Оптимизация распределенного комплекса обработки спутниковых данных

Кихтенко В.А. Смирнов В.В. Чубаров Д.Л.
Институт вычислительных технологий СО РАН
{vladimir.kikhtenko, dmitri.chubarov, valentin.smirnov}@gmail.com

Работа выполняется при поддержке проекта IV.31.2.1. Программы фундаментальных исследований СО РАН на 2010 - 2012 гг., Российского фонда фундаментальных исследований (гранты 11-07-12048-офи-м-2011, 12-07-00545-а); Программы интеграционных фундаментальных исследований Президиума СО РАН (междисциплинарные проекты No. 131, Программы поддержки ведущих научных школ (грант НШ-931.2008.9).

Аннотация: В ИВТ СО РАН поддерживается комплекс глубокой обработки спутниковых данных, принимаемых в режиме прямой передачи. При модернизации комплекса для выполнения вычислений на кластере в параллельном режиме основным фактором ограничивающим производительность стала централизованная система хранения данных. В работе описывается оптимизация комплекса, снижающая нагрузку на СХД за счет использования локальных ресурсов каждого узла кластера.

Одной из задач центра мониторинга социально-экономических процессов и природной среды ИВТ СО РАН является обработка спутниковых данных, принимаемых в НИЦ "Планета" в режиме прямой передачи со спутников Terra и Aqua [1]. Суть комплекса обработки заключается в запуске цепочек программ-обработчиков, предоставляемых NASA. Каждая из программ выполняет генерацию одного продукта, например маски облачности или индекса вегетации, и в качестве исходных данных принимает продукты более низких уровней. В связи с наращиванием числа поддерживаемых продуктов и увеличением объема исходных данных (с 2012 года мы стали получать данные с приёмного комплекса в Хабаровске) возникла необходимость наращивания вычислительных мощностей для обработки.

Для ускорения цикла обработки, мы реализуем этот процесс на вычислительном кластере ИВТ в распределенном параллельном режиме. Это возможно благодаря тому, что процесс обработки обладает существенным параллелизмом. С одной стороны исходные данные разбиваются на независимые фрагменты (витки спутника, 5-ти минутные гранулы витка, квадраты координатной сетки), а с другой стороны многие программы обработчики могут быть запущены параллельно, так как не зависят от результатов друг друга.

Эксплуатация этого параллелизма возможна с применением потоковой модели вычислений. Сценарий обработки описывается в виде графа потока данных. Вершинами этого графа являются программы обработчики, а дуги описывают зависимости между ними. Этот подход реализован с использованием системы управления процессами Taverna [2]. Этот выбор обусловлен с одной стороны моделью вычислений, хорошо подходящей для описания процесса обработки спутниковых снимков, а с другой стороны богатыми возможностями для расширения её функционала через подключаемые модули. Алгоритм в Taverna

представляется в виде набора «процессоров» - черных ящиков с некоторым количеством входных и выходных портов. Порты процессоров связаны друг с другом и образуют ациклический граф. Этот граф описывает передачу параметров между процессорами и определяет ограничения на порядок запуска процессоров.

Для того, чтобы воспользоваться этими возможностями для Taverna был написан плагин, позволяющий использовать в качестве её процессоров Bash-скрипты (https://github.com/kikht/ict-taverna-modules/wiki/Bash-activity-plugin). В настоящее время поддерживаются выполнение скриптов на локальной машине, удаленной с доступом по SSH, а также системы управления очередями кластера Torque и Slurm. Единственным требованием является наличие общей файловой системы для всех узлов, участвующих в обработке.

Анализ производительности новой реализации показал, что использование общей сетевой СХД существенно ограничивает масштабируемость комплекса. Увеличение числа узлов не приводит к уменьшению времени обработки. Причинами такого поведения являются несколько факторов:

- Большой объем продуктов, вплоть до нескольких гигабайт на один снимок. Передача таких объемов данных по сети занимает значительное время.
- Из-за большого объема продукты не помещаются в кэше файловой системы в оперативной памяти.
- Кэш сетевых файловых систем не очень эффективен из-за необходимости его синхронизации между узлами.
- Случайный характер доступа при генерации и чтении продуктов. В сочетании с малой эффективностью кэша это приводит к большому числу мелких обращений к СХД. Это плохо как с точки зрения сетевых накладных расходов, так и с точки зрения режима работы дисковых массивов.

Для преодоления этой неэффективности сетевых файловых систем, мы стараемся максимально полно использовать возможности каждого узла по отдельности, в частности его локальные диски. Время доступа к ним значительно меньше времени обращения к сетевой СХД. Кроме того, для них возможен агрессивный кэш в локальной памяти, так как нет необходимости в его синхронизации между узлами.

Этот подход реализуется следующим образом, программы-обработчики запускаются на локальном диске, а затем результаты асинхронно копируются на сетевое хранилище. При запуске каждого модуля, мы стараемся выбрать узел для него так, чтобы максимальное количество данных лежало у него на локальном диске. В результате достигаются сразу несколько положительных эффектов. Во-первых, продукты копируются на СХД последовательно и целиком. Это наилучший режим работы для практически любой системы хранения. Во-вторых, по возможности, чтение данных производится с быстрого локального диска. И в-третьих, на локальном диске нужно хранить только данные участвующие в текущем процессе обработки, а не весь архив данных.

Все вместе это позволяет обрабатывать объемы данных возможные только для СХД с производительностью локальных файловых систем. С учетом этой оптимизации удалось добиться линейной масштабируемости комплекса.

Литература

1. **Ю.И. Шокин, Н.Н. Добрецов, В.В. Смирнов, А.А. Лагутин, В.Н. Антонов, А.В. Калашников** Система информационной поддержки задач оперативного мониторинга на основе данных дистанционного зондирования // Тезисы докладов Восьмой открытой Всероссийской конференции "Современные проблемы дистанционного зондирования земли из космоса" (Москва, 15 - 19 ноября 2010 г.). М.: ИКИ РАН, 2010. – С. 40-41.
2. **Duncan Hull и др.** Taverna: a tool for building and running workflows of services. // Nucleic Acids Research, vol. 34, 2006.