Использование сетей хранения данных и кластерных файловых систем в центре обработки геномных данных

Науменко С.А.[1,2,3], Арифулов Р.Н.[3]

[1] *Институт проблем передачи информации РАН, Большой Каретный переулок 19 стр.1, Москва, Российская Федерация*
[2] *Факультет биоинженерии и биоинформатики, Московский государственный университет Ленинские горы МГУ 1, стр. 73, Москва, Российская Федерация*
[3] *Факультет информационных технологий и управления, Российский химико-технологический университет, Миусская пл. 9, Москва, Российская Федерация*

naumenko@iitp.ru, arifulovrenat@gmail.com

Abstract. *Modern biomedical research requires sequencing which constantly produces terabytes of data to be processed. This advances the requirements to datacenter hardware and software. In our datacenter we use storage systems having 400TB of total capacity, computing servers and SAN infrastructure. The problem we faced was the choice of an open source clustered filesystem among RHEL GFS2 and Oracle OCFS2 (both released under GPL). In the report I will provide the general introduction to the field of genomic data analysis and bioinformatics, the description of our computer system's architecture, and, finally, our experience with GFS2/OCFS2. Our tests of performance and stability have shown that OCFS2 significanlty outperforms GFS2.*

Key words

Next generation sequencing, datacenter, storage area network, clustered filesystem

Современная биомедицинская наука не обходится без секвенирования (прочтения последовательностей) полных геномов, транскриптомов и другой геномной информации. Высокопроизводительные секвенаторы способны прочитать за две недели несколько полных человеческих геномов. В процессе работы секвенатор генерирует до 5TB изображений с детекторов, которые по окончании процесса преобразуются в десятки миллионов строковых последовательностей ДНК – коротких чтений длиной 100-150 символов.

Полученная информация проходит множество стадий обработки до того, как она может быть использована в научном исследовании. В случае геномного проекта по организму, геном которого ранее не был прочитан, это следующие этапы: оценка качества чтений, фильтрация чтений по качеству и по длине, сборка протяженных областей (контигов) из коротких чтений, соединение контигов в лестницы (скаффолды) при помощи парных чтений, заполнение пробелов между контигами, верификация сборки, аннотация генома.

Каждая из стадий требует большого объема вычислений, операций ввода-вывода и свободного дискового пространства. При наличии нескольких геномных проектов в лаборатории это поднимает планку требований к архитектуре компьютерной системы: серверы должны одновременно иметь доступ к большому объему дискового пространства на большой скорости.

Традиционная организация вычислительной системы в научном учреждении (вычислительные серверы – файл-сервер с хранилищем, с доступом по NFS4) с трудом выдерживает нагрузки при обработке геномных данных. Узкие места возникают как на клиентских серверах (несколько расчетных процессов с интенсивным вводом-выводом, попав на один сервер, исчерпывают пропускную способность сетевого интерфейса, что резко снижает среднее время расчета), так и на файл-сервере: при возрастании потока запросов на операции ввода-вывода с разных клиентских серверов файл-сервер переходит в режим высокой загрузки, и время отклика резко сокращается, что также снижает среднее время расчета.

Выход из данной ситуации лежит в использовании систем хранения данных, а не серверов с дисками, инфраструктуры SAN (Storage area network – сеть хранения данных), промышленных дистрибутивов OS Linux, открытых кластерных и распределенных файловых систем.

SAN позволяет серверам одновременно монтировать один и тот же дисковый ресурс по блочному протоколу SCSI. Аппаратно возможны 2 варианта: организация сети Fiber Channel (FC), либо использование IP сети для передачи данных. Сеть FC выигрывает по скорости – возможно бюджетное построение сети на скорости 4Gb/s, 8Gb/s, в то время как типичная IP сеть работает на скорости 1Gb/s. В случае IP сети блочный протокол реализуется либо при помощи FCoE (Fiber Channel over Ethernet), либо протокола iSCSI. Также необходимо, чтобы система хранения поддерживала соответствующий протокол.

В случае SAN каждый сервер монтирует дисковый том по блочному протоколу таким же образом, как и собственный локальный диск, поэтому использование обычной файловой системы XFS или EXT4 невозможно, поскольку приводит к конфликтам и потере данных при одновременном доступе. Поэтому необходимо объединение серверов в кластер и развертывание кластерной файловой системы, которая обладает механизмом параллельного доступа и соответствующими блокировками, предотвращающими потерю данных.

По лицензии GPL доступны файловые системы GFS2 (Global File System) от RedHat и OCFS2 (Oracle Cluster File System) от Oracle. GFS2 доступна в дистрибутивах RHEL6, CentOS6, Scientific Linux6. OCFS2 доступна в дистрибутивах, производных от RHEL5 (начиная со RHEL6 Red Hat исключила код OCFS2 из ядра), и в дистрибутиве Oracle Linux 6 с Unbreakable Linux Kernel (модифицированным ядром, в которое включен код OCFS2).

RedHat не рекомендует запуск GFS2 на кластерах более 16 узлов, максимальный рекомендуемый размер тома 100TB на 64 битной архитектуре.

Запуск GFS2 включает: обновление до последних версий ядра (в ядре RHEL 6.0 при переходе с GFS на GFS2 была допущена ошибка, которая приводила к нестабильности работы GFS2), синхронизацию времени, настройка кластерного окружения, запуск кластерных служб (cman), кластерного менеджера томов (clvmd), создание логического тома, форматирование файловой системы, монтирование файловой системы.

Запуск OCFS2 включает обновление ядра и утилит ocfs-tools, настройку конфигурации кластера (файл cluster.conf), запуск службы o2cb, форматирование файловой системы, монтирование файловой системы.

Мы развернули обе кластерные файловые системы в сетях хранения, состоящих из трех вычислительных узлов и сравнили скорости последовательного чтения-записи для одного-трех узлов. Результаты тестов показаны на рис. 1.

Все скорости доступа оценивались снизу, для наихудшего случая, с отключением использования кэша дисков и RAID-контроллера. Это означает, что средняя скорость доступа к данным в режиме обычной эксплуатации за счет эвристик, встроенных в RAID-контроллер, будет выше приведенных оценок.

По результатам тестов и по стабильности работы кластерных служб для дальнейшей эксплуатации в центре обработки данных выбрана файловая система OCFS2.

dd (write), oflag=direct

Nodes	OCFS2	GFS2	NFS4
1	120 MB/s	59 MB/s	57 MB/s
2	78 MB/s	54 MB/s	53 MB/s
3	64 MB/s	45 MB/s	51 MB/s

dd (read), iflag=direct

Nodes	OCFS2	GFS2	NFS4
1	274 MB/s	174 MB/s	29 MB/s
2	178 MB/s	90 MB/s	29 MB/s
3	124 MB/s	73 MB/s	29 MB/s

Рис. 1. Скорость последовательной записи и чтения на файловых системах OCFS2, GFS2, NFS4

This work was supported by Russian Foundation for Basic Research (grant № 12-07-31261).