать непрерывный периодический сигнал, занимающий ограниченную полосу частот. Также ДПФ широко применяется для цифровой фильтрации сигналов.

Кроме ДПФ для анализа сигналов может также применяться вейвлет-преобразование. Оно имеет широкий круг использования, при этом наиболее часто используемым способом его реализации является быстрое дискретное вейвлет-преобразование, которое используется БВП используется для выполнения цифровой фильтрации сигналов и спектрального анализа.

В **третьей главе** рассмотрены возможности применения модели распределённой обработки данных MapReduce для реализации основных алгоритмов анализа цифровых сигналов в распределённых системах.

Модель распределённых вычислений MapReduce хорошо подходит для вычисления таких характеристик сигналов как пиковые значения, размах колебаний, среднее квадратическое и пик-фактор. Это обусловлено тем, что перечисленные выше вычисления могут быть выполнены над произвольными кусочками входных значений сигнала, а затем, полученные результаты могут быть легко объединены между собой для получения итогового конечного результата вычислений.

Проанализирована возможность применения модели MapReduce для вычисления дискретного преобразования Фурье путём рекурсивной декомпозиции с помощью алгоритма Кули-Тьюки. Вместо вычисления дискретного преобразования Фурье для данного вектора длинны N, числа стоящие на чётных и нечётных позициях вектора, преобразуются в два отдельных вектора длинны N/2. Эта операция проводится далее рекурсивно, в результате чего вся последовательность разбивается на более мелкие кусочки. Итоговый результат подсчитывается путём комбинирования полученных преобразований Фурье для составных векторов. Главной идеей данного алгоритма вычисления быстрого преобразования Фурье является так называемая рекурсивная декомпозиция. Множество мелких преобразований вычисляется легче чем одно большое преобразование.

Предложена схема организации многофазной обработки map и reduce функциями для программной реализации алгоритма Кули-Тьюки для выполнения в распределённых системах. Также модель MapReduce может применяться при необходимости последовательной обработки (фильтрации) относительно небольших частей исходных данных, являющихся составной частью общей большой последовательности.

В **четвертой главе** рассмотрена практическая реализация ПО для распределённой обработки сигнальных данных (вычисление числовых характеристик и спектральный анализ), представлены результаты экспериментального исследования разработанной системы. На базе фреймворка Apache Hadoop был организован вычислительный кластер состоящего из одного главного узла и переменного числа подчинённых, предназначенный для распределённых вычислений с использованием модели MapReduce. Клиентский модуль работает с кластером посредством обращения к сервисным интерфейсам.

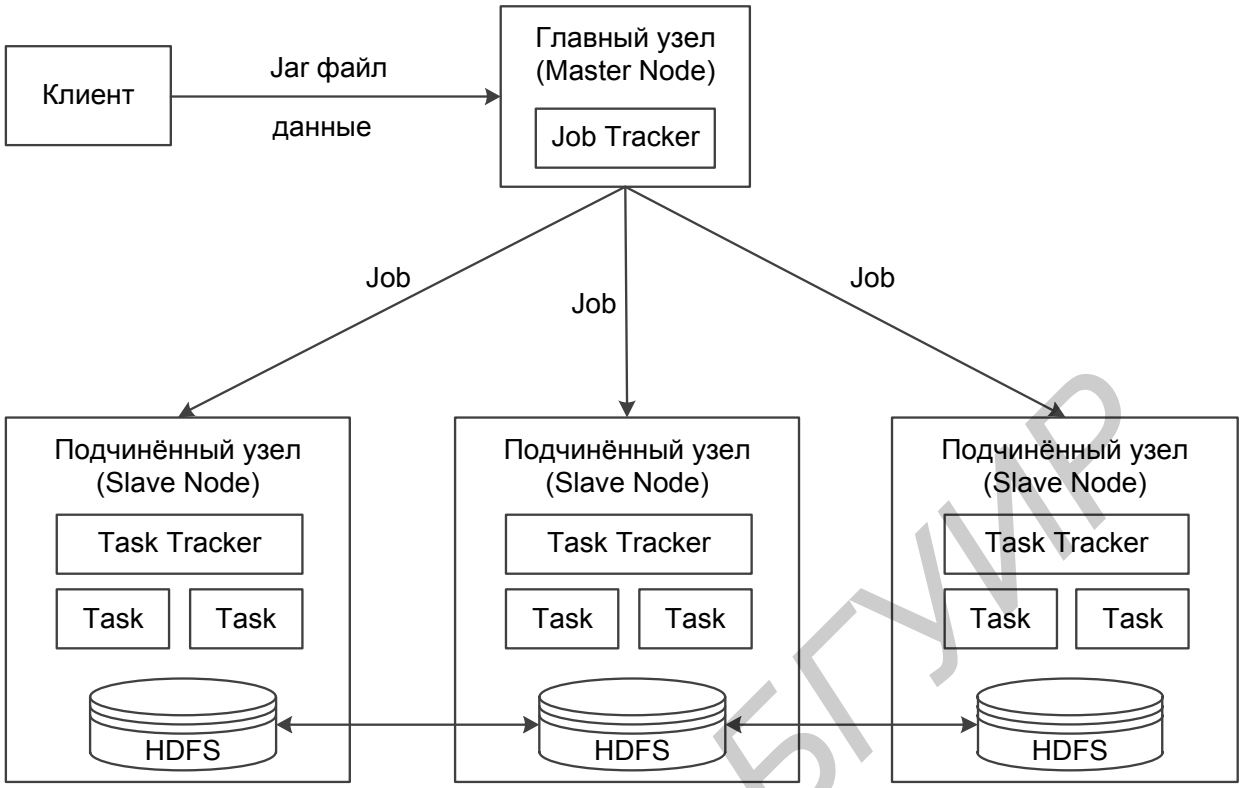

Рисунок 1 – Схема архитектуры системы

Для проведения экспериментальных наблюдений был использован вычислительный кластер на базе платформы Amazon EC2. Экспериментальный кластер состоял из пяти машин. В качестве главного узла использовался EC2 (Elastic Compute Cloud) сервер m3.xlarge с 4-х ядерным центральным процессором Intel Xeon E5-2670, 16 Гб оперативной памяти и SSD накопителем. В качестве подчинённых узлов выступали четыре EC2 сервера m1.large с 2-х ядерными процессорами Intel Xeon, 8 Гб оперативной памяти и HDD накопителями. Были произведены замеры времени выполнения для MapReduce реализаций алгоритмов нахождения пиковых значений и вычисления быстрого преобразования Фурье.

ЗАКЛЮЧЕНИЕ

Основные научные результаты диссертации

1. Проведён анализ существующих на сегодняшний день подходов к программному параллелизму и распределённой обработке данных, в результате которого выделены основные применяемые архитектуры и модели разработки распределённых систем. Обоснованы основные преимущества подхода MapReduce для использования в анализе возможностей организации распределённой обработки экспериментальных данных.

2. Предложены и программно реализованы алгоритмы распределённого вычисления основных числовых характеристик цифровых дискретных сигналов

с использованием модели распределённой обработки данных MapReduce. Проведён экспериментальный анализ реализованных алгоритмов при их применении к обработке последовательностей исходных сигнальных данных разных размеров. Сделан вывод о целесообразности применения таких алгоритмов при обработке массивов данных больших объёмов, так как при обработке небольшого количества данных данный подход зачастую проигрывает простой обработке на одной машине из-за дополнительных затрат на синхронизацию и коммуникации между узлами распределённой системы.

3. Предложен алгоритм организации распределённого вычисления быстрого преобразования Фурье с помощью программной модели MapReduce. Разработана его программная реализация с использованием фреймворка Apache Hadoop и проведён ряд тестов для анализа целесообразности использования данного подхода в зависимости от параметров используемого аппаратного обеспечения и размера преобразуемой последовательности. В результате чего сделан вывод о целесообразности вычисления ДПФ с применением MapReduce лишь в случае обработки больших последовательностей или для последовательной обработки (фильтрации) большого количества преобразуемых последовательностей.

4. Разработано программное средство для упрощения взаимодействия с вычислительным кластером построенным с помощью фреймворка Apache Hadoop, которое предоставляет пользовательский интерфейс для решения задач анализа цифровых сигнальных данных. Также разработаны программные скрипты для быстрого конфигурирования тестового кластера при использования различного числа вычислительных узлов.

Рекомендации по практическому использованию результатов

1. Полученные результаты формируют теоретическую и практическую базу для разработки ПО компьютерных систем для решения задач распределённой обработки цифровых сигнальных данных на вычислительных кластерах с применением компьютеров общего назначения.

2. Реализованные алгоритмы распределённой обработки могут применяться для вычисления числовых характеристик и спектрального анализа больших массивов сигнальных данных.

СПИСОК ОПУБЛИКОВАННЫХ РАБОТ

1. Казимирчик, Д.В. Использование технологии MapReduce для организации распределённой обработки сигнальных данных / Д.В. Казимирчик // Компьютерные системы и сети: материалы 50-ой научной конференции аспирантов, магистрантов и студентов. – Минск: БГУИР, 2014. – с. 36-37.