Система параллельной обработки спутниковых данных

Кихтенко Владимир Андреевич
e-mail: kva911@gmail.com

С 2009 года в ИВТ СО РАН ведется прием и обработка данных дистанционного зондирования Земли с сенсоров MODIS установленных на спутниках Terra и Aqua [1][2]. В настоящее время довольно остро стоит проблема наращивания производительности комплекса. Это вызвано увеличением объема входных данных и необходимостью в более глубокой обработке. Для ускорения цикла обработки предлагается реализовать процесс вычислений на кластере, получив при этом возможность конфигурирования потока вычислений.

Обработка данных производится отдельными независимыми вычислительными модулями [3]. Каждый из них принимает на вход параметры запуска и набор файлов с исходными данными («продуктами») и по-рождает на выходе другой набор продуктов. Для распараллеливания предлагается использовать модуль-ность алгоритма обработки. Отдельные обработчики могут быть запущены одновременно, если для них готовы все необходимые входные продукты. Кроме того, они обрабатывают не весь снимок целиком, а тре-буют его разбиения на гранулы по 5-ти минутной сетке. Это позволяет обрабатывать различные гранулы параллельно. Основные проблемы при реализации такого распараллеливания это: специфицирование алго-ритма обработки в виде, позволяющем выделить возможности для распараллеливания, и реализация его выполнения в распределенной среде кластера.

В качестве управляющего ядра используется система Taverna [4]. В ней алгоритм обработки представляется в виде набора «процессоров»с некоторым количеством входов и выходов, связанных зависимостями по данным. В процессе интерпретации алгоритма процессоры активируются, как только у них появляются данные на всех входных портах. Как только процессор отрабатывает, значения выходных портов передаются по дугам графа на вход другим процессорам, которые в свою очередь активируются. Также, в Taverna присутствует поддержка списковых портов у процессоров. Можно указать, что некоторый выходной порт процессора возвращает список значений, и подключить его к входному порту, принимающему только одно значение. Это приведет к параллельному или последовательному (по выбору разработчика) исполнению копий этого процессора с входными данными, соответствующими каждому элементу списка. Если же список поступает на вход целой иерархии процессоров, то отдельные элементы движутся по графу независимо по принципу конвейера – как только копия процессора получает свои данные, она немедленно исполняется. Возможна и обратная ситуация, порт, порождающий одиночные значения можно связать с портом, принимающим список. В этой ситуации интерпретатор дождется завершения выполнения всех копий первого процессора, соберет все результаты в список и передаст второму. Поддерживается настраиваемая итерация по нескольким спискам сразу. Подробнее модель представления алгоритмов и её семантика описаны в [5].

Описанная модель представления и исполнения алгоритмов очень хорошо подходит для задачи обра-ботки спутниковых снимков. В этой модели вычислительные модули-обработчики напрямую проецируются на процессоры, входные и выходные продукты на соответствующие порты, а разбивка на гранулы соответствует работе со списками данных. При исполнении алгоритма, описанного в таком виде, интерпретатор Taverna автоматически распараллелит его исполнение, учитывая как модульную структуру алгоритма, так и нарезку исходных данных на гранулы. Таким образом, при наличии достаточного количества узлов время обработки сокращается до времени обсчета критического пути в графе обработчиков. Сама по себе Taverna не имеет средств управления кластером и запуска на нем внешних программ, но предоставляет большие возможности для расширения собственного функционала через систему плагинов [6].

Выполнение задач на кластерах обычно управляется специализированным менеджером ресурсов, таким как SLURM [7], Torque/PBS [8] или GRAM (из Globus toolkit) [9]. Общая схема работы с этими менеджерами такова: клиент делает запрос на выполнение некоторой программы на определенном объеме машинных ресурсов (например, количестве процессоров), а менеджер в соответствии со своими политиками ставит поступающие задачи в очередь и при освобождении ресурсов исполняет их. Для интеграции этих менеджеров с Taverna была написана java-библиотека Executor API, которая подключается к ней как плагин. Она позволяет при активации процессора сформировать на основе входных данных скрипт для исполнения на узле и поставить его в очередь задач кластера, а после его завершения передать результаты через выходные порты процессора. Разработанная библиотека предоставляет абстрактный API, не зависящий от конкретного менеджера ресурсов, что позволяет использовать весь комплекс на кластерах различной конфигурации. В настоящее время реализована поддержка менеджера SLURM, а также удаленный доступ к кластеру по SSH.

Предложенный подход позволяет адаптировать вычислительный комплекс не только к одному подкон-трольному кластеру, но и к более крупным архитектурам, построенным по технологии GRID [10]. Это воз-можно благодаря тому, что управляющее ядро полностью отделено от вычислительной части и оперирует лишь указателями на файлы с данными. Можно расширить формат указателей (сейчас это просто положение файла на общей для кластера файловой системе) и добавить автоматическую загрузку входных данных на целевой узел. Основной проблемой при таком расширении оказывается распространение вычислительных модулей между системами участвующими в вычислении.

Разработанная система находится на стадии внедрения. Результаты тестов показывают уменьшение времени обработки до 3-4 раз. В то же время существенно увеличивается нагрузка на систему хранения, и именно она становится узким местом в работе комплекса.

Список литературы

1. **Шокин Ю.И., Пестунов И.А., Смирнов В.В.** *Корпоративная информационная система СО РАН для сбора, хранения и обработки спутниковых и наземных данных* // Труды X Всероссийской конференции с участием иностранных ученых "Проблемы мониторинга окружающей среды (ЕМ-2009)". Кемерово, 27-30 октября 2009. Горный информационно-аналитический бюллетень. 2009. Вып. 2, т. 1 (т. 2).

2. **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 конферен-ции "Современные проблемы дистанционного зондирования земли из космоса"(Москва, 15 - 19 ноября 2010 г.). М.: ИКИ РАН, 2010. – С. 40-41.

3. **Лагутин А.А., Никулин Ю.А., Жуков А.П., Резников А.Н., Синицин В.В., Шмаков И.А.** *Математические технологии оперативного регионального спутникового мониторинга характеристик атмосферы и подстилающей поверхности ч. 1. MODIS* // Вычислительные технологии. – 2007. – Т. 12. – № 2.
4. **Duncan Hull и др.** *Taverna: a tool for building and running workflows of services.* // Nucleic Acids Research, vol. 34, 2006.
5. **Jacek Sroka, Jan Hidders, Paolo Missier, и Carole Goble** *Formal semantics for the Taverna 2 work?ow model.* // Journal of Computer and System Sciences, 2009.
6. **Paolo Missier и др.** *Taverna, reloaded.* // Scientific and Statistical Database Management, Lecture Notes in Computer Science, 2010.
7. **Morris Jette и Mark Grondona***SLURM: Simple Linux Utility for Resource Management* // Pro-ceedings of ClusterWorld Conference and Expo, San Jose, California, 2003.
8. **Garrick Staples** *TORQUE resource manager* // SC 39;06 Proceedings of the 2006 ACM/IEEE confer-ence on Supercomputing.
9. **Martin Feller, Ian Foster и Stuart Martin***GT4 GRAM: A functionality and performance study* // Teragrid 2007 conference, Madison.
10. **Ian Foster, Carl Kesselman и Steven Tuecke***The anatomy of the grid: Enabling scalable virtual organizations.* // International J. Supercomputer Applications, 15(3), 2001.