Инженерный вестник Дона, №6 (2022)
ivdon.ru/ru/magazine/archive/n6y2022/7688

Средства ускорения выполнения задач с большим объёмом операций ввода/вывода в гетерогенной вычислительной системе.

Д.Л. Абдрахманов, Д.Н. Жариков, Д.В. Завьялов

ФГБОУ ВО «ВолгГТУ»

Аннотация: Данная статья посвящена возможности использования оперативного запоминающего устройства, как подключаемого дискового массива с целью ускорения вычислений в рамках гетерогенной вычислительной системы. Рассмотрены проблемы применения RAID-массива жёстких дисков, как основного дискового пространства для хранения данных, генерируемых приложениями для гетерогенной системы задач. Представлены принципы использования оперативного запоминающего устройства как дискового пространства, а также описаны особенности подключения к полученному таким образом хранилищу данных посредством сети.
Ключевые слова: параллельное вычисление, файловая система, оперативная память, монтирование, гетерогенность, вычислительный комплекс.

1. Введение

Математическое моделирование прогнозируемых механических процессов будущего продукта в различных специализированных программных средах до начала его фактической реализации является одним из наиболее важных аспектов современных процедур разработки и производства высокотехнологичного оборудования. Основные факторы, определяющие эффективность данного процесса – точность модели и скорость вычислений, проводимых на основе различных вычислительных комплексов.

Уровень качества и степень точности созданной модели, отражающий действительность, обуславливают широту её использования для обоснования фактической работоспособности конечного продукта, поступающего в производство. Качественная аппроксимация предполагает существенное усложнение модели и, как следствие, увеличение объёма необходимых ресурсов и значительный рост временных затрат на её разработку.

Возможности современных персональных компьютеров не соответствуют описанным ранее требованиям, так как не обеспечивают

построение адекватных симуляций. Для достижения данных целей требуется использовать серверные платформы и созданные на их основе вычислительные комплексы, обеспечивающие решение задач в параллельном режиме, имеющие большой объём памяти и связанные высокоскоростной сетью [1].

Распределение нагрузки между узлами вычислительной серверной платформы предполагает постоянное обращение к дисковым массивам как в целях получения новых данных о модели, так и для записи и/или считывания промежуточных и конечных итогов проведённого моделирования [2].

2. Хранилища данных

Доминирующие позиции в современном рынке средств хранения больших объёмов данных принадлежат жёстким дискам (далее HDD) и твердотельным накопителям (далее SSD).

Широкое использование жёстких дисков обуславливается их надёжностью. Существенные недостатки – наличие движущихся частей, которые могут прийти в негодность вследствие механического износа, сравнительно малая скорость работы и возможность существенной фрагментации хранящихся данных.

Твердотельные накопители не имеют описанных выше недостатков, так как при их изготовлении используется Flash-память. Основное негативное явление в этом случае – высокая стоимость производства ячеек памяти, которая обуславливает использование настоящих накопителей для записи не 1 бита на ячейку, а 2 – 4 бит на ячейку. Данное положение усложняет работу контроллера и снижает скорость операций ввода-вывода, а также отрицательно влияет на сохранность данных. При этом обеспечивается увеличение объёма хранимой информации при неизменной стоимости готового продукта.

2.1 Сетевые хранилища

Скорость, с которой вычислительные узлы могут принимать данные от сервера, ограниченна пропускной способностью сетевого соединения. Например, в широко используемой InfiniBand FDR она достигает 54 Гбит в секунду. Пропускная способность HDD или SSD недостаточна для организации передачи данных при обращении 1 узла к серверу посредством данной сети.

При увеличении количества узлов, получающих доступ к ресурсу, возникает проблема невозможности их одновременной работы, что будет приводить к значительным потерям производительности при решении задач с большим объёмом ввода/вывода данных [3].

Описанные выше противоречия между требованиями по решению задач и возможностями организации процесса их решения подтверждают актуальность проблемы повышения скорости работы хранилища данных.

2.2 RAID

Процедура удешевления SSD посредством записи в ячейку памяти большего количества бит данных приводит к снижению надёжности Flash накопителя. Данная проблема может быть частично решена посредством использования избыточного массива независимых дисков (далее RAID).

Использование RAID-массивов обеспечивает выполнение части требований к повышению надёжности и скорости работы.

В частности, применение RAID 5 позволяет увеличить скорость доступа к данным, но наблюдаемый рост недостаточен для работы с большими числовыми массивами [4].

2.3 TMPFS

Описанные выше условия определяют преимущества применения SLC SSD, в которых каждая ячейка содержит 1 бит информации.

SLC-накопители относительно дороги, то есть, хранение больших объёмов данных предполагает использование особых серверных платформ с большим количеством «посадочных мест» для накопителей.

Решение задач в параллельном режиме подразумевает использование схожих по производительности узлов. При применении гетерогенной системы для решения большинства таких задач не всегда возможно тотальное использование узлов с максимальной эффективностью. Вычислительные среды преимущественно не имеют собственных RAID-массивов или SSD, но обладают большим объёмом оперативной памяти (далее ОЗУ).

ОЗУ – тип хранилища информации, предназначенный для чтения и записи больших объёмов данных на скоростях, существенно превышающих RAID-массивы на основе SSD. ОЗУ при множественных операциях записи значительно надёжнее, чем Flash-память.

В операционных системах (далее ОС) семейства Linux возможно подключение оперативной памяти в качестве дискового пространства посредством использования модуля системы монтирования tmpfs.

Монтирование ОЗУ позволяет получить накопитель, превышающий по скорости обмена InfiniBand FDR [5]. Современные серверные процессоры интегрируют среду для построения высокоскоростного накопителя объёмом до 2 Тб. Новые стандарты памяти (тип DDR5) поддерживают объёмы памяти до 8 Тб. В частности, для ускорения работы большинства современных вычислительных задач достаточно 180 Гб ОЗУ [6].

Надёжность такого хранилища обеспечивается следующей особенностью: в случае, если доступный объём ОЗУ недостаточен для выполнения поставленных задач, система автоматически позиционирует файл подкачки в качестве части хранилища [7]. Данная возможность

обеспечивает сохранение работоспособности предоставленного накопителя, сопровождающееся уменьшением скорости проведения операций [7, 8].

Рассмотрим процедуры монтирования на примере узлов с процессорами Intel Xeon Phi KNL, имеющими низкую эффективность для большинства вычислительных задач, (объём ОЗУ – 208 Гб).

3. Предлагаемая конфигурация

3.1 Первоначальная настройка

Для начала работы с присоединением ОЗУ как дискового хранилища требуется в ОС семейства Linux выполнить несколько настроек.

На первом этапе необходимо описать монтирование оперативной памяти в файле конфигурации /etc/fstab, представленное на рис. 1. Директория /cache должна существовать и не содержать файлов.

```
tmpfs /cache tmpfs nodev,nosuid,size=192G 0 0
```

Рис. 1. – Конфигурация fstab

На втором этапе требуется смонтировать дополнительный раздел в файловую систему, используя команду sudo mount -a.

Затем следует проверить подключение дискового массива, обратившись к команде df -h.

После подтверждения подключения нового массива (в данном случае на 192 Гб), следует инициализировать его в сети, прописав следующую строку в /etc/exports, приведённую на рис. 2.

```
/cache          192.168.100.0/24(rw,fsid=1,sync,no_root_squash,no_all_squash)
```

Рис. 2. – Конфигурация exports

В приведённой записи 192.168.100.0/24 – сеть, в которой реализуется доступ к указанному массиву.

Осуществляется перезапуск демона nfs посредством команды sudo systemctl restart nfs.service.

Производится подключение инициализированного сетевого диска к любому компьютеру в сети. Например, IP-адрес машины, на котором смонтирована ОЗУ – 192.168.100.34, тогда в файле /etc/fstab на целевой машине получим запись, указанную на рис. 3:

```
192.168.100.34:/cache /cache nfs rsize=65536,wsize=65536,timeo=14,intr 0 0
```

Рис. 3. – Конфигурация fstab целевой машины

После этого необходимо выполнить команду mount -a. Параметры rsize и wsize подбираются в зависимости от объёма данных, определённых решаемой задачей.

Эмпирически для сети InfiniBand FDR наибольшую производительность в синтетических тестах обеспечивают значения rsize=65536 и wsize=65536.

Директория /cache на целевой машине содержит ссылку на директорию /cache узла с IP адресом 192.168.100.34, которая является частью оперативной памяти этого узла.

3.2 Синтетика

В синтетическом тесте были осуществлены следующие процедуры:

1. Создание файла большого объёма (Single File) и нескольких малых файлов того же объёма (Multiple File).

2. Сравнение скорости создания Single File и Multiple File (представлены средние результаты по 10 различным запускам).

3. Передача данных по сети с обращением одного или нескольких узлов к сетевым хранилищам.

4. Сравнение времени приёма данных с хранилищ различных видов.

Анализ результатов реализации описанных операций позволил сделать ряд заключений.

При создании файлов в подключённом по сети ОЗУ время работы составило 42.8 секунд для Single File и 44.3 секунды для Multiple File.

В случае использования HDD время создания составило 252.8 секунд и 373.7 секунды соответственно.

Для Single File преимущество ОЗУ над HDD примерно в 6 раз, для Multiple File до 8 раз, что подтверждается гистограммой на рис. 4.

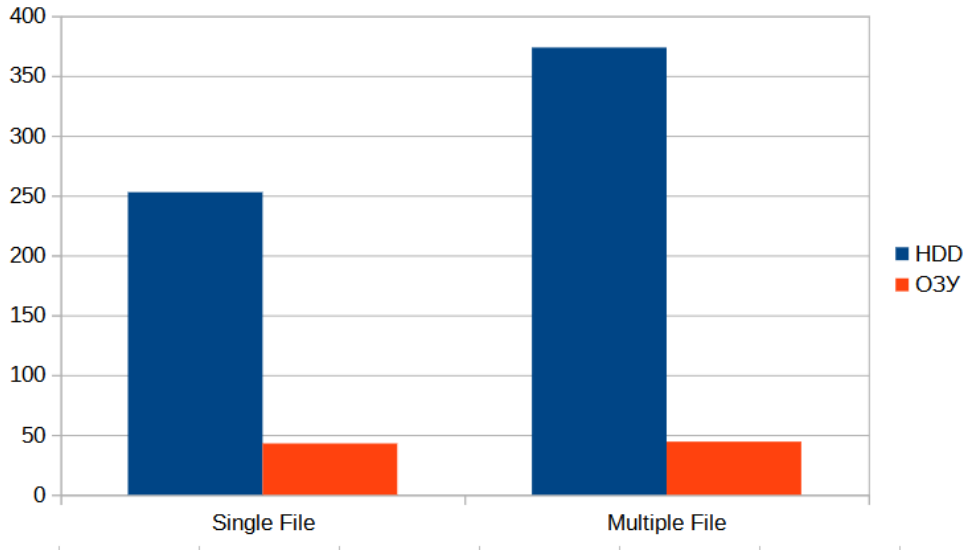

Рис. 4. – Сравнение производительности

В случае HDD заметны временные задержки, характерные для оперирования малыми файлами [9-10]. В случае с ОЗУ, аналогичные явления обусловлены, предположительно, задержкой обращения по сети.

При передаче созданных файлов по сети между узлами производится 2 типа измерений: передача данных с RAID массива и передача из оперативной памяти для вариантов с 1 или 2 принимающими узлами.

Для варианта с 1 узлом время выполнения передачи из ОЗУ составило 68 секунд, для RAID массива 384 секунды.

Для варианта с 2 узлами результаты составили 87 и 489 секунд соответственно, согласно рис 5.

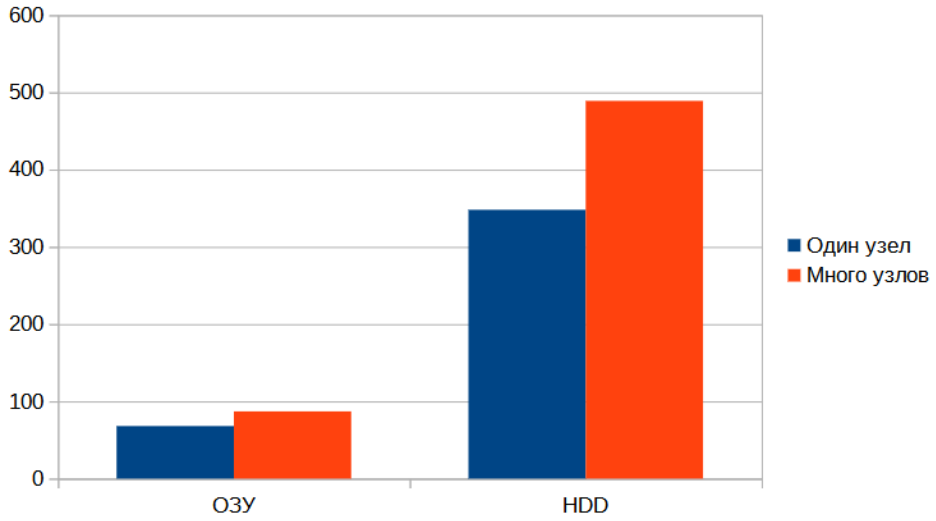

Рис. 5. – Время работы с единым и множественными файлами

Наблюдается падение скорости передачи данных ОЗУ на 27%. Соответствующие потери RAID составляют 40%. ОЗУ характеризуется более высокой скоростью.

3.3 Молекулярная динамика

Целью следующего тестирования являлось сравнение времени выполнения дисковых операций в программном комплексе Quantum Espresso (использовались запись и чтение большого массива промежуточных данных). Рассматривалось решение вычислительной задачи посредством использования пакета CP для расчётов молекулярной динамики. Аналогично предыдущему тестированию, использовались 2 типа хранилища данных.

В случае с удалённым RAID-массивом на дисках HDD время работы с файлами составило 16 минут 9 секунд; при использовании подключенного ОЗУ диска время работы с данными не превысило 7 минут 14 секунд, согласно рис. 6.

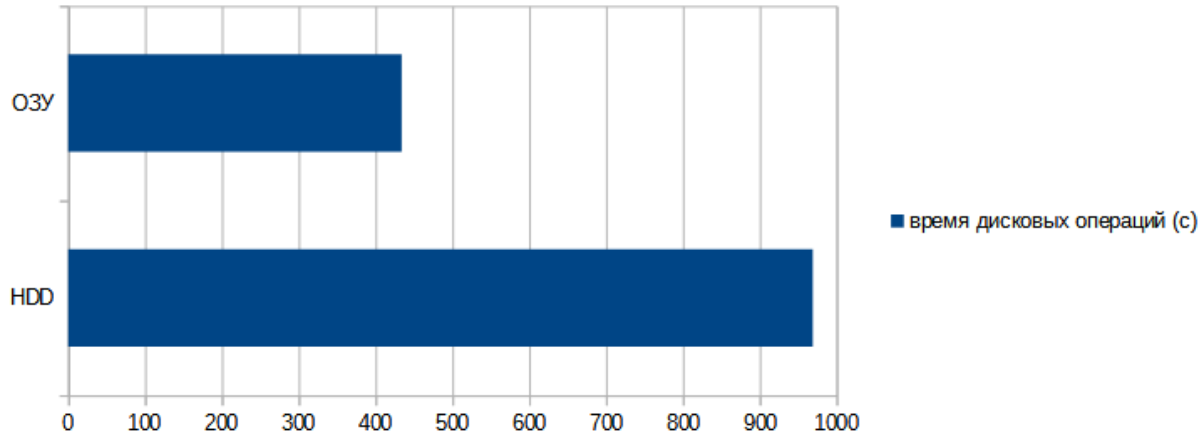

Рис. 6. – Время работы СР в различных вариантах

Наблюдалось увеличение в 2,23 раза скорости производимых файловых операций при переходе от использования HDD RAID массива к ОЗУ.

Заключение

Представленные в работе результаты тестирования на основе разработанных настроек подтверждают возможность использования ОЗУ в качестве запоминающих массивов, подключаемых посредством внутренних программных компонентов ОС семейства Linux и сетевых коммуникаций на основе сетевого протокола Network File System (NFS). Также ОЗУ может позиционироваться как средство хранения, получения и использования данных.

Положительный прирост скорости работы с данными, получаемый в этих условиях, обеспечивает минимизацию простоя ресурсов распределённой вычислительной системы. Данный эффект возникает вследствие позиционирования части узлов, как хранилищ данных и последующего использования получившихся накопителей для введения в работу дополнительных мощностей.

Недостатком системы является энергозависимость носителя информации, обуславливающая потерю данных при внезапном отключении энергии, вследствие чего сохранение данных на стандартных носителях не

может быть произведено в удовлетворительный временной период в случае потери питания.

Литература

1. Бегаев, А. А., Благовещенский Ф.Ф., Сальников А.Н. Анализ влияния сетевой топологии на задержку при передачах сообщений в вычислительном кластере // Параллельные вычислительные технологии (ПаВТ'2018) Короткие статьи и описания плакатов. 2018. С. 397-397.

2. Андреев А.Е., Егунов В.А., Завьялов Д.В., Жариков Д.Н. Развитие направления параллельных и высокопроизводительных вычислений в ВолгГТУ // Параллельные вычислительные технологии (ПаВТ'2021). Короткие статьи и описания плакатов. XV международная конференция. Челябинск, 2021. С. 151-161.

3. Олзоева, С.И. Об организации вычислительных кластеров в высших учебных заведениях // Вестник Бурятского государственного университета. 2012. № SB. С. 217-220.

4. Абдрахманов Д.Л., Жариков Д.Н., Завьялов Д.В. Разработка и внедрение системы мониторинга вычислительного кластера ВолгГТУ // Параллельные вычислительные технологии (ПаВТ'2021). Короткие статьи и описания плакатов. XV международная конференция. Челябинск, 2021. С. 277-277.

5. Горелов А.А., Майсурадзе А.И., Сальников А.Н. Анализ структуры задержек передачи информации в вычислительном кластере // Суперкомпьютерные дни в России. Труды международной конференции. Суперкомпьютерный консорциум университетов России, Федеральное агентство научных организаций России. 2015. С. 546-551.

6. Савостин И.А., Горбунов А.Н. Тестирование вычислительного кластера на базе учебной лаборатории университета // Вестник образовательного

консорциума Среднерусский университет. Информационные технологии. 2018. № 2 (12). С. 45-47.

7. Шульман В.Д., Волхонцева П.Д., Максименко О.Е. Перспективы экстенсивного роста производительности вычислительных кластеров // Вопросы устойчивого развития общества. 2022. № 2. С. 293-298.

8. Романов А.М., Попов Д.С., Стрельников О.И. Запуск задач на вычислительном кластере ВолгГТУ // Молодой ученый. 2011. № 6-1. С. 130-133.

9. Otwagin A., Pynkin D. The virtualization of computing cluster resources for integration in grid environment // MIPRO 2009 - 32nd International Convention Proceedings: Microelectronics, Electronics and Electronic Technology, MEET and Grid and Visualizations Systems, GVS. 2009. С. 261-264.

10. Overeinder B.J., Sloot P.M.A., Heederik R.N., Hertzberger L.O. A dynamic load balancing system for parallel cluster computing // Future Generation Computer Systems. 1996. Т. 12. № 1. С. 101-115.

References

1. Begaev, A. A., Blagoveshenskij F.F., Salnikov A.N. Parallelnye vychislitelnye tehnologii (PaVT'2018). Korotkie stati i opisaniya plakatov. 2018. pp. 397-397.

2. Andreev A.E., Egunov V.A., Zavyalov D.V., Zharikov D.N. Parallelnye vychislitelnye tehnologii (PaVT'2021). Korotkie stati i opisaniya plakatov. XV mezhdunarodnaya konferenciya. Chelyabinsk, 2021. pp. 151-161.

3. Olzoeva, S.I. Vestnik Buryatskogo gosudarstvennogo universiteta. 2012. № SB. pp. 217-220.

4. Abdrahmanov D.L., Zharikov D.N., Zavyalov D.V. Parallelnye vychislitelnye tehnologii (PaVT'2021). Korotkie stati i opisaniya plakatov. XV mezhdunarodnaya konferenciya. Chelyabinsk, 2021. pp. 277-277.

5. Gorelov A.A., Majsuradze A.I., Salnikov A.N. Trudy mezhdunarodnoj konferencii. Superkompyuternyj konsorcium universitetov Rossii, Federalnoe agentstvo nauchnyh organizacij Rossii. 2015. pp. 546-551.

6. Savostin I.A., Gorbunov A.N. Vestnik obrazovatelnogo konsorciuma Srednerusskij universitet. Informacionnye tehnologii. 2018. № 2 (12). pp. 45-47.

7. Shulman V.D., Volhonceva P.D., Maksimenko O.E. Voprosy ustojchivogo razvitiya obshestva. 2022. № 2. pp. 293-298.

8. Romanov A.M., Popov D.S., Strelnikov O.I. Molodoj uchenyj. 2011. № 6-1. pp. 130-133.

9. Otwagin A., Pynkin D. The virtualization of computing cluster resources for integration in grid environment MIPRO 2009 - 32nd International Convention Proceedings: Microelectronics, Electronics and Electronic Technology, MEET and Grid and Visualizations Systems, GVS. 2009. pp. 261-264.

10. Overeinder B.J., Sloot P.M.A., Heederik R.N., Hertzberger L.O. A dynamic load balancing system for parallel cluster computing Future Generation Computer Systems. 1996. T. 12. № 1. pp. 101-115.