

РАСПРЕДЕЛЕННАЯ ПАРАЛЛЕЛЬНАЯ ФАЙЛОВАЯ СИСТЕМА Lustre ДЛЯ ОБРАБОТКИ И АНАЛИЗА ДАННЫХ МЕГАСАЙЕНС-ПРОЕКТА NICA

Д. В. Беляков¹, А. А. Кокорев^{1,*}, Д. В. Подгайный¹

¹ Объединенный институт ядерных исследований, Дубна, Россия

Предложено решение на основе распределенной параллельной файловой системы Lustre для быстрого копирования данных и выполнения одновременных расчетов на суперкомпьютере «Говорун», расположенном в ЛИТ ОИЯИ, и вычислительном кластере NCX, находящемся в ЛФВЭ ОИЯИ, включающее режим отказоустойчивости на основе компонентов Lustre и программных пакетов Pacemaker/Corosync. Разработанная архитектура построена на основе современного серверного оборудования с высокоскоростным подключением к сети Ethernet 400 Gbps между разными корпусами Института. Приведены результаты тестирования производительности распределенной параллельной файловой системы Lustre, полученные на основе работы программной утилиты IOR, работающей на основе технологии MPI.

This paper proposes a solution based on the distributed parallel file system Lustre for fast data copying and simultaneous calculations on the Super Computer Govorun located at LIT JINR and the computing cluster NCX located at LHEP JINR, included a fault-tolerance mode based on Lustre components and Pacemaker/Corosync software packages. The developed architecture is built on modern server equipment with a high-speed connection to the Ethernet 400 Gbps network between different buildings of the Institute. The results of testing the performance of the distributed parallel file system Lustre are presented, obtained on the basis of the IOR software utility, which operates on the basis of MPI technology.

PACS: 89.20.Ff; 07.05.Kf

ВВЕДЕНИЕ

Участники коллабораций MPD, BM@N и SPD мегасайенс-проекта NICA при выполнении расчетов активно задействуют различные вычислительные ресурсы ОИЯИ: Tier-1, Tier-2, суперкомпьютер «Говорун» [1], входящие в состав Многофункционального информационно-вычислительного комплекса (МИВК) ОИЯИ [2], и вычислительный кластер NCX ЛФВЭ [3]. При этом не редки ситуации, когда для выполнения

* E-mail: kaa@jinr.ru

расчетов пользователи вынуждены переносить данные между площадками в ручном режиме, что занимает большое количество времени.

Для решения данной проблемы была применена и развернута распределенная параллельная файловая система Lustre, которая включает в свой функционал возможность объединять географически расположенные серверы хранения и создавать единую систему хранения данных, что поможет пользователям решить их проблему с переносом своих данных и распараллеливанием задач на серверах.

В данной работе представлены результаты проектирования и реализации высокоскоростного обмена данных, построенного на основе параллельной файловой системы Lustre [4], объединяющей ресурсы вычислительного кластера NCX (ЛФВЭ) и суперкомпьютера «Говорун» (ЛИТ).

РАСПРЕДЕЛЕННАЯ ПАРАЛЛЕЛЬНАЯ ФАЙЛОВАЯ СИСТЕМА Lustre

Архитектура файловой системы 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.

В основе файловой системы Lustre лежит файловая система ZFS [5], которая обеспечивает надежное и эффективное управление дисковым пространством. ZFS отличается от традиционных файловых систем своей архитектурой, ориентированной на целостность данных и простоту управления. Одной из ключевых особенностей ZFS является ее метод записи данных: вместо перезаписи существующих блоков она всегда записывает новые блоки. Это позволяет избежать проблем, связанных с частичной записью данных, которые могут привести к повреждению файловой системы. Благодаря этому подходу ZFS не нуждается в журнале, как большинство других файловых систем, что упрощает ее структуру и повышает производительность.

В рамках проекта по развертыванию и настройке распределенной файловой системы Lustre между двумя площадками ОИЯИ, а именно ЛИТ и ЛФВЭ, было выделено серверное оборудование с каждой площадки. На этом оборудовании была развернута распределенная параллельная файловая система Lustre. Распределенная архитектура Lustre обеспечивает высокую доступность и отказоустойчивость на основе программных пакетов Pacemaker/Corosync [6], так как в случае сбоя одного из компонентов система продолжает работать, используя другие компоненты.

На рис. 1 представлена схема распределенной файловой системы Lustre, в которой было задействовано серверное оборудование.

Со стороны ЛИТ выделено серверное оборудование, входящее в состав суперкомпьютера «Говорун», — два сервера для управления файловой системой Lustre и два сервера хранения, которые имеют следующие технические характеристики:

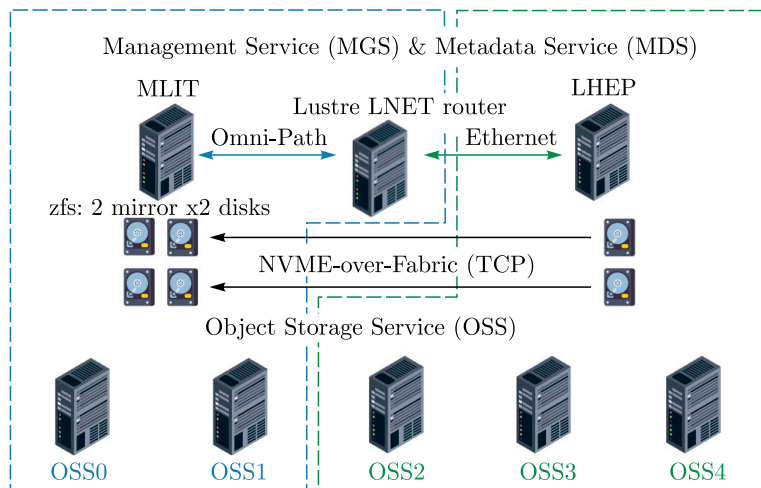


Рис. 1. Схема реализации файловой системы Lustre

— два сервера управления ПСК «Торнадо» TDN711 [7]: 2 процессора Intel(R) Xeon(R) Platinum 8268 CPU @ 2.90GHz, 96 core, 192 Gb оперативной памяти, Intel Omni-path 200 GiB/s;

— два сервера хранения ПСК «Торнадо» TDN551 AFS [8]: 2 процессора Intel(R) Xeon(R) Gold 6248R CPU 3.00GHz, 48 core, 376 Gb оперативной памяти, Intel Omni-path 200 GiB/s, 32× твердотельных накопителя Intel 30,73 TB, формфактор E1.L общей емкостью порядка 1 PB.

На площадке ЛФВЭ было выделено серверное оборудование, сервер для управления файловой системой Lustre, которые имеют следующие технические характеристики.

Сервер управления Supermicro — 2 процессора Intel(R) Xeon(R) CPU E5-2690 v4 @ 2.60GHz, 56 core, 256 Gb оперативной памяти, Mellanox 200 GiB/s.

Три сервера хранения Supermicro — 2 процессора Intel(R) Xeon(R) Gold 6230R CPU 2.10GHz, 52 core, 786 Gb оперативной памяти, Mellanox 200 GiB/s, 16× твердотельных накопителей Intel 15,36 TB, формфактор E1.L общей емкостью 245,76 TB.

МЕТОДИКА ТЕСТИРОВАНИЯ ПРОИЗВОДИТЕЛЬНОСТИ ФАЙЛОВОЙ СИСТЕМЫ

Для изучения функционирования и оценки быстродействия файловой системы Lustre применялись методы тестирования с помощью инструментария IOR. IOR использует MPI для координации операций чтения и записи, выполняемых параллельно на множестве вычислительных узлов.

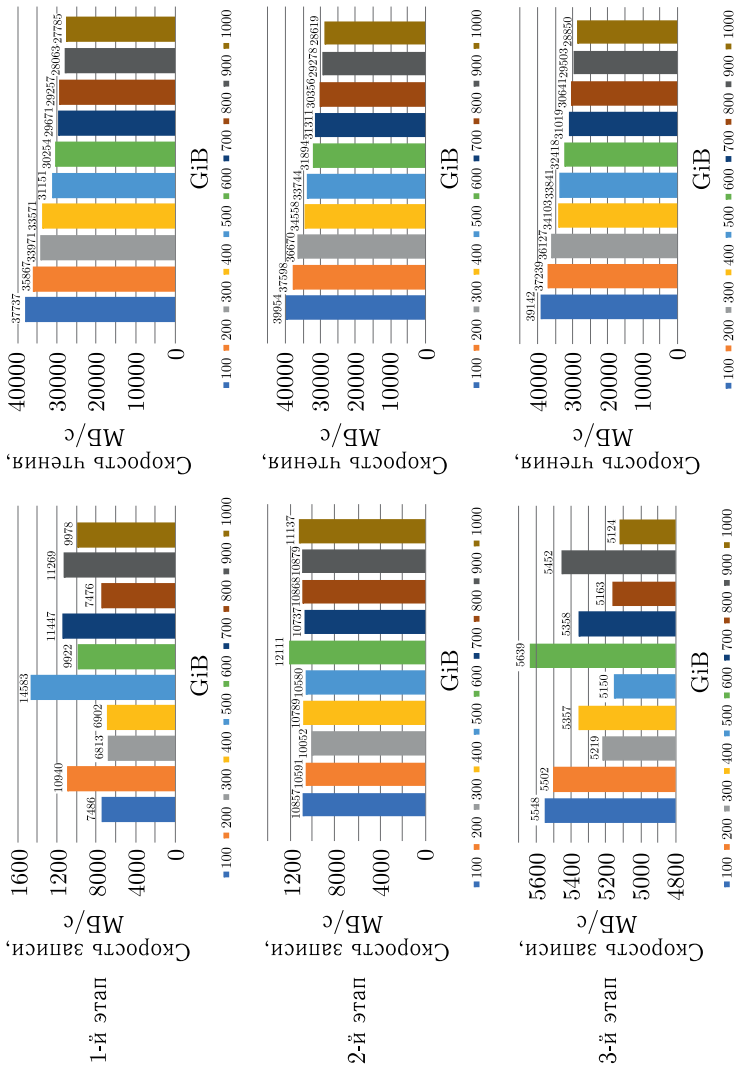


Рис. 2. Результаты тестирования файловой системы Lustre

Процесс тестирования был организован следующим образом: для программы IOR [9] был создан конфигурационный файл, размещенный на управляющем вычислительном узле. Этот файл содержал перечень всех узлов, задействованных в работе IOR. Для параллелизации процессов ввода-вывода IOR опирается на пакет OpenMPI [10], который, в свою очередь, основан на технологии MPI. В процессе компиляции IOR из исходного кода был установлен пакет разработки `openmpi-devel`. Запуск тестов осуществлялся из командной строки при помощи команды `mpirun`.

После завершения тестовых процедур собранные данные записываются в файл и преобразуются в графическое представление для упрощения дальнейшего анализа и интерпретации результатов.

РЕЗУЛЬТАТЫ ТЕСТИРОВАНИЯ ФАЙЛОВОЙ СИСТЕМЫ Lustre

Тестирование проводилось в три этапа инструментарием IOR, который использует программный пакет OpenMPI. Результаты тестирования файловой системы Lustre приведены на рис. 2.

1-й этап включал в себя тестирование файловой системы, когда запись и чтение происходили на всех серверах хранения, т. е. на серверах ЛИТ и ЛФВЭ. Из графиков на рис. 2 видно, что максимальная скорость записи составила 14583 МБ/с, а минимальная — 6813 МБ/с.

2-й этап включал в себя тестирование файловой системы, когда запись и чтение происходили на всех серверах хранения ЛИТ. Из графиков видно, что скорость записи данных на сегмент ЛИТ происходит быстрее и стабильнее, чем на остальных двух этапах тестирования.

3-й этап включал в себя тестирование файловой системы, когда запись и чтение происходили на всех серверах хранения, расположенных на площадке ЛФВЭ. Из графиков видно, что запись данных происходит медленнее, чем на графиках 1-го и 2-го этапов.

Скорость чтения при выполнении трех этапов тестирования почти ничем не отличается.

Распределенная параллельная файловая система Lustre позволила создать единое пространство хранения данных, доступное с обеих площадок, что значительно упростило обмен данными и совместную работу ученых.

Финансирование. Работа выполнена при поддержке Министерства науки и высшего образования Российской Федерации, проект «Фундаментальные и прикладные исследования на мегасайенс-экспериментальном комплексе NICA (ОИЯИ)» FSWU-2025-0014.

СПИСОК ЛИТЕРАТУРЫ

1. Суперкомпьютер «Говорун». http://hlit.jinr.ru/supercomputer_govorun/ (дата обращения 20.07.2025).
2. Многофункциональный информационно-вычислительный комплекс ОИЯИ. <https://micc.jinr.ru/> (дата обращения 22.07.2025).

3. Лаборатория физики высоких энергий. <https://lhep.jinr.ru/> (дата обращения 29.07.2025).
4. Файловая система Lustre. <https://www.lustre.org/> (дата обращения 02.08.2025).
5. Файловая система ZFS. <https://zfsonlinux.org/> (дата обращения 10.08.2025).
6. Pacemaker/Corosync. <https://clusterlabs.org/projects/> (дата обращения 15.08.2025).
7. РСК «Торнадо» TDN711. <https://rscgroup.ru/product/rsc-tornado-servers/tdn711/> (дата обращения 20.08.2025).
8. РСК «Торнадо» TDN551 AFS. <https://rscgroup.ru/product/rsc-tornado-servers/tornado-afs/> (дата обращения 20.08.2025).
9. IOR. <https://github.com/hpc/ior> (дата обращения 25.08.2025).
10. OpenMPI. <https://www.open-mpi.org/> (дата обращения 30.08.2025).