Секция «Биоинженерия и биоинформатика»

Применение кластерных файловых систем в центре обработки данных высокопроизводительного секвенирования.

Арифулов Ренат Надирович
Аспирант
Российский химико-технологический университет им. Д.И. Менделеева, Информационных технологий и управления, Москва, Россия
E-mail: arifulovrenat@gmail.com

В настоящий момент секвенирование (прочтение последовательностей) является неотъемлемой частью биомедицинской науки. Производительность современных секвенаторов такова, что на прочтение нескольких полных человеческих геномов требуется две недели. В результате работы секвенатора получается порядка 5 TB изображений, которые затем преобразуются в строковые последовательности ДНК (короткие чтения) длиной 100-150 символов. Затем данные проходят несколько этапов обработки, таких как: 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. Следует отметить что данные операции помимо вычислительных мощностей являются требовательными к операциям ввода-вывода (так как происходит обработка больших массивов данных). Вследствие этого т.н. «узким местом» становится доступ к файловому хранилищу центра обработки данных. Традиционная схема организации доступа к файловому хранилищу в научном учреждении – файл-сервер с хранилищем с доступом к нему по протоколу NFS не справляется с нагрузками при обработке геномных данных. Выходом в данном случае может являться использование систем хранения данных SAN (Storage area network), промышленных дистрибутивов Linux, открытых кластерных файловых систем. SAN обеспечивает доступ клиентов к системе хранения данных на блочном уровне (как к локальному диску). Вследствие этого необходимо использование специальной кластерной файловой системы, которая позволит избежать конфликтов и потерю данных при одновременном доступе к ней. В ходе исследования в сети хранения данных были развернуты открытые кластерные файловые системы (GFS2 от RedHat и OCF2 от Oracle) файловые системы и были изучены их показатели производительности и стабильности. По результатам исследования для внедрения в качестве основной кластерной файловой системы в сети хранения данных выбрана файловая система OCF2 как показавшая лучшие характеристики, и имеющая лучший потенциал масштабируемости.

Литература

1. http://docs.redhat.com/docs/ruRU/Red_Hat_Enterprise_Linux/5/html/Global_File_Syste (Red Hat GFS 2)

2. Troppens U., Muller-Friedt W., Wolafka R., Erkens R., Haustein N. Storage Networks Explained. Basics and Application of Fibre Channel SAN,NAS, iSCSI, InfiniBand and FcoE, Second Edition.

3. Shepler S., Eisler M., Noveck D. Network File System (NFS) Version 4 Minor Version 1 Protocol. IETF RFC 5661

4. https://oss.oracle.com/projects/ocfs2/(Project: OCFS2)