Опыт эксплуатации центра обработки данных и вычислительного кластера в лаборатории эволюционной геномики

Арифулов Р.Н.
ФББ МГУ, РХТУ
arifulovrenat@gmail.com

Науменко С.А.
ИППИ РАН, ФББ МГУ, РХТУ
sergey.naumenko@yahoo.com

Аннотация

В лаборатории эволюционной геномики для обработки данных и проведения расчетов используется кластерная вычислительная система. За полтора года её работы приняты данные от 19 запусков секвенатора Illumina Hiseq 2000 общим объемом около 18T, произведена сборка десятков геномов и транскриптомов, обработано более 50000 вычислительных задач, работает более 50 пользователей. Основные трудности связаны с обработкой потоков данных, объем которых увеличивается в результате доступа новых групп исследователей к возможностям высокопроизводительного секвенирования. Предлагается решение по оптимальному управлению потоками данных, состоящее из использования распределенной файловой системы lustre для основной массы расчетов на кластере, и выделения специальных томов, доступных по протоколу fiber channel для самых крупных проектов.

1. Введение

В октябре 2011 года в лаборатории эволюционной геномики ФББ МГУ был запущен вычислительный кластер и центр обработки данных высокопроизводительного секвенирования [1].

Потребности в дисковой памяти и к полосе пропускания к ней постоянно растут, особенно в связи с появлением проектов обработки транскриптомов, в которых данные проекта могут занимать 10-20T.

2. Кластерные файловые системы

Тестирование кластерных файловых систем OCFS2 [2], GFS2 [3] под нагрузкой показало, что файловая система блокируется с вычислительного сервера так же, как бы блокировалась локальная файловая система, что затрудняет просмотр и редактирование файлов с других серверов.

Данные файловые системы не могут быть использованы для создания томов, с которых происходит обработка данных, они подходят для быстрого и редкого доступа к большим объемам данных (например, для получения чтений из рид-архива, или для обмена данными на стадиях обработки: передача данных после анализа качества на сборку генома, или после сборки генома для аннотирования).

3. Сетевые и распределенные файловые системы

В отличие от OCFS2, GFS2, сетевая файловая система NFS [4] и распределенная файловая система lustre [5] работают путем передачи запросов и получения частей файлов по сети, поэтому даже при большой нагрузке с одного или нескольких вычислительных серверов (которая естественно ограничивается пропускной способностью сетевого интерфейса сервера в 100 МБ/с).

С другой стороны, скорость обработки потоков данных ограничивается способностью системы хранения читать и записывать файлы. В нашем случае эта скорость равна 500MB/с в случае одного файл-сервера с доступом по NFS и

1GB/с на распределенной файловой системе lustre с участием двух хранилищ.

4. Оптимальное управление потоками данных

Среди десятков проектов, обработка данных по которым идет одновременно, обычно выделяются 1-3 проекта, которые порождают наибольшие потоки данных (объем всех данных проекта составляет несколько терабайт, потоки данных сотни мегабайт в секунду), иногда сравнимые по величине со всеми остальными проектами, вместе взятыми (это сборка больших эукариотических геномов, проекты с десятками с сотнями транскриптомов).

К сожалению, существующие планировщики задач, насколько известно авторам, не позволяют осуществлять эффективное планирование ресурсов ввода-вывода. Исторически задачи планирования решались по отношению к ресурсам процессора и памяти. Возможно тестировать в этих целях нестандартное применение прокси серверов, которые используются для управления сетевым траффиком.

Выход из данной ситуации видится в разделении идущих проектов по величине порождаемых потоков данных на большие (сотни МБ/с), средние (десятки МБ/с) и маленькие (1МБ/с) и выделении для больших проектов отдельных локальных дисков и серверов, динамически выделяя и подключая их на время проекта по протоколу FiberChannel (рис.1). Для средних и маленьких потоков следует использовать распределенную файловую систему lustre, которая позволяет справедливо распределить ресурсы файл-серверов по всем вычислительным серверам.

5. Направления развития центра обработки данных

Необходим еще один узел с большим объемом оперативной памяти (1Т).

В случае увеличения количества масштабных проектов, порождающих большие потоки данных, например, экзомного секвенирования для медицинских исследований следует добавить хранилище, способное экспортировать диски по протоколу fiber channel.

В случае роста количества проектов, порождающих средние и маленькие потоки данных следует добавлять новые хранилища для распределенной файловой системы lustre.

6. Благодарности

Работа осуществлялась при поддержке Министерства образования и науки РФ (гранты 11.G34.31.0008 and 8814) и РФФИ (грант 12-07-31261).

Литература

[1] Арифулов Р.Н., Науменко С.А. Опыт создания центра обработки данных и вычислительного кластера для лаборатории эволюционной геномики. //Сборник трудов конференции «Информационные технологии и системы» (ИТиС'12). Петрозаводск, 19-25 августа 2012г. ISBN 978-5-901158-19-7.

[2]https://oss.oracle.com/projects/ocfs2/(Project: OCFS2)

[3]https://access.redhat.com/site/documentation/ru-RU/Red_Hat_Enterprise_Linux/6/html/Global_File_System_2 (Red Hat GFS 2)

[4]http://nfs.sourceforge.net/

[5]http://www.whamcloud.com/lustre/

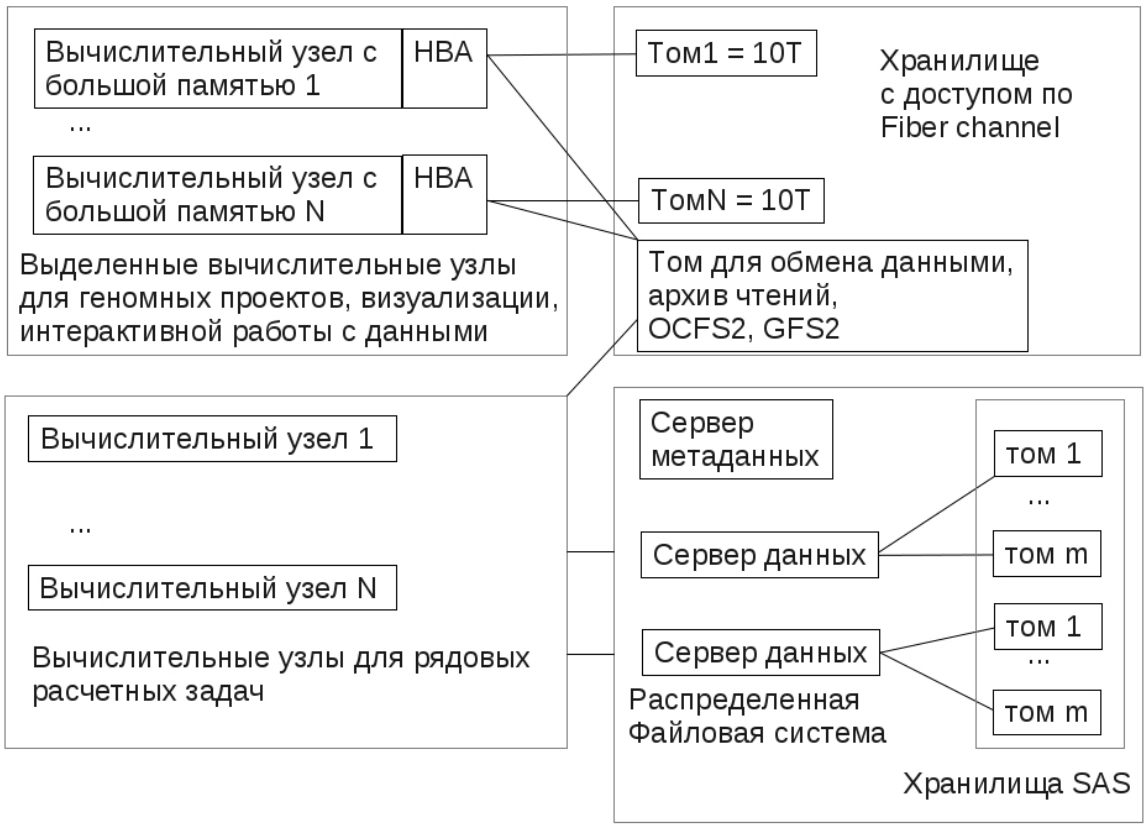

Рис.1. Выделение дисков для крупных потоков данных и общего распределенного хранилища для средних и небольших потоков.