РАСПРЕДЕЛЕННАЯ СИСТЕМА ХРАНЕНИЯ ДАННЫХ НА ОСНОВЕ DCACHE

Е.Ю. Куклин

Институт Математики и Механики УрО РАН, Екатеринбург, Россия

1. Введение. В Уральском отделении Российской академии наук ведутся работы по созданию распределенной среды высокопроизводительных вычислений на основе ГРИД-технологий [1]. Одним из главных компонентов такой среды является распределенная система хранения данных (РСХД), которая предусматривает объединение систем хранения в регионах присутствия УрО РАН. К РСХД подключаются экспериментальные установки институтов УрО РАН, вычислительные кластеры и суперкомпьютеры. Участники данного проекта системы хранения — это Институт Математики и Механики г. Екатеринбурга и Институт Механики Сплошных Сред г. Пермь.

Представлен подход к созданию территориально распределенной системы хранения данных на основе промежуточного программного обеспечения dCache [2] из проекта European Middleware Initiative [3]. dCache ориентирован на большие объёмы экспериментальной информации. Он позволяет объединять различные типы накопителей для построения хранилища в несколько сотен терабайт, при этом все файлы в нем логически упорядочены в одно дерево; для хранения метаинформации используется база данных. Также, dCache предоставляет возможность простого расширения посредством добавления новых узлов и может работать как вместе с ленточными библиотеками, так и без них. dCache разрабатывается с 2000 года и является результатом совместной работы DESY (Hamburg) и FermiLab (Chicago).

2. Выбор ППО. EMI включает три варианта программного обеспечения для создания распределенных систем хранения: dCache, Disk Pool Manager и Storage Resource Manager. Для создания РСХД УрО РАН был выбран dCache, так как он обеспечивает поддержку ГРИД и Интернет протоколов, имеет качественную документацию и активное сообщество, а также удобен в установке и администрировании. Дополнительной причиной выбора dCache является то, что на его основе построен единственный распределенный центр Tier1 WLCG, узлы хранения которого расположены в разных Скандинавских странах [4]. После проведенного тестирования файловых систем, данные на узлах было решено хранить в ФС XFS [5] компании Silicon Graphics. Она показала хорошие результаты в тестах на запись/чтение/доступ к данным, и так же, как EXT4, является встроенной в Linux активно развивающейся файловой системой.

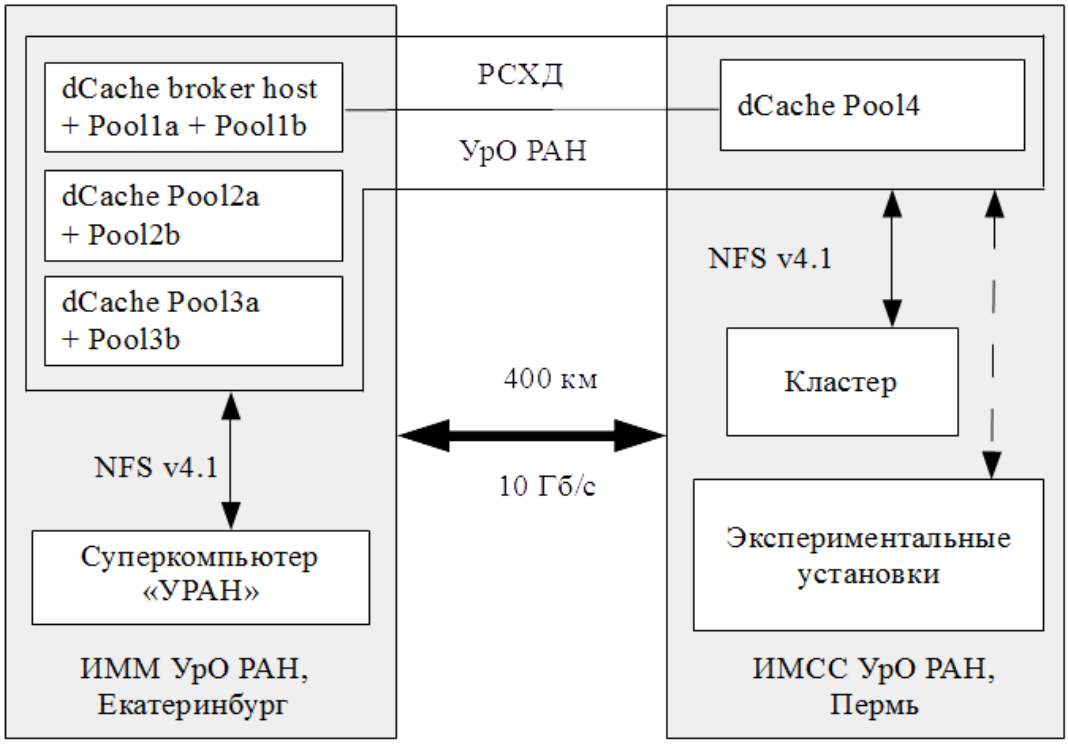

Рис. 1. Схема реализации первой очереди РСХД УрО РАН

3. Текущая реализация. В настоящее время выполнена первая очередь реализации РСХД на четырех серверах Supermicro с ОС Scientific Linux 6.5 и dCache 2.6 из репозитория EMI. Все сервера аппаратно идентичны; полезная емкость РСХД составляет 210 ТБ. Узлы хранения расположены в двух вычислительных

центрах: в Институте Математики и Механики УрО РАН г. Екатеринбурга (3 шт.) и в Институте Механики Сплошных Сред УрО РАН г. Пермь. Схема реализации представлена на рисунке 1. Вычислительные центры находятся на расстоянии 400 км друг от друга и соединены выделенным каналом связи. Работоспособность канала обеспечивает DWDM оборудование компании ECI-Telecom; комплектация установленных платформ позволяет передачу двух λ-каналов по 10 Гбит/с. Подключение к РСХД выполнено по протоколу NFS версии 4.1 с использованием Parallel NFS, полная поддержка которого появилась в последней версии Scientific Linux 6.5 Carbon.

ИММ УрО РАН имеет в своем распоряжении суперкомпьютер «УРАН», занимающий 7 место в рейтинге Топ50 по СНГ; в ИМСС УрО РАН также есть свой кластер. Согласно идеологии dCache, один из серверов РСХД выступает в качестве посредника: к его хранилищу монтируются кластеры и удаленные каталоги пользователей, и через него же они связываются с другими узлами хранения. Тестирование показало, что узким местом являлась гигабитная сеть РСХД — суперкомпьютер «УРАН». Новое 10-гигабитное оборудование для внутренней сети ИММ УрО РАН позволяет это исправить и значительно увеличить скорость обмена данными. Тем не менее, пока не удается преодолеть порог в 400 мб/с при передаче данных между вычислительными центрами. Эффективность работы установленных в узлах хранения сетевых карт также оставляет желать лучшего, несмотря на проведенные оптимизации TCP стеков ОС. Самый эффективный алгоритм (bic) показал суммарную скорость передачи данных лишь 6 Гбит/с.

dCache позволяет создавать несколько копий файлов, что выравнивает нагрузку на узлы, повышает надежность и оптимизирует использование пространства хранилища. Так, репликация даст возможность узлу в ИМСС хранить локальную копию необходимых данных, уменьшая время на доступ к ним и снижая нагрузку на выделенный канал между вычислительными центрами. С другой стороны, разбрасывание данных по разным серверам позволяет избежать перегрузки одного узла. Ленточных библиотек для архивного хранения данных в институтах УрО РАН не имеется.

4. Планы на будущее. Сейчас в ИММ УрО РАН проводится тестовая эксплуатация рассмотренной РСХД на предмет производительности и отказоустойчивости. Позже планируется уделить внимание безопасности и расширить возможности мониторинга системы. Следующим этапом реализации будет подключение экспериментальных установок в ИМСС УрО РАН и УрФУ. Данные от этих установок будут записываться на РСХД и обрабатываться удаленно на суперкомпьютере «УРАН», в том числе с возможностью обработки в режиме реального времени и управлением ходом экспериментов. В дальнейшем планируется подключение РСХД к международной ГРИД-инфраструктура WLCG [6].

Проект поддержан грантами УрО РАН 12-П-1-2012 и РФФИ 14-07-96001-р_урал_а.

ЛИТЕРАТУРА:

1. М.Л. Гольдштейн, А.В. Созыкин, Г.Ф. Масич, А.Г. Масич. "Вычислительные ресурсы УрО РАН. Состояния и перспективы" // Параллельные вычислительные технологии (ПаВТ'2013): труды международной науч. конференции. Челябинск, издат. центр ЮУрГУ, 2013. С. 330-337.
2. dCache: http://www.dcache.org.
3. European Middleware Initiative | EMI: http://www.eu-emi.eu.
4. G. Behrman, P. Fuhrmann, M. Gronager and J. Kleist. "A distributed storage system with dCache" // Journal of Physics: Conference Series 119. 2008.
5. XFS: http://oss.sgi.com/projects/xfs.
6. Worldwide LHC Computing Grid | WLCG: http://wlcg.web.cern.ch.