DATA STORAGE SYSTEMS OF «HYBRILIT» HETEROGENEOUS COMPUTING PLATFORM FOR SCIENTFIC RESEARCH CARRIED OUT IN JINR: FILESYSTEMS AND RAIDS PERFORMANCE RESEARCH

A.A. Kokorev[2,a], D.V. Belyakov[1], M.A. Lyubimova[2]

[1] *Laboratory of Information Technologies, Joint Institute for Nuclear Research, Dubna, Russian Federation*

[2] *State budgetary educational institution of higher education of the Moscow region University "Dubna", Dubna, Russian Federation*

E-mail: [a] kaa@jinr.ru

"HybriLIT" Heterogeneous platform is a part of the Multifunctional Information and Computing Complex (MICC) of the Laboratory of Information Technologies named after MG Meshcheryakov of JINR, Dubna. Heterogeneous platform consists of Govorun supercomputer and HybriLIT education and testing polygon. Data storage and processing system is one of the platform components. It is implemented using distributed and parallel filesystems (NFS, EOS, Lustre). Platform performance depends on many factors, including performance of storage and file systems.

The best storage performance for wide variety of user jobs may be obtains with optimal filesystem parameters. The number of tests of local filesystems (EXT family and XFS) was carried out. There were empirically obtained an optimal parameters o data storage system at which the performance have been high results.

The new methodology was developed for analyzing the obtained measurements of IOPS (input-output operations per second) and Latency (milliseconds) for results evaluations.

Various filesystems were analyzed by the developed methodology. The conclusion was drawn about of optimal parameters of the investigated filesystems.

Keywords: distributed, high-performance computing, big data, storage systems, file systems.

Alexandr Kokorev, Dmitriy Belyakov, Maria Lyubimova

СИСТЕМЫ ХРАНЕНИЯ И ОБРАБОТКИ ДАННЫХ НА ГЕТЕРОГЕННОЙ ВЫЧИСЛИТЕЛЬНОЙ ПЛАТФОРМЕ «HYBRILIT»: ИССЛЕДОВАНИЕ ПРОИЗВОДИТЕЛЬНОСТИ ФАЙЛОВЫХ СИСТЕМ И ДИСКОВЫХ МАССИВОВ

А.А. Кокорев[2,a], Д.В. Беляков[1], М.А. Любимова[2]

[1] *Лаборатория информационных технологий, Объединенный институт ядерных исследований, г. Дубна, Российская Федерация*

[2] *Государственное бюджетное образовательное учреждение высшего образования Московской области Университет «Дубна», г. Дубна, Российская Федерация*

E-mail: [a] kaa@jinr.ru

Гетерогенная платформа «*HybriLIT*» является частью Многофункционального информационно-вычислительного комплекса (МИВК), Лаборатории информационных технологий имени Мещерякова М.Г. ОИЯИ, г. Дубна. Гетерогенная платформа состоит из суперкомпьютера «Говорун» и учебно-тестового полигона «*HybriLIT*». Одним из компонентов платформы является система хранения и обработки данных, которая реализована с помощью распределенных файловых систем (*NFS, EOS, Lustre*). Производительность платформы зависит от многих факторов, в том числе и от работы системы хранения данных и файловых систем.
Для получения максимальной производительности системы хранения данных для широкого спектра пользовательских задач необходимо определить оптимальные параметры файловых систем. В проведенном ряде исследований файловых систем (семейства *EXT*, *XFS*), дисковых массивов (*RAID* 5, 6, 10) и программных массивов (*ZFS*) был определен набор параметров, при котором производительность системы хранения данных показала высокие результаты.
Для оценки результатов была разработана новая методика анализа полученных измерений величин *IOPS* (количества операций ввода-вывода в секунду) и *Latency* (задержки, в миллисекундах).
В соответствии с разработанной методикой были проанализированы различные файловые системы и сделан вывод о наборе оптимальных параметров исследуемых файловых систем.

Ключевые слова: распределённые, высокопроизводительные вычисления, большие данные, системы хранения данных, файловые системы.

Александр Кокорев, Дмитрий Беляков, Мария Любимова

1. Введение

Системы хранения и обработки данных гетерогенной платформы «*HybriLIT*»[1] обеспечивают пользователей широким набором возможностей: распределённые файловые системы *EOS* и *Lustre* применяются для размещения больших объёмов данных; динамические хранилища, создаваемые по технологии «*Storage-on-Demand*», для временных файлов при выполнении расчётов; дисковые серверы с файловой системой *ZFS*, подключённые по протоколу *NFS v*4 ко всем вычислительным узлам и пользовательским интерфейсам, для хранения домашних директорий пользователей. В основе применяемых файловых систем используются технологии — *RAIN* (*Redundant Array of Independent Nodes*) и/или *RAID* (*Redundant Array of Independent Disks*), что обеспечивает высокую степень надёжности хранения данных.

Пользователи гетерогенной платформы применяют различные математические алгоритмы для обработки и анализа данных, создавая одновременно различные виды нагрузок на систему хранения данных гетерогенной платформы. Высокопроизводительные алгоритмы предъявляют жёсткие требования к скорости доступа и производительности систем хранения данных, что в условиях многопользовательских систем является трудной задачей.

Современная парадигма хранения данных разделяет данные пользователей на несколько типов в зависимости от необходимости использования — «холодные», «тёплые», «горячие» и «ультра-горячие». Для каждого типа данных требуется применять свою систему хранения, построенную на определённом типе носителей информации, и подключённую в локальную сеть многопользовательской системы наиболее подходящей сетевой технологией.

Вопросы выбора, настройки и оптимизации систем хранения и обработки данных являются особо актуальными для многопользовательских вычислительных систем.

2. Производительность файловой системы

Измерение производительности дисковой системы сервера обычно выполняют с помощью запуска синтетических тестов, создавая различные виды нагрузки, искусственно эмулируя разные роли, например, файлового сервера или базы данных. Применяют тесты на последовательные и случайные операции чтения и записи, а также комбинации этих операций в разных процентных соотношениях. Результатом измерений обычно являются — величина пропускной способности дисковой системы, измеряемой в единицах операций в секунду (*IOPS*) или в количестве байт в секунду (например, *Mbit/sec*, *Kbits/sec*), а также величина задержки дисковой системы (*Latency*) в миллисекундах, отражающей скорость обслуживания очереди дисковых операций. Существует ряд факторов, в некоторой степени влияющих на сами результаты измерений — к ним относятся использование системного буфера и дискового кэша. На короткие очереди дисковых операций влияние этих факторов оказывает большое влияние, при постоянной нагрузке влияние снижается.

Фактически, применение синтетических тестов показывает результаты работы самих тестов в неких «условных» ситуациях, не всегда совпадающих с реальными нагрузками. Другой задачей является стремление достигнуть производительности, заявленной фирмой-изготовителем. Однако методика тестирования и программный код «фирменных» тестов являются закрытой информацией и не доступны для применения.

Таким образом, применяемая методика (наборы тестов) и получаемые результаты не совсем достоверно отражают реальные значения производительности, более того, сравнение с данными, предоставляемыми фирмой-изготовителем, носит в некой степени условный характер.

Классическая методика измерения производительности оперирует с носителем информации в монопольном режиме, измеряя количество операций (или байтов) в секунду и время задержки в миллисекундах в разных режимах чтения и/или записи. В то время как для многопользовательских систем существенен режим работы многих пользователей, создающих одновременно разные виды нагрузок: последовательные и случайные операции чтения и/или записи разных объёмов данных в разные области файловой системы, не зависимо от того представлена она одним лишь дисковым сервером (*RAID*) или группой серверов (*RAIN*).

Исходя из этих условий решающими факторами становятся не только время обслуживания на уровне единичного носителя информации и задержки на сетевом уровне, но и одновременно конкурирующие различные виды нагрузок, создаваемые задачами многих пользователей.

Поведение файловой системы под нагрузкой, т.е. её производительность, зависит от многих факторов. Одна часть которых — представляет параметры файловой системы и устанавливается администратором при её создании, другая часть которых — определяет различные виды нагрузки и задаётся в конфигурационных файлах используемых тестов. Множество всех факторов образует своеобразные «степени свободы», фиксируя которые, можно исследовать поведение файловой системы и определить наиболее «удачную» комбинацию, т.е. при каких условиях будет получена максимальная производительность.

Составив комбинации параметров по всем «степеням свободы», будет подготовлен полный набор тестов, в соответствии с которым необходимо выполнить тестирование. Полученные значения пропускной способности и величины задержки дисковой системы представляют собой огромный массив данных, который необходимо проанализировать. Однако сравнение полученных результатов представляет задачу повышенной сложности — как определить наиболее «удачную» комбинацию?

При одном наборе параметров — файловая система самая «быстрая» по операциям чтения и не очень «быстрая» по операциям записи, при другом наборе — наоборот, самая «быстрая» по операциям записи и не очень «быстрая» по операциям чтения. Кроме того, присутствует широкий спектр комбинаций, при которых файловая система даёт приемлемую производительность по обеим операциям одновременно, не значительно выигрывая по одной операции и проигрывая по другой. Для многопользовательских вычислительных систем эта область параметров представляет особый интерес, поскольку одновременно работающие задачи разных пользователей предъявляют различные требования к производительности файловой системы.

Для анализа полученных данных, сравнения между собой и с результатами тестирования другой файловой системы — требуется разработка новой методики анализа производительности.

3. Методика анализа производительности файловой системы

На основе проведённого ряда исследований файловых систем семейства (*EXT, XFS*)[2] была разработана новая методика анализа производительности систем хранения и обработки данных.

В ходе выполнения тестирования файловой системы с заданной комбинацией параметров измерялись следующие величины: количество операций чтения/записи в секунду (*IOPS*) и величина задержки дисковой системы в миллисекундах, т.е. латентность (*Latency*):

$$r = IOPS\ read\ ,\ L_r = Latency\ read$$
$$w = IOPS\ write\ ,\ L_w = Latency\ write \quad (1)$$

где r и w — количество *IOPS*, соответственно, для операций чтения и записи, а L_r и L_w — значение латентности, мс.

Для сравнения результатов исследования производительности файловой системы с разными комбинациями параметров, а также для последующего сравнения с результатами, полученными в ходе тестирования других файловых систем, размещённых на тех же дисковых носителях, были выполнены следующие действия:

Результаты измерений производительности были представлены в виде пары чисел:

$$(r, w) = (IOPS\ read, IOPS\ write) \quad (2)$$

Для каждой пары чисел было вычислено отклонение от данных, предоставленных фирмой-изготовителем и выполнено нормирование на собственное значение:

$$(x, y) = \left(\frac{r - r_i}{r}, \frac{w - w_i}{w}\right) * 100\% \quad (3)$$

где r_i и w_i — количество *IOPS*, соответственно, для операций чтения и записи, предоставленные фирмой-изготовителем; формируют так называемую линию тренда.

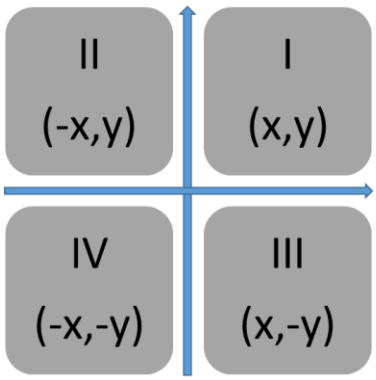
Полученные значения (*x, y*) представляют собой отклонение результатов измерений от линии тренда в процентном отношении. Пары чисел (*x, y*) были размещены на координатной плоскости. В I квадранте (*x, y*) оказались результаты, которые располагаются выше линии тренда. В IV квадранте (*-x, -y*) оказались результаты, которые располагаются ниже линии тренда. В II квадранте (*-x, y*) и III квадранте (*x, -y*) оказались результаты, у которых *IOPS write* и *IOPS read*, соответственно, выше линии тренда.

Для анализа полученных значений (*x, y*) были вычислены модули расстояний:

$$M_{IOPS} = \sqrt{x^2 + y^2} \tag{4}$$

$$M_L = \sqrt{L_r^2 + L_w^2} \tag{5}$$

где L_r и L_w — значения латентности для соответствующей пары значений (*x, y*).



Для I квадранта:
Чем дальше от (0,0), тем лучше, то есть сортировка по возрастанию модуля M_{IOPS}.

Для II и III квадрантов:
Чем больше y(II) или x(III), тем лучше, то есть сортировка по возрастанию для y (*IOPS write*), для x (*IOPS read*).

Для IV квадранта:
Чем ближе к (0,0), тем лучше, то есть сортировка по убыванию модуля M_{IOPS}.

Контрольное правило для любого квадранта:
Чем меньше модуль M_L, тем лучше.

Рисунок 1. Критерии отбора наиболее «удачной» комбинации параметров файловой системы

Критерии отбора наиболее «удачной» комбинации параметров файловой системы (рис. 1) были получены на основе анализа расположения пары значений (*x, y*) в зависимости — 1) номера квадранта, 2) значения модуля M_{IOPS}; при этом в качестве контрольного правила учитывалось значение модуля M_L.

Для I квадранта: чем дальше от линии тренда (0,0), тем лучше; применялась сортировка по возрастанию модуля M_{IOPS}.

Для II и III квадрантов: чем больше *y* (II) или *x* (III), тем лучше; применялась сортировка по возрастанию для *y* (*IOPS write*), для *x* (*IOPS read*).

Для IV квадранта: чем ближе к линии тренда (0,0), тем лучше; применялась сортировка по убыванию модуля M_{IOPS}.

Контрольное правило для всех квадрантов: чем меньше модуль M_L, тем лучше.

Разработанная методика анализа производительности файловой системы позволила успешно проанализировать другие файловые системы — дисковые массивы (*RAID 5,6,10*) и программные массивы (*ZFS*).

4. Инструмент тестирования и тестовые стенды

В качестве инструмента тестирования использовалась специализированная утилита *FIO* (*Flexible I/O Tester*)[3] для измерения производительности блочных устройств, разработанная Дженсом Эксбо (*Jens Axboe*). Набор тестов включал в себя три вида нагрузки — случайное чтение, случайная запись и одновременное случайное чтение, и запись. Модель доступа была определена асинхронным режимом ввода-вывода и задействованным или выключенным буфером операционной системы и дисковым кэшем.

При тестировании были использованы файловые системы: *EXT2, EXT3, EXT4, XFS*; дисковые массивы *RAID 5,6,10* и программный массив *ZFS*; размер кластера файловой системы — 1, 2, 4, 8, 16, 32, 64 *KB*; размер блока записи — от 1 до 64 *KB*; глубина очереди — 32 и 64

последовательных пакета.

В качестве тестовых стендов были использованы:

1) Вычислительный модуль *Dell PowerEdge FC430* (*2x Intel Xeon E5-2680, 12 Cores @ 2.5GHz, 256 GB DDR4-2133 ECC*) с тестируемым диском *Intel DataCenter S3610 SSD*, объёмом *400 GB*.

2) Дисковый сервер *SuperMicro SuperStorage 6027R-E1R12N* (*2x Intel Xeon E5-2630 v2, 6 Cores @ 2.6 GHz, 16 GB DDR3L-1333 ECC*) с тестируемыми дисками *12x Seagate Constellation ES.3 SAS*, объёмом *2 TB*, подключённые к системе с помощью *RAID*-контроллера *SuperMicro SMC2108* на основе чипа *LSI/Broadcom SAS 2108* (*9260-8i*). Были созданы дисковые массивы *RAID 5* (3 диска, *3.6 TB*), *RAID 6* (4 диска, *3.6 TB*) и *RAID 10* (4 диска, *3.6 TB*); на каждом дисковом массиве был создан тестовый файл размером *400 GB* (для сравнения результатов со стендом 1).

3) Вычислительные сервера-лезвия *RSC Tornado TDN421* (*2x Intel Xeon Platinum 8268, 24 Cores @ 2.9 GHz, 192 GB DDR4-3200 ECC*)[4], установленные в стойку *Tornado TCC153B* и интерконнектом *Intel Omni-Path 100 Gbit/sec*, с тестируемыми дисками *2x Intel DataCenter P4511 NVMe*, 1-2 *TB*. Были созданы распределённые хранилища по технологии «*Storage-on-Demand*» с файловой системой *ZFS* и объёмами *10.9 TB* и *29 TB* (6 дисков, */nfs/bltp-test* и 32 диска, */nfs/mpd*), подключёнными по протоколу *NFS* по высокоскоростной сети *Intel Omni-Path 100 Gbit/sec*; на каждом хранилище был создан тестовый файл размером *400 GB* (для сравнения результатов со стендом 1).

5. Примеры результатов анализа производительности файловых систем и дисковых массивов

На рис. 2-3 представлены примеры сравнительного анализа производительности файловых систем (рис. 2) и дисковых массивов (рис. 3), полученных методом объединения графиков по каждому квадранту в соответствии с разработанной методикой. Для упрощения графиков каждая файловая система изображена одним цветом вне зависимости от блока записи и размера кластера.

Заметим, что при некоторых комбинациях размеров блока и кластера результаты разных файловых систем образуют «общие» структуры — кластеры, которые показывают, что разные файловые системы ведут себя почти одинаково при соответствующих комбинациях параметров.

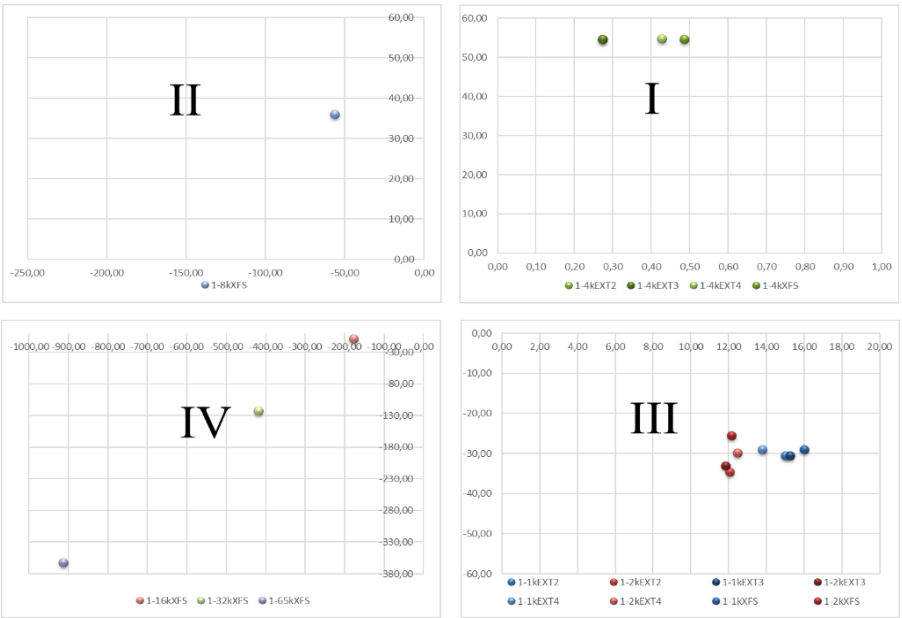


Рисунок 2. Результаты анализа теста чтение/запись файловых систем семейства *EXT* и *XFS*, дисковый кэш и буфер выключены, глубина очереди 32.

Исследование дисковых массивов после сравнения и анализа результатов тестирования локальных файловых систем, поэтому в качестве файловой системы для тестирования дисковых массивов была выбрана *EXT4* с размером кластера 4096 байт — как оптимальная (самая быстрая по операциям ввода/вывода) файловая система.

При исследовании дисковых массивов *RAID 5, 6, 10* все полученные данные попали в IV квадрант, что соответствует ситуации, когда количество операций ввода/вывода существенно меньше, чем можно получить на одном диске (данное значение было получено экспериментальным путём, т.к. фирма-изготовитель не указала эти данные).

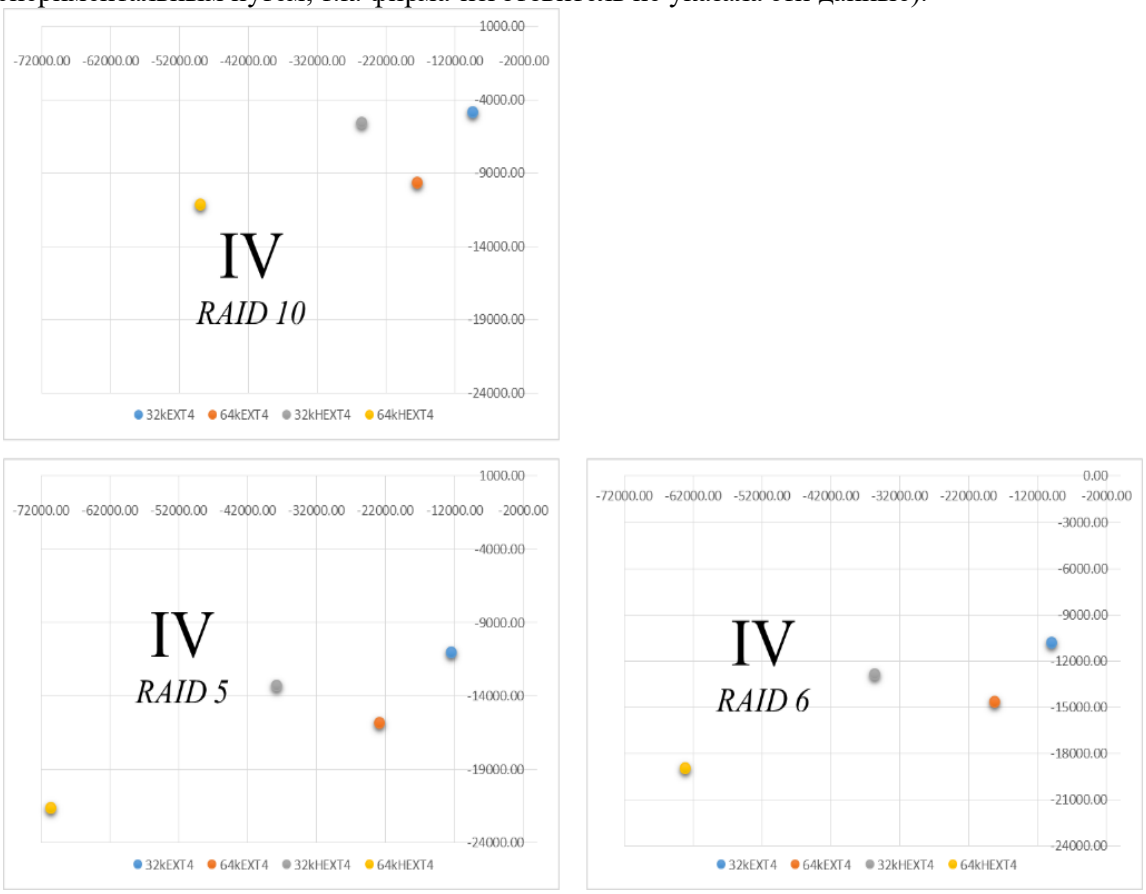


Рисунок 3. Результаты анализа тестов чтения/записи и одновременного чтения/записи дисковых массивов *RAID 5,6,10*, дисковый кэш включён, буфер используется, глубина очереди 32.

6. Заключение

В ходе выполнения работы было проведение исследование систем хранения и обработки данных на гетерогенной платформе «*HybriLIT*» — локальных файловых систем семейства *EXT* и *XFS*, дисковых массивов *RAID 5, 6, 10* и систем хранения данных, построенных по технологии «*Storage-on-Demand*». Разработана новая методика анализа результатов тестирования производительности файловой системы, которая была успешно применена для анализа полученных результатов. На основе выполненного сравнительного анализа были определены оптимальные параметры файловых систем, при которых наблюдается максимальная производительность.

7. Благодарность

Исследование выполнено при финансовой поддержке РФФИ в рамках научного проекта № 18-02-40101.

Литература

[1] Гетерогенная платформа «HybriLIT». Веб-сайт URL: http://hlit.jinr.ru/

[2] Кокорев А.А., Беляков Д.В. Системы хранения и обработки данных на Гетерогенной вычислительной платформе «HybriLIT»: исследование производительности файловых систем. Российский университет дружбы народов. Веб-сайт URL: https://events.rudn.ru/event/107/book-of-abstracts.pdf //Information and Telecommunication Technologies and Mathematical Modeling of High-Tech Systems. 2021. С. 62 – 63.

[3] Специализированная утилита FIO Веб-сайт URL: http://freshmeat.sourceforge.net/projects/fio

[4] Решения компании РСК. Веб-сайт URL: https://rscgroup.ru/project/jinr-hpc-cluster-govorun/