УДК 004.272.26
DOI 10.53403/9785951505200_571

ПОИСК ОПТИМАЛЬНОЙ КОНФИГУРАЦИИ СЕТЕВОЙ ФАЙЛОВОЙ СИСТЕМЫ НА МНОГОПРОЦЕССОРНОЙ ВЫЧИСЛИТЕЛЬНОЙ СИСТЕМЕ

Емельянова Светлана Сергеевна (SSEmelyanova @vniief.ru),
Ильченко Евгений Александрович

ФГУП «РФЯЦ-ВНИИЭФ», г. Саров Нижегородской обл.

В докладе представлены материалы научно-исследовательской работы по оптимизации файловой системы многопроцессорной вычислительной системы (МВС). Представлены и обоснованы разработанные конфигурации, описан опыт внедрения разработанных конфигураций и приведены результаты их тестирования. Сделаны выводы о применимости выбранной файловой системы в конкретной кластерной системе.

Ключевые слова: файловые системы, многопроцессорная вычислительная система, производительность, масштабируемость, отказоустойчивая конфигурация, Lustre, NFS, оптимизация.

RESEARCH OF OPTIMAL NETWORK FILE SYSTEM OPTIMAL CONFIGURATION AT MULTIPROCESSOR COMPUTE SYSTEM

Emelyanova Svetlana Sergeevna, Ilchenko Evgeny Aleksandrovich

FSUE “RFNC-VNIIEF”, Sarov Nizhny Novgorod region

This paper contains research materials on optimization of the multiprocessor computer system. Authors present and substantiate developed configurations, describe experience of introduction of the developed configurations to the computing system, as well as present the test results. Authors conclude the applicability of the selected file system in specific clustered system.

Key words: file systems, multiprocessor computing system, performance, scalability, fault-tolerant configuration, Lustre, NFS, optimization.

Введение

Одним из этапов проектирования сложных многокомпонентных изделий является расчетное обоснование их характеристик и параметров. Для решения этих задач применяются методы инженерного анализа, предполагающие проведение ряда вычислительных экспериментов. Такие эксперименты требуют для получения результатов большого количества вычислительных ресурсов. Поэтому, в таких случаях широко используются высокопроизводительные кластеры и суперкомпьютерные системы.

Высокопроизводительные кластеры представляют собой сложные технические системы. Важным компонентом высокопроизводительного кластера является распределенная сетевая файловая система, которая обеспечивает совместное использование файловых ресурсов узлами кластера.

В одном из расчетных кластеров ФГУП «РФЯЦ-ВНИИЭФ» для организации совместного доступа к файловым ресурсам, размещенным на выделенной системе хранения данных, используется протокол сетевого доступа NFS. Эксплуатация МВС и исследование мировых практик построения высокопроизводительных вычислительных комплексов позволили выявить возможности повышения эффективности использования исследуемого расчетного комплекса за счет совершенствования распределенной сетевой файловой системы, обеспечивающей совместное использование файловых ресурсов узлами кластера.

Перед началом работы были поставлены задачи:

– разработать варианты улучшения конфигурации, представленной в 2019 году;

– обеспечить рост производительности МВС на операциях ввода/вывода не менее чем на 10 %;

– обеспечить отказоустойчивость и расширить аппаратное обеспечение МВС без потери производительности.

Для решения поставленных задач, проведено исследование возможностей модернизации разрабо-

танной в 2019 году конфигурации с файловой системой (ФС) Lustre[1], разработаны и исследованы тестовые конфигурации, выбран оптимальный вариант.

Исходная конфигурация

МВС состоит из управляющего узла (УУ) и 14 вычислительных узлов (ВУ), распределенных по трем разделам в соответствии с конфигурацией используемого в них оборудования [1]:

– Раздел 1 (9 ВУ);

– Раздел 2 (4 ВУ);

– Раздел 3 (1 ВУ).

Структурная схема МВС представлена на рис. 1

Сетевая файловая подсистема МВС предназначена для решения следующих задач:

– доступ с ВУ к установленным программам для расчета,

– доступ с ВУ к исходным данным для расчета,

– сохранение результатов при параллельном счете.

Учитывая масштаб МВС и простоту развертывания, для организации совместного доступа к файловым ресурсам был выбран протокол сетевого доступа NFS версий 3 и 4, дополнительно усиленный функционалом системы хранения данных[2] (СХД).

Физическая схема МВС представлена на рис. 2.

ВУ выступают в роли клиентов NFS, а УУ в роли сервера NFS.

Конвергентный коммутатор объединяет две технологии: 10 Gigabit Ethernet и Fibre Channel.

СХД предоставляет ресурсы для всех операций дискового ввода/вывода на узлах МВС:

– для хранения данных операционных систем ВУ;

– для хранения исходных данных и результатов вычислений;

– для хранения используемых для вычислений пакетов программ (расчетное программное обеспечение, библиотеки MPI и т. д.).

Несмотря на то, что данная конфигурация легка в развертывании, она также имеет следующие недостатки [1]:

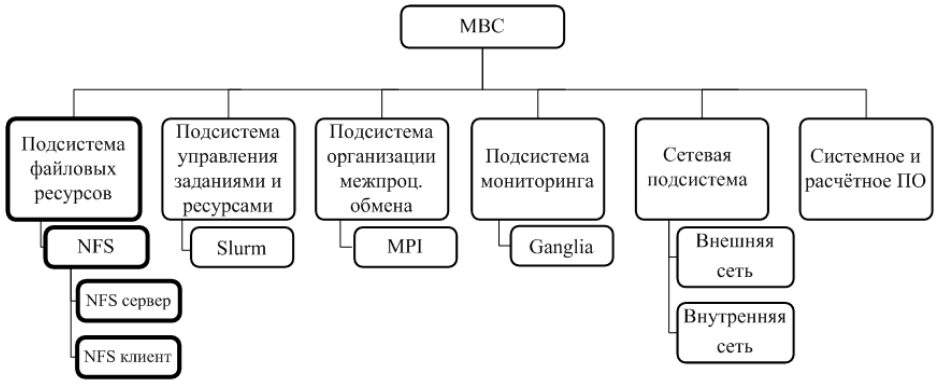

Рис. 1. Структурная схема МВС

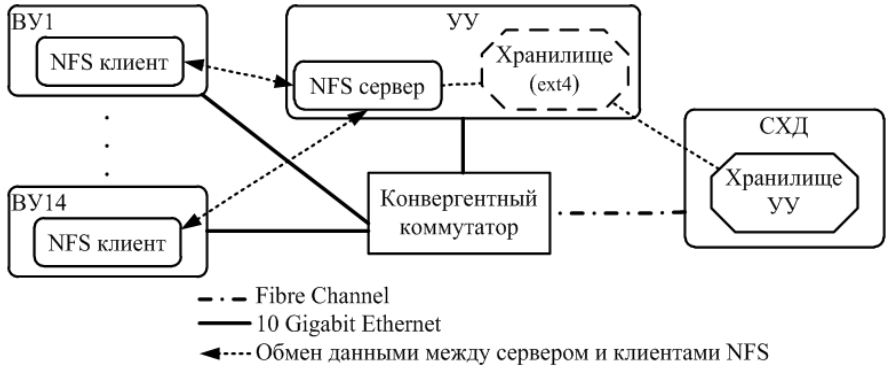

Рис. 2. Физическая схема МВС с компонентами NFS

[1] ФС Lustre является свободно распространяемой файловой системой с открытым исходным кодом.

[2] Технология UltraPath – разработка компании Huawei, предоставляемая совместно с оборудованием компании, позволяет обращаться к данным на СХД в 3–4 раза быстрее за счет одновременного использования обоих контроллеров СХД и многопоточным обращениям через каждый контроллер к этому дисковому массиву.

– Единая точка отказа. При отказе NFS сервера, все вычислительные серверы теряют данные, хранящиеся на данном ресурсе;

– Недостаточная производительность. Протокол сетевого доступа NFS является последовательным, соответственно способен предоставить меньшую производительность, чем параллельные файловые системы;

– Недостаточная масштабируемость. Для организации одного общего файлового ресурса использовать больше одного NFS сервера невозможно, а при увеличении количества клиентов, использующих данный ресурс, сильно снижается производительность такой системы.

Предыдущее исследование

В 2019 г. было проведено исследование по разработке оптимальной конфигурации на данной МВС. В ходе проведения исследования было разработано три схемы расположения компонент ФС Lustre на МВС. Данные конфигурации были внедрены и протестированы. Проведено сравнение результатов тестирования новых конфигураций с используемой конфигурацией ФС Lustre на МВС. В результате проведенного исследования была разработана конфигурация, которая дает преимущество в производительности приблизительно на 10 % по сравнению с используемой конфигурацией.

Конфигурация МВС с ФС Lustre

Конфигурация МВС с ФС Lustre состоит из активно/активных пар и является отказоустойчивой, однако под серверы ФС Lustre задействованы не все ВУ МВС. Состав данной конфигурации представлен в табл. 1. Структурная схема конфигурации МВС с ФС Lustre представлена на рис. 3 [1].

Таблица 1

Составные части серверных компонент ФС Lustre второй конфигурации

№	Серверы Lustre	Выделенные хранилища
1	MDS_0	MDT_0, емкостью 400ГБ
2	MDS_1	MDT_1, емкостью 400ГБ
3	OSS_0	OST_0, емкостью 2 ТБ
4	OSS_1	OST_1, емкостью 2 ТБ
5	OSS_2	OST_2, емкостью 2 ТБ
6	OSS_3	OST_3, емкостью 2 ТБ
7	MGS_0	MGS, емкостью 100ГБ
8	MGS_1	пассивный сервер для хранилища MGT_0

Преимущества конфигурации МВС с ФС Lustre:

– Отказоустойчивость. Задействована только часть вычислительных серверов под активно/активные пары таким образом, чтобы при необходимости отключения любого из разделов, не произошло потери данных, и расчеты не были остановлены;

– Производительность. Ожидалось повышение производительности за счет распределения серверных компонент ФС Lustre на разные узлы МВС.

Недостатки данной конфигурации:

– Сложность первоначальной настройки.

Конфигурация МВС с ФС Lustre разработана и тестировалась в ходе исследования, проводимого в 2019 году. На тот момент эта конфигурация показала, в большинстве своем, наилучшие результаты при тестировании типовыми задачами.

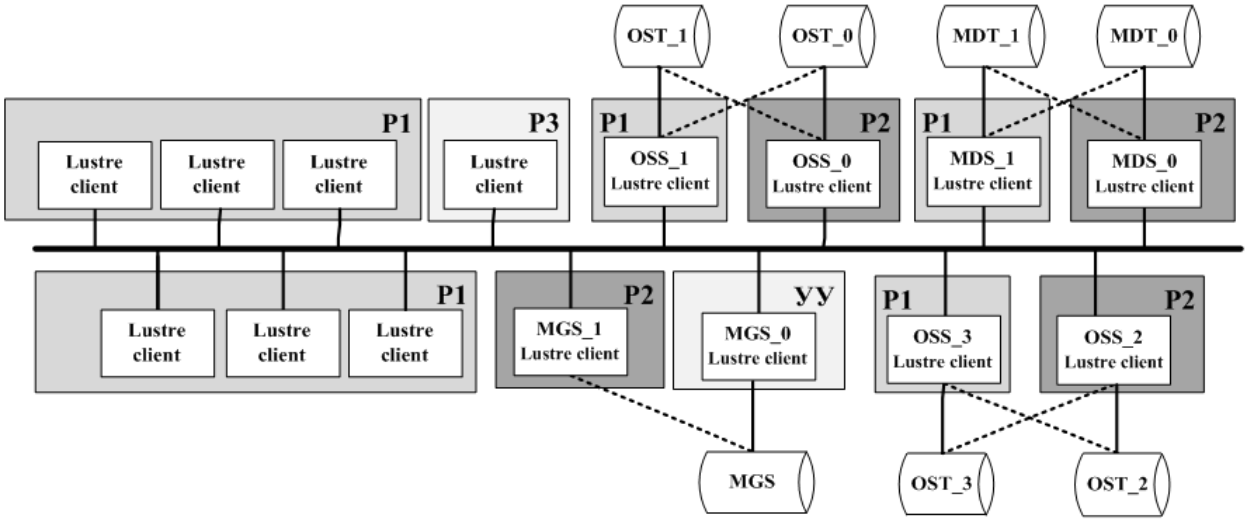

Рис. 3. Структурная схема конфигурации МВС с ФС Lustre

Возможные варианты улучшения конфигурации МВС с ФС Lustre

Выбранную в ходе исследования в 2019 году конфигурацию МВС с ФС Lustre можно улучшить за счет использования дополнительных технологий, таких как MultiPath и Stripping.

Multipath – способ доступа к устройству массового хранения, при котором к нему от центрального процессора ведет несколько путей (физических и/или логических). Типичный пример – хранилище, одновременно подключенное к нескольким SCSI-портам компьютера.

Multipath позволяет объединить несколько маршрутов ввода-вывода между серверами и дисковыми массивами в единое целое. Маршруты в этом случае представляют собой физические SAN-соединения, которые могут включать отдельные кабели, переключатели и контроллеры. В результате агрегации будет создано новое устройство.

Striping – дополнительная технология, используемая в кластерах с ФС Lustre. Данная технология обеспечивает чередование данных между хранилищами в циклическом режиме. Пользователи могут опционально сконфигурировать для каждого файла количество полос, размер полосы, и хранилища, которые будут использоваться. По существу, файлы могут быть расщеплены на множество частей, которые затем будут храниться на разных хранилищах ФС Lustre.

Чередование может быть использовано для повышения производительности, когда совокупная пропускная способность к одному файлу превышает пропускную способность одного хранилища. Возможность чередования также полезна, когда одно из хранилищ не имеет достаточно свободного места для размещения всего файла.

Чередование позволяет сохранять сегменты или «куски» данных файла на различных хранилищах. В файловой системе Lustre используется модель RAID 0, согласно которой «куски» чередуются между определенным числом объектов. Число объектов в одном файле указывается через stripe_count.

Каждый объект содержит «кусок» данных из файла. Когда «кусок» данных, подлежащих записи в определенный объект превышает stripe_size, следующий «кусок» данных в файле сохраняется в следующий объект.

При внедрении разработанных конфигураций, чередование было выключено. Включалось чередование после прогона всех тестов. После включения чередования, все разработанные конфигурации показали результаты хуже, чем конфигурации без чередования. Причину данного поведения пока выявить не удалось.

В связи с тем, что чередование не показало должных результатов, дальнейшее тестирование проводилось только с использованием технологии Multipath.

Конфигурация МВС с ФС Lustre с использованием MultiPath

Технология Multipath внедрялась в конфигурацию МВС с ФС Lustre, которая является оптимальной по результатам исследования 2019 года. Структурная схема конфигурации МВС с ФС Lustre с использованием MultiPath [2] представлена на рис. 4.

Расположение компонент в структурной схеме конфигурации МВС с ФС Lustre с MultiPath аналогично схеме конфигурации МВС с ФС Lustre конфигурации без Multipath. Единственное отличие в том, что выделенные диски для данных показывались в нескольких экземплярах на серверах.

Преимущества конфигурации МВС с ФС Lustre с MultiPath остаются такими же, как и в конфигурации без использования MultiPath: отказоустойчивость и производительность. Недостатком данной конфигурации является сложность первоначальной настройки.

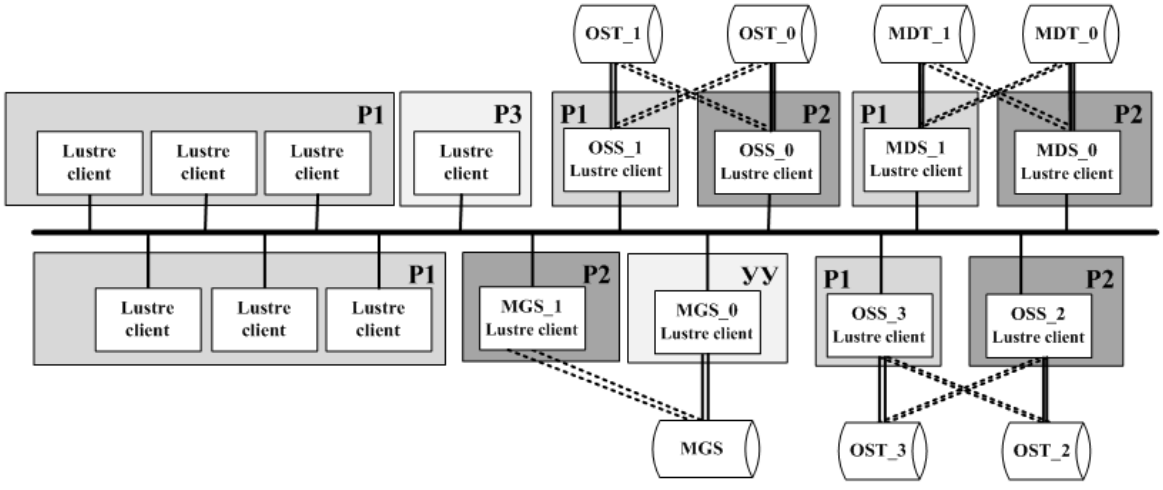

Рис. 4. Структурная схема МВС с ФС Lustre с MultiPath

Сравнительное тестирование разработанных конфигураций МВС

Цель тестирования: выявление конфигурации МВС с большей производительностью.

Тестирование проводилось на следующих конфигурациях МВС:

1. Конфигурация с NFS и UltraPath;
2. Конфигурация с ФС Lustre (4 группы по 2 сервера);
3. Конфигурация с ФС Lustre+MultiPath (4 группы по 2 сервера).

Для каждой конфигурации МВС тестирование проводилось в два этапа:

– тестирование синтетическими тестами;
– тестирование типовыми задачами.

Тестирование синтетическими тестами

Данное тестирование осуществлялось с помощью тестовых утилит: IOZone и Bonnie++. Каждый вид теста проделывался в общей сложности пять раз. Все синтетические тесты запускались на МВС, не занятой другими вычислительными задачами. В разных конфигурациях МВС тесты запускались на одинаковых узлах.

Тесты IOZone запускались со следующими ключами: –i0 –i1 –s300g –r64k.

-i0 -i1 – данные параметры выполняют операции чтения/записи файла, а так же повторного чтения и перезапись того же файла;

-s300g – выбран файл, размером в 300 ГБ, так как данный размер превышал доступный объем оперативной памяти на ВУ и был достаточным для корректного проведения теста;

-r64k – выбран размер блока в 64 КБ.

По окончанию теста IOZone выводит значения скорости записи, перезаписи, чтения и повторного чтения в КБ/с. Так же использовался ключ – O, для вывода результатов в iops[3].

Тест Bonnie++ выводит значения скорости создания, чтения и удаления при последовательной записи файлов и при записи файлов в случайном порядке.

Тест Bonnie++ запускался со следующими параметрами:

1500:128:0:40 – создание 1500*1024 файлов, максимальный размер 128 КБ, минимальный размер 0 байт, 40 поддиректорий.

Тестирование типовыми задачами (нагрузочное тестирование)

В качестве нагрузочного тестирования использовалось тестирование типовыми задачами. Сравнивалось время счета задач при полной загрузке МВС.

Для нагрузочного тестирования подобраны задачи для одного из часто используемого решателя в подразделениях РФЯЦ ВНИИЭФ. Выбраны такие задачи, которые генерируют большое количество файлов с результатами, так, чтобы вычислительные узлы постоянно обращались к хранилищу для записи файлов.

Запуск задач проводился с нагрузкой на весь кластер[4] по 5 прогонов для каждой конфигурации. Распределение ядер на тестовые задачи представлено в табл. 2.

Таблица 2

Распределение ядер на тестовые задачи

№	Раздел	Кол-во ядер	№	Раздел	Кол-во ядер
1	Раздел 1	64	5	Раздел 3	16
2	Раздел 1	32	6	Раздел 2	64
3	Раздел 1	32	7	Раздел 2	32
4	Раздел 1	16			

Выбрано именно такое разделение процессорных мощностей для того, чтобы можно было сравнивать не только конфигурации МВС между собой, но также сравнивать время расчета задач на одинаковом количестве ядер на разных разделах внутри каждой конфигурации МВС.

Результаты тестирования синтетическими тестами

Результаты тестирования синтетическими тестами представлены в табл. 3–5 (больше – лучше).

[3] iops – количество операций ввода-вывода в секунду.

[4] Задачи распределены так, что занимают все процессорные мощности МВС.

Таблица 3

Результаты тестирования тестом bonnie++

		Последовательная запись			Запись в случайном порядке		
№	ВУ	Запись (КБ/с)	Чтение (КБ/с)	Удаление (КБ/с)	Запись (КБ/с)	Чтение (КБ/с)	Удаление (КБ/с)
NFS+UltraPath							
1	Раздел 1	515	60	1561	515	130	1501
2	Раздел 2	532	75	1603	603	105	1625
3	Раздел 3	363	56	1416	563	56	1553
Lustre							
4	Раздел 1	505,6	261,8	1516,4	508,4	247,8	2456,8
5	Раздел 2	512,2	281,2	1534,4	510,6	269,6	2431,2
6	Раздел 3	397,8	103,4	1480	490,8	107,8	2019,2
Lustre + MultiPath							
7	Раздел 1	542	710	1387	543	867	2685
8	Раздел 2	554	403	1189	553	389	2245
9	Раздел 3	526	408	1349	533	377	2619

Таблица 4

Результаты тестирования тестом IOZone (в iops)

№	ВУ	Запись (iops)	Повторная запись (iops)	Чтение (iops)	Повторное чтение (iops)
NFS+UltraPath					
1	Раздел 1	2749	2798	3793	2913
2	Раздел 2	3810	3490	2093	2101
3	Раздел 3	3105	3201	2109	2239
Lustre					
4	Раздел 1	8718	8776	5838	5657
5	Раздел 2	9846	9737	6419	6643
6	Раздел 3	8514	8664	4431	4314
Lustre + MultiPath					
7	Раздел 1	5841	6240	5792	5857
8	Раздел 2	8382	8507	5277	5464
9	Раздел 3	8062	8163	4750	4671

Таблица 5

Результаты тестирования тестом IOZone (в KB/s)

№	ВУ	Запись (КБ/с)	Повторная запись (КБ/с)	Чтение (КБ/с)	Повторное чтение (КБ/с)
NFS+UltraPath					
1	Раздел 1	162356	460921	110354	223496
2	Раздел 2	136521	432961	190689	210641
3	Раздел 3	69782	150365	125684	121635
Lustre					
4	Раздел 1	556539	554429	297637	292529
5	Раздел 2	580412	627216	497795	419338
6	Раздел 3	530777	546663	261620	261988
Lustre + MultiPath					
7	Раздел 1	403372	420547	355980	330321
8	Раздел 2	569295	580732	335248	337113
9	Раздел 3	569310	567094	355214	366447

Результаты тестирования типовыми задачами

Результаты тестирования типовыми задачами представлены в табл. 6.

Таблица 6

Результаты тестирования типовыми задачами

№	Кол-во ядер	Раздел	NFS+UltraPath	Lustre	Lustre + MultiPath
1	64	Раздел 1	542м.	683м 36с	540м 30с
2	32	Раздел 1	918м. 52с	879м 2с	738м 18с
3	32	Раздел 1	988м. 44с.	977м 36с	943м 36с
4	16	Раздел 1	1426м. 6с.	1329м 1с	1075м 54с
5	16	Раздел 3	2199м. 35с.	2979м 7с	2321м 18с
6	64	Раздел 2	744м.17с.	723м 18с	676м 42с
7	32	Раздел 2	1303м. 24с.	1026м 4с	886м

Выводы по результатам тестирования

По результатам тестирования ФС Lustre + Multipath было отмечено, что такая конфигурация имеет значительный прирост в производительности при тестировании типовыми задачами, небольшой прирост при тестировании с помощью синтетического теста bonnie++. Однако, по тестированию тестом iozone улучшенная конфигурация уступает предыдущим конфигурациям.

Таким образом, можно сделать вывод, что конфигурация с технологией Multipath является оправданным выбором для вычисления задач.

В табл. 7 показано отклонение (в процентах) от результатов тестирования типовыми задачами исходной конфигурации.

Таблица 7

Сравнительная таблица тестирования типовыми задачами (меньше-лучше)

№	Кол-во ядер	Раздел	Lustre	Lustre+MultiPath	NFS+UltraPath
1	64	Раздел 1	+26 %	-0,5 %	542м.
2	32	Раздел 1	-4 %	-19 %	918м. 52с
3	32	Раздел 1	-1 %	-4,5 %	988м. 44с.
4	16	Раздел 1	-6 %	-24 %	1426м. 6с.
5	16	Раздел 3	+35 %	+5 %	2199м. 35с.
6	64	Раздел 2	-2 %	-9,1 %	744м.17с.
7	32	Раздел 2	-21 %	-32 %	1303м. 24с.

Заключение

В результате проведенных работ:

– проанализирована текущая конфигурация МВС в части сетевой файловой системы;

– проанализирована конфигурация МВС с ФС Lustre, выбранная в результате предыдущего исследования;

– исследованы варианты улучшения производительности конфигурации МВС с ФС Lustre;

– спроектирована улучшенная конфигурация МВС с ФС Lustre;

– выполнено сравнительное тестирование конфигураций МВС:

– Конфигурация с NFS и UltraPath;

– Конфигурация с ФС Lustre (4 группы по 2 сервера);

– Конфигурация с ФС Lustre+MultiPath (4 группы по 2 сервера).

В итоге можно сделать вывод, что применение ФС Lustre на МВС целесообразно. Разработанная в рамках данного исследования конфигурация МВС с ФС Lustre+MultiPath может быть внедрена в МВС для повышения производительности, отказоустойчивости и масштабируемости файловой системы при дальнейшем использовании.

Список литературы

1. Емельянова С. С., Ильченко Е. А. Поиск оптимальной конфигурации сетевой файловой системы на многопроцессорной вычислительной системе № 1.2 / 17-я научно-техническая конференция «Молодежь в науке» // Сборник докладов. Саров: ФГУП «РФЯЦ-ВНИИЭФ», 2019.