Опыт создания центра обработки данных и вычислительного кластера для лаборатории эволюционной геномики

Арифулов Р.Н.
РХТУ им. Д.И. Менделеева
arifulovrenat@gmail.com

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

Аннотация

В рамках проекта создания лаборатории эволюционной геномики на ФББ МГУ возникла подзадача построения вычислительной системы, которую можно использовать для обработки данных, поступающих с высокопроизводительного секвенатора Illumina Hiseq 2000. Созданный центр обработки данных состоит из 30 вычислительных узлов общего назначения, одного узла с большим объемом оперативной памяти (512Gb), трех систем хранения данных по 144TB каждая, сетей обмена данными, вычислительной и служебной сети, графического ускорителя вычислений, кондиционеров и источника бесперебойного питания.

1. Введение

Перед создаваемой вычислительной системой были поставлены следующие требования: достаточный объем дискового пространства и ширина канала связи для приема данных высокопроизводительного секвенирования, достаточный объем оперативной памяти для задач, связанных со сборкой геномов, достаточное количество вычислительных ядер для решения задач аннотирования и анализа геномов.

2. Прием данных

Высокопроизводительный секвенатор Illumina Hiseq 2000 [1] в ходе секвенирования производит большое количество изображений, которые по окончании прогона в результате процедуры определения нуклеотидов (base calling) преобразуются в чтения в формате fastq, после чего исходные изображения могут быть удалены. В ходе прогона изображения должны передаваться по сети в систему хранения данных. Объем информации, которую необходимо хранить в ходе прогона оценивался представителями поставщика секвенатора в 50TB (как показали реальные прогоны, этот объем не превышает 5TB).

Анализ пропускной способности гигабитной сети ethernet показал, что передача 50T данных за время прогона (10 дней) практически полностью загружает этот канал, не оставляя в будущем возможности увеличения пропускной способности. Поэтому от управляющего компьютера Illumina до системы хранения было проложено 4 кабеля для обеспечения избыточности и надежности. В будущем при возникновении необходимости 4 канала связи могут быть объединены в один с суммарной пропускной способностью до 400 MB/s (теоретический максимум).

Для хранения данных были выбраны системы хранения, состоящие из двух параллельно работающих RAID-контроллеров и трех дисковых полок, несущих 2TB диски SATA2. Суммарный объем дисков трех таких систем хранения составил 432TB. Две системы хранения соединяются с серверами посредством интерфейса SAS, третья входит в сеть хранения данных (рис. 1).

3. Узел для сборки геномов

Задача сборки геномов из коротких чтений требует большого объема оперативной памяти, поскольку лидирующие на сегодняшний день

программы [2,3,4] используют алгоритм с графовым представлением всего генома и всех чтений. При выборе конфигурации ориентиром послужил узел, который использовался для *de novo* сборки геномов человека и мыши ассемблером Allpaths-LG [2]: четырехпроцессорный сервер, 48 ядер на общей памяти 512GB.

4. Сеть хранения данных (SAN)

При доступе к данным посредством сети 1Gb ethernet по протоколу NFS скорость передачи данных составляет не более 100 MB/s. При работе с большими объемами данных эта скорость недостаточна. Для узла сборки геномов и части вычислительных серверов была организована сеть хранения данных на основе протокола Fiber Channel 4Gb/s (400MB/s). В будущем при обновлении системы хранения можно перейти на использование протокола FC 8Gb/s без замены FC адаптеров и FC коммутатора.

Для использования общего дискового пространства рядом серверов необходима организация симметричной кластерной файловой системы. Используется кластерная файловая система GFS2 [5].

5. Сеть обмена данными

Помимо обеспечения высокоскоростного доступа к данным для малого количества вычислительных узлов необходимо обеспечить гарантированную (неснижаемую) скорость доступа к данным для большого количества расчетных процессов.

Для решения этой задачи была построена сеть обмена данными (рис. 1) в составе двух систем хранения, двух серверов данных и одного сервера метаданных. Серверы данных и метаданных обмениваются информацией на скорости 10Gb/s, расчетные серверы передают информацию по каналу 1Gb/s.

В этой сети предполагается использовать распределенную файловую систему lustre [6], которая физически распределяет файлы по обоим системам хранения таким образом, что вычислительные серверы загружают каналы связи от серверов данных до систем хранения равномерно.

В настоящее время из-за необходимости как можно скорее ввести вычислительную систему в строй, неопределенности статуса поддержки файловой системы lustre (в начале 2011 года фирма Oracle прекратила разработку и передала полномочия сообществу разработчиков) и сложности её настройки и тестирования, сеть обмена данными функционирует по протоколу NFS. Вторая система хранения используется для резервного копирования. С увеличением количества пользователей системы развертывание lustre остается приоритетной задачей. За 2011 год свободное сообщество разработчиков сформировалось и в конце первого квартала 2012 года готовится релиз lustre 2.2 для систем класса RHEL 6 [6].

6. Вычислительные узлы

Для решения задач эволюционной геномики, аннотации геномов, картирования чтений вычислительная система содержит 30 двухпроцессорных узлов (24 ядра в режиме hyperthreading, 48GB RAM).

10 вычислительных узлов входят в сеть хранения данных.

Все вычислительные узлы подключены к сети обмена данными, вычислительной сети и служебной сети.

Узлы 31 и 32 подключены к графическому ускорителю вычислений.

7. Программное обеспечение

Используется свободная операционная система Scientific Linux 6.1 [7], которая является клоном промышленной системы Red Hat Enterprise Linux, выпускаемым силами научного сообщества (главным образом сотрудниками Cern и Fermilab).

Для управления конфигурациями серверов используется puppet [8], для управления вычислительными заданиями – torque [9], для мониторинга – nagios [10].

8. Выводы

При создании сложного вычислительного комплекса силами научной лаборатории естественно делегировать эту задачу крупной компании, для которой этот проект стал бы имиджевым. К сожалению, в нашем случае это оказалось невозможным.

У создаваемой системы не было прототипа, требования к ней были трудно формализуемыми, их необходимо было разрабатывать в ходе итеративного процесса проектирования. Крупные

вендоры предлагали типовые решения из области научных вычислений, в которых упор делался на процессорную мощь (большое количество вычислительных ядер, вычислительная сеть с большой пропускной способностью). Оказалось, что крупные вендоры не обладают отделами, осуществляющими проектирование в нашем случае (поскольку это небольшая система для крупной компании), а pre-sale специалисты такого проектирования не предоставляют.

После сравнения различных предложений мы остановились на сравнительно небольшой, но известной своей надежностью компании, которая смогла вложить значительные силы в совместную работу по проектированию системы.

В нашем случае система представляет скорее центр обработки данных, наподобие тех, что используются в киноиндустрии, банковской деятельности, обработке геологических данных, а вычислительный аспект имеет вторичный приоритет.

Не имея опыта сборки геномов, трудно было представить объем оперативной памяти, который нужен для задач такого рода. В первых версиях проекта мы планировали 2 узла с памятью по 196GB каждый. Однако выяснилось, что для работы с геномами эукариот необходимо не менее 512GB. В настоящее время на этом узле обработаны геномы длиной 300-400MB и 1.1 GB. При обновлении системы в будущем возможно расширение памяти узла до 768GB.

Больших усилий потребовала организация работ, не связанных непосредственно с вычислительной системой: выбор помещения, проведение дополнительных линий электропитания, ремонт помещения, установка кондиционеров.

Мы использовали источник бесперебойного питаний на 40KVA, который способен запитать от батарей всю систему в течение 10-15 минут, хотя зачастую ИБП используют только для критичных узлов. Этот подход себя оправдал, т.к. за полгода эксплуатации мы пережили несколько кратковременных отключений питания, в том числе во время работы секвенатора Illumina.

Литература

[1] http://www.illumina.com/systems/hiseq_systems.ilmn

[2] S. Gnerre et al., *High-quality draft assemblies of mammalian genomes from massively parallel sequence data*, Proc. Natl. Acad. Sci. USA (2011), 108(4), 1513-1518.

[3] D.R. Zerbino, E. Birney, *Velvet: Algorithms for de novo short read assembly using de Bruijn graphs*, Genome Research (2008), 18: 821-829.

[4] Li et al., *De novo assembly of human genomes with massively parallel short read sequencing*, Genome Res (2010),20 (2), 265-72.

[5] S. Whitehouse, *The GFS2 Filesystem*, Proceedings of the Linux Symposium (2007), Ottawa, Canada, 253-259.

[6] http://wiki.whamcloud.com/display/PUB/Lustre+2.2

[7] http://www.scientificlinux.org/

[8] http://puppetlabs.com/

[9] http://www.adaptivecomputing.com/products/open-source/torque/

[10] http://www.nagios.org/

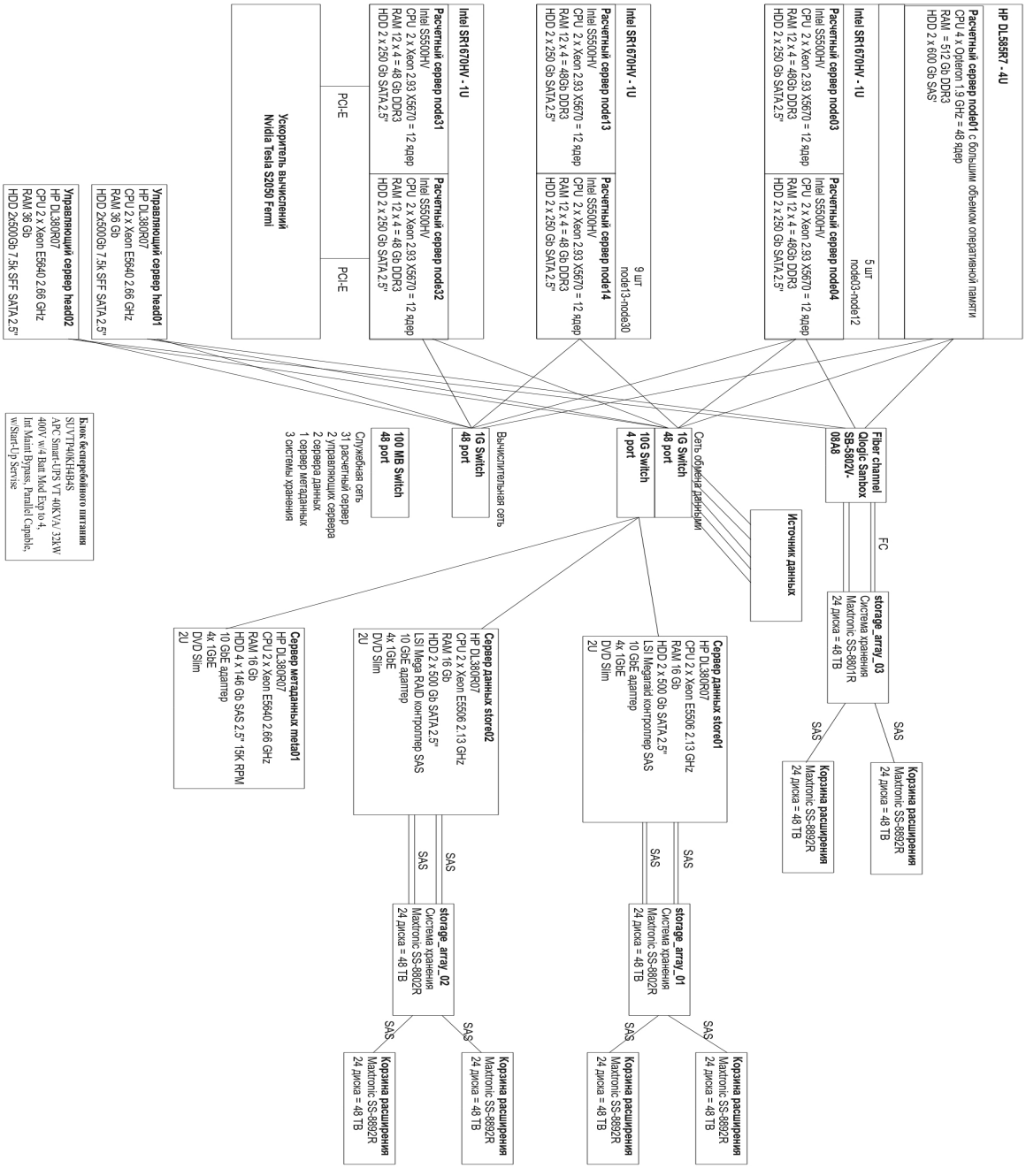

Рис. 1. Схема вычислительной системы