УДК 004.4

Методика тестирования файловой системы Lustre на суперкомпьютере «Говорун»

Д. В. Беляков, А. А. Кокорев, Д. В. Подгайный

Лаборатория информационных технологий,
Объединённый институт ядерных исследований,
ул. Жолио-Кюри 6, Дубна, Московская область, Россия, 141980

Email: dmitry@jinr.ru, kaa@jinr.ru, podgainy@jinr.ru

В настоящее время большинство высокопроизводительных кластеров или суперкомпьютеров имеют гибридную вычислительную архитектуру, как правило CPU и графические ускорители, высокоскоростные низколатентные межузловые сетевые коммуникации. Узким местом современных HPC систем, то есть ограничением производительности, является скорость работы с системами хранения данных, для преодоления этого разрыва используются параллельные файловые системы (ФС). Одной из самых популярных среди свободно распространяемых ФС является Lustre, которая также используется на суперкомпьютере «Говорун». Основными параметрами любой системы хранения данных (СХД) являются объем, скорость чтения/записи и отказоустойчивость, которую можно трактовать как время надежного хранения данных. При этом, для заданного объема дискового пространства ФС Lustre может быть сконфигурирована различным образом и основные ее параметры могут сильно варьироваться. Для выбора оптимальной конфигурации ФС, обеспечивающей наибольшую производительность задач пользователей необходимо иметь эффективный инструментарий.

В докладе представлена методика тестирования ФС Lustre на основе использования инструментария Interleaved or Random (IOR) Benchmark, использующего технологию MPI для синхронизации процессов чтения/записи запущенных параллельно на нескольких узлах на СК «Говорун».

Ключевые слова: параллельные, высокопроизводительные кластеры, отказоустойчивость, системы хранения данных, файловая система

1. Введение

Суперкомпьютер «Говорун» [1] является основной вычислительной частью гетерогенной платформы HybriLIT [2] Лаборатории информационных технологий им. М.Г. Мещерякова (ЛИТ ОИЯИ, г. Дубна). Суперкомпьютер «Говорун», предназначен для проведения ресурсоемких массивно-параллельных расчетов и обработки больших объемов данных. СК «Говорун» является инновационной гиперконвергентной программно-определяемой системой и обладает уникальными свойствами по гибкости настройки под задачу пользователя, обеспечивая максимально эффективное использование вычислительных ресурсов суперкомпьютера. В состав суперкомпьютера входят — CPU-компонента, построенная на базе решения компании РСК [3] «Торнадо», содержащая 117 вычислительных узлов, и GPU-компонента, построенная на базе 5 серверов компании NVIDIA DGX-1 с графическими ускорителями Tesla V100 и 5 серверами Niagara с графическими ускорителями NVIDIA A100.

Основными направлениями исследований в ОИЯИ являются задачи физики высоких энергий, включая задачи моделирования для экспериментов BM@N, MPD, SPD мега-сайенс проекта NICA, применение квантовых симуляторов, задачи биоинформатики, машинного обучения и глубокого обучения и т.д. Кроме больших вычислительных ресурсов данный спектр задач требует использования высокоскоростных систем работы с данными, связанных с вычислительными узлами высокоскоростной внутренней сетью с низкой латентностью от 100 Gbit/sec и выше. С этой

целью, на СК «Говорун» была разработана и внедрена иерархическая система обработки и хранения данных, представляющая собой единую централизованно управляемую систему, имеющую несколько уровней хранения данных: «очень горячие» данные, «горячие» данные и «теплые» данные. Использование этого решения позволило сформулировать и реализовать концепцию работы с Большими данными на СК «Говорун» как реализацию отображения (mapping) основных характеристик больших данных V3 (Volume – большие объемы данных для обработки и хранения, Velocity – необходимость в высокоскоростной их обработки, Variety – данные различных типов) на программно-аппаратные характеристики суперкомпьютера H3 (Heterogenity – набор вычислителей разного типа, Hierarchy – многоуровневая организация доступа к данным, Hyperconvegence – динамичная организация систем хранения данных) [4].

Основным элементом иерархической системы обработки и хранения данных является файловая система Lustre [5], которая задействована как на «теплом», так и на горячем слоях. При этом, следует отметить, что ФС Lustre может иметь разные настройки, что влияет на скорость операций чтения/записи, объем дискового пространства, доступного пользователям и на надежность хранения данных.

В данной работе представлена методика тестирования файловой системы Lustre на основе использования инструментария Interleaved or Random (IOR) Benchmark [6] с целью выбора оптимальной конфигурации ФС Lustre для обеспечения работы с данными для задач пользователей СК «Говорун».

2. Параллельная распределенная файловая система Lustre

Архитектура ФС Lustre включает в себя службу управления Lustre (MGS), хранилище метаданных (MDS) и службу объектного хранения (OSS). Файловая система Lustre может обслуживать сетевые подключения по протоколам – RDMA (InfiniBand, Omni-Path) и TCP/IP.

На нижнем уровне ФС Lustre находиться комбинированная файловая система zfs [7] с менеджером логических томов, представляющая инструменты для простого управления дисковыми массивами. Zfs не перезаписывает данные, а всегда оперирует новыми блоками, для обеспечения консистентности данных не нужен журнал, как в большинстве других файловых систем. К ее сильным сторонам можно отнести безопасное хранение информации, упор на целостность, наличие работы с большим объемом данных. ФС zfs работает с различными конфигурациями пулов таких как raidz1 – 2, draid1 – 2 и другими.

В ходе настройки файловой системы Lustre на всех задействованных узлах были созданы пулы с использованием файловой системы zfs при создании пулов использовались следующие настройки zfs:

– `canmount=off` – не позволяет монтировать пул при его создании;

– `cachefile=none` – позволяет отключить файл кэша, для аварийного переключения, где пул всегда должен быть импортирован программным обеспечением;

– `multihost=on` – свойство предназначено для использования в конфигурациях отработки отказоустойчивости.

При настройке Lustre использовался параметр Lustre Network (LNET) [8]

– `options ksocklnd_tx_buffer_size=* r_buffer_size=*` с помощью которого можно настроить буфер передачи и приема данных.

– `oss_num_threads` позволяет указать количество потоков обслуживания одним OSS.

Увеличение размера потоков может помочь, когда:

– несколько операционных систем экспортируются из одной операционной системы.

– серверное хранилище работает синхронно.

– завершение ввода-вывода занимает слишком много времени из-за медленного хранения.

Уменьшение размера потоков может помочь, если:

– клиенты перегружают объем хранилища

– есть много сообщений о «медленном вводе-выводе»

– `mds_num_threads` – аналогичный параметр используется для количества потоков обслуживания MDS. Максимальное количество потоков (MDS_MAX_THREADS) равно 1024.

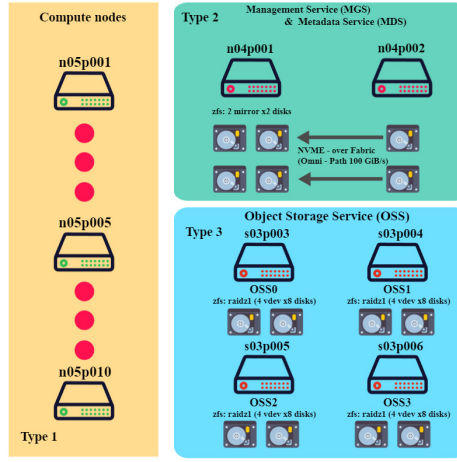

Рис. 1. Схема реализации ФС Lustre

На Рис. 1 представлена схема файловой системы Lustre в которой были задействованы два типа вычислительных узлов и один тип узлов хранения данных, входящих в состав суперкомпьютера «Говору», которые имеет следующие технические характеристики:

Тип 1 вычислительный узел РСК [3] Торнадо: 2 процессора Intel(R) Xeon(R) Platinum 8368Q 2.60GHz, 76 ядер, 2 ТВ оперативной памяти, построенной на базе Intel Optane Persistent Memory, Intel Omni-path 200 GiB/sec, 4х твердотельных накопителя Intel 4 TB форм-фактор M2 2280 общей емкостью 16 ТБ.

Тип 2 вычислительный узел РСК Торнадо: 2 процессора Intel(R) Xeon(R) Platinum 8268 CPU 2.90GHz, 48 ядер, 192 Gb оперативной памяти, Intel Omni-path 200 GiB/sec, 2 твердотельных накопителя Intel 1 TB форм-фактор M2 2280 общей емкостью 2 TB.

Тип 3 узел хранения данных: процессор Intel(R) Xeon(R) Gold 6248R CPU 3.00GHz, 48 ядер, 376 Gb оперативной памяти, Intel Omni-path 200 GiB/sec, 32х твердотельных накопителя Intel 30.73 TB форм-фактор E1.L общей емкостью порядка 1 PB.

3. Методика тестирования производительности файловой системы

Исследование режимов работы и тестирование производительности файловой системы Lustre инструментарием IOR, который использует технологию MPI [9] для синхронизации процессов чтения/записи запущенных параллельно на нескольких узлах.

Тестирование проводилось следующим образом: для инструментария IOR был создан хост файл на вычислительном узле содержащий список узлов, на которых запускался IOR. IOR использует программный пакет OpenMPI базирующийся на технологии MPI для распараллеливания процессов чтения/записи. При сборке из исходных кодов инструментария IOR был установлен программный пакет openmpi-devel, а сами тесты запускались с командной строки командой mpirun.

После выполнения тестирования полученные данные сохраняются в файл и визуализируются для удобства дальнейшего анализа.

4. Результаты тестирования

Тестирование проводилось для десяти различных конфигураций сборки zfs пулов с тремя типами рейдов – raidz1, raidz2 и draid2.

raidz1 – аналог массива raid5, который используется в системах хранения данных, основанных на технологии zfs. Хотя пул дисков, объединенных в raidz, имеет ту же отказоустойчивость в один диск, как и raid5, механизм реализации этого сильно отличается от классической технологии raid5.

raidz2 – большое, высоконадежное и относительно дорогое хранилище, для которого требуется минимум 3 диска. raiddz2 похож на raid6 – позволяет пережить сбой двух дисков за счет сохранения двух разных функций четности. Все соображения относительно схемы размещения блоков и размера блока такие же, как и для raidz.

draid – вариант raidz, который обеспечивает интегрированные распределенные «горячие» резервы, что позволяет ускорить восстановление данных, сохраняя при этом преимущества raidz. Виртуальное устройство draid состоит из нескольких внутренних групп raidz, каждая из которых имеет устройства данных D и устройства четности P. Эти группы распределяются по всем дочерним элементам, чтобы полностью использовать доступную производительность диска.

Также при создании z-пулов варьировались следующие параметры:

– `vdev` – виртуальное устройство содержащие в себе физические устройства;

– `disk` – количество дисков в одном vdev

Для обозначения собранных рейдов используется следующее представление:

draid2:0d:0s:0c:

– `d (data)` – количество устройств передачи данных на группу резервирования;

– `s (spares)` – количество отведенных дисков для горячей замены;

– `c (children)` – количество дисков, включенных в развертывание dRAID.

Результаты тестирования представлены на Рис. 2 и 3.

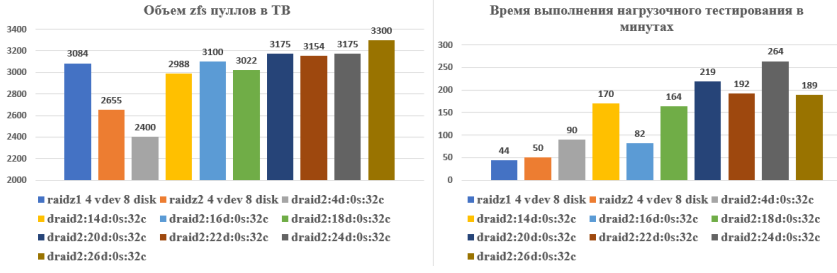

Рис. 2. Объем доступного пользователям дискового пространства (слева) и время выполнения нагрузочного тестирования (справа) для различных конфигураций zfs пулов

Как видно из левой диаграммы Рис. 2, объем дискового пространства, доступного пользователям сильно варьируется от выбора параметров zfs пулов – от минимального значения 2400 ТБ для draid2:4d:0s:32c до максимального 3300 ТБ для draid2:26d:0s:32c. Разница между этими значениями составляет 900 ТБ, что соответствует 22,5% от общего дискового пространства. На правой диаграмме представлено время выполнения нагрузочного тестирования, при этом и здесь видна сильная вариабельность от выбора z-пула – минимальное значение составляет 44 минуты для raidz1 4 vdev 8 disk, а максимальное – 264 минуты draid2:24d:0s:32c, т.е. разница составляет 220 минут, что также как и для объема дискового пространства имеет существенное значение и влияет на результаты представленные выше.

На Рис. 3 представлены диаграммы скорости чтение/запись во время выполнения тестов. Из диаграммы слева видно, что скорость чтения варьируется от 15239 МБ в секунду для draid2:14d:0s:32c и 21705 МБ в секунду для пула raidz2 4 vdev 8 disk, разница составляет 6466 МБ в секунду или 29,8%. Для скорости записи данных минимальное значение составляет 655 МБ в секунду для draid2:24d:0s:32c, а максимальное – 4549 МБ в секунду для radiz1 4 vdev 8 disk, разница составляет 3894 МБ в секунду или 85,6%. При этом, как видно из приведенных диаграмм максимальные и минимальные значения скоростей чтения и записи приходятся на разные zfs пулы, это обстоятельство нужно учитывать при выборе оптимальной конфигурации файловой системы.

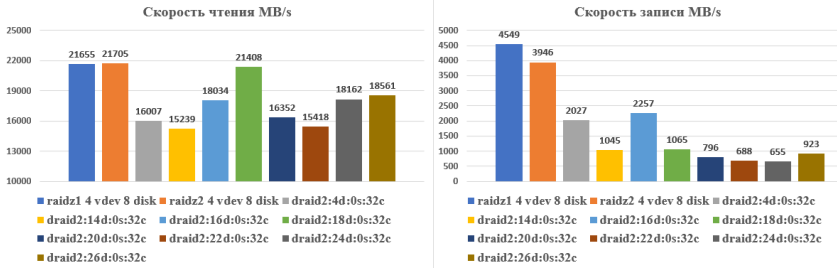

Рис. 3. Результаты тестирования на скорость чтения (слева) и скорость записи (справа)

Проанализировав полученные данные, можно сделать вывод, что zfs пул radiz1 подходит для использования как универсальный рейд, подходящий под большинство классов задач пользователей, требующих максимальную скорость записи данных, при этом не требующих самой большой скорости чтения и имеющего вполне удовлетворительный объем дискового пространства, доступного пользователям. В частности, таким требованиям удовлетворяют задачи эксперимента MPD на комплексе NICA [10].

5. Заключение

На платформе HybriLIT, включающий в себя суперкомпьютер «Говорун» решаются вычислительные задачи, предъявляющие различные требования к работе с данными. Следствием этого, является невозможность создания универсальной системы обработки и хранения данных, позволяющей всем типам задач выполняться одинаково эффективно. На основе представленной в этой работе методики тестирования файловой системы Lustre с использованием инструментария IOR получены результаты, позволяющие выбрать наиболее оптимальную конфигурации ФС в зависимости от требований задачи пользователя. В частности, таким образом, была выбрана

и введена в эксплуатацию ФС Lustre для генерации, реконструкции и физического анализа событий для эксперимента MPD NICA.

Отметим, что разработанная методика подходит не только для тестирования ФС Lustre, но и любой другой файловой системы, на нижнем уровне которой лежит zfs, на пример, BeeGFS, EOS и др.

Благодарности

Исследования в данном направлении были поддержаны специальным грантом РФФИ («Мегасайенс – NICA») № 18-02-40101.

Литература

1. Суперкомпьютер «Говорун», URL: http://hlit.jinr.ru/ (Дата обращения: 01.02.2024).
2. Беляков Д.В., Бутенко Ю.А., Валя М., et al. Гетерогенная платформа HybriLIT // Бюллетень «Новости ОИЯИ». — 2019. — №2. — С. 19-25. URL: http://www1.jinr.ru/News/Novosti_2-2019_url_mini.pdf (Дата обращения: 01.01.2024).
3. Группа компаний РСК, URL: http://www.rscgroup.ru/ru/ (Дата обращения: 22.01.2024).
4. Belyakov D.V., Dolbilov A.G., Moshkin A.A., et al.: “Using the “Govorun” Supercomputer for the NICA Megaproject” // CEUR Workshop proceedings, 2018, Vol. 2507, pp. 316-320.
5. Файловаясистема Lustre, URL: https://www.lustre.org/ (Дата обращения: 02.02.2024).
6. Interleaved or Random (IOR) Benchmark, URL: https://wiki.lustre.org/IOR (Дата обращения: 10.02.2024).
7. OpenZFS, URL: https://openzfs.org/wiki/Main_Page (Дата обращения: 20.02.2024).
8. Lustre Tuning, URL: https://wiki.lustre.org/Lustre_Tuning#Sizing_the_MDT (Дата обращения: 12.02.2024).
9. OpenMPI (Intel version), URL: https://docs.hpc.shef.ac.uk/en/latest/decommissioned/sharc/software/parallel/openmpi-intel.html#gsc.tab=0 (Дата обращения: 15.01.2024).
10. Nuclotron-based Ion Collider fAcility, URL: URL:https://nica.jinr.ru/ru/ (Дата обращения: 25.01.2024).

UDC 004.4

The methodology of testing the Lustre file system on the «Govorun» supercomputer

D. V. Belyakov, A. A. Kokorev, D. V. Podgainy

Laboratory of Information Technologies
Joint Institute for Nuclear Research
Joliot-Curie 6, Dubna, Moscow region, 141980, Russia

Email: dmitry@jinr.ru, kaa@jinr.ru, podgainy@jinr.ru

Currently, most high-performance clusters or supercomputers have a hybrid computing architecture, as a rule CPU and graphics accelerators, high-speed low-latency internodal network communications.

The bottleneck of modern HPC systems, that is, the performance limitation, is the speed of working with data storage systems, to bridge this gap, parallel file systems (FS) are used. One of the most freely distributed file systems is Luster, which is also used on the «Govorun» supercomputer. The main parameters of any data storage system area volume, read/write speed and fault tolerance, which can be handled as the time of reliable data storage. At the same time, for a given amount of disk space, the Lustre FS can be configured in various way and her main parameters can vary greatly. To select the optimal configuration of the FS, which ensures the greatest productivity of tasks users, it is necessary to have effective tools. The report presents a methodology for testing the Luster FS based on the use of the Interleaved or Random (IOR) Benchmark toolkit, which uses MPI technology to synchronize read/write processes running in parallel on several nodes on the Govorun supercomputers.

Key words and phrases: parallel, high-performance clusters, fault tolerance, data storage systems, file system.