ПОИСК ОПТИМАЛЬНОЙ КОНФИГУРАЦИИ СЕТЕВОЙ ФАЙЛОВОЙ СИСТЕМЫ НА МНОГОПРОЦЕССОРНОЙ ВЫЧИСЛИТЕЛЬНОЙ СИСТЕМЕ КБ-2

С. С. Емельянова, Е. А. Ильченко

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

В докладе представлены материалы научно-исследовательской работы по оптимизации файловой системы многопроцессорной вычислительной системы КБ-2 (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.

Введение

Одним из этапов проектирования сложных многокомпонентных изделий является расчетное обоснование характеристик и параметров изделий. Для решения этих задач применяются методы инженерного анализа, предполагающие проведение ряда вычислительных экспериментов. Вычислительные эксперименты являются ресурсоемкими и требуют для получения результатов большого количества вычислительных ресурсов. 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 КБ-2 разработан и эксплуатируется высокопроизводительный кластер МВС. Расчетный комплекс МВС был разработан [1] и введен в постоянную эксплуатацию в 1 квартале 2017 года. Эксплуатация МВС и исследование мировых практик построения высокопроизводительных вычислительных комплексов позволили выявить возможности повышения эффективности использования расчетного комплекса КБ-2 за счет совершенствования распределенной сетевой файловой системы, обеспечивающей совместное использование файловых ресурсов узлами кластера.

Перед началом работы была поставлена задача:

– разработать конфигурацию сетевой файловой системы, которая обеспечит рост производительности МВС на операциях ввода/вывода не менее чем на 8 %;

– конфигурация должна быть отказоустойчивой и позволять расширять аппаратное обеспечение МВС без потери производительности.

Для решения поставленной задачи, проведено исследование возможности повышения отказоустойчивости, производительности и масштабируемости распределенной файловой подсистемы МВС, разработаны и исследованы тестовые конфигурации, выбран и внедрен в эксплуатацию оптимальный вариант.

Исходная конфигурация

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

– раздел 1 (9 вычислительных узлов);

– раздел 2 (4 вычислительных узла);

– раздел 3 (1 вычислительный узел).

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 (технология UltraPath от компании Huawei позволяла обращаться к СХД в 3–4 раза быстрее за счет одновременного использования обоих контроллеров СХД и многопоточным обращениям через каждый контроллер к дисковому массиву).

Структурная схема МВС представлена на рис. 1.

ВУ выступают в роли клиентов NFS, а УУ в роли сервера NFS. Физическая схема МВС представлена на рис. 2. 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 ОС ВУ;

– для хранения исходных данных и результатов вычислений;

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

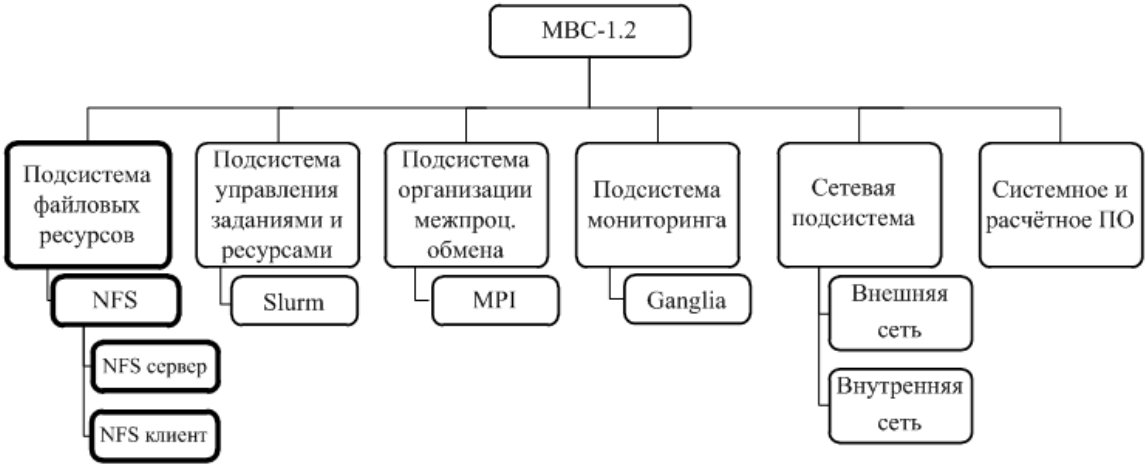

Рис. 1. Структурная схема МВС

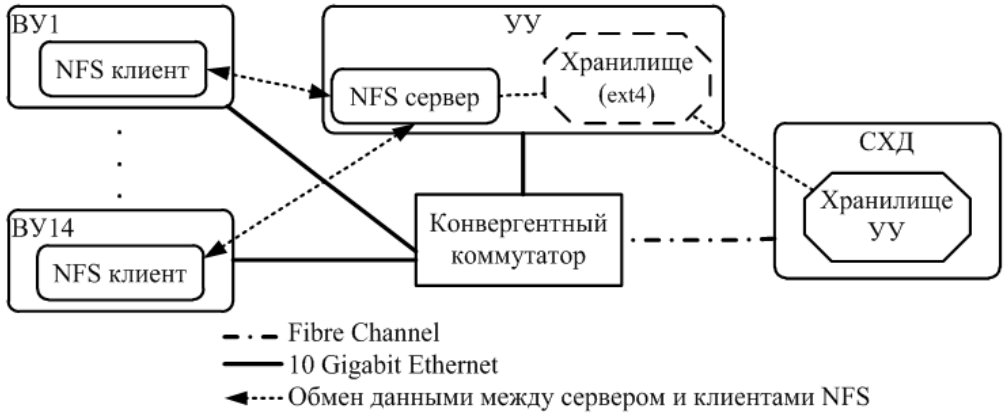

Рис. 2. Физическая схема МВС с компонентами NFS

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

– 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.

Выбор файловой системы

При выборе сетевой файловой системы для внедрения в существующую МВС предъявлялись следующие основные требования:

– высокая отказоустойчивость;

– возможность расширения готовой системы;

– открытый исходный код;

– высокая эффективность использования ресурсов.

В рамках предыдущей работы [2] были рассмотрены следующие файловые системы (далее – ФС):

– Ceph;

– GPFS;

– Gluster;

– NFS;

– Lustre.

Из всех рассмотренных вариантов файловых систем, ФС Lustre в наибольшей степени удовлетворяет всем перечисленным выше требованиям.

ФС Lustre имеет ряд преимуществ [2, 3, 4, 5]:

– высокая доступность сетевой ФС Lustre включает надежный механизм отказоустойчивости и восстановления, обеспечивающий прозрачную перезагрузку серверов при неисправности.

– масштабируемость. Увеличение серверных компонент ФС Lustre положительно влияет на производительность файловой системы.

– производительность сетевой файловой системы МВС по предварительным оценкам, основанным на изученной литературе [4, 5], возрастет на 10-15% за счет распределенной архитектуры и параллельной работы нескольких клиентов ФС Lustre.

После детального изучения возможностей ФС Lustre и МВС были разработаны три схемы расположения компонент ФС Lustre на МВС. Далее описаны все разработанные конфигурации, а также выделены плюсы и минусы разработанных конфигураций.

Конфигурация № 1 с ФС Lustre

Данный вариант конфигурации предполагал размещение всех серверных компонент на управляющем узле МВС. Схема конфигурации №1 представлена на рис. 3.

Преимущества данной конфигурации:

– простота развертывания.

Недостатки данной конфигурации:

– единая точка отказа.

Конфигурация № 2 с ФС Lustre

В данном варианте задействованы все ВУ из раздела 1 и раздела 2 (Р1 и Р2 на рис. 4 и 5). Данная конфигурация строилась на основе активно/активных пар. Всего 7 пар, из них 5 активно/активных пар серверов хранения данных, одна активно/активная пара серверов метаданных и одна активно/пассивная пара серверов управления (см. табл. 1).

Преимущества данной конфигурации:

– отказоустойчивость. Данную конфигурацию можно считать отказоустойчивой, за счет построения сервисов в парах активный/активный. Если один узел выходит из строя, его хранилищем станет управлять второй. Так же, сервисы метаданных расположены на самых последних узлах в нумерации СУЗ, вследствие чего меньше загружены вычислениями.

– производительность. В данном случае, повышение производительности файловой системы ожидалось за счет большого количества серверных компонент, распределенных по разным серверам МВС.

Недостатки данной конфигурации:

– сложность первоначальной настройки.

Таблица 1

Составные части серверных компонент ФС Lustre первой 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	OSS_4	OST_4, емкостью 2 ТБ
8	OSS_5	OST_5, емкостью 2 ТБ
9	OSS_6	OST_6, емкостью 2 ТБ
10	OSS_7	OST_7, емкостью 2 ТБ
11	OSS_8	OST_8, емкостью 2 ТБ
12	OSS_9	OST_9, емкостью 2 ТБ
13	MGS_0	MGT_0, емкостью 100ГБ
14	MGS_1	пассивный сервер для хранилища MGT_0

Структурная схема данной конфигурации представлена на рис. 4.

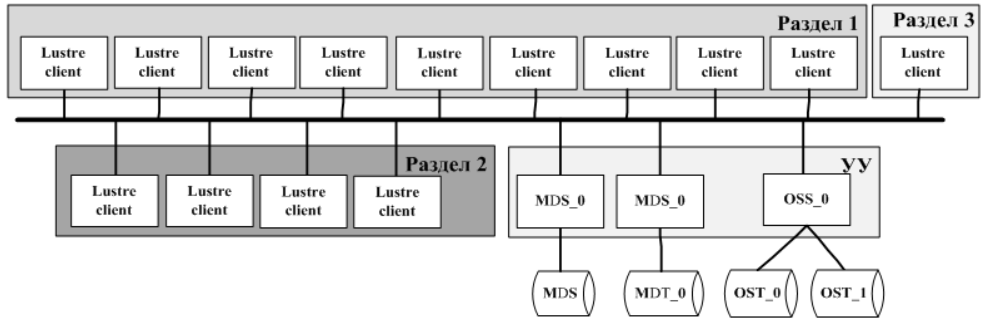

Рис. 3. Структурная схема конфигурации № 1 с ФС lustre

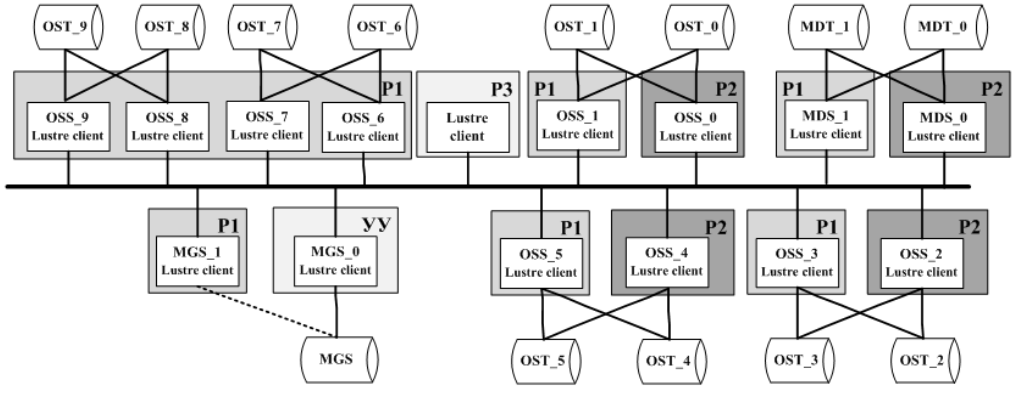

Рис. 4. Структурная схема конфигурации № 2 МВС с ФС Lustre

Конфигурация № 3 с ФС Lustre

Данная конфигурация состоит из активно/активных пар и является отказоустойчивой, однако под серверы ФС Lustre задействованы не все вычислительные серверы МВС. Состав данной конфигурации представлен в табл. 2.

Таблица 2

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	MGT_0, емкостью 100ГБ
8	MGS_1	пассивный сервер для хранилища MGT_0

Структурная схема данной конфигурации представлена на рис. 5.

Преимущества данной конфигурации:

– 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.

Сравнительное тестирование разработанных конфигураций

Цель тестирования: выявление наиболее производительной конфигурации.

Тестирование проводилось на следующих конфигурациях.

1. Конфигурация с NFS и UltraPath;
2. Конфигурация № 1 с ФС Lustre (все серверные части ФС Lustre на управляющем узле);
3. Конфигурация № 2 с ФС Lustre (7 групп по 2 сервера);
4. Конфигурация № 3 с ФС Lustre (4 группы по 2 сервера).

Для каждой конфигурации тестирование проводилось в два этапа:

– тестирование синтетическими тестами;

– тестирование типовыми задачами.

Тестирование синтетическими тестами

Данное тестирование осуществлялось с помощью тестовых утилит: IOZone и Bonnie++ (см. табл. 3–5). Каждый вид теста проделывался в общей сложности пять раз. Все синтетические тесты запускались на МВС, не занятой другими вычислительными задачами. В разных конфигурациях 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[1].

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.

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

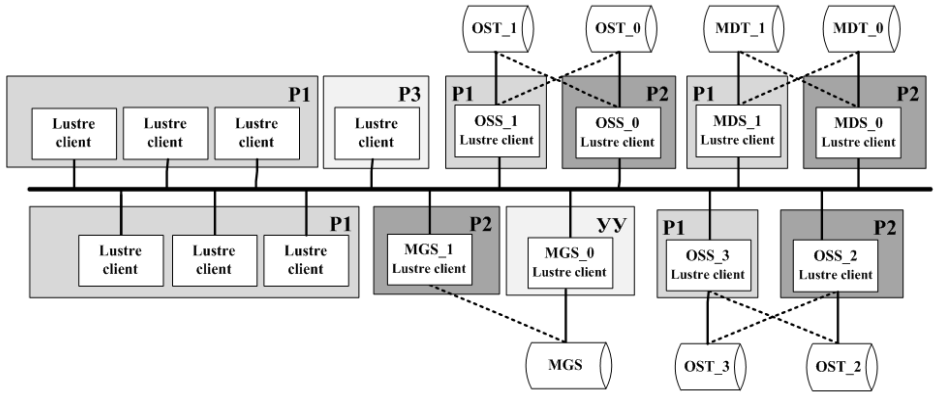

Рис. 5. Структурная схема конфигурации № 2 МВС с ФС Lustre

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 (1 вариант конфигурации)							
4	Раздел 1	506	45	1444	497	56	1476
5	Раздел 2	525	46	1528	574	53	1495
6	Раздел 3	351	44	1385	494	43	1519
Lustre (2 вариант конфигурации)							
7	Раздел 1	481	629	1281	498	622	2356
8	Раздел 2	525	306	1137	484	311	2099
9	Раздел 3	416	282	1090	425	290	2035
Lustre (3 вариант конфигурации)							
10	Раздел 1	505,6	261,8	1516,4	508,4	247,8	2456,8
11	Раздел 2	512,2	281,2	1534,4	510,6	269,6	2431,2
12	Раздел 3	397,8	103,4	1480	490,8	107,8	2019,2

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 (1 вариант конфигурации)					
4	Раздел 1	2593	2786	3708	2818
5	Раздел 2	3663	3392	1964	1999
6	Раздел 3	2845	3083	2078	2184
Lustre (2 вариант конфигурации)					
7	Раздел 1	8341	8704	6158	5638
8	Раздел 2	9433	10023	7396	7356
9	Раздел 3	8349	8495	7155	6793
Lustre (3 вариант конфигурации)					
10	Раздел 1	8718	8776	5838	5657
11	Раздел 2	9846	9737	6419	6643
12	Раздел 3	8514	8664	4431	4314

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 (1 вариант конфигурации)					
4	Раздел 1	154269	458615	208966	213265
5	Раздел 2	139230	432825	188419	201863
6	Раздел 3	43021	78897	111946	118412
Lustre (2 вариант конфигурации)					
7	Раздел 1	540937	541185	403108	452205
8	Раздел 2	573539	598109	321370	311963
9	Раздел 3	497317	521161	415445	409880
Lustre (3 вариант конфигурации)					
10	Раздел 1	556539	554429	297637	292529
11	Раздел 2	580412	627216	497795	419338
12	Раздел 3	530777	546663	261620	261988

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 КБ-2. 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[2] по 5 прогонов для каждой конфигурации (см. табл. 6).

Таблица 6

Распределение ядер на тестовые задачи

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

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

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

Результаты тестирования типовыми задачами

Результаты тестирования типовыми задачами представлены в табл. 7.

Выводы по результатам тестирования

По результатам тестирования, как синтетическими тестами, так и типовыми задачами, разработанные конфигурации с ФС Lustre показывают лучшие результаты в счете задач и в синтетических тестах. В частности, конфигурация № 3 с ФС Lustre показала небольшое преимущество при счете тестовых задач, а результаты синтетических тестов конфигурации № 3 с ФС Lustre превосходят результаты тестов других конфигураций приблизительно на 30 % (см. табл. 8, 9).

Таблица 7

Результаты тестирования типовыми задачами

№	Кол-во ядер	Раздел	Конф-я № 1 Lustre	Конф-я № 2 Lustre	Конф-я № 3 Lustre	NFS+UltraPath
1	64	Раздел 1	796 м 12 с	731 м 47 с	683 м 36 с	542 м
2	32	Раздел 1	887 м 36 с	1070 м 53 с	879 м 2 с	918 м 52 с
3	32	Раздел 1	1412 м 24 с	1067 м 28 с	977 м 36 с	988 м 44 с
4	16	Раздел 1	1821 м 36 с	1213 м 58 с	1329 м 1 с	1426 м 6 с
5	16	Раздел 3	2886 м 48 с	2510 м 19 с	2979 м 7 с	2199 м 35 с
6	64	Раздел 2	853 м	781 м 44 с	723 м 18 с	744 м 17 с
7	32	Раздел 2	1263 м 48 с	1169 м 21 с	1026 м 4 с	1303 м 24 с

Таблица 8

Сравнительная таблица теста IOZone в iops (больше – лучше)

№	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 (1 вариант конфигурации)					
4	Раздел 1	–5 %	–0,4 %	–2 %	–3 %
5	Раздел 2	–3 %	–2 %	–6 %	–4 %
6	Раздел 3	–8 %	–3 %	–1 %	–2 %
Lustre (2 вариант конфигурации)					
7	Раздел 1	+203 %	+211 %	+62 %	+93 %
8	Раздел 2	+147 %	+187 %	+253 %	+250 %
9	Раздел 3	+168 %	+165 %	+239 %	+203 %
Lustre (3 вариант конфигурации)					
10	Раздел 1	+217 %	+213 %	+53 %	+94 %
11	Раздел 2	+158 %	+178 %	+206 %	+216 %
12	Раздел 3	+174 %	+170 %	+110 %	+92 %

Таблица 9

Сравнительная таблица тестирования типовыми задачами (меньше-лучше)

№	Кол-во ядер	Раздел	Конф-я № 1 Lustre	Конф-я № 2 Lustre	Конф-я № 3 Lustre	NFS+UltraPath
1	64	Раздел 1	+47 %	+35 %	+26 %	542 м
2	32	Раздел 1	–3 %	+16 %	–4 %	918 м 52 с
3	32	Раздел 1	+43 %	+8 %	–1 %	988 м 44 с
4	16	Раздел 1	+26 %	–15 %	–6 %	1426 м 6 с
5	16	Раздел 3	+31 %	+14 %	+35 %	2199 м 35 с
6	64	Раздел 2	+14 %	+5 %	–2 %	744 м 17 с
7	32	Раздел 2	–3 %	–10 %	–21 %	1303 м 24 с

Возникшие проблемы с конфигурациями ФС Lustre

Ресурсы системы хранения данных используются одновременно двумя подсистемами: вычислительной подсистемой (МВС) и подсистемой виртуализации.

Во время проведения тестирования второй конфигурации ресурсы системы хранения данных были нагружены так, что наблюдались существенные задержки при работе с подсистемой виртуализации. На момент написания данного доклада, исследование данной проблемы еще не завершено.

Заключение

В результате проведенных работ было выполнено:

– проанализирована текущая конфигурация МВС в части сетевой файловой системы;

– проанализированы варианты замены существующей сетевой файловой системы МВС на более производительную сетевую файловую систему;

– спроектировано решение по внедрению ФС Lustre на МВС;

– выполнено сравнительное тестирование конфигураций:

– конфигурация № 1 с ФС Lustre (все серверные части ФС Lustre на одном узле);

– конфигурация № 2 с ФС Lustre (7 групп по 2 сервера);

– конфигурация № 3 с ФС Lustre (4 группы по 2 сервера);

– конфигурация с NFS+UltraPath.

Исходя из всего вышеизложенного, можно сделать вывод, что применение файловой системы Lustre на МВС целесообразно. Разработанная в рамках исследования конфигурация № 3 с ФС Lustre может быть внедрена в МВС для повышения производительности, отказоустойчивости и масштабируемости файловой системы при дальнейшем использовании.

Поскольку во время тестирования, как ожидалось, наиболее производительной конфигурации 2 ФС Lustre были зафиксированы значительные задержки в отклике подсистемы виртуализации, функционирующей на базе той же СХД, что и МВС, было принято решение провести дополнительное исследование данного явления с целью определения узкого места в системе и возможной ее оптимизации.

Литература

1. Иванушкин В. В., Ильченко Е. А. Совершенствование технологии проведения высокопроизводительных расчетов в контуре СТ КБ-2. Создание Многопроцессорной вычислительной системы // Отчет об ОКР. Саров: ФГУП «РФЯЦ-ВНИИЭФ», 2017.

2. Емельянова С. С., Иванушкин В. В., Ильченко Е. А. Внедрение распределенной файловой системы массового параллелизма Lustre на МВС-1.2 // Отчет о НИР. Саров: ФГУП «РФЯЦ-ВНИИЭФ», 2018.

3. Беляева А. А., Биряльцев Е. В., Галимов М. Р., Демидов Д. Е., Елизаров А. М., Жибрик О. Н. Кластерная архитектура программно-технических средств организации высокопроизводительных систем для нефтегазовой промышленности. 2017.

4. Назаренко Е. В. Сравнение архитектуры распределенных файловых систем. 2016.

5. Lustre Software Release 2.x Operations Manual. 2011.