Об использовании кластеров из одноплатных компьютеров для анализа данных в Интернете вещей

И. М. Никольский[1,a], К. К. Фурманов[2]

[1]Московский государственный университет им М.В. Ломоносова,
факультет вычислительной математики и кибернетики,
Россия, 119991, г.Москва, ГСП-1, Ленинские горы, д. 1, стр. 52

[2]Национальный исследовательский университет «Высшая школа экономики»,
факультет экономических наук,
Россия, 101000, г. Москва, ул. Мясницкая, д. 20

E-mail: [a]haifly@rambler.ru

В данной работе рассматриваются возможности кластеров из одноплатных компьютеров при решении задач анализа данных. Подобные кластеры исследуются во многих академических организациях (университеты Саутгемптона и Эдинбурга, суперкомпьютерный центр Барселоны). Однако эти разработки носят в основном экспериментальный характер. В данной статье предложена практическая сфера применения кластеров из одноплатных компьютеров, а именно Интернет вещей (Internet of Things, IoT). Показано, что в условиях новых тенденций, когда обработка данных IoT частично перемещается из облачных инфраструктур ближе к устройствам сбора данных (концепции edge computing, fog computing), рассматриваемые кластеры могут стать важной частью локальной инфраструктуры IoT. В рамках предлагаемой работы построен небольшой кластер из RaspberryPi. Вычислительные возможности машины протестированы на задаче оценки длин интервалов для наблюдений в регрессионной модели. Задача решалась методом бутстрап. Показано, что даже будучи собранным из недорогих комплектующих, кластер может весьма эффективно проводить параллельную обработку данных.

Ключевые слова: одноплатный компьютер, интернет вещей, анализ данных.

Введение

Данная работа посвящена исследованию возможностей одноплатных компьютеров в области научных вычислений. Этот класс устройств стал популярен после того как фирма Broadcom выпустила в 2011 году знаменитый Raspberry Pi [RaspberryPi]. Он рекламировался как полноценный Linux-компьютер размером с кредитную карту и стоимостью в 35 долларов. Изначально планировалось, что он будет использоваться для обучения школьников программированию. Однако (неожиданно для создателей устройства) RaspberryPi вызвал большой интерес у широкого круга энтузиастов. В короткие сроки было предложено огромное количество проектов с использованием этого утройства.

Под одноплатным компьютером (single-board computer) понимается компьютер, в котором все необходимые для функционирования модули (процессор, ОЗУ и т.д.) размещены на одной плате. На сегодняшний день спектр таких изделий весьма широк. Это и более дешевые аналоги RaspberryPi (Orange Pi, Banana Pi), и изделия солидных компаний (Intel Galileo, Samsung Artik). Отдельно стоит выделить Parallela Board. Производитель - фирма Adapteva - заявляет следующие характеристики: 16-ядерная версия - 32 Гфлопс, 64-ядерная - 102 ГФлопс.

Такое разнообразие этих миниатюрных компьютеров можно объяснить, в частности, распространением концепции Интернета вещей (Internet of Things, IoT). Суть ее состоит в объединении в единую сеть разнообразных электронных устройств - датчиков, микроконтроллеров и т.д. Такие системы находят свое применение в сферах безопасности, здравоохранения, сельского хозяйства и т.д.

Как правило, данные, генерируемые в IoT, загружаются в некоторую облачную инфраструктуру, где происходит их обработка. Однако такая схема неприемлема, если анализ данных должен происходить быстро. Предложенная компанией Cisco концепция "туманных вычислений" (fog computing, [Fog computing…]) предполагает, что при необходимости обработка данных будет происходить локально, там же, где они собраны. Одноплатные компьютеры хорошо подходят на роль "капли" в этом "вычислительном тумане" благодаря своим компактным размерам, низкому энергопотреблению и достаточно большим (по сравнению с микроконтроллерами) вычислительным возможностям.

Поскольку объем данных, генерируемый компонентами IoT, может быть огромным, для их обработки может потребоваться не один, а целый кластер из одноплатных компьютеров. На данный момент существует несколько проектов построения таких кластеров, например, Iridis-Pi (университет Саутгемптона, 64 Raspberry Pi, см. [Simon, 2013]), проект resin.io (128 Raspberry Pi). Такие кластеры хорошо подходят для образования в области распределенных вычислений, но в IoT пока применения не нашли.

В данной работе исследуется применимость кластеров из одноплатных компьютеров для задач анализа данных. Используется кластер, построенный одним из авторов. В качестве тестовой взята задача из области регрессионного анализа. Далее в п2 представлено описание устройства кластера, п3 - формулировка тестовой задачи, в п4 - результаты вычислительных экспериментов.

Кластер из Raspberry Pi

В рамках данной работы был построен кластер из трех RaspberryPi 2 model B. Приведем основные характеристики этой модели RaspberryPi: процессор - ARM Cortex-A7, частота 900Мгц, четыре ядра, ОЗУ-1 Гб. Отметим, что на Raspberry Pi есть аппаратный генератор случайных чисел, что может быть очень полезно для задач математической статистики и криптографии. В качестве ПЗУ используются microSD карты. Питание осуществляется через USB-хаб, который подключен к сети через блок питания с выходным напряжением 5В.

Вычислительные узлы соединены с помощью роутера Asus RT-N12, на каждом из них установлена ОС Raspbian. Стоимость кластера (три Raspberry Pi, провода, USB-хаб, роутер, microSD карты) укладывается в 17 тыс руб по ценам 2016 года.

Тестовая задача

Применим наш кластер для исследования регрессионной модели. В качестве тестовых данных будем использовать набор данных о моллюсках Galiotis из коллекции UCI Machine Learning [Machine Learning Repository]. Исследуемая регрессионная модель задается следующей формулой

$$log(N_r) \sim s + sz_{mean} + w_s + w_{sh},$$

где N_r - число колец на раковине моллюска, s - пол, sz_{mean} - усредненный размер, w_s - вес тела моллюска, w_{sh} - вес раковины. Рассмотрение данной модели целесообразно, так как по числу колец можно определить возраст моллюска.

Оценим длины интервалов для наблюдений с помощью метода бутстрап. Этот метод близок по своей сути к методу Монте-Карло. Из исходной выборки делаются повторные выборки (перевыборки, resamples), для каждой вычисляется доверительный интервал, затем результаты агрегируются. Его распараллеливание не представляет труда. Каждый узел кластера делает N перевыборок, вычисляется доверительный интервал. Полученные на всех узлах интервалы усредняются.

Описанный алгоритм был реализован в виде программы на языке R. Входные параметры - количество наблюдений, доверительный уровень, количество перевыборок бутстрапа. Кластерные вычисления (т.е. запуск рабочих процессов на каждом из узлов) реализованы с помощью встроенных возможностей R.

Результаты вычислительных экспериментов

В таблице 1 показаны результаты вычислительных экспериментов. Первая колонка(nboot) - количество перевыборок бутстрапа, вторая - количество наблюдений, третья - время (в секундах) вычислений на трех узлах, четвертая - время выполнения на одном узле. Доверительный уровень во всех экспериментах равен 0,95.

Рассмотрим первые пять строк таблицы. Мы видим, что последовательная версия выполняет N перевыборок быстрее, чем параллельная версия выполняет расчет с N перевыборками на каждом из трех узлов. Это происходит из-за высоких накладных расходов на запуск процессов и коммуникации.

Попробуем взять в параллельной версии на каждом узле меньше перевыборок на каждом узле, чем в последовательной версии. И здесь мы уже получаем ускорение - параллельная версия программы выполняет 3900 перевыборок (по 1300 на каждом узле) быстрее, чем последовтельная 1500 перевыборок. Соответствующие результаты представлены в последних двух строках табл.1.

Таблица 1. Результаты вычислительных экспериментов

nboot	sample.len	паралл3узла	послед
1000	50	42.099	27.527
1000	80	41.708	27.926
1000	200	45.225	29.534
1500	200	43.884	67.292
2000	200	89.659	59.166
1500	200		67.605
1300	200	58.901	

Заключение

Одноплатные компьютеры лишь недавно вышли за пределы сегмента устройств для автоматизации производства и начали завоевывать массовый рынок. Их небольшие размеры, малое энергопотребление и достаточно высокие вычислительные возможности способны сделать их основой для реализации проектов в рамках новой парадигмы интернета вещей. Научное сообщество также проявляет интерес к компьютерам, в надежде создать компактные энергоэффективные кластеры для высокопроизводительных вычислений. Полученные в данной работе результаты показывают, что даже будучи построенным из недорогих компонент, кластер из одноплатных компьютеров может давать ускорение при решении задач математической статистики.

Список литературы

Raspberry Pi. [Electronic resource]: raspberrypi.org.

Fog Computing and the Internet of Things: Extend the Cloud to Where the Things Are. [Electronic resource]. URL: https://www.cisco.com/c/dam/en_us/solutions/trends/iot/docs/computing-overview.pdf, (accessed 30.10.2016).

Simon J. Cox, James T. Cox, Richard P. Boardman, Steven J. Johnston, Mark Scott, Neil S. O'Brien Iridis-pi: a low-cost, compact demonstration cluster // Cluster Computing, June 2013, Volume 17, Issue 2, pp 349-358 .

Machine Learning Repository: archive.ics.uci.edu/ml/.

On usage of singleboard computer clusters for Internet of Things data analysis

I. M. Nikolsky[1,a], K. K. Furmanov[2]

[1]Lomonosov Moscow State University, Faculty of Computational Mathematics and Cybernetics, 1, b.52, GSP-1, Lenin hill, Moscow, Russia, 119991

[2]National Research University "Higher School of Economics", Faculty of Economic Sciences,20, Myasnitskaya st., Moscow, Russia, 101000

E-mail: [a]haifly@rambler.ru

In present work we study capabilities of a single-board computer cluster to solve data analysis poblems. Such clusters are investigated in several academic organisations (universities of Southhampton and Edinburgh, Barcelona supercomputer center). But these are mainly research projects. We propose a way of practical application of single-board computer clusters, namely usage in Internet of Things. We point out that in presence of new trends when IoT data analysis is partially moved from clouds closer to data collection devices (edge computing, fog computing) the considered clusters may grasp an important role in local IoT infrastructers. In present paper we build a small cluster of Raspberry Pi computers. We test computational efficiency of our "supercomputer" on a problem of estimation of confidence intervals for observations in a regression model. The problem is solved by bootstrapping method. It is shown that even a custer built of inexpensive hardware may be efficient in parallel data analysis.

Keywords: singleboard computer, Internet of Things, data analysis