ISBN 978-5-901795-13-2 ПРОГРАММНЫЕ СИСТЕМЫ: ТЕОРИЯ И ПРИЛОЖЕНИЯ. Переславль-Залесский, 2008

Е. О. Сергеева

Эффективное использование распределенных хранилищ данных*)

Научный руководитель: к.х.н. А. А. Московский

Аннотация. Данная работа посвящена разработке и реализации технологии для эффективного хранения и обработки данных. Представлено решение, базирующееся на файловой системе Lustre. Также описан способ разработки эффективных параллельных приложений для обработки данных, расположенных в Lustre.

1. Введение

За последние годы стало привычным, что данные дистанционного зондирования Земли (наблюдения поверхности Земли авиационными и космическими средствами, оснащенными различными видами съемочной аппаратуры) широко используются в самых различных областях человеческой деятельности — для выявления пожаров, наблюдения за облачными массами, температурой поверхности суши и поверхности моря и т. д. Данные дистанционного зондирования Земли преимущественно являются высококачественными мультиспектральными изображениями. Такие изображения нужно:

- хранить, при этом накапливаются большие объемы информации (до нескольких петабайт), так как съемки отличаются высокой периодичностью;
- обрабатывать, при этом задачи обработки аэрокосмических изображений характеризуются необходимостью выполнений различных вычислительных операций над большими объемами данных, что требует значительных вычислительных мощностей.

Обработка данных на вычислительных системах кластерного типа позволит использовать кластер не только для обработки, но и в качестве оперативного хранилища данных. При этом данные равномерно распределяются по узлам кластера. Очевидно, что наиболее

*) Представлено по тематике: *Программное обеспечение для суперЭВМ.*

эффективным будет назначение задач по обработке данных на те узлы кластера-хранилища, на которых непосредственно находятся данные. В противном случае, если обрабатываемая информация будет достаточно большого объема, накладные расходы от передачи данных между узлами окажутся слишком большими, и мы не получим прироста производительности от распараллеливания.

2. Распределенная кластерная файловая система

В качестве распределенной файловой системы использовалась Lustre [1], разработанная Cluster File Systems [2]. Эта файловая система обладает следующими достоинствами:

(1) масштабируемость;
(2) производительность;
(3) открытый исходный код;
(4) поддержка отказоустойчивости;
(5) возможность получения информации о том, на каком из узлов кластера расположен тот или иной файл или его часть;
(6) возможность принудительно размещать файлы на заданных узлах кластера.

Последние два пункта имеют особое значение для выполнения поставленной задачи — совмещения хранилища данных и средств их эффективной обработки.

3. Использование возможностей Lustre при выравнивании нагрузки

Чтобы наиболее эффективно обрабатывать данные, хранящиеся в файловой системе Lustre, необходимо задачу, связанную с обработкой какого-либо фрагмента файла, отправить именно на тот узел, на котором эти данные физически расположены. В качестве средства реализации был выбран TSim [3] — библиотека для параллельных вычислений, реализующая основные концепции OpenTS [4]. Для определения узла расположения данных разработана следующая схема:

- определяется смещение в файле (В реализованной задаче обрабатывались снимки в формате TIFF [5], при помощи интерфейса библиотеки libtiff [6] удается получить точные данные о смещении);

- определяется схема расположения файла в кластере при помощи функции интерфейса Lustre. Функция возвращает, помимо другой информации число и размер порций (stripes), на которые разбит файл;
- по данному смещению, определяется номер порции файла, которую обрабатывает данная задача;
- параллельная файловая система Lustre использует барабанную (round-robin) стратегию назначения порций файлов на узлы, что позволяет по номеру порции определить номер узла в группе узлов кластерной установки, хранящих данный файл. По номеру определяется IP адрес узла;
- по IP адресу узла среди всех узлов, участвующих в расчете в TSim, находится IP который указан в конфигурации Lustre. Выполнение задачи назначается на этот узел.

Таким образом, в ходе обращения клиента к файлу в Lustre лишь обращение за метаданными вызовет обращение по сети, поскольку чтение производится из файла, расположенного на том же узле.

4. Классификация космических снимков по метрике Махаланобиса

Метрика Махаланобиса [7] (Mahalanobis Distance) — это особый вид расстояния, который активно применяется в статистике. Расстояние Махаланобиса основано на корреляции, благодаря которой можно анализировать различные сложные структуры данных. Эта метрика отличается от Евклидового расстояния тем, что учитывает корреляции внутри анализируемого множества. Параллельный классификатор использует расстояние Махаланобиса для определения принадлежности текущего обрабатываемого элемента множества к одному из выбранных экспертом подмножеств (ROI[1]) из всего множества входных данных. Такой выбор осуществляется с помощью определения минимального из расстояний Махаланобиса от обрабатываемого элемента до каждого из ROI. В случае если найденное минимальное расстояние оказывается больше выбранного экспертом порога, текущий элемент считается неклассифицированным. Контролируемый классификатор на базе метрики Махаланобиса [8] был разработан в ИПС РАН с использованием системы динамического автоматического распараллеливания OpenTS.

[1]ROI — Region Of Interest

Для демонстрации эффективности разработанной схемы была реализована версия описанного выше классификатора с использованием TSim, шаблона Map [9] и возможностей параллельной файловой системы Lustre. Технически, в основе программы лежит независимая обработка множества пикселей — то есть в данной задаче очень удобно использовать шаблон Map. Для эффективной работы с файловой системой Lustre в программу был добавлен планировщик. При реализации планировщика узел, на который назначается выполнение задачи, выбирался в строгом соответствии с описанным выше способом определения узла расположения данных в кластерном хранилище по данному смещению в файле. В связи с тем, что обрабатываемые снимки были представлены в формате TIFF, смещение в файле определялось при помощи средств библиотеки libtiff. Также следует подчеркнуть, что информация о схеме расположения файла в Lustre определяется один раз, при запуске программы. Далее она добавляется в параметры шаблона и используется планировщиком. Схема расположения всех обрабатываемых файлов в одном каталоге считается идентичной, в соответствии с документацией к файловой системе.

Таким образом, была реализована параллельная версия классификатора на TSim, учитывающая расположение данных на узлах и использующая шаблон параллельного программирования Map.

5. Результаты

При тестировании классификатора, данные располагались в хранилище (в параллельной файловой системе Lustre). При тестовых запусках на одном узле все данные располагались на этом же узле; при тестировании на двух узлах, данные, соответственно, располагались на этих двух узлах. Такая особенность расположения обрабатываемых данных связана с планировщиком алгоритма, который отправляет задачу, обрабатывающую данные на тот узел, на котором они расположены. Конфигурационные параметры кластера, на котором проводилось тестирование, показаны в таблице 1.

Было проведено многократное тестирование системы, усредненные результаты обработки шести снимков объемом приблизительно по 37 и 151 Mb показаны соответственно в таблицах 2 и 3.

Место расположения	ИПС РАН
Число вычислительных узлов	2
Тип процессора	Intel(R) Xeon(TM) 2.80GHz
Оперативная память узла	1 GB
Дисковая память установки	250+80 GB
Тип системной сети	Gigabit Ethernet
Конструктив узла (форм-фактор)	2U

Таблица 1. Установка, на которой проводились исследования — demo.botik.ru

Количество узлов	1 узел	2 узла
Время, сек	24.133	15.171
Процент	100%	62.86%

Таблица 2. Обработка данных общим объемом около 222 Mb

Количество узлов	1 узел	2 узла
Время, сек	1650.234	924.166
Процент	100%	56,002%

Таблица 3. Обработка данных общим объемом около 906 Mb

6. Вывод

В данной работе были поставлены и выполнены следующие задания:

- Был разработан и реализован механизм распределения заданий на основе информации, получаемой планировщиком от прикладного программного интерфейса Lustre. Решены технические сложности, возникающие при использовании файлов в формате TIFF
- Реализованный механизм был использован в программе обработки космических снимков по метрике Махаланобиса

- Было проведено тестирование полученной программы, демонстрирующее прирост производительности работы программы

Таким образом, при помощи файловой системы Lustre и реализованной схемы продемонстрирован способ использования кластерной установки, как хранилища и эффективного обработчика данных дистанционного зондирования Земли.

Благодарности

Данная работа была частично поддержана и выполнялась в рамках следующих проектов:

- «метапрограммирование на основе шаблонных классов C++ как средство создания высокопроизводительных распределенных приложений»;
- «разработка средств параллельной обработки изображений дистанционного зондирования Земли с динамическим распределением нагрузки. Разработка прототипа распределенного архива изображений с единым каталогом информации».

Список литературы

[1] Cluster File Systems Lustre, Эл. ресурс: `http://wiki.lustre.org`.

[2] Cluster File Systems, Эл. ресурс: `http://en.wikipedia.org/wiki/Cluster_File_Systems`.

[3] Московский А. T-Sim — библиотека для параллельных вычислений на основе подхода T-системы: «Международная конференция «Программные системы: теория и приложения», 2006.

[4] OpenTS, Эл. ресурс: `http://opents.botik.ru/`.

[5] Википедия TIFF, Эл. ресурс: `http://ru.wikipedia.org/wiki/TIFF`.

[6] LibTIFF - TIFF Library and Utilities, Эл. ресурс: `http://www.libtiff.org/`.

[7] Wikipedia Расстояние Махаланобиса, Эл. ресурс: `http://en.wikipedia.org/wiki/Mahalanobis_distance`.

[8] Абрамов С.М., Московский А.А., Первин А.Ю. Разработка высокопроизводительных клиент-серверных приложений для работы с данными дистанционного зондирования Земли. — Москва: Научно-техническая конференция ФГУП «РНИИ КП», 10.

[9] Московский А., Первин А., Сергеева Е. Первый опыт реализации шаблона параллельного программирования на основе T-подхода: «Международная конференция «Программные системы: теория и приложения», 2006.

E. O. Sergeeva. *Efficient using Data Warehousing* // Proceedings of Progranm Systems institute scientific-practical conference "Program systems: Theory and applications", devoted to the 15^{th} anniversary of Pereslavl University named A. K. Ailamazyan. — Pereslavl-Zalesskij, 2008. — p. 213 — 219. — ISBN 978-5-901795-13-2 *(in Russian).*

Abstract. This paper is dedicated to development and implementation of technology for efficient data storage and processing using cluster system. We present solution based on Lustre file system. As well, this paper describes this method to design parallel applications using Lustre.

Перевод проверен: кандидат хим. наук А. А. Московский